

# Improving Biomedical Literature Search Engines for Medical Professionals

Der Fakultät für Informatik  
der Universität Duisburg-Essen

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte Dissertation

von

**Sameh Frihat**

aus  
Jenin

1. Gutachter: Prof. Dr.-Ing. Norbert Fuhr
2. Gutachter: Prof. Dr. Christin Seifert

Tag der mündlichen Prüfung: 22.11.2024

# DuEPublico

Duisburg-Essen Publications online

UNIVERSITÄT  
DUISBURG  
ESSEN

*Offen im Denken*

ub

universitäts  
bibliothek

Diese Dissertation wird via DuEPublico, dem Dokumenten- und Publikationsserver der Universität Duisburg-Essen, zur Verfügung gestellt und liegt auch als Print-Version vor.

**DOI:** 10.17185/duepublico/82698

**URN:** urn:nbn:de:hbz:465-20241204-103603-2



Dieses Werk kann unter einer Creative Commons Namensnennung - Nicht kommerziell 4.0 Lizenz (CC BY-NC 4.0) genutzt werden.

UNIVERSITÄT  
DUISBURG  
ESSEN

*Open-Minded*

DISSERTATION

---

**Improving Biomedical Literature  
Search Engines for Medical  
Professionals**

---

Faculty of Computer Science  
University of Duisburg-Essen

to obtain the degree  
Dr.-Ing.

**M. Sc. Sameh Frihat**  
from Jenin

Advisor:  
Prof. Dr.-Ing. Norbert Fuhr

Duisburg, Germany  
2024



# Acknowledgments

I finally reached the end of this journey in which Information Retrieval was just a small part of all I lived and learned.

I would like to express my deepest gratitude to all those who supported me throughout the journey of completing this dissertation. First and foremost, I am incredibly thankful to my advisor, Prof. Dr. Norbert Fuhr, for his invaluable discussions and his encouragement. From the beginning of my PhD journey, the assurance that I was in good hands whenever I mentioned your name has proven profoundly true over the last three and half years. I deeply appreciate the insightful discussions that helped refine and shape my research. I am also grateful to the examination committee members for taking the time to read and evaluate my work. Furthermore, I would like to acknowledge the support and encouragement of my colleagues and friends at Duisburg-Essen University and WisPerMed RTG. I am also grateful to Dr. Ahmet Aker and Dr. Emina Kurtic for their continuous support and encouragement throughout this process. Their continuous feedback and motivation played a significant role in completing this work. To my family, thank you for your endless support, patience, and understanding over the years. Your unwavering belief in me provided the strength to persevere during the most challenging times. Finally, I extend my gratitude to everyone who, in one way or another, contributed to the completion of this dissertation. Your support and encouragement will always be remembered.



# Abstract

Medical search engines for experts are concerned with retrieving relevant information to support medical professionals' information-seeking tasks. However, current search engines for medical research publications often fail to provide a quick access to accurate and reliable information because they rely on a bag-of-words approach that treats documents as collections of words to be matched with the keywords in the search query.

This thesis seeks to improve the state-of-the-art search techniques in biomedical research publication repositories by exploring methods of semantic information retrieval, and for context modeling by considering the characteristics of professional users performing the search. To this end, this work adopts the notion of multidimensional relevance in biomedical publication search and evaluates the performance of the retrieval systems in this context by addressing the following factors: personalization, credibility, semantic relevance, interactivity and integration with Large Language Models.

Specifically, as a way to address personalization we evaluate the methods for identifying document difficulty levels, expressed as readability and technicality of a document, as well as methods for classification of medical documents into medical sub-fields they address. Credibility is tackled by proposing methods of classifying biomedical publications according to the Level of Evidence (LoE) they are based on and evaluating how this affects medical document retrieval. Semantic relevance is addressed by identifying bio-concepts, i.e. genes, diseases and chemicals in the research papers and evaluating multiple methods of incorporating them into the document retrieval. Bio-concepts are also used to enhance the interactivity capability of a medical search engine, where the extracted bio-concepts are made available to the user during search by visualization. Finally, conversational search engine settings, where information retrieval is combined with the capabilities of a Large Language Model (LLM) in a Retrieval Augmented Generation (RAG) setup are investigated.

The results of these investigations are implemented into WisPerMed, a new search engine for the Medline database. A user study conducted with 131 medical practitioners demonstrated that this search engine reduces the time spent searching and the number of queries needed to finish a particular task, supports quicker and more accurate decision-making in medicine and increases user satisfaction with the search process.

Our findings show that (1) incorporating semantic relevance significantly improves the quality of retrieved information from medical literature, (2) the use of bio-concepts and LoE

improves both the precision and trustworthiness of search results and (3) the developed models for predicting user-specific parameters allows for personalizing search results on aspects as document difficulty and medical sub-fields for more relevant outcomes. As a result, this work contributes to creating more efficient and effective tools for medical professionals, facilitating better patient care, and advancing medical research.

**Keywords:** Biomedical Information Retrieval, Level of Evidence, Biomedical Concepts, WisPerMed, Large Language Models, Personalized Search Engines, Contextual Medical Search.



# Zusammenfassung

Medizinische Suchmaschinen für Experten konzentrieren sich darauf, relevante Informationen zur Unterstützung der Informationssuche von medizinischen Fachkräften bereitzustellen. Allerdings bieten aktuelle Suchmaschinen für medizinische Forschungspublikationen oft keinen schnellen Zugang zu genauen und zuverlässigen Informationen, da sie auf einem Bag-of-Words-Ansatz beruhen, der Dokumente als Sammlungen von Wörtern betrachtet, die mit den Schlüsselwörtern in der Suchanfrage abgeglichen werden.

Diese Arbeit zielt darauf ab, die Suchmaschinen für biomedizinische Forschungspublicationsdatenbanken zu verbessern, indem Methoden der semantischen Informationssuche und zur Kontextmodellierung untersucht werden. Zu diesem Zweck führt diese Arbeit das Konzept der multidimensionalen Relevanz in der Suche nach biomedizinischen Publikationen ein und bewertet die Leistung der Suchsysteme in diesem Kontext unter Berücksichtigung der folgenden Faktoren: Personalisierung, Vertrauenswürdigkeit, semantische Relevanz, Interaktivität und Integration mit Large Language Models (LLMs).

Konkret werden zur Personalisierung die Methoden zur Identifizierung des Schwierigkeitsgrades von Dokumenten (ausgedrückt durch die Lesbarkeit und den technischen Anspruch eines Dokuments), sowie Methoden zur Klassifizierung medizinischer Dokumente in die medizinischen Fachgebiete, die sie behandeln, bewertet. Vertrauenswürdigkeit wird durch die Klassifizierung biomedizinischer Publikationen nach dem Level of Evidence (LoE), auf dem sie basieren, angegangen und deren Einfluss auf die medizinische Dokumentensuche wird untersucht. Die semantische Relevanz wird durch die Identifizierung von Biokonzepten (Genen, Krankheiten und Medikamenten) in den Forschungsarbeiten, sowie durch die Bewertung verschiedener Methoden zu deren Integration in die Dokumentensuche berücksichtigt. Biokonzepte werden auch verwendet, um die Interaktivität einer medizinischen Suchmaschine zu unterstützen, wobei die extrahierten Biokonzepten den Nutzern während der Suche durch Visualisierung zugänglich gemacht werden. Schließlich wird das Szenario der konversationalen Suche untersucht, bei dem die Informationssuche mit den Fähigkeiten eines großen Sprachmodells (LLM) im Rahmen der Retrieval-Augmented-Generation (RAG) kombiniert wird.

Die Ergebnisse dieser Untersuchungen wurden in WisPerMed implementiert, eine neue Suchmaschine für die Medline-Datenbank. Unsere Benutzerstudie mit 131 medizinischen Fachkräften zeigte, dass diese Suchmaschine die Suchzeit und die Anzahl der benötigten

Suchanfragen reduziert, eine schnellere und genauere Entscheidungsfindung in der Medizin unterstützt und die Zufriedenheit der Nutzer mit dem Suchprozess erhöht.

Unsere Ergebnisse zeigen, dass (1) die Einbeziehung der semantischen Relevanz die Qualität der aus der medizinischen Literatur abgerufenen Informationen erheblich verbessert, (2) die Verwendung von Biokonzepten und LoE sowohl die Genauigkeit als auch die Vertrauenswürdigkeit der Suchergebnisse erhöht und (3) die entwickelten Modelle zur Vorhersage nutzerspezifischer Parameter die Suchergebnisse in Bezug auf Dokumentschwierigkeit und medizinische Fachgebiete für relevantere Ergebnisse personalisieren würden. Diese Arbeit trägt somit zur Entwicklung effizienterer und effektiverer Werkzeuge für medizinische Fachkräfte bei, erleichtert eine bessere Patientenversorgung und fördert die medizinische Forschung.

**Schlüsselwörter:** Biomedizinische Informationssuche, Evidenzlevel, Biomedizinische Konzepte, WisPerMed, Large Language Models (LLM), Personalisierte Suchmaschinen, Kontextsensitive medizinische Suche.

# List of Abbreviations

<b>Abbreviation</b>	<b>Description</b>
BERT	Bidirectional Encoder Representations from Transformers
BioNLP	Biomedical Natural Language Processing
BM25	Best Matching 25
BM25F	BM25 Fielded
EBM	Evidence-Based Medicine
ICC	Intraclass Correlation Coefficient
IR	Information Retrieval
LIME	Local Interpretable Model-Agnostic Explanations
LLM	Large Language Model
LoE	Level of Evidence
LTR	Learning to Rank
MeSH	Medical Subject Headings
NDCG	Normalized Discounted Cumulative Gain
NER	Named Entity Recognition
NLP	Natural Language Processing
RCT	Randomized Controlled Trial
RAG	Retrieval-Augmented Generation
RMSE	Root Mean Square Error
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SMOG	Simple Measure of Gobbledygook
Std	Standard Deviation
TF-IDF	Term Frequency-Inverse Document Frequency
TREC PM	Text Retrieval Conference Precision Medicine
UMLS	Unified Medical Language System
WPM	Words Per Minute
API	Application Programming Interface



# Contents

<b>Acknowledgments</b>	<b>i</b>
<b>Abstract</b>	<b>iii</b>
<b>Zusammenfassung</b>	<b>v</b>
<b>List of Abbreviations</b>	<b>vii</b>
<b>1. Introduction</b>	<b>1</b>
1.1. Motivation . . . . .	6
1.2. General Research Questions . . . . .	8
1.3. Thesis Contribution . . . . .	10
1.4. Thesis Statement and Summary . . . . .	11
1.5. Publications . . . . .	12
<b>2. Related Work</b>	<b>13</b>
2.1. Information Retrieval . . . . .	13
2.1.1. Evaluating Search Systems . . . . .	15
2.1.2. Multidimensional Relevance . . . . .	16
2.1.3. Interactive Information Retrieval . . . . .	17
2.2. Search in the Health Domain . . . . .	18
2.2.1. Health Search using General Search Engines . . . . .	18
2.2.2. Health Search using Specialized Search Engines . . . . .	19
2.2.3. Medical Search Engines Challenges . . . . .	21
2.3. Natural Language Processing in Medical IR . . . . .	22
2.3.1. Biomedical NLP . . . . .	22
2.3.2. PubMed-BERT: Enhancing Biomedical Text Understanding . . . . .	23
2.3.3. Large Language Models (LLMs) in Biomedical NLP . . . . .	24
2.4. Personalization Aspects in Medical Search Engines . . . . .	25
2.5. Summary . . . . .	27
<b>3. Document Difficulty Estimation for Medical Search Engine Personalization</b>	<b>29</b>
3.1. Introduction . . . . .	29

---

3.2. Related Work . . . . .	31
3.2.1. Readability Formulas . . . . .	32
3.2.2. Machine Learning Approaches . . . . .	33
3.3. Materials and Methods . . . . .	34
3.3.1. Data Collection . . . . .	35
3.3.2. The Model of Document Difficulty . . . . .	37
3.3.3. Evaluation . . . . .	38
3.4. Results . . . . .	39
3.4.1. User Agreement on Document Difficulty . . . . .	39
3.4.2. Interlingual Differences in Document Difficulty . . . . .	39
3.4.3. Modelling Document Difficulty . . . . .	42
3.5. Discussion . . . . .	43
3.6. Summary . . . . .	45
<b>4. Medical Document Subfield Classification for Search Engine Personalization</b>	<b>47</b>
4.1. Introduction . . . . .	47
4.2. Materials and Methods . . . . .	49
4.2.1. Data Collection . . . . .	49
4.2.2. Experimental Setup . . . . .	52
4.2.3. Evaluation . . . . .	53
4.3. Results . . . . .	53
4.3.1. Performance Evaluation . . . . .	53
4.3.2. Statistical Test Results . . . . .	54
4.4. Discussion . . . . .	56
4.5. Summary . . . . .	57
<b>5. Bio-Concepts for Enhanced Precision in Medical Search</b>	<b>59</b>
5.1. Introduction . . . . .	59
5.2. Materials and Methods . . . . .	61
5.2.1. Dataset . . . . .	61
5.2.2. Documents Annotation with Bio-concepts . . . . .	63
5.2.3. Experimental Setup . . . . .	64
5.2.4. Evaluation Metric . . . . .	67
5.2.5. Parameter tuning . . . . .	67
5.3. Results . . . . .	68
5.3.1. Sparse Retrieval . . . . .	69
5.3.2. Hybrid Retrieval . . . . .	70
5.4. Discussion . . . . .	71
5.5. Summary . . . . .	73

---

<b>6. Level of Evidence Classification for Improving Medical Search Engine Credibility</b>	<b>75</b>
6.1. Introduction	75
6.2. LoE Classifier	78
6.2.1. Data	79
6.2.2. Experimental Setup	80
6.2.3. Classifier Evaluation	81
6.2.4. Classifier Discussion	84
6.3. Levels of Evidence as a filter in medical IR	85
6.3.1. Data	86
6.3.2. Experimental Setup	86
6.3.3. Results	87
6.4. Discussion	88
6.5. Summary	89
<b>7. User Evaluation of WisPerMed</b>	<b>91</b>
7.1. Introduction	91
7.2. Methodology	93
7.2.1. Implementation	93
7.2.2. User Study Setup	94
7.2.3. Data Collection	98
7.3. Results	100
7.3.1. Quantitative Analysis	100
7.3.2. Qualitative Feedback	103
7.4. Discussion	105
7.5. Summary	106
<b>8. Retrieval Augmented Generation</b>	<b>109</b>
8.1. Introduction	109
8.2. Materials and Methods	111
8.2.1. Dataset	111
8.2.2. Experimental setup	112
8.3. Results	116
8.3.1. System Evaluation	116
8.3.2. Application: Medical LLMs Search Engine	118
8.4. Discussion	119
8.5. Summary	121
<b>9. Discussion and Conclusion</b>	<b>123</b>
9.1. Discussion of Research Questions	124
9.1.1. Personalization	125
9.1.2. Bio-concepts	125

---

9.1.3. Levels of Evidence . . . . .	126
9.2. Implications for Practice . . . . .	127
9.3. Limitations . . . . .	128
9.4. Future Work . . . . .	129
9.5. Conclusion . . . . .	130
<b>A. Appendix</b>	<b>133</b>
A.1. Complete Version of The Annotation Tool . . . . .	133
A.2. Complete Version of the User Study of Chapter 7 . . . . .	137
A.3. WisPerMed Search Result Page . . . . .	146
<b>Bibliography</b>	<b>149</b>



# 1. Introduction

In the dynamic and fast-paced environment of modern healthcare, the ability to access accurate, relevant and trustworthy medical information is paramount, as it can lead to decisions that directly impact human lives. Traditional methods of consulting medical literature or databases are often time-consuming and impractical due to the large and fast-growing volume of relevant literature in this field. On one hand, this need for fast and efficient information access in medical settings parallels the information search demands in other fields and everyday contexts. On the other hand, information search by healthcare professionals has unique and specific requirements.

The necessity to access information quickly and efficiently has existed since the inception of the internet as the increased connectivity and data availability impacted developments in many fields [251]. As the online information space became increasingly complex and overwhelming for information searchers [286], search engines were developed to allow users to describe the desired results using word representations [22]. The development of general search engines set the foundation for more specialized efficient search tools in different domains and fields like medicine and healthcare.

The primary function of a search engine is to provide users with search results that are as relevant and useful as possible at the moment of the search. To achieve this, various personalization methods have been developed, including sophisticated algorithms for personalization based on user profiles, personalization based on contextual information and interactive information retrieval (IR).

The personalization based on user profile involves the creation of user profiles and continuous refinements. These profiles are set manually or constructed automatically from different data points, including past search queries, click behavior, browsing history, and even interactions across different devices and platforms. This way of personalization is best known from E-commerce platforms. For instance, when a user repeatedly searches for organic skincare products in an E-commerce platform, the system identifies this pattern and tailors the search results to highlight relevant items, such as organic face creams, serums, and cleansers. This dynamic adjustment not only improves the shopping experience by providing more accurate and desired product recommendations but also enhances customer satisfaction and loyalty. This allows the system to continuously refine its suggestions and interface based on ongoing interactions, creating a highly personalized shopping environment that evolves with the

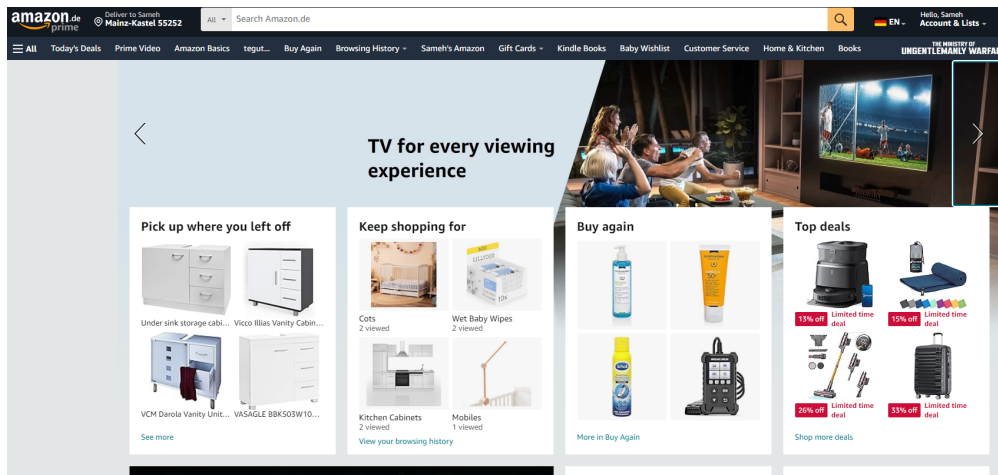


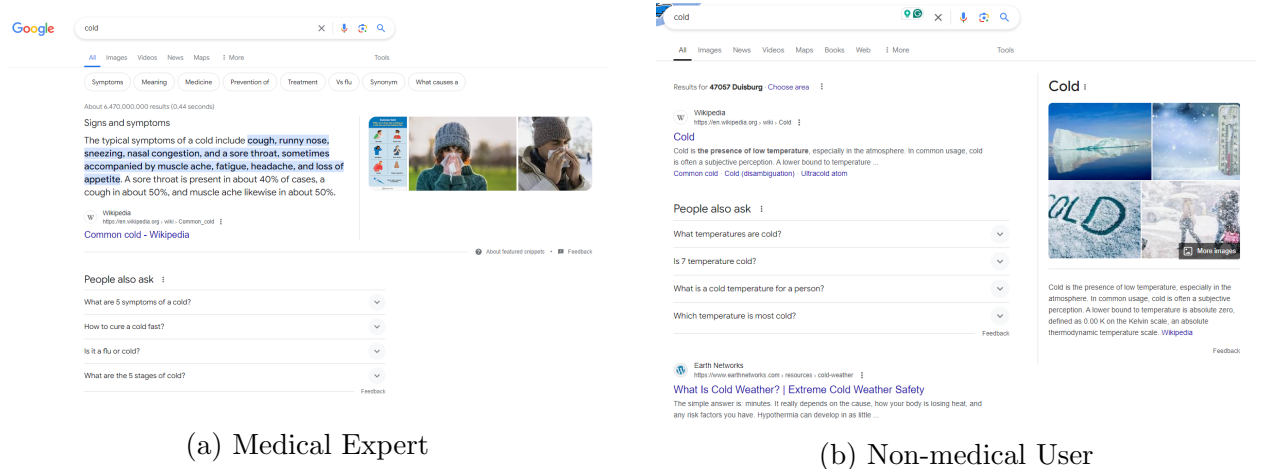
Figure 1-1.: Screenshot of Amazon homepage showing the personalized page based on the user interaction.

user's changing preferences, as shown in Figure 1-1.

However, personalization based on user profiles takes place even in less obvious contexts. For instance, a user who frequently searches for medical journals and clinical studies will have a profile indicating a preference for academic and professional content. Consequently, when this user searches for a term like “cold”, the search engine prioritizes results related to medical research and clinical information about medical treatment for cold, as shown in Figure 1-2a. On the other hand, search results for a user with no medical search history to retrieve documents about the “cold”, result would be in a different context, like cold weather, as shown in Figure 1-2b. This personalized approach ensures that users receive results that are most pertinent to their interests and needs, enhancing the overall search experience by reducing the effort required to find relevant information.

In addition to personalization based on user profiles, search engines also rely heavily on contextual aspects to refine search results. These aspects include factors such as time, location, previous search activities, and the device being used. For example, a user searching for “Pizza” will receive restaurants to order pizza based on the location and time as shown in Figure 1-3. This increases the chances that retrieved restaurants will be able to deliver the pizza as they are open at the time of search and in geographical proximity to the user. Contextual signals such as these help search engines understand the immediate environment and the intent behind search queries, allowing them to retrieve results that are not only personalized but also contextually appropriate.

Search engines also integrate interactive IR techniques, aiming to put the user at the center of the retrieval cycle. These techniques include real-time user feedback, personalized query refinement, and adaptive learning mechanisms. This approach enhances the efficiency of the search process and ensures that users receive the most relevant information according



(a) Medical Expert

(b) Non-medical User

Figure 1-2.: Screenshots of searching for “cold” on Google search engine by two users, medical doctor and non-medical person, on their personal computers

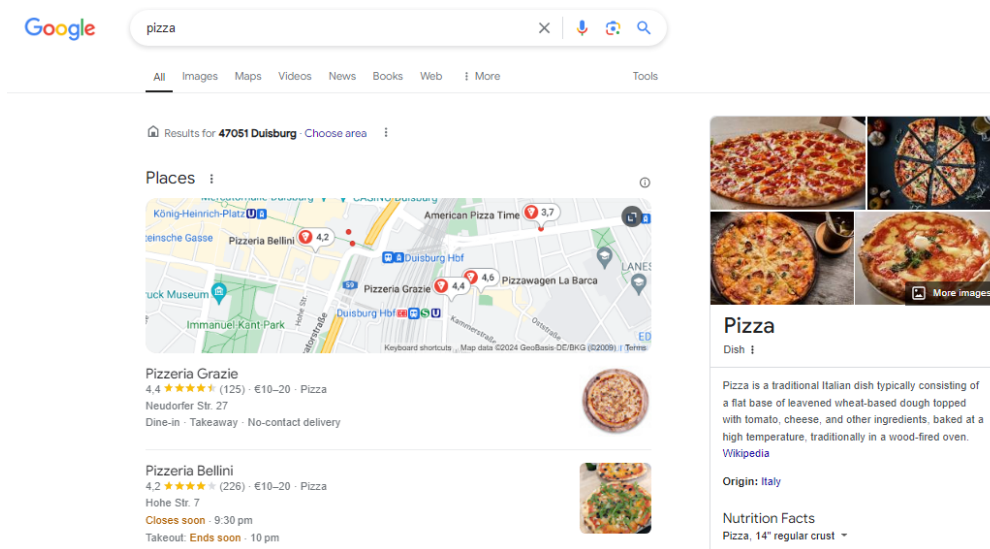


Figure 1-3.: Screenshot of searching for “Pizza” on Google search engine showing the different aspects considered such as time, location, and task.

to their information needs.

Unlike general domains, searching for medical information presents distinct challenges due to the complexity and specificity of the content involved. Addressing the demands of this highly specialized field medical search engines emerged to help filter and access the huge amount of online health-related information effectively [14, 10]. However, most innovative medical search engines focus on empowering health consumers to access health-related information rather than health experts and professionals [161, 291, 141].

A number of techniques have been implemented to simplify healthcare consumers’ access

to medical information. For example, personalization techniques have been proposed, where users' language skills are stored in user profiles and considered during the search to increase readability [44]. Furthermore, the complexity of medical jargon can be tackled by integrating technologies like the Unified Medical Language System (UMLS), [228, 6] as implemented in MedicoPort. MedicoPort is a medical search engine which utilizes the UMLS to semantically improve search effectiveness for users without medical training, thereby enabling the general public to access more complex medical knowledge [33]. Finally, medical search engines for health consumers also use interactive retrieval methods, which are known to significantly improve search efficacy and effectiveness [303].

While these techniques allow to retrieve most relevant search results, they generally do not address the trustworthiness or credibility of the results. In health-related search settings, search results often lead to decisions that directly impact human lives. For this reason, it is crucial that medical search engines address credibility, whether they are aimed at health consumers or experts. Credibility refers to a situation, where findings are consistent under similar conditions [249]. In medical settings, relevance and credibility contribute to identifying significant information, which implies that findings have a practical and meaningful impact that is not due to chance in terms of its effect on patient care or outcomes [215].

Some studies propose to tackle the credibility of search results in search engines for health consumers by utilizing the interaction between health-expert and non-expert searches. According to Schwarz et al. [224] for example, the popularity of a webpage among experts is an important factor in helping non-experts identify credible websites. Search engines therefore prioritize pages visited by medical experts to improve search result credibility.

Credibility, however, is an essential requirement for search results retrieved by medical search engines for experts too. To address the substantial need for a credible evidence base, a framework for evaluating the credibility of medical documents has been developed within the field of evidence-based medicine (EBM). This Level of Evidence (LoE) framework includes hierarchies of evidence, ranking clinical guidelines and systematic reviews at the top, followed by randomized controlled trials (RCTs), cohort studies, case reports, expert opinions, and finally, animal studies (cf. Chapter 6). These evidence levels could serve as credibility measures in medical search engines for experts, yet to date, they have not been utilized for this purpose.

Thus, despite complex solutions to personalization and credibility, medical search engines for health consumers do not directly address the information needs in health experts' search. Numerous studies confirm that there is a significant difference in information needs of health professionals as opposed to those of health consumers and that distinguishing between health consumers and health experts can significantly improve their satisfaction with search engine results [44, 193, 224, 287, 284].

Access to precise and up-to-date medical information is crucial for healthcare profes-

sionals, impacting clinical decision-making, patient care and outcomes [184], but medical literature or specialist database search engines must navigate a rapidly growing volume of data. Healthcare professionals face problems that are multidisciplinary in nature and often require interdisciplinary collaboration and evidence-based information. Moreover, search engines for medical experts need to address multilingual needs and potential biases [228], as well as accurately interpret highly specialized medical jargon of the medical literature that general search engines and laypeople often struggle with [24, 21].

Although specialized medical search engines for professionals exist (cf. Chapter 2, Section 2.2), they have several limitations. PubMed, for example, the best known general purpose specialist search engine for biomedical literature, ignores several implicit aspects of the search context, like task, previous interaction and level of evidence. It also does not consider the user personality features, such as expertise or level of interest. Therefore, it does not sufficiently address medical experts' information needs. For example, without a credibility assessment content of low evidence credibility might be retrieved, such as the notion that eating apricot seeds will cure cancer [252], which is proven to be inaccurate and even can cause cyanide poisoning [270]. Additionally, the lack of personalization for language difficulty might retrieve evident results with complex language that does not meet the user's literacy needs. Not considering a user's level of expertise can also lead to irrelevant search results, although they might be both high in credibility and low in difficulty. For instance, a cardiologist might benefit from search results that prioritize the latest research in cardiology, while a general practitioner might require a broader spectrum of medical information [235, 17]. Thus, the limitations arising from ignoring factors beyond query similarity in expert medical search engines are substantial [222]. As medical knowledge continuously expands, medical experts' ability to accurately retrieve only the most relevant specialist information is becoming increasingly essential [182].

PubMed, however, only supports the decision on similarities between the query and documents and considers some document features like citation count and publication date [69]. This *bag-of-words* approach treats queries and documents as collections of words, ignoring deeper semantic relationships between them and the wider context of the search and user characteristics. Considering semantic relationships is essential in the medical domain because the terminology and the specific information needs of healthcare professionals demand more sophisticated search techniques. In many cases, medical experts are not exactly sure about the task formulation in query text. A bag-of-words approach pushes users to perform exploratory searches, which is often not viable on their typically tight schedules. These are some of the reasons why modern IR systems integrate the concept of *semantic relevance* — an approach that considers the meaning and context of search terms, as well as the complex features of the search context.

Semantic IR methods which utilize Natural Language Processing (NLP) and domain-specific ontologies offer a promising solution to the challenges posed by traditional bag-of-

words search methods. These systems can deliver more accurate and relevant results by understanding the relationships between different concepts within the biomedical literature. In addition, all information activities that occur within the search context can affect and influence how individuals access information, engage with the system, make decisions about retrieved results, and assess the retrieval system [92, 109]. Semantic IR methods strive to consider the context and intent behind a query, as well as user knowledge. This approach not only improves the efficiency of information retrieval but also enhances the overall search experience by reducing the cognitive load on users, enabling them to find the most pertinent information quickly and effectively [258].

Semantic IR methods have not been applied to expert search in biomedical research publication repositories. This work argues that developing a medical search engine, which health professionals can use to search biomedical literature and which implements techniques for personalization, credibility and interactivity can substantially improve search efficacy and efficiency compared to the currently available search systems.

## 1.1. Motivation

The main motivation behind this work derives from the above observation that search engines for biomedical literature still need substantial improvement before they can successfully address medical professionals' information search needs. In this work improvement is defined as a user-centric, task-focused category along with the observation in [53] that information retrieval research should focus on helping users understand the information they retrieve, in addition to improving the retrieval process itself. This requires a transition from a bag-of-words approach as currently available through PubMed to semantic IR methodologies that seek to consider more complex search aspects. Specifically, an approach to search is needed that addresses the problems of personalisation and information credibility essential in medical settings, as well as maximizes search efficiency necessary in a highly fast-paced environment.

Previous studies investigated and proposed several factors that can be integrated into medical search engines to support medical experts. An expert's field of expertise or interest, user-specific context, as well as interface design and information visualization for easy interpretation and understanding have been investigated. In addition, query formulation, evidence credibility and tools that support quick title and abstract screening in systematic reviews have been proposed [90, 91, 142, 308, 220].

More recently, conversational search engines have transformed the way users search for information in general and in the medical field in particular. Conversational search engines are systems based on Large Language Models (LLMs) that enable users to interact and retrieve information through natural language dialogue. While LLM's ability to generate coherent

and contextually appropriate responses is impressive, the accuracy and reliability of these responses in the highly specialized and critical field of medicine need a thorough evaluation. LLMs, despite their sophistication, are not immune to errors. They can produce convincing but factually incorrect information, a phenomenon known as “hallucination” [242]. In the context of healthcare, where misinformation can have serious, even life-threatening consequences, this issue is of particular concern [264].

A search engine for medical experts requires transforming the retrieved document space into a multidimensional one, where aspects like document readability, document technicality level, information evidence level and medical sub-fields of expertise are critical for matching user needs [108, 273, 232, 267]. Methods that allow personalized access to documents, such as a document’s difficulty, the medical subfield the document addresses, the most important medical concepts in the document, as well as credibility level estimation are highly needed in medical search engines for experts [73, 195]. Such a search engine would provide more precise and relevant results, improving efficiency and reducing search time [267]. Additionally, interactive IR techniques can be employed to allow medical experts to interact with automatically extracted medical concepts, thus refining search results based on expert users’ specific needs and contexts and saving time needed to access relevant information.

Personalizing expert medical search engines by integration of user profiles can address some of the limitations of the state-of-the-art search engines for medical experts. A user’s field of expertise, language proficiency, and search behavior can be stored in the profile and used at retrieval time to streamline the information retrieval process and reduce the cognitive load on medical professionals by filtering out irrelevant or overly complex information. Moreover, integrating user-specific factors such as prior search history, interaction patterns, and explicit user preferences can further personalize the search experience [174]. This focus on context and personalized user profiles would offer a tailored, user-centric approach to information retrieval that aligns with medical practitioners’ specific needs and workflows, ultimately contributing to better patient care and clinical outcomes [129].

The integration of credibility aspects in document relevance evaluation can address the limitation of retrieving unsubstantiated content. Medical search engines for general health consumers ensure their credibility and ease of understanding by analyzing pages that have been reviewed by medical experts or incorporating user profiles into the search process. This approach helps ensure that the information presented is both reliable and accessible to users. However, when targeting different audiences, these aspects need to be adjusted [278]. For medical experts, document difficulty must align with their advanced knowledge [53], and credibility should be based on sources trusted by professionals [217], which often differ from those trusted by general consumers.

This thesis investigates several ways to improve on the state of the art information retrieval systems for medical literature search. To this end it considers the multidimensional

definition of relevance during information search by medical experts. It evaluates crucial personalization features, specifically, a document’s level of difficulty (i.e. its readability and technicality) and the medical specialization field the document addresses. Credibility assessment in medical literature search is studied by incorporating the strength of scientific evidence into the IR process, as defined in the Level of Evidence framework already well established in the field of evidence based medicine. Levels of evidence are also investigated in a retrieval augmented generation (RAG) setting, which combines the medical publication retrieval with the generative power of LLMs. In addition, this work evaluates semantic search, in which automatically extracted bio-concepts such as genes, mutations, diseases, etc, are highlighted and displayed during the search, allowing medical professional users to easily further personalize their search queries.

The credibility and semantic aspects are incorporated into WisPerMed, an advanced search engine tailored to the requirements of research literature search for healthcare professionals. The effectiveness of WisPerMed is evaluated in realistic search scenarios with medical professionals to demonstrate the benefits of semantic relevance aspects for the efficiency and effectiveness of medical literature search. By addressing these critical factors, WisPerMed provides a robust solution that meets the sophisticated demands of medical experts, ultimately improving their ability to deliver high-quality patient care in high demand environments.

## 1.2. General Research Questions

The primary aim of this work is to improve medical literature search engines by transitioning from traditional bag-of-word approaches to those considering semantic relevance, which requires addressing several under-researched areas in medical information retrieval.

First, while personalization using user profiles is a well-researched field, there is a gap in evidence of how accurately medical publications themselves can be tagged, so that they can be matched with medical expert user profiles effectively. To this end, this work addresses the following research question:

**RQ1: Can we categorize medical research publications into their difficulty levels and medical subfields?**

In this context, it is important to ask how recent Large Language Models (LLMs) perform on these personalization-related tasks relative to substantially more resource-intensive fine-tuned models. We address this question on the task of classification of publications into medical subfields. Specifically, the second research question we address is:

**RQ2: Can LLMs classify medical literature into its corresponding medical subfield?**



Second, including bio-concepts like genes, diseases, and chemicals in the search for medical literature has the potential to make the search more precise and the results more relevant to the users. Unlike the current bag-of-words approach in medical publication search engines such as PubMed, including bio-concepts allows for document retrieval based on semantic relevance. It is as yet not understood however, to what extent this benefits both the retrieval process and the user's experience. Our next research questions address this gap:

**RQ3: Does query expansion with bio-concepts improve the relevance of retrieved medical publications?**

**RQ4: How do bio-concepts affect user search experience when included as interactive features, such as abstract highlights, word clouds, or query expansions?**

Third, credibility and trustworthiness of the search results is a key dimension in the multidimensional space that defines the relevance of publication search results for medical experts. The credibility standards are well established within the professional community and defined in frameworks, such as Level of Evidence (LoE) Framework. Yet, the evidence base of medical publications is not considered in current approaches to medical publication search. This work explores the benefits of integration of existing professional guidelines into the search process by addressing the following research questions:

**RQ5: Can Level of Evidence be identified in medical research papers?**

**RQ6: Does the inclusion of Levels of Evidence as a filter improve the efficiency and effectiveness of medical publication retrieval?**

**RQ7: How useful is the indication of a publications Level of Evidence to medical professionals when searching for relevant research publications?**

Finally, with the current popularity of LLMs, it is almost imperative to investigate whether conversational search engines using LLMs can replace traditional search engines and generate relevant and trustworthy responses to prompts posed by medical expert users. To this end, the medical publication retrieval and the generative power of LLMs can be combined in the so-called Retrieval-Augmented Generation (RAG) framework. We investigate, whether the indication of a publication's evidence base improves the performance of LLMs in professional medical search tasks. To this end, the final research question this work addresses is:

**RQ8: What is the impact of Levels of Evidence in a retrieval augmented generation (RAG) setting?**

### 1.3. Thesis Contribution

This thesis contributes to the field of medical information retrieval for medical experts by addressing the limitations of traditional keyword-based search methods and exploring the integration of multidimensional and semantic relevance aspects in medical search engines presented by WisPerMed, the search engine developed in the light of this work. The key contributions of this work are as follows:

- **Extracting Personalization Aspects from Documents:** This thesis focuses on developing methods to identify and extract personalization values from medical documents. It involves creating specialized datasets for categorizing medical subfields and assessing document difficulty, alongside evaluating whether zero-shot LLMs can outperform traditional fine-tuned methods in these tasks.
- **Integration of Bio-Concepts for Enhanced Precision:** By expanding traditional search queries with bio-concepts such as genes, diseases, and chemicals, this thesis explores how semantic relevance can be further improved. The work evaluates the impact of including these bio-concepts on the precision of retrieval, demonstrating their potential to make search results more relevant to the specific needs of healthcare professionals.
- **Credibility Assessment through Level of Evidence (LoE):** The thesis integrates the concept of Level of Evidence (LoE) into the medical search process, offering a framework for assessing the credibility of retrieved documents. By incorporating LoE as a filter in search engines, the research shows how this approach can improve both the efficiency and trustworthiness of information retrieval, supporting better decision-making in medical practice.
- **Development and Evaluation of WisPerMed:** As a practical outcome of this research, the thesis presents WisPerMed, a medical search engine that integrates the above advancements in semantic relevance, personalization, bio-concept expansion, and credibility assessment. The effectiveness of WisPerMed is evaluated through a user study involving medical professionals, highlighting its potential to enhance the efficiency and utility of medical literature search compared to existing tools like PubMed. We also present the first investigation of integrating WisPerMed into an LLMs-based conversational search engine using RAG architecture, with the aim of potentially improving generated responses.

Overall, this thesis makes significant contributions to the development of more sophisticated and user-centric medical search engines, addressing both the technical challenges and the practical needs of medical professionals in accessing relevant, reliable, and timely information.

## 1.4. Thesis Statement and Summary

This thesis argues that integration of semantic relevance, personalization, and credibility factors into medical search engines can significantly improve the effectiveness and efficiency of information retrieval for healthcare professionals. By moving beyond traditional keyword-based methods and embracing NLP techniques, bio-concepts expansion, and LoE framework, the thesis aims to address the medical practitioners' information needs not only when dealing with patients but also when conducting systematic reviews.

In response to the challenges faced by medical practitioners in retrieving relevant and credible medical literature, the thesis proposes the development and evaluation of a medical search engine, WisperMed, that considers the above factors. These are extracted using machine learning models and evaluated to enhance the retrieval process, ultimately improving the effectiveness and efficiency of medical information retrieval. Evaluation results demonstrate that incorporating these contextual factors not only enhances retrieval effectiveness but also significantly improves the efficiency of medical literature searching compared to the PubMed search engine. A user study involving 131 medical experts indicated a reduction in the number of queries and time to retrieve information, with strong effect sizes. Additionally, users praised the simplicity and clear visualization of the search engine components, including interactive elements and Nutri-score-like level of evidence indicators. This thesis contributes to the field by offering a medical search engine designed to meet the specific needs of medical practitioners, integrating contextual aspects to enhance both efficiency and utility.

The work presented in this thesis is organized as follows:

**Chapter 1 - Introduction** presents the motivation, context, high-level research questions, and the main contributions of the thesis.

**Chapter 2 - Related Work** reviews the existing literature on medical search engines and the different aspects integrated into the search engine.

**Chapter 3 - Document Difficulty** presents a method to personalize medical search engines by extracting readability and technicality level as difficulty aspects using a dataset annotated by medical experts.

**Chapter 4 - Field of Expertise** presents a method to personalize medical search engines by extracting the medical field discussed in the medical article.

**Chapter 5 - Level of Evidence** presents a method to extract the level of evidence from medical documents and evaluate its impact on retrieval efficiency.

**Chapter 6 - Bio-concepts** describes the impact of extending the search query with bio-concepts on the retrieval efficiency in an interactive setting.

**Chapter 7 - User evaluation** compares the search engine performance with PubMed

as a baseline in a user study with clinical practitioners.

**Chapter 8 - Level of Evidence based LLMs** investigates the integration of LoE into LLMs and examines the incorporation of WisPerMed within Retrieval-Augmented Generation.

**Chapter 9 - Discussion and Conclusion** discusses the findings, limitations and future work, and concludes the thesis.

## 1.5. Publications

This work is published in the following publications:

- 2023 Frihat, S., Beckmann, C. L., Hartmann, E. M., & Fuhr, N. (2023). Document Difficulty Aspects for Medical Practitioners: Enhancing Information Retrieval in Personalized Search Engines. *Applied Sciences*, 13(19).
- 2024 Frihat, S., & Fuhr, N. (2024). Enhancing Medical Literature Retrieval with Biomedical Concepts. Submitted to the *International Journal of Data Science and Analytics*
- 2024 Frihat, S., & Fuhr, N. (2024). Supporting Evidence-Based Medicine by Finding Both Relevant and Significant Works. arXiv preprint arXiv:2407.18383.
- 2021 Frihat, Sameh, and Norbert Fuhr. "TREC 2021 Clinical Trials Retrieval, Duisburg-Essen University submission." TREC. 2021.
- 2024 Frihat, S., & Fuhr, N. (2024). Enhancing Biomedical Literature Retrieval with Level of Evidence and Bio-Concepts: A Comparative User Study. Submitted to ACM/IEEE Joint Conference on Digital Libraries (JCDL) IEEE.

## 2. Related Work

This chapter reviews previous research essential to developing WisPerMed, a context-sensitive, personalized, medical search engine designed for efficient use by practitioners at the point of care. The chapter begins with a review of research in Information Retrieval (IR) in Section 2.1 and covers general recent developments in this research area, as well as multidimensionality issues and their relation to the evaluation of IR systems. We also address interactive IR. Section 2.2 reports on previous work in search within the healthcare domain and the use of general and specialized search engines. It outlines the state-of-the-art in the field that WisPerMed search engine directly contributes to. In Section 2.3 recent developments in Natural Language Processing (NLP) and Medical IR are discussed, in particular Large Language Models (LLMs) and Biomedical NLP methods, whose development has direct bearing on this work. Finally in Section 2.4 personalization issues in medical search engines are discussed.

### 2.1. Information Retrieval

Information retrieval (IR) is the science of finding relevant information from large and diverse datasets based on user queries [16, 36]. The main goal of an IR system is to satisfy the user's information needs efficiently and effectively [2]. To achieve this, the IR system should match the user's textual description of the needed information (query) with the stored representations of documents and produce a ranked list of documents that are relevant to the query [16]. Researchers identified the core issues of IR as information need, relevance, and evaluation [255].

Information need represents the starting point for someone seeking information. The need can be explicit, where the user articulates the textual questions, or implicit, where the user has a vague idea of what they need [162]. Representing information needs through textual queries can be challenging, particularly for implicit queries, often resulting in the retrieval of irrelevant information for the user [283]. A piece of retrieved information is considered relevant if the user perceives it to be valuable in terms of their personal information needs [163]. To find relevant information, researchers proposed several retrieval models and systematically evaluated their performance. Some earlier models were boolean models [136], vector space models [239], probabilistic models [74], and learning to rank [154].

Learning to Rank (LTR) is a central technique in IR systems, particularly in web search engines and now also applied in the medical domain, where the precision and relevance of retrieved information are of utmost importance [154]. LTR leverages machine learning algorithms to train models that can automatically rank documents or search results according to their relevance to a given query. This capability is crucial in biomedical contexts, where accurate and relevant information can significantly impact clinical decision-making and research outcomes.

LTR algorithms are designed to optimize the order of search results so that the most relevant documents appear at the top. These algorithms learn from various features, such as the content of the documents, metadata, user interaction data, number of clicks on a document, and other context features, to predict the relevance of each document to a particular query [154].

LTR approaches can be categorized into three main types: pointwise, pairwise, and listwise ranking. Pointwise ranking approaches treat the ranking problem as a regression or classification task where each document is scored individually [34]. In contrast, pairwise ranking models learn to differentiate between pairs of documents, predicting which document in a pair is more relevant to the query [34]. The listwise ranking, considered the most effective, takes into account the entire list of documents during training and optimizes the order of the entire list using metrics like Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), or NDCG.

More recent developments in the field of IR modelling and the rapid advancement in machine learning and neural networks led to the development of more capable models such as Word2Vec, BERT, and other transformer-based models [72]. These developments have significantly improved IR accuracy by capturing model complex semantic relationships in data [276]. Moreover, there has been a shift towards personalized and context-aware IR systems to leverage user behavior and preferences to search results, making significant contributions in the field of interactive information retrieval [153] and multidimensional relevance [194].

IR systems are now more integrated with different types of applications such as search engines, digital libraries, and enterprise search systems. This allows users to engage in dialogue with systems and refine their queries using conversational and interactive retrieval to improve search outcomes [56, 299]. The systems now evaluate relevance to users' queries beyond mere topicality, incorporating factors such as evidence, readability, and technicality. This approach renders the document multidimensional by tagging it with various dimensions in addition to the original textual content representing the topic.

Despite these considerable developments in the field of IR, evaluating the effectiveness of IR systems remains a complex challenge [164]. The challenge extends beyond developing new metrics and methodologies to better capture user satisfaction and relevance - they also

encompass ethical considerations, such as ensuring user privacy, mitigating bias in retrieval algorithms and promoting fairness and transparency [68, 305, 260].

### 2.1.1. Evaluating Search Systems

Evaluating search engines and information retrieval models is a multifaceted process that aims to ensure their effectiveness in retrieving results that meet users' information needs [133]. This process can be broadly categorized into offline and online evaluations, each employing different methods and metrics [42, 204].

#### Offline Evaluations

Offline evaluations often involve creating a test collection with three main components: a (very large) document collection, a sample of queries, and a set of relevance judgments (whether a document is relevant to the queries) [212]. Search systems are evaluated using pre-defined quantitative metrics such as precision, recall, F1 score, and more advanced metrics like NDCG (Normalized Discounted Cumulative Gain), and R-precision [212]. These metrics allow researchers to empirically determine the effectiveness of different retrieval models. Several open-source test collections are widely used in offline evaluations, such as TREC<sup>1</sup>, CLEF<sup>2</sup>, and NTCIR<sup>3</sup>, providing a standardized benchmarking environment [212].

#### Online Evaluations

Online evaluations often involve comparing how actual users use different versions of a search engine in real-time (known as A/B testing). The main goal of this method is to assess the impact of changes on user behaviour and satisfaction [223]. Depending on the goal of the study, several metrics can be included, such as click rate, task time, and bounce rate [42]. Feedback surveys usually follow user studies after completing the A/B testing [231]. This approach aims to ensure balancing between quantitative data with qualitative insights to optimize search engine performance [79, 304].

#### Other Considerations in IR Evaluations

Evaluating search engines can go beyond the online and offline evaluations by considering other aspects, such as user intents and query ambiguity, with the aim of ensuring the user's overall satisfaction [40, 113, 107]. Biases in search algorithms ensure fair and unbiased results

---

<sup>1</sup><https://trec.nist.gov/>

<sup>2</sup><http://www.clef-initiative.eu/>

<sup>3</sup><https://research.nii.ac.jp/ntcir/index-en.html>

while maintaining user privacy and transparency in data usage [180, 60, 28]. Scalability aims to simulate realistic search scenarios, especially in domains like healthcare, where the quality of evidence and specific biomedical concepts are important [39, 31, 178]. This is similar to the techniques integrated into the WisPerMed search engine, where the level of evidence (Chapter 6) and bio-concepts (Chapter 5) enhance the precision and relevance of search results.

### 2.1.2. Multidimensional Relevance

In the field of information retrieval, understanding and defining relevance has changed significantly over time. One of the first definitions of relevance was straightforward: a document is considered relevant to a query if its topic matches the topic requested [67]. This concept, known as topical relevance, was widely adopted due to its simplicity and clear definition [219]. Major IR evaluation tasks such as TREC and CLEF still rely on metrics that measure this type of relevance [27, 271].

As the field developed, it became clear that relevance can not be limited to just topicality. Rather, relevance is a multidimensional concept [194]. Imagine a medical expert searching for information on recent literature on skin cancer treatment. Although topical relevance is important, what if the user is an expert in medical pathology? This means that the oncology related documents, although topically relevant, are not for this particular case as pathology content is needed. Therefore, other aspects could be included in the multidimensional space of relevance, such as user knowledge, document timeline and availability [148, 265]. Schamber published a non-exhaustive list of 80 relevance criteria [218]. However, accounting for multidimensional relevance criteria makes the assessment of the IR systems more complex as the evaluation needs to include personal judgments and preferences across multiple criteria and consider the variability of users' needs.

To consider these multidimensional aspects of documents in information retrieval, researchers have proposed different frameworks. Fuhr et al. [76] suggested extending the document dimension with automatically generated information (similar to the nutritional label on food packaging). This label would help score documents on several dimensions as well as topical relevance, such as credibility, readability, actuality, and opinion. This would allow all users not only to assess the document's relevance quickly based on multiple aspects but also for all search engines to filter out results based on these aspects [83].

An additional critical dimension of relevance is task-based evaluation, which assesses how well the system helps users complete their specific tasks [227]. In medical information retrieval, the true measure of a system's effectiveness is not just in retrieving relevant documents but in supporting the clinician's decision-making process or aiding researchers in their studies. Task-based evaluation focuses on the end goal: how much the retrieved information



contributes to solving the user's problem or completing their task. For instance, a clinician looking for the latest treatment guidelines for a particular condition needs information that not only matches the query terms but is also up-to-date, credible, and practically applicable to patient care. Similarly, a researcher conducting a literature review requires documents that provide comprehensive coverage of the topic, high-quality evidence, and are relevant to their specific research questions. In these cases, task-based evaluation looks at whether the search results help the user make accurate clinical decisions, form a solid basis for research, or gain the necessary insights to proceed with their work.

Task-based evaluation is implemented by measuring the outcomes of using the information retrieval system in real-world scenarios as in Chapter 7 of this work. This can involve user studies where participants complete specific tasks using the system, and their performance is measured in terms of accuracy, efficiency, and satisfaction. For example, in a medical setting, task-based evaluation might involve scenarios where clinicians use the system to find information needed for diagnosis or treatment and then assess how this information impacts patient outcomes or clinical workflow efficiency.

This approach ensures that the information retrieval system is not only theoretically sound in terms of topical and multidimensional relevance but also practically useful in real-world applications. By focusing on the actual tasks users need to accomplish, task-based evaluation provides a more holistic understanding of a system's effectiveness, highlighting areas for improvement that might not be evident through traditional relevance metrics alone. This makes task-based evaluation a vital component in developing and refining information retrieval systems that truly meet the needs of their users, particularly in critical fields like medicine.

In summary, multidimensional document relevance in information retrieval recognizes that relevance is not just about matching topics but also involves understanding the user's context, the credibility and timeliness of information, and the ease with which users can understand the content. This comprehensive view helps create search engines that deliver more useful, reliable, and user-friendly results, ultimately enhancing the user's ability to achieve their specific objectives.

### 2.1.3. Interactive Information Retrieval

According to Kelly et al., the foundations of Interactive Information Retrieval can be traced back to diverse fields, such as traditional information retrieval, library and information sciences, psychology, and human-computer interaction [120]. This field primarily focuses on users' behaviors, tasks, and information needs rather than the system's requirements [294]. This allowed interactive information retrieval to serve as a bridge between system-oriented and user-oriented approaches, ensuring the IR system's effectivity for users [294].

Interactive information retrieval research considers aspects from both the user's and the system's perspectives [294]. For example, a researcher might conduct a study in which physicians use an interactive retrieval system to search for medically relevant literature on treatment options. The analysis could reveal insights into how physicians formulate their search queries, how they interact with the system's suggestions, and how the system's iterative feedback mechanism helps refine their search results to better match their clinical needs [290].

Interactive IR is considered as an area within the IR field that focuses on the interaction between the user and the retrieval systems. Unlike "traditional" IR systems, which focus on results returning static queries, interactive IR systems involve continuous dialogue with the user to improve search results [283]. The process aims to better understand the users' information needs and try to meet them, which can be complex [153].

Moreover, interactive IR systems usually integrate context-aware retrieval techniques that take into account the user's current context, including their search history, user profile, and the task at hand [153]. By leveraging contextual information, these systems can provide more relevant and personalized search results, enhancing the user's ability to find useful and relevant information efficiently.

In summary, Interactive Information Retrieval represents a significant development in IR by facilitating the connection between the user and the system. This approach not only enhances the retrieval process through continuous feedback and interaction but also tailors search results to better meet the user's dynamic information needs [294].

## 2.2. Search in the Health Domain

Numerous models have been developed in the field of information retrieval to describe different search behaviours, such as the underlying causes, information-seeking activities, and the intricate relationships underpinning these search behaviours. Search engines in the health domain can be investigated based on the type of users and the specific needs they address. Various search engines are specifically designed to cater to the distinct needs of different user groups, whether they are health consumers (laypeople and the general public) or health professionals (physicians, nurses, and researchers).

### 2.2.1. Health Search using General Search Engines

General search engines like Google, Bing, and Yahoo are designed to serve the general public in finding information and are commonly used by health consumers to find health-related information. These search engines are easy to use, providing a broad range of information

from different sources like news articles, blogs, medical websites, and forums.

Toms and Latter observed 48 users searching for four health-related topics using Google. They used logs, video screen capture, verbal protocols, and self-reported questionnaires to study user behavior. Their results indicated significant problems in query formulation (on average 4.2 keywords were used per query, but out of those, 3.2 were stopwords and thus not processed by the search engine) and in making efficient selections from result lists [261].

Several studies have studied aspects that can influence the search and improve it [244, 261, 284, 101]. Out of these, Hersh suggested two important aspects that should be considered to improve the user's satisfaction and search efficiency: readability and trustworthiness [101]. Several researchers proposed techniques to integrate the document understandability into account for better personalization in general search engines [256, 44, 125, 45, 238, 155]. Initial efforts to incorporate document understandability into search engines have shown promising results. For example, researchers at Yahoo developed a system that personalized search content based on user familiarity with a topic, using a classifier trained on articles from Simple Wikipedia and standard Wikipedia [256]. This approach improved content ranking by tailoring results to the user's reading level. Another important study by Wang et al. compared the quality of the results from the major search engines (Google, Yahoo!, Bing, and Ask.com) for the query "breast cancer" [278]. They concluded that the search engines provided high overlap between the results, with more than 50% overlap between any two search engines. Thus, if any search engine provides poor quality results, it is likely that results will also be poor for other search engines. Therefore, there is a clear need to consider the understandability and trustworthiness of general search engines with respect to health-related content search.

### 2.2.2. Health Search using Specialized Search Engines

Specialized search engines such as OpenMD<sup>4</sup>, Embase<sup>5</sup>, PsycINFO<sup>6</sup>, and PubMed<sup>7</sup> (see Table 2-1 for a comprehensive list) are designed specifically for health experts' needs and often provide access to more reliable resources, which is critical for medical decision-making and research.

Out of these medical search engines for health experts, PubMed stands out it is considered the most popular biomedical search engine, managing the largest biomedical database in the world (Medline) [102]. Medline includes 30+ million abstracts, from which 10 million available with full text. It is also considered the base for several biomedical IR collections

---

<sup>4</sup><https://openmd.com/>

<sup>5</sup><https://www.embase.com/>

<sup>6</sup><https://psycnet.apa.org>

<sup>7</sup><https://pubmed.ncbi.nlm.nih.gov/>

Table 2-1.: List of current search engines for medical experts and brief description.

Search Engine	Goal
PubMed	General-purpose biomedical literature search engine on abstracts with 30+ million articles, including 10+ million open access articles [188].
LitVar	Searching relevant information for all synonyms to a given variant [187].
PICO	Search engine which retrieves articles for a given “Medical condition”, “Intervention”, “Compare to”, and “Outcome” [30].
DiGSeE	Text mining search engine which provides evidence sentences for a given (“genes”, “disease”, “biological events”) triplet [124].
OncoSearch	Text mining search engine which provides sentences describing gene expression changes in cancers [137].
LitSuggest	Search engine which finds and recommends biomedical literature based on topic of interest [5].
PubTator	Search engine which highlights bio-concepts [281, 280].
COVID-19 Challenges & Directions	Search engine for exploring medical terms mentioned in potential research directions or scientific problems [135].

dealing with scientific publications, including TREC PM<sup>8</sup>, TREC Genomics<sup>9</sup>, BioASQ<sup>10</sup>, iSearch [159], and OHSUMED<sup>11</sup>.

Dogan et al. [110] analysed PubMed log data for one whole month. Their main finding is that PubMed users are more likely to reformulate the search query rather than investigate more documents on the results page. The most frequent types of search are search by name, search by gene/protein, and search by disease [110]. Also, the use of known abbreviations in queries is very frequent. These aspects are discussed in more detail in Chapter 5.

The impact of experience on search behavior has been evident since the early use of specialized search engines [23, 106]. White et al. performed a log-based analysis across several domains, including medicine, to develop a field expertise classifier based on user behavior. They showed that experts had higher success rates within their domains of expertise but not outside them, highlighting the domain-specific nature of search expertise [284].

In summary, while the public widely uses general search engines for health information, specialized search engines offer targeted, reliable resources for health experts. Both types of search engines play important role in the field of health IR, capturing different users needs.

<sup>8</sup><https://trec.nist.gov/data/precmed.html>

<sup>9</sup><https://trec.nist.gov/data/genomics.html>

<sup>10</sup><http://bioasq.org/>

<sup>11</sup><https://paperswithcode.com/dataset/ohsumed>

### 2.2.3. Medical Search Engines Challenges

The medical domain presents unique challenges for search engines which are not present in many other fields.

- **Rapidly Evolving Knowledge:** Medical knowledge and guidelines are constantly updated with new research, clinical trials, and evolving best practices, requiring advanced search engines to continuously incorporate the latest and relevant information [301, 168].
- **High Stakes of Accuracy:** Inaccurate or outdated information can have serious consequences, including misdiagnosis, incorrect treatment, and adverse patient outcomes, making accuracy and reliability paramount [116].
- **Diverse User Base:** Medical search engines cater to a wide range of users, including patients, general practitioners, specialists, researchers, and students, each with different information needs and levels of expertise [48].
- **Language and Jargon:** Medical terminology is highly specialized and often includes complex jargon and abbreviations that can be difficult for non-specialists to understand, necessitating sophisticated natural language processing capabilities [111, 35, 234].
- **Multidisciplinary Nature:** Medicine encompasses various subfields such as cardiology, oncology, neurology, etc. requiring search engines to understand and differentiate between these diverse areas [46].
- **Evidence-Based Information:** Medical professionals rely on evidence-based information, including clinical guidelines, systematic reviews, and meta-analyses, which must be accurately indexed and prioritized [171].
- **Contextual Relevance:** The relevance of medical information can depend heavily on context, such as patient age, medical history, and specific symptoms, requiring advanced algorithms to deliver contextually appropriate results. This makes formulating the right query complex when searching for unstructured information [247, 246].
- **Complex Queries:** Medical queries often involve complex, multi-faceted questions that require search engines to understand and process detailed and nuanced requests [95].
- **Ethical Considerations:** Medical information is subject to strict regulations and ethical guidelines, particularly regarding patient confidentiality, data security, and the dissemination of treatment recommendations. This could make most relevant data sources and applications private and not open source [119, 3].
- **Bias and Credibility:** Ensuring the credibility and unbiased nature of medical information is critical, as biased or commercial content can lead to misinformation and

potential harm [250].

These factors make developing effective medical search engines a complex and challenging task. They require specialized approaches and continuous updates to meet the medical community's high standards. In the following chapters, we discuss the “Language and Jargon” and the “Multidisciplinary Nature” in Chapter 3 and Chapter 4, respectively, by integrating these aspects into personalization. We also tackled the “Evidence-Based Information” by integrating the LoE framework into biomedical IR in Chapter 6. Finally, we considered the “Complex Queries” using the integration of bio-concepts in Chapter 5.

## 2.3. Natural Language Processing in Medical IR

Natural Language Processing (NLP) has revolutionized the field of information retrieval, including the medical domain. The unique contribution of NLP to understanding, interpreting, and generating human language makes it an invaluable tool for enhancing the retrieval and analysis of biomedical information. This section explores the advancements in NLP, focusing on Biomedical NLP (BioNLP), which aims to bridge the gap between biomedical text and knowledge, with a detailed discussion on PubMed-BERT and Large Language Models (LLMs) and their applications in medical information retrieval.

### 2.3.1. Biomedical NLP

This specialized area within NLP deals with processing and analyzing text in the biomedical and clinical domains. BioNLP's primary goal is to convert unstructured biomedical texts into structured and actionable knowledge. This is achieved by several key tasks, such as Named Entity Recognition (NER), relation extraction, and document classification, which are considered the most popular applications in BioNLP.

BioNLP techniques are used to identify and classify key biological terms such as genes, diseases, mutations, proteins, and chemicals within texts as presented in Chapter 5. For instance, in a sentence discussing the effects of a new drug on cancer cells targeting protein, NER algorithms can accurately identify and categorize “drug”, “cancer cells”, “protein”, and other relevant terms. On the other hand, relation extraction applications focus on understanding the relationships between these terms. For example, relation extraction techniques can identify and categorize the interaction between the drug and the protein, which is critical for understanding biomedical research literature. Document classification focuses on classifying and categorizing biomedical documents based on their content, such as document readability in Chapter 3, document level of evidence in Chapter 6, and medical subfields in Chapter 4.

One significant advancement in NLP that has impacted BioNLP is the development of BERT (Bidirectional Encoder Representations from Transformers). Before BERT, models like Word2Vec and GloVe (Global Vectors for Word Representation) laid the groundwork by representing words as vectors in a continuous vector space, capturing semantic relationships between words based on their context in large text corpora. Word2Vec, introduced by Mikolov et al. [175], enabled efficient computation of word embeddings through skip-gram and continuous bag-of-words models, while GloVe, developed by Pennington et al. [200], leveraged global word-word co-occurrence statistics to produce word vectors.

BERT has set a new standard for NLP tasks by providing a framework for understanding the context and relationships within texts. BERT is a transformer-based model developed by Google designed to pre-train deep bidirectional representations by joint conditioning on both left and right contexts in all layers. This allows BERT to achieve state-of-the-art results on a wide range of NLP tasks.

Unlike traditional models that read and process text sequentially, BERT processes entire sequences of words simultaneously, allowing it to understand a word's context based on its surroundings. Moreover, as BERT is pre-trained on a huge amount of text data from several fields, it allows it to capture a wide range of linguistic information. Finally, fine-tuning the model for different downstream tasks such as NER, question answering, and text classification in various domains such as biomedical, chemistry, and others makes it highly adaptable.

The introduction of BERT has revolutionized NLP by significantly improving the performance of models on a variety of tasks. Its ability to understand context and relationships within text has paved the way for more sophisticated and accurate models in many domains, including biomedicine.

### 2.3.2. PubMed-BERT: Enhancing Biomedical Text Understanding

Building on the success of BERT in analyzing and understanding the text, PubMed-BERT [86] is a domain-specific variant pre-trained on a large corpus of biomedical literature from PubMed, as explained in Section 2.2.2. This enables PubMed-BERT to handle the unique language and terminology of the biomedical field with greater accuracy. The motivation behind using PubMed data is to leverage the vast, specialized vocabulary and context found within millions of biomedical abstracts and articles, ensuring the model is well-equipped to process and understand complex biomedical texts.

PubMed-BERT's specialized vocabulary, derived exclusively from biomedical texts, ensures better representation and understanding of domain-specific terms. This targeted pre-training significantly enhances performance on various biomedical NLP tasks such as NER, relation extraction, and question answering, setting new benchmarks in these applications.

Unlike models pre-trained on mixed-domain corpora, PubMed-BERT's in-domain training allows it to interpret and process biomedical data more effectively.

The development and implementation of PubMed-BERT have led to substantial improvements in several fields of biomedicine, including medical search engines. These enhancements enable search engines to interpret and respond to complex biomedical queries with higher relevance and precision. By more effectively understanding the context and semantics of user queries, PubMed-BERT supports medical professionals by providing accurate and contextually relevant information, thereby facilitating better clinical decision-making processes.

### 2.3.3. Large Language Models (LLMs) in Biomedical NLP

While PubMed-BERT has set a new standard for domain-specific language models, the advent of LLMs has further revolutionized the field of natural language processing, including biomedical NLP. LLMs, such as GPT-3 and beyond, have demonstrated remarkable capabilities in understanding and generating human language, offering new possibilities for processing and analyzing biomedical texts on an even larger scale.

LLMs are trained on extensive datasets comprising billions of words from diverse sources, including articles, books, and internet content. This extensive training allows LLMs to develop deep associative relationships between words, enabling them to perform complex language tasks with minimal or no specific fine-tuning. In healthcare, LLMs like ChatGPT have shown potential by passing medical licensing exams and providing quality responses to medical queries. These models' ability to understand and generate text with near-human accuracy makes them valuable tools for clinical, educational, and research applications in medicine [259].

However, the use of LLMs in medicine is not without limitations. One significant concern is confabulation (the accuracy of the information they generate). LLMs are trained on vast datasets that may contain inaccuracies or outdated information, leading to the generation of incorrect or misleading responses. This is particularly critical in the medical field, where inaccurate information can have serious consequences. Moreover, LLMs often lack transparency, making it difficult to understand how they arrive at specific conclusions, which complicates trust and accountability.

Another limitation is the ethical concern related to the deployment of LLMs in clinical settings. Issues such as data privacy, bias in training data, and the potential for misuse of AI-generated information pose significant challenges. Ensuring that LLMs do not perpetuate biases present in their training data is crucial to avoid reinforcing health disparities. Additionally, there is a need for robust governance and oversight to manage the ethical implications of using AI in healthcare.

Despite these limitations, the integration of LLMs in healthcare settings holds great



promise. They can enhance the efficiency of clinical workflows, support medical education through interactive tutoring systems, and aid in research by automating data analysis and summarization tasks. The ongoing development and refinement of LLMs promise to bring even greater advancements to the field of biomedical NLP, paving the way for more sophisticated and accurate applications while addressing the associated challenges.

## 2.4. Personalization Aspects in Medical Search Engines

Personalization in medical search engines can be achieved by modelling various user-related factors that can enhance the retrieval of information for medical experts. These factors include document difficulty estimation addressed in Chapter 3 and identifying a medical subfield a document addresses as in Chapter 4. Other important aspects include:

- **Knowledge Levels:** adapting search results based on the user's knowledge level ensures that the information provided is neither too simplistic nor overly complex. For example, medical students may require more foundational and explanatory content, while experienced specialists might need advanced and detailed information [114]. This customization helps in making the search experience more efficient and relevant, allowing users to access information that matches their expertise.
- **Field of Interests:** Personalizing content based on the user's specific interests within the medical field can significantly improve the relevance of the information retrieved. For instance, a user specializing in oncology would benefit from search results that prioritize cancer research, treatments, and recent advancements in the field [97]. This aspect ensures that users receive updates and relevant information aligned with their professional focus.
- **Age:** Incorporating the user's age can help adapt the search results to provide information that is more relevant to their stage in career and life. For example, younger experts might seek more educational content and recent research developments, whereas older, more experienced experts may prioritize practical applications and advanced medical techniques [230].
- **Languages:** Providing information in the user's preferred language is crucial for comprehension and application. Multilingual support in search engines ensures that users receive information in the language they are most comfortable with, which is particularly important in diverse, multicultural medical settings. This can improve the accuracy and usability of the information retrieved [300].
- **User Preferences and Behavior:** Analyzing user behavior and preferences, such as frequently accessed types of documents (e.g., clinical trials, review articles) and preferred sources, can personalize the search experience. This helps in prioritizing results

that match the user's established patterns and preferences, enhancing the relevance and efficiency of information retrieval [117].

These personalization aspects in medical search engines can significantly enhance the efficiency and accuracy of information retrieval. Considering multiple personalization aspects allows search engines to provide results that are more relevant to the user's needs and context. This ensures that medical professionals receive information that is both accurate and directly applicable to their area of practice [61]. Integrating personalization in search engines reduces the time and effort required to find relevant information, which is essential at the point of care [300].

By integrating user profiles that include detailed information about the user's speciality, role, clinical context, and patient-specific details, search engines can deliver contextually relevant results [117]. It also supports complex queries, which are often complex and multifaceted. Personalized search engines can better handle these queries by understanding the user's specific needs and providing comprehensive, relevant results that address multiple aspects of the query [97].

While personalization in medical search engines offers significant benefits, it also presents challenges, such as ensuring the accuracy and completeness of user profiles, which is critical for effective personalization. This requires continuous updating and refinement based on user behavior and feedback. Several studies demonstrated the effectiveness of analyzing search behavior and logs to build automatic user profiles. However, extracting the aspect values on the document level remains an open question in the medical field due to the limited labeled datasets. Moreover, training a learning-to-rank algorithm that handles all these personalization aspects is considered a challenge due to the extensive annotated data required, as explained in Section 2.1.

By addressing these challenges, personalized medical search engines can significantly improve the quality and relevance of information retrieval, ultimately enhancing decision-making processes for medical professionals and improving patient outcomes. In Chapter 3 and Chapter 4, we present the importance of information difficulty and field of expertise, respectively, as well as automatic methods to estimate these in medical documents. We focus on document difficulty estimation and on incorporating users' field of interest by identifying medical subfields to which medical documents belong. These two personalisation methods are foundational for effective medical search engine personalization. All the remaining personalization aspects lie out of scope of this study and remain an area for future work.

## 2.5. Summary

This chapter reviewed previous research relevant to the concepts used in this thesis. In Section 2.1, we overviewed the core concepts of the information retrieval field used in this thesis. In particular, we introduced the methods for evaluating search systems and the concept of multidimensional document retrieval, which is important to understand how contextual aspects can be integrated into IR systems. Also, we presented the field of interactive information retrieval, which refers to the user aspect of IR systems. Section 2.2 introduced the search engines for medical consumers and experts in medical search engines. It also presented the medical search engine challenges. Section 2.3 focused on introducing the NLP methods and models used in this thesis and presenting the importance of NLP techniques in improving medical IR. Finally, Section 2.4 explained the personalization aspects in medical search engines for experts. It also presented the application of personalization techniques and expected challenges.



# 3. Document Difficulty Estimation for Medical Search Engine Personalization

As detailed in the previous chapters, personalized search engines adapt results based on user profiles, created from manual configurations or by analyzing logs automatically. Such consideration of document features enhances the contextual relevance and accuracy of information retrieval. This chapter develops methods to estimate document difficulty, specifically the readability and technicality, of medical publications for personalized medical search engines. It covers data collection and labeling by medical experts, a transformer-based model for readability and technicality prediction, and a comparison with traditional methods. In this chapter we investigate the following research questions:

- Do users agree on readability and technicality of a document?
- Are there differences in readability and technicality between German and English texts?
- What are good methods for estimating readability and technicality of a text?

## 3.1. Introduction

As argued in the previous chapters, considering users' characteristics and search behaviour by personalizing search engines is crucial to building search systems that deliver relevant and useful results to the individual users [61, 300]. In general, search engine personalization involves creating user profiles using manual configuration or automatically by analyzing user interactions and search logs [114, 97, 230]. Personalization is a well studied topic in medical IR research. However, while previous studies [156, 81, 189] have primarily focused on developing personalized search engines for health information consumers, such as laypeople and patients, there is a clear gap in adapting to the specific requirements of medical practitioners. Unlike laypeople, medical practitioners possess specialized expertise and language proficiency in their respective fields. In the healthcare domain, documents range from highly technical and specialized articles to more general and easily readable texts. Addressing this range of document difficulty levels is essential for creating search engines such as WisPerMed

that meet the specific needs of medical practitioners at the point of care. To do so efficiently, it is necessary to clearly define first what the term document difficulty refers to.

Document difficulty refers to the ease of comprehension of the information presented to a user. Previous research has discussed *readability* and *technicality* as criteria that define document difficulty. Readability primarily pertains to text understandability, while technicality is more focused on the concepts and domain-specific knowledge within the text [272, 37, 96]. Research conducted by Entin and Klare [63] revealed a significant influence of readability of a text among other factors on users' comprehension of the material. In the health domain, Hedman [98] proposed to use the readability for accepting or revising health-related documents.

It is important to note that readability and technicality are not mutually exclusive aspects [37]. Texts can exhibit varying degrees of both characteristics, leading to different combinations that may arise between these aspects. For instance, a text can be easy to read while still containing high technicality, or it can be difficult to read with low technicality, as shown in the following examples<sup>1</sup>.

- **Easy to read and high technicality:**

Autologous hematopoietic stem cell transplantation has emerged as a promising therapeutic intervention for individuals with refractory multiple myeloma. This treatment approach has shown remarkable advancements in terms of progression-free survival and overall response rates, signifying its potential in improving patient outcomes.

- **Hard to read and high technicality:**

The pathophysiological mechanisms underlying idiopathic pulmonary fibrosis involve aberrant activation of transforming growth factor-beta signaling pathways, leading to excessive deposition of extracellular matrix components and subsequent progressive scarring of lung tissue.

- **Easy to read and low technicality:**

Regular physical exercise has been widely recognized as a key lifestyle intervention for the prevention of cardiovascular diseases, with numerous studies demonstrating its positive impact on reducing the risk of heart attacks, stroke, and hypertension.

- **Hard to read and low technicality:**

Carcinogenesis is a multifactorial process characterized by the dysregulation of cellular homeostasis, involving intricate interactions between oncogenes and tumor suppressor genes that disrupt normal cell growth control mechanisms, resulting in uncontrolled proliferation and the formation of malignant tumors.

Considering the readability and technicality of documents in medical search engines in-

---

<sup>1</sup>All examples have been reviewed and approved by a senior medical practitioner. Examples of easy-to-read content are from [268], while examples of harder-to-read content are from [293].

creases the likelihood of tailored, relevant search results based on medical practitioners' expertise and language proficiency. Without such customization, search results may include highly technical papers or articles with varying readability, requiring manual sifting and wasting valuable time. The lack of customization based on technicality and ease of reading hinders precision, relevance, and quick access to necessary information. Furthermore, complex language and dense scientific jargon can impede comprehension for practitioners without specialized expertise, hindering decision-making [190].

Integrating readability and technicality into the search engine can be achieved in two ways: (1) by **filtering** out search results that exceed a certain threshold of readability or technicality or by incorporating readability and technicality scores into user profiles ensuring that the retrieved documents are filtered according to the user's preferred level of comprehension; or (2) by **ranking** documents higher in search results, if they match the desired readability and technicality criteria as indicated by a relevance score calculated using these criteria [36].

By incorporating document difficulty aspects, search engines improve information accessibility, enhance comprehension, and optimize decision-making, thus facilitating the efficient use of practitioners' time and expertise [190, 18]. These are the reasons why document difficulty estimation is a highly researched topic in IR. The following section summarizes relevant previous work on modelling document difficulty.

## 3.2. Related Work

Considering document difficulty criteria in IR, particularly in domain-specific contexts, has received significant attention. Researchers have recognized the challenges faced by both domain experts and average users when searching for domain-specific information, such as medical and health-related content, from online resources [226].

A common issue encountered by users in IR systems is the presence of search results that encompass documents with varying levels of readability [296, 112, 192]. This poses a challenge, particularly for users with limited domain knowledge or lower education levels, as well as those facing physical, psychological, or emotional stress [296, 236]. Consequently, there is a need for IR systems that not only retrieve relevant documents but also prioritize those with higher readability, adapting to the diverse needs of users [240].

A large body of literature proposed techniques for personalizing general search engines by taking document difficulty into account [256, 44, 125, 45, 238, 155]. In addition, Becker reported that most health content is not well designed for the elderly [19], and Stossel et al. found health education content readability level is not at accessible level [248]. Similar findings of these studies confirmed by others on the content of the top-ranked documents by

search engines [70, 62, 13, 82, 173, 198, 287].

To address this challenge, IR and linguistics researchers explored different approaches to develop models for automatically estimating text readability, such as [118, 257]. These approaches can be divided into two categories: (1) readability formulas and (2) machine learning. The importance of domain-specific readability computation in IR has been emphasized [158]. By integrating concept-based readability and domain-specific knowledge into the search process, researchers aim to enhance the accessibility and relevance of search results. These efforts contribute to empowering users, including both domain experts and average users, with efficient and reliable IR tools [296].

### 3.2.1. Readability Formulas

Since the early 20th century, researchers in this field have developed a variety of readability formulas aimed at laypeople [59]. Many of these formulas are still widely used today [77]. Dubay et al. listed the most widely used formulas, such as the Simple Measure of Gobbledygook (SMOG), the Dale-Chall Readability formula, the Flesch Reading Ease formula, the Fog Index, and the Fry Readability Graph.

These formulas typically analyze syntactic complexity and semantic difficulty. Syntactic complexity is often evaluated by examining sentence length, while semantic difficulty is measured using factors such as syllable count or word frequency lists. Other factors that have been found to influence readability include the presence of prepositional phrases, the use of personal pronouns, and the number of indeterminate clauses. However, it is important to note that readability formulas specifically targeting medical practitioners are currently lacking. Further research is needed to develop readability formulas tailored to the unique needs and expertise of medical professionals.

#### Simple Measure of Gobbledygook (SMOG)

The SMOG Index was introduced by clinical psychologist G. Harry McLaughlin in 1969 [169]. It is designed to estimate the years of education required to comprehend a piece of written text accurately by counting the words of three or more syllables in three ten-sentence samples. The formula calculates the reading grade level based on a simple mathematical equation that incorporates the count of polysyllabic words within a sample text.

The SMOG formula has been widely used in various fields, including education, health-care, and IR [277]. Its simplicity and ease of application make it a popular choice for estimating readability levels. However, it is important to note that the SMOG formula may have limitations when applied to specific domains, such as technical or scientific texts, as it does not consider the domain-specific terminology and nuances that might impact compre-



hension [288].

Nonetheless, SMOG stands out as a well-suited formula for healthcare applications. It consistently aligns results with expected comprehension levels, employs validation criteria, and maintains simplicity in its application [277]. These factors make it a reliable choice for assessing the readability of healthcare-related documents. Researchers frequently use the SMOG formula as a reference point when evaluating alternative readability models or proposing new formulas tailored to specific domains. However, it remains not designed to tackle the text's technical difficulties nor designed for domain experts.

### **Dale-Chall Readability Formula**

The Dale-Chall Readability Formula is a widely used readability measure that provides an estimate of the comprehension difficulty of a given text. Developed by Edgar Dale and Jeanne Chall in 1948 [49], this formula takes into account both the length of sentences and the familiarity of words to determine the readability level.

The Dale-Chall readability formula calculates its final score by examining the proportion of words in a given text that do not belong to a predefined list of commonly known words. This list comprises 3000 words that fourth-grade students generally understand. The formula calculates the readability score by incorporating the average sentence length and the percentage of unfamiliar words.

The Dale-Chall formula's notable advantage lies in its focus on word familiarity, enhancing its ability to assess readability, especially for less experienced readers or those with limited vocabulary. This practical and accessible approach considers both word familiarity and sentence length, offering insights into comprehension difficulty for readers of different proficiency levels. However, it is crucial to acknowledge the formula's limitations and its suitability for specific contexts.

### **3.2.2. Machine Learning Approaches**

Machine learning approaches have a more precise assessment of text complexity compared with traditional formulas [238]. One of the first models, developed by Si and Callan [238], used word frequency from a unigram language model and sentence length distribution, providing a more accurate prediction than one of the components alone. However, this task required a substantial corpus of training documents for accurate predictions.

Building on this foundation, researchers have incorporated a wider array of features to improve readability prediction models. For instance, Zeng et al. integrated the concept of consumer health vocabulary (CHV) [302]. The CHV maps consumer vocabulary to technical terms and assigns difficulty scores to concepts. This approach is useful for bridging the gap

between consumer terms (relatively easier to understand) and specialised medical terms (relatively harder to understand).

Further studies expanded the feature sets used in machine learning models to include more syntactic (such as Part-of-Speech tags and word length) and semantic attributes (such as Named-Entities and word frequency) with the aim of capturing more contextual information about the words in the text [123]. However, more recent developments in natural language processing (NLP) have further enhanced the capability of machine learning models for readability prediction. Techniques such as transformer-based models (e.g., BERT, GPT) have been employed to capture deep contextual relationships within texts, providing more accurate readability assessments [55]. These models can be fine-tuned on domain-specific corpora to improve their performance in specialised areas, such as medical or legal texts. This technique is used in this work to predict the readability and technicality level of medical scientific abstracts.

Integrating readability and technicality into the search engine can be achieved in one of the following three ways: (1) by filtering out search results that exceed a certain threshold of readability or technicality; (2) by incorporating readability and technicality scores into user profiles ensuring that the retrieved documents align with the user's preferred level of comprehension; or (3) by ranking documents higher in search results, if they match the desired readability and technicality criteria as indicated by a relevance score calculated using these criteria [36].

### 3.3. Materials and Methods

We leverage the power of pre-trained language models and fine-tuning techniques to predict the difficulty aspects, i.e. the readability and technicality, of a given document. Generally, most commonly used traditional readability formulas were not developed for technical materials [127] and are oversimplified to deal with specialized data [296]. Nevertheless, readability formulas for medical literature have been developed and used [292, 66]. Therefore, we evaluate the performance of our machine learning approach against these selected readability formulas and can compare both approaches to document difficulty estimation. Specifically we evaluate the capability of language models to capture and classify the readability and technicality levels of medical documents. We also investigate the differences in language complexity and technicality between English and German medical abstracts.

Our process begins with creating a dataset by extracting PubMed articles and obtaining readability and technicality annotations from medical experts. We then analyzed the dataset's general characteristics, difficulty levels, and language variations. Finally, we used this data to refine our pretrained BERT model.

### 3.3.1. Data Collection

We encountered a challenge in finding datasets specifically designed for medical practitioners, as most existing datasets target laypeople. Therefore, we compiled a dataset of medical abstracts from PubMed and sought the expertise of medical doctors and medical students to annotate each article with readability and technicality scores.

The dataset creation process started with the extraction of 10,000 articles from PubMed, specifically targeting those with available abstracts in both German and English. The data were then stored in a MongoDB and afterwards extended with information about the readability and technicality of those abstracts. This was completed in an annotation process by 216 medical students and practitioners using an annotation tool developed specifically for this purpose. In this process, a total number of 209 annotated articles could be gathered. An overview of the data acquisition is shown in Figure 3-1.

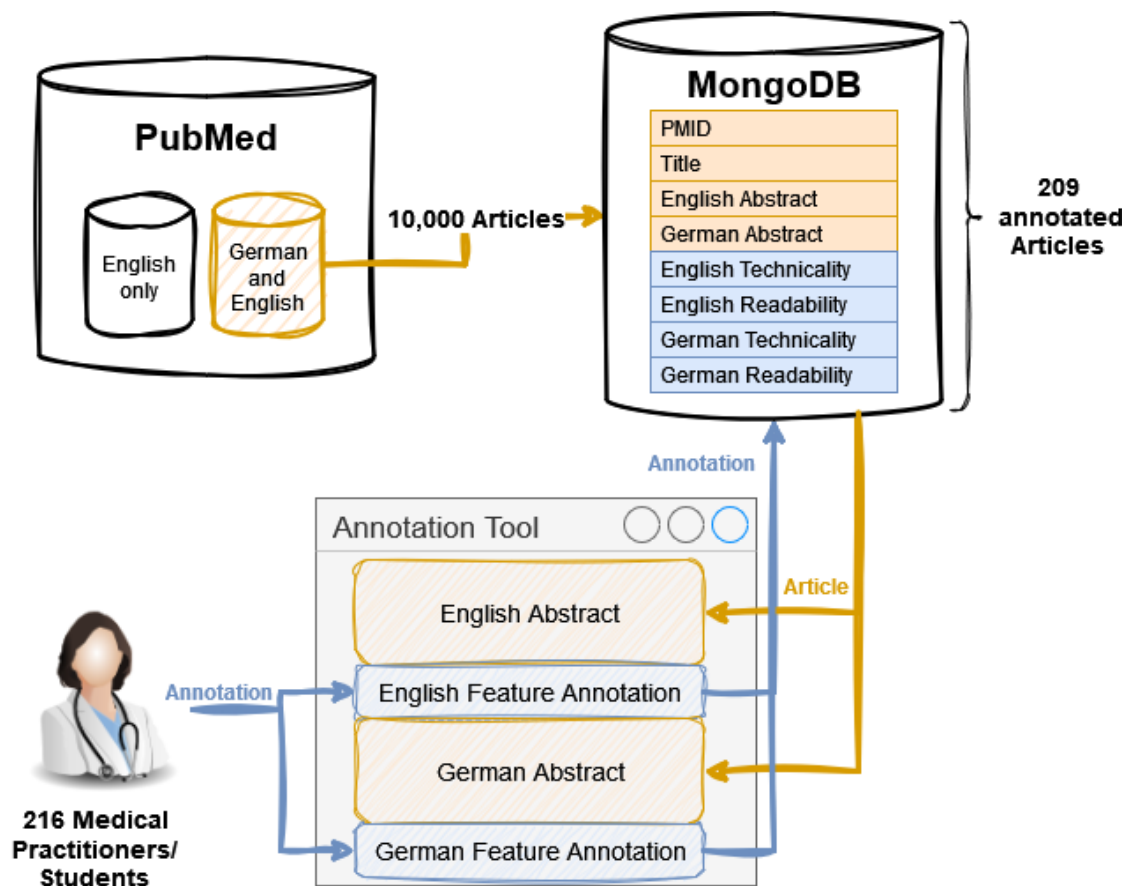


Figure 3-1.: The dataset creation process includes extracting articles from PubMed and expertly annotating abstracts for technicality and readability assessment.

Table 3-1.: Annotation process participants' distribution among four categories.

Category	Med. Student		Med. Doctor		Total
	Junior	Senior	Junior	Senior	
Participants	70	59	59	28	216
	32.4%	27.3%	27.3%	13%	100%

## Documents

We chose to focus on research articles' abstracts as they serve as concise summaries of the main research findings and are widely utilized for initial screening and IR purposes. In light of this, the PubMed database proves to be an ideal resource for our specific use case. PubMed consists of an extensive collection of scientific literature spanning various medical disciplines, making it a comprehensive repository of valuable medical information. To ensure a manageable dataset, we opted to download a subset of articles from the vast PubMed database. As selection criteria, we chose 10,000 random articles that had abstracts available and were written in both German and English.

## Participants

A total of 216 participants were recruited from university hospitals for the annotation process. Those can be divided into four categories, as shown in Table 3-1: 70 medical students up to the 6th semester (junior students), 59 medical students in the 7th semester or higher (senior students), 59 medical doctors with up to 2 years of experience (junior doctor), and 28 doctors with more than 2 years of experience (senior doctor). All participants either studied at a German university or worked in a German hospital, ensuring they possessed the necessary domain knowledge. Additionally, all participants were asked to assess their German and English language proficiency based on the Common European Framework of Reference (CEFR)<sup>2</sup>, ranging from B1 (Intermediate level) to C2 (Native level). Furthermore, in accordance with institutional regulations and to maintain complete anonymity, no further questions were asked.

## Annotation

In the initial phase of the study, we developed a web-based application using Python and MongoDB to facilitate the evaluation process. This application allowed participants to log in

<sup>2</sup>CEFR is a European standard for describing language ability. It describes language ability on a six-point scale, from A1 for beginners, up to C2 for those who have mastered a language.

anonymously, access clear guidelines, and review the criteria for rating abstracts' readability and technicality.

Each participant assessed 2–3 different abstracts, rating them for ease of reading and technicality on a scale from 0 to 100 in 5-point increments (0, 5, 10, ..., 100). They also identified relevant medical disciplines addressed in the abstracts.

Participants had no time constraints, providing flexibility in completing the evaluation. The average annotation time for each document was 176 (Std=73) seconds, highlighting variability in annotation durations. The annotation process was also conducted at an average rate of 113 (Std=12) words per minute (WPM), emphasizing diverse annotation speeds, which is consistent with Klatt et al.'s study [128].

To ensure reliability and minimize bias, three independent participants evaluated each abstract, and their scores were averaged to represent text complexity and technicality fairly. Annotations were conducted independently, enhancing the validity of ratings with respect to reliability and consistency.

## Dataset Overview

The annotation process proved to be a resource-intensive task, primarily due to the challenges associated with securing time from busy medical professionals, including both practicing physicians and medical students. With their commitments ranging from extended working hours and on-call duties to direct patient care responsibilities and rigorous exam preparation, allocating time for annotation was constrained. As a result, only 209 abstracts were annotated for technicality and readability aspects. Each annotated abstract included both English and German versions, ensuring comprehensive coverage of the research literature across languages. The dataset encompassed various medical disciplines, creating a representative subset of medical research publications.

### 3.3.2. The Model of Document Difficulty

Using this dataset for training we developed regression models that estimate the readability and technicality of medical abstracts/documents. We split the dataset into 70% for the training set and 30% for testing.

To establish a baseline for our dataset, we employed pretrained BERT models designed for medical text, "PubMedBERT" for English abstracts [86] and "German-MedBERT" for German abstracts [51]. BERT, known for its impressive performance in various Natural Language Processing (NLP) tasks, was a fitting choice for our project.

The fine-tuning process for the pretrained BERT models, specifically "BERT-Readability"

for assessing ease of reading and “BERT-Technicality” for evaluating technicality, involved using our dataset, which includes annotated medical abstracts, each assigned scores on a scale from 0 to 100 for ease of reading and technicality. Our main objective was to train these models to predict these scores based on the textual content within the abstracts.

To assess the performance of “BERT-Readability” and “BERT-Technicality,” we employed the root mean square error (RMSE) metric. RMSE measures the average difference between predicted and actual scores, with lower RMSE values indicating better alignment between predictions and actual scores.

By leveraging pretrained BERT models, we established a foundational framework for predicting readability and technicality in medical abstracts. “BERT-Readability” and “BERT-Technicality” serve as baseline models and also as reference points for future analyses. This enables us to assess the effectiveness of any forthcoming advancements or novel techniques introduced into our work.

### 3.3.3. Evaluation

To judge the performance of our models relative to an alternative approach for document difficulty estimation, we specifically selected readability formulas that have been utilized in the domain of medical literature [292, 66]. These commonly used readability formulas were tested on the same test set as our models.

Prior to evaluation, we took the necessary steps to ensure compatibility between the outputs of the readability formulas and the ground truth values of our dataset. To achieve this, we employed rescaling/normalization techniques to align the results of each formula with the range and distribution of the dataset’s ground truth scores. This approach allowed us to establish a fair and consistent basis for comparison. Subsequently, we evaluated the performance of each readability formula using the same evaluation metrics employed for our models (“BERT-Readability” and “BERT-Technicality”).

By incorporating a comparison with these readability formulas, we are able to gain a broader perspective on the strengths and limitations of our models. This comparative analysis allows us to assess whether our models outperform or align with the established readability formulas in the specific context of medical documents. The objective is to position the performance of our models within the broader landscape of readability assessment methods utilized in the medical domain.

## 3.4. Results

Given our data set and modelling methodology for document difficulty, we present the results with respect to each research question.

### 3.4.1. User Agreement on Document Difficulty

When attempting to model document difficulty for medical search engines it is necessary to establish to what extent professional search engine users themselves possess a notion of document difficulty and its defining aspects, readability and technicality. These difficulty criteria may be very subjective or alternatively, they could be more widely shared in the professional community. To establish to what extent this is the case, we analyse the mean intraclass correlation coefficient (ICC) as a way of measuring annotator agreement on readability and technicality scores assigned to the PubMed abstracts by our annotators. According to Koo et al. guideline [131], ICC values can be interpreted as follows: below 0.5: poor annotator agreement, 0.5 - 0.75: moderate agreement, 0.75 - 0.9: good agreement and exceeding 0.90 excellent agreement.

The mean ICC for annotations was found to be 0.81 (Std=0.08) This result indicates a substantial consistency and agreement among the annotations provided by medical professionals, suggesting that professional users generally agree on the level of readability and technicality of medical publications.

### 3.4.2. Interlingual Differences in Document Difficulty

The next research question concerns the differences between medical documents in German and English in terms of readability and technicality.

Table 3-2 shows general differences between German and English medical documents and their abstracts. The average length of the abstracts was found to be 215 (Std=90) words for English and 190 (Std=77) words for German abstracts. It is noteworthy that, while English abstracts have a higher word count, German abstracts tend to be longer in terms of character count. This observation is due to the nature of the German language, where words often contain more characters compared to English [20]. Specifically, the average character count for English abstracts was 1480 (Std=585), while, for German abstracts, it was 1587 (Std=616) characters. The German language has many compound words corresponding to English phrases, which partially explains the differences.

Figure 3-2 shows a visual representation of the readability score distribution in English and German and identifies any potential outliers or patterns. The average readability scores across all abstracts were found to be 64.21 (Std=21.54) and 61.50 (Std=21.67) for English

Table 3-2.: Descriptive statistics on the annotation process.

Word count Eng.	215 (Std=90)
Word count Ger.	190 (Std=77)
Char. count Eng.	1480 (Std=585)
Char. count Ger.	1587 (Std=616)
WPM	113 (Std=12)
Docu. annot. time in sec.	176 (Std=73)
ICC	0.81 (Std=0.08)

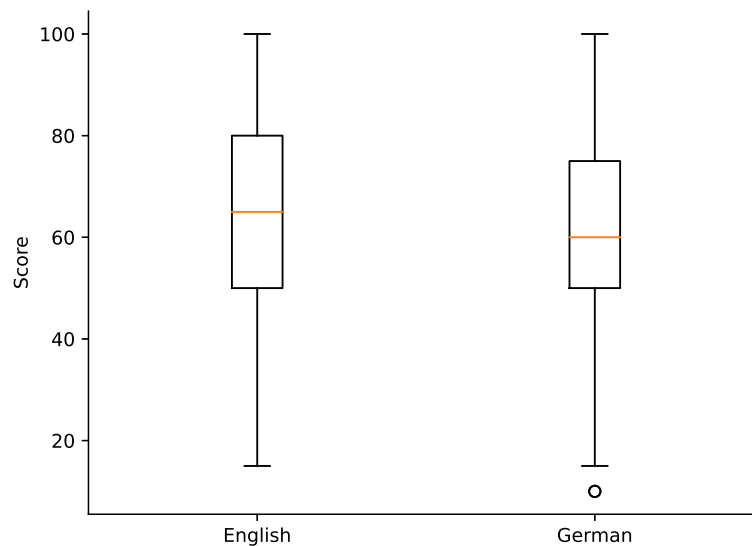


Figure 3-2.: Readability score distributions for English and German abstracts.

and German respectively.

Figure 3-3 shows the difference between English and German in readability of medical article abstracts. The positive side of the graph shows the articles that are easier to read in English, and articles that are harder to read in German on the negative side. As the figure indicates English medical abstracts are slightly easier to read compared to German abstracts. This disparity can be attributed to several factors. First, the English language generally exhibits a more straightforward and concise writing style, which may enhance readability for a wider audience. Second, English has a larger presence in the global scientific community, leading to greater standardization and familiarity among medical practitioners. Consequently, English abstracts may be tailored to a broader readership, including non-native English speakers.



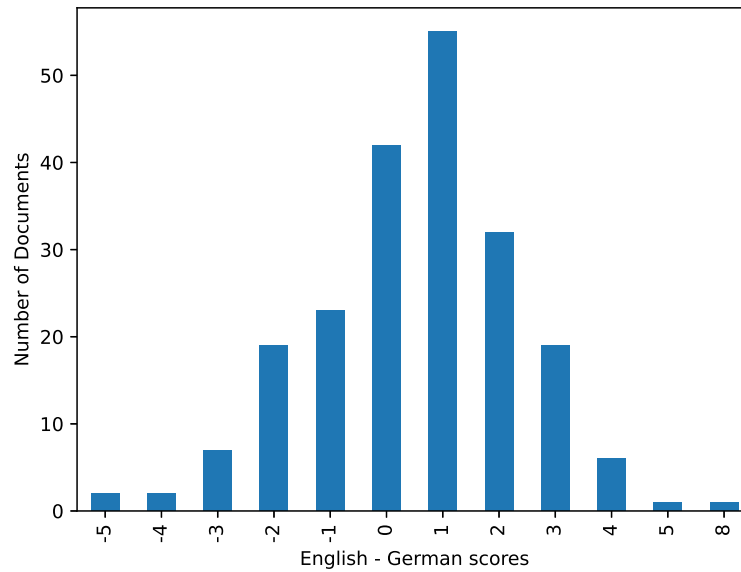


Figure 3-3.: Readability score difference per article between English and German versions of an article. Scores are between 0 and 20.

Figure 3-4 shows a visual representation of the technicality score distribution and identifies any potential outliers or patterns. The average technicality scores for the dataset were 30.55 (Std=10.02) and 26.72 (Std=12.81) for English and German respectively.

Figure 3-5 shows the difference between English and German technicality, where the positive side of the graph shows articles with higher technicality in German, and articles with higher technicality in English on the negative side. As Figure 3-5 shows German medical abstracts tend to exhibit higher levels of technicality compared to their English counterparts. This finding aligns with previous studies [104, 84] highlighting the inherent complexity of the German language, particularly in the medical domain. The higher technicality level of German abstracts can be attributed to the frequent usage of specialized medical terminology and the structural intricacies of the German language itself.

In answer to our second research question above we conclude that there are differences in readability and technicality between German and English medical documents. As expected given findings reported in previous work, medical documents written in German are rated as slightly less readable and more technical than the same documents in English.

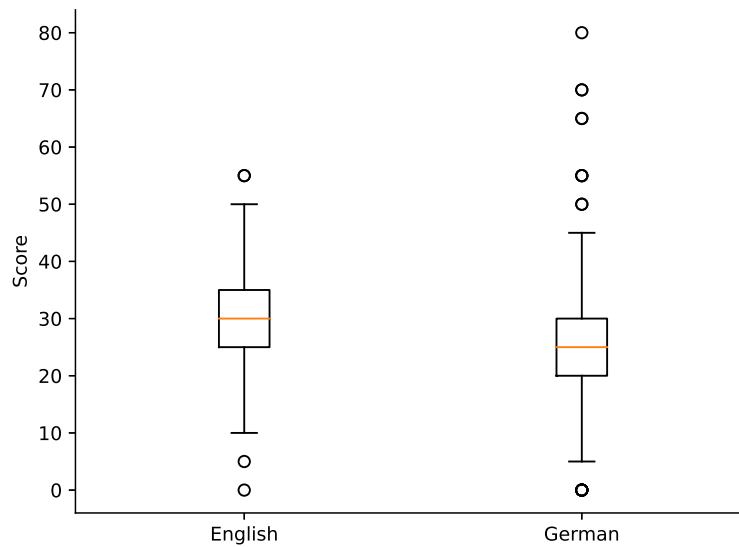


Figure 3-4.: Technicality score distributions for English and German abstracts.

### 3.4.3. Modelling Document Difficulty

The last research question this chapter aims to address is what method performs best on the task of automatic estimation of readability and technicality scores of a document. This is the key question given the overall goal of this work because an automatic prediction of a document’s readability and technicality is a necessary component of a personalised medical search engine for medical professionals such as WisPerMed.

Using our annotated data set described above (Section 3.3.1) we developed BERT-Readability and BERT-Technicality regression models to predict readability and technicality levels in scientific research abstracts on a prediction scale between 0 and 100. We evaluate the performance of BERT-Readability and BERT-Technicality using the RMSE metric.

As shown in Table 3-3, the RMSE for readability is 10.61 for English abstracts and 11.80 for German abstracts. For technicality, the RMSE is 9.42 for English abstracts and 10.07 for German abstracts. These results demonstrate the models’ effectiveness in predicting readability and technicality scores.

To address our research question, which automatic method of estimating document difficulty performs best, we conduct a comparative assessment, pitting our models against widely adopted and well-established readability formulas commonly utilized in healthcare literature [202, 253]. In Section 3.3.3 we outlined our set, so that RMSE scores are indeed comparable between BERT-models and readability formulas. Table 3-3 shows the RMSE scores for each of the methods. The results demonstrate that our BERT models consistently

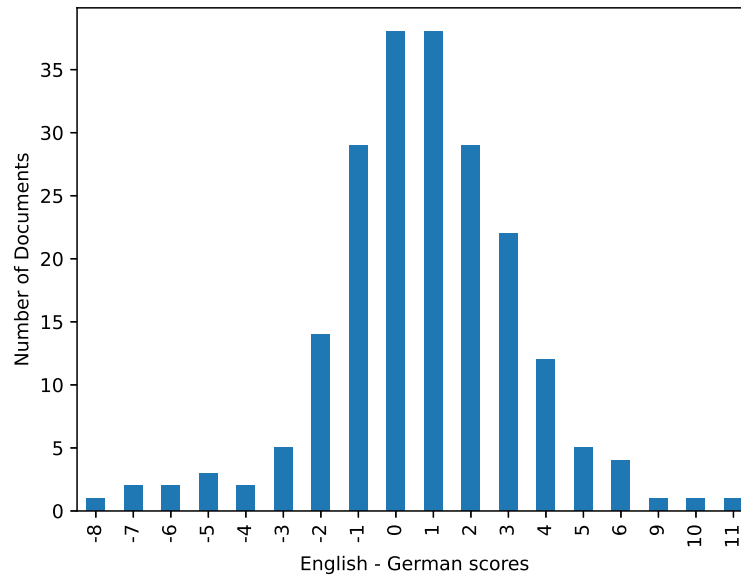


Figure 3-5.: Technicality score difference per article between English and German versions of an article. Scores are between 0 and 20.

outperform the traditional formulas in terms of prediction accuracy. This underscores the critical advantage of employing specialized models tailored explicitly for medical experts.

The existing readability formulas assessed in our comparison rely on a set of general features, such as sentence length, syllable count, and word complexity, which are designed for assessing text readability across various domains. In contrast, our models have been meticulously fine-tuned to account for the specific needs and nuances of medical research abstracts, offering a more precise and effective solution for this specialized context.

Based on these results we conclude in response to our third research question that specialized models tailored explicitly for medical experts are the best performing method for automatically estimating the readability and technicality of medical documents.

## 3.5. Discussion

The main aim of this chapter was to develop a tool for automatic estimation of document difficulty levels in medical document. We defined document difficulty as readability and technicality of a document and evaluated a transformer-based machine learning approach against readability formulas, which are a traditional approach to document difficulty estimation. We developed BERT based readability and technicality models using a novel dataset we created specifically for the purpose, which consists of 209 scientific research abstracts from

Table 3-3.: Formulas and our models' performance on the dataset: RMSE scores comparison.

Formula	English	German
BERT-Technicality	<b>9.42</b>	<b>10.07</b>
BERT-Readability	<b>10.61</b>	<b>11.80</b>
Coleman-Liau Index	19.96	20.22
SMOG Index	21.28	23.59
Gunning Fog Index	26.13	30.66
Dale-Chall Readability Score	26.77	23.61
Flesch-Kincaid Grade Level	27.93	28.38
Automated Readability Index	30.95	30.26
Gutierrez de Polini Index	33.90	31.98
Szigriszt-Pazos Index	32.93	30.94
Fernandez-Huerta Index	32.30	31.17
Flesch Reading Ease	34.34	32.60
Gulpease Index	34.46	35.48

All formulas are available in Textstat Python library. <https://github.com/textstat/textstat>.

diverse medical disciplines, available in both English and German. Each abstract underwent annotation by medical practitioners, who assigned ease of reading and technicality scores. BERT models substantially outperformed readability formulas on the task of readability and technicality score prediction, suggesting that this is the best way forward to personalize medical search engines by document difficulty estimation.

The integration of readability and technicality aspects into search engines has practical implications for medical practitioners. This personalization based on language proficiency and expertise enhances the precision and relevance of search results, leading to more efficient and effective IR, optimizing decision-making and patient care. Therefore, the readability and technicality models developed in this chapter serve as valuable tools to integrate into user profiles for personalized search engines in order enable the use of filtering-based techniques or contribute to the ranking algorithm used in the retrieval process. Additional research is required to examine the feasibility of this integration. A future study could focus on implementing text difficulty aspects into IR ranking algorithms. The study should address algorithmic refinement, adaptability, and user experience assessment for practical implementation.

Some limitations need to be considered when interpreting the results and generalizing findings from this chapter. Although the dataset offered useful insights, its limited size may restrict the generalizability of our findings to a wider medical literature context and various medical disciplines. Our research focused on English and German abstracts, which

may not represent the full linguistic diversity of medical literature. Future studies could expand to include a more extensive range of languages to enhance the scope of applicability. While our regression models demonstrated superior performance, their complexity may pose challenges in real-world implementation. Future research should address ways to streamline these models for practical use.

There are also further promising avenues for future research. For example, future work could explore more sophisticated models and advanced transfer learning techniques, with the aim to improve the accuracy and applicability of readability assessment in the context of medical research abstracts.

Furthermore, another compelling area of future study would involve investigating the direct influence of readability and technicality assessment on the decision-making processes and patient care outcomes of medical professionals.

## 3.6. Summary

This chapter outlines methods for estimating document difficulty with the aim of enhancing search engine personalization. In Section 3.1 we defined document difficulty as the readability and the technicality of a text and further explained these two key concepts. Section 3.2 reviewed relevant previous work on readability and technicality estimation, introducing the two methods considered in this work: traditional readability formulas and machine learning approaches. In Section 3.3, we presented the data set annotated by medical professionals. Our results presented in Section 3.4 and discussed in Section 3.5 indicate that human annotators achieve a good degree of agreement when assigning technicality and readability scores to documents. The results also highlight the differences between German and English medical texts, with German documents rated somewhat less readable and more technical than the same documents in English. Finally, we found that BERT-based models of document difficulty outperform traditional readability formulas in assigning difficulty scores to documents, making them a best choice for difficulty estimation for the purpose of personalizing medical search engines.



# 4. Medical Document Subfield Classification for Search Engine Personalization

This chapter investigates a further method for personalization of search engines for medical experts: the automatic identification of the medical subfield to which a document belongs. This task aims to increase the relevance of retrieved documents by offering professional users an option to select search results which directly address the medical subfield of their expertise. The chapter covers the development of classification models including the dataset creation for their training and the evaluation of a fine-tuned PubMedBERT model and a GPT-4 LLM against a baseline Random Forest Classifier. In this chapter we address the following research questions:

- How can we classify a medical document into its medical subfield?
- Is the automatically collected dataset effective for this classification task?
- Can we replace fine-tuned models with zero-shot learning via LLMs?

## 4.1. Introduction

Medical experts often seek highly specialized information related to their field of expertise. For instance, while an oncologist searches for the latest research on treatments for these types of cancer, a radiologist might look for imaging techniques for these types of cancer. Thus, relevant results for a search query by an oncologist might not be completely relevant for radiologists. Understanding these differences and considering a user's field of expertise in the search is critical for developing effective personalized medical search engines [61, 284].

Studies have explored different techniques for predicting the user's field of expertise. Most approaches involve automated query log analysis to identify patterns that differentiate between expert groups [285]. However, to achieve a match between a document's thematic field and a user's field of specialty, it is essential to not only estimate the user's field of expertise, but also to develop a method of identifying a document's targeted users based on the field of expertise that the document addresses. This would allow the ranking of retrieved

documents based on the multidimensional document relevance techniques, resulting in more relevant and useful information.

This task essentially belongs to a well-established IR area of document classification, which aims to categorize documents into predefined classes based on their content. Considering a document’s thematic field in relevance evaluations [126] significantly increased effectiveness in the IR fields such as sports, news, or health, leading to faster document relevance judgments [298, 214]. Traditional document classification techniques include machine learning models like support vector machines and random forests, which have been applied successfully in various domains [78, 130]. However, these methods often struggle with the complexity and specificity of documents in highly specialized fields [130]. Biomedical document classification raises unique challenges due to the immense and continuously growing volume of medical documents, intensively interrelated context between documents, as well as due to the very specialized nature of the medical subfields, such as “Internal Medicine”, “Radiology”, “Dermatology”, etc.

To consider users’ specialized fields of interest in a medical search engine and retrieve documents relevant to those fields, three key steps need to be completed : (1) extracting users’ interests, specialties, or fields of study; (2) categorizing documents according to the subfields they address; (3) define a multidimensional relevance score considering both the user’s field of interest and the documents thematic field. Several studies investigated the idea of creating user profiles where user-related information is stored, such as interest, age, location, and others, either by analyzing user behavior or manual configuration [152, 132, 44]. Methods of probabilistic retrieval and learning to rank made it easy to consider these aspects in the ranking algorithm by handling structured document retrieval with different data types [126].

In this experiment, we aim to tackle the problem of categorizing biomedical documents into their respective subfields by analyzing document content. By investigating advanced classification techniques, including traditional machine learning, state-of-the-art transformer-based models, and LLMs, we seek to develop a robust method for medical document classification. A similar study on classifying medical publications into cancer types applied traditional machine learning and deep learning approaches and showed robust performance [130]. Therefore, we are leveraging similar classification techniques and other state-of-the-art transformer-based and LLM-based models, seeking to develop a robust method for categorizing and classifying medical documents into medical subfields.

The following sections describe this model development and their evaluation in detail.



Table 4-1.: Medical subfields and specialties. The subfields are used as classes for the different classifiers.

Anesthesiology	Emergency Medicine	Cardiology
Dermatology	Forensic Medicine	Family Practice
Geriatrics	Medical Genetics	Internal Medicine
Neurology	Otolaryngology	Ophthalmology
Pathology	Physical Medicine and Rehab.	Pediatrics
Podiatry	Radiology	Public Health
Surgery	Urology	Psychiatry

## 4.2. Materials and Methods

Our overall methodology involved creating two datasets: one with manual annotations by experts and the other automatically created dataset similar to Kolukisa et al.’s [130] method of creating the dataset using queries in PubMed [188] including 21 distinct medical subfields. Then, we developed three different architectures of classifiers (Random forest, fine-tuned PubMedBERT, and zero-shot GPT-4 LLM) using the automatic dataset and evaluated them using the manual dataset. This dual dataset approach allows for a comprehensive evaluation of the classifiers, ensuring that the models are not only accurate but also generalizable across different sources of medical literature.

### 4.2.1. Data Collection

There is no known open-source dataset available online that we can use. Therefore, we create new datasets to perform for medical document classification. This required setting up a list of medical subfields as labels and then collecting documents within the same medical field. Our approach encompassed a comprehensive strategy to ensure the dataset’s diversity and representativeness, addressing the complexity of medical literature. In this experiment, we collect two datasets: (1) a manual dataset and (2) automatic dataset as described above.

#### Classes

The list of classes represents the main medical subfields. The list was compiled based on an extensive review of relevant previous work, such as Lett et al. [143] and by consulting field experts on the topic. Table 4-1 lists the medical topics under which any medical document can be classified. These classes also represent the majority of medical specialties; any other subfield can be added under one of these fields, such as “Pulmonary Medicine” which can be added under “Internal Medicine”.

## Manual Dataset

During the annotation round of medical abstracts from PubMed with readability and technicality scores as described in Chapter 3, we included two dropdown lists on each article’s medical subfield with the same pre-defined list of classes. It was mandatory to select a subfield from the first dropdown list, and the second was optional, allowing experts to assign a secondary subfield if applicable.

Each article was presented to three experts. Therefore, we merged the results for each article by assigning consensus subfield labels, i.e. subfields selected by at least two experts. Applying this resulted in no article left without consensus subfield assignment in our manual dataset.

The total number of annotated articles by experts is 209, which is a small number for developing and validating classifiers with a large number of classes, as it is the case in our set up. With 21 different classes in the dataset, on average, each class received a total of 9.95 (max=21, min=1) documents per class. This small sample size necessitated a creation of an additional and larger dataset to train the models effectively. However, due to the busy schedule of medical experts, it is prohibitive to create a larger manual dataset. Therefore, we utilized the Medline database [172] to create an automatic dataset.

## Automatic Dataset

The Medline biomedical publication database [172] indexes more than 30 million articles, which is accessible using PubMed search engine [188]. The database is considered one of the largest databases representing research articles from all medical specialties. Therefore, we utilized PubMed to create a dataset automatically. With the help of medical experts, we created a list of sub-topics within the medical subfields, such as “Gastroenterology”, “Endocrinology”, and “Obstetrics” under ‘Internal Medicine’.

We used these sub-topics and subfields as queries on PubMed search engine to retrieve the top 500 documents per query if retrieved. Using this iteration, we collected a total of 24K unique documents on all subfields after dropping duplicates within the same subfield, including 300 documents per “Podiatry” subfield as the smallest number of documents per subfield. This provided a diverse dataset representative of each subfield. Similar data collection processes were used and evaluated in other studies such as [130], showing the effectiveness of such methods in building datasets for content-based classification.

To develop our classifiers, we kept the dataset balanced by performing the SMOTE undersampling technique on 300 instances per class, resulting in a dataset of 6300 documents. Then, we split the dataset into 60% for training and  $2 \times 20\%$  for validation and testing.

One observation from the dataset is the multi-subfield documents, where 28% of the

manually annotated documents by domain experts have two subfields, and 0.8% of the automatically annotated documents have two subfields. This suggests the existence of interconnected documents with multiple fields. These interconnected medical documents suggest the need to consider documents with multiple labels. For instance, a document discussing the “Pediatric Neurology” topic can be classified under both “Pediatrics” and “Neurology” subfields due to the overlapping nature of these specialties. This interconnection reflects the complexity and collaborative nature of medical practice, where specialties often converge to provide comprehensive patient care [143].

The interconnection between medical fields suggests that the medical document classifier needs to handle not only multiple classes but also the possibility of multiple labels per document. However, the limited number of documents in the automatic dataset with multiple labels makes the task of a multilabel classifier more complex. Therefore, we decided to focus on single-label classification for this study and dropped the documents with multiple labels from the automatic dataset. Alternatively, we rely on using the prediction confidence values to assign one or two subfields to the respective document. For example, if the classifier returned a confidence value for a document D of 51% for “Internal Medicine” and 40% for “Surgery” then D can be classified into the two subfields. However, the confidence value threshold to consider multiple labels needs to be evaluated.

## Dataset Evaluation

To assess the quality and relevance of the automatically generated dataset, we employed GPT-4 LLM to assess the accuracy of the documents-subfield pairings. Each abstract was paired with its respective medical subfield and fed into the LLM with a prompt requesting an accuracy score on a scale between 0 and 10, as shown in Figure 4-1. This score reflects how well the abstracts’ content aligns with the specified medical subfield and aims to provide a quantitative measure of automatic labeling precision.

To have a baseline for judging the results, we used the same prompt on the manual dataset to compare results between the two datasets. To handle the multiple labels issue in the manual dataset, we selected the one with the majority of experts’ annotations. If there was no majority, we submitted two prompts, each with different labels, assigned the label with a higher score to document in the manual dataset and used it for classifier evaluation.

This automated evaluation process using LLMs offered a rapid and scalable method to measure the overall dataset integrity and validate the consistency and accuracy of our data collection methods. This part is inspired by the work of Faggioli et al. [65] showing the potential of utilizing LLMs for relevance judgment. This approach provides an extra reliability check and could also provide insights into refining the dataset collection and labeling process, which ensures the robustness of the classifier’s training foundation.

**Instruction:** You are an expert medical doctor interested in medical research. Given the subfield and abstract pair below, evaluate how accurately the abstract is classified under the subfield. Provide a score between 0 and 10, where 0 means completely inaccurate and 10 means perfectly accurate.

**Subfield:** {subfield}  
**Abstract:** {title} {abstract}

**Evaluation:**

Figure 4-1.: LLM prompt for evaluating the automatic dataset quality.

We evaluated the entire manual dataset and 1050 documents from 6300 documents in the automatic dataset, selecting 50 documents per class. The results using the GPT-4 prompt shown in Figure 4-1, we revealed comparable scores: 7.6 (Std=2.10) on the automatic dataset and 7.82 (Std=2.23) out of 10 on the manual dataset. These results indicated a high level of agreement between manually and automatically annotated datasets, supporting the reliability of our automatic dataset.

## 4.2.2. Experimental Setup

For the task of medical subfield classification, we focused on developing traditional machine learning techniques, fine-tuning a BERT-based transformer model, and utilizing an LLMs zero-shot classification to compare the classifiers' performance. Our goal was to assess the performance across different modeling approaches to determine the most effective method for this classification task.

We developed the following classifiers:

**Random Forest (RF)** RF servers as our baseline model. It is trained on the training set for multi-class classification. As a preprocessing step, we removed stop words and applied a stemmer. We use TF-IDF vectorization, chi-squared feature selection, and K-Fold cross-validation using the validation dataset, evaluating its performance with the macro-F1 score.

**PubMedBERT** PubMedBERT is a natural choice for this domain-specific classification task, as it is a transformer-based model pre-trained using abstracts sourced directly from PubMed [86]. We fine-tuned this classifier directly on the training set to classify texts into

specific medical subfields, with the macro-F1 score as the evaluation metric.

**GPT-4** We utilized the power of the GPT-4 LLM with zero-shot learning to predict the targeted class (medical subfield) using the prompt shown in Figure 4-2. This approach aimed to evaluate the capabilities of modern LLMs in classifying biomedical documents without extensive fine-tuning and compare it to fine-tuned transformer-based models.

**Instruction:** You are an expert medical doctor interested in medical research. Classify the following abstract into one of the medical subfields: {list\_of\_subfields}. Abstract: {title} {abstract}.

**Answer:**

Figure 4-2.: LLM prompt for classifying document into its medical sub-field.

### 4.2.3. Evaluation

We evaluated our classifiers using the macro-F1 score, using the test split of the automatic dataset and the manual dataset. This dual evaluating strategy ensured that our models were tested on both the larger, diverse automatic dataset and the smaller, high-quality manual dataset. Testing on the manual test set would highlight how well the model generalizes to independently sourced data.

## 4.3. Results

### 4.3.1. Performance Evaluation

Tabel 4-2 summarizes the performance of each classifier on the automatic tests and the manual testset, showing the relative effectiveness of each approach. Figure 4-3 shows the confusion matrix of each classifier per testset. This also allows for highlighting the performance of each class. One general observation from the confusion matrix is that misclassifications in the automatic test set are distributed across all classes, whereas in the manual dataset, misclassifications are concentrated in a subset of the classes.

**Random Forest** The RF model's performance with a macro-F1 score of 0.52% did not surpass the BERT-based model. However, it surprisingly achieved higher performance by

Table 4-2.: Medical subfields classifiers performance on the automatic and manual datasets.  
Macro F1 score

	Automatic testset	Manual dataset
Random Forest (RF)	0.52	0.54
PubMed-BERT	<b>0.58</b>	<b>0.60</b>
GPT-4	0.50	0.51

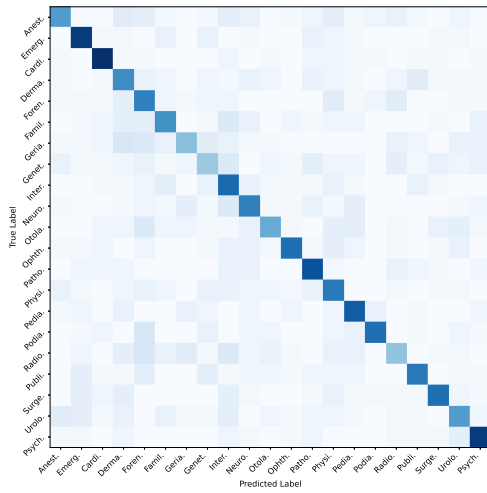
2% than the GPT-4 model, which is considered the state-of-the-art LLM. This highlights that traditional machine learning models can still hold their ground against more advanced models in certain contexts. The model also showed consistent performance on the manual dataset, with a 54% macro-F1 score, demonstrating its robustness across different datasets from different sources.

**PubMedBERT** The BERT-based classifier performed best with a 0.58% macro-F1 score and showed a consistent score with 0.60 macro-F1 in performance on the manual dataset. The PubMedBERT model outperformed zero-shot LLM predictions, emphasizing the importance of fine-tuning for achieving high accuracy in specific domains. This highlights the strength of domain-specific fine-tuned models in handling specialized classification tasks.

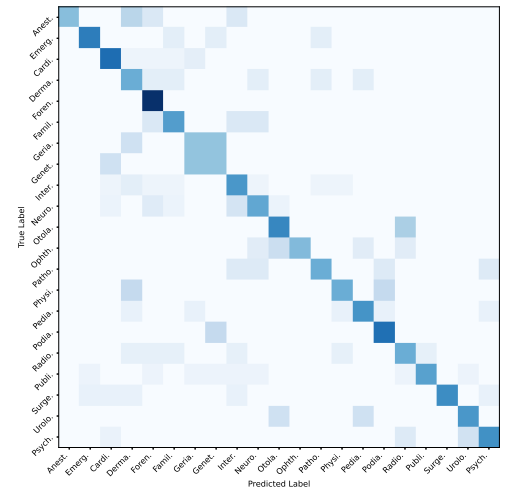
**GPT-4** The zero-shot LLM-based classifier achieved the worst performance with 0.50% on the automatic dataset and 0.51% on the manual dataset. This demonstrated the potential of using LLMs in scenarios where labeled training data is scarce or unavailable. However, despite its lower performance, the LLM's ability to classify documents without prior training on the specific dataset is noteworthy and suggests improvements through techniques like few-shot learning or supervised fine-tuning.

### 4.3.2. Statistical Test Results

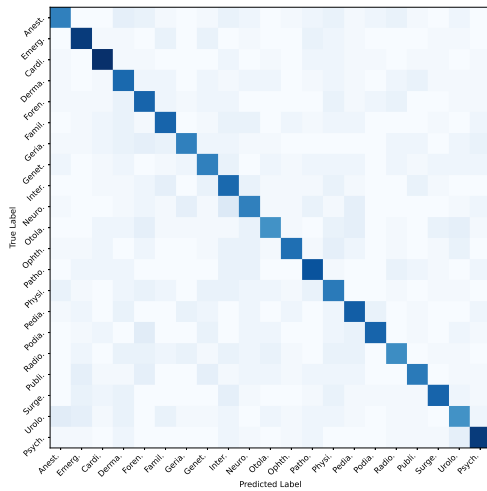
To determine whether the observed differences in performance are statistically significant, we performed the Wilcoxon Signed-Rank Test for each pair of classifiers on both datasets at  $\alpha = 0.05$  with Bonferroni correction. The statistical tests confirm that PubMedBERT significantly outperforms both Random Forest and GPT-4 on both datasets. However, the performance differences between Random Forest and GPT-4 are not statistically significant.



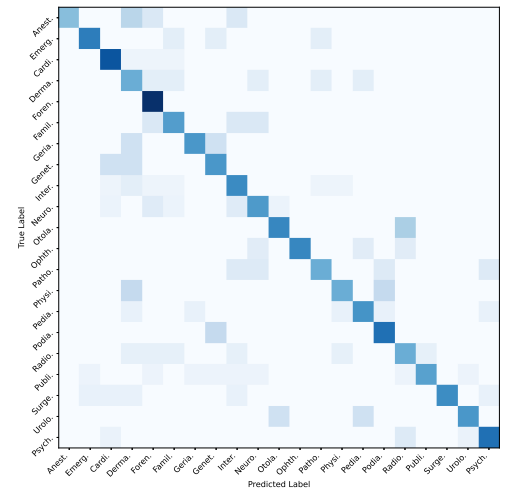
(a) Random Forest automatic testset



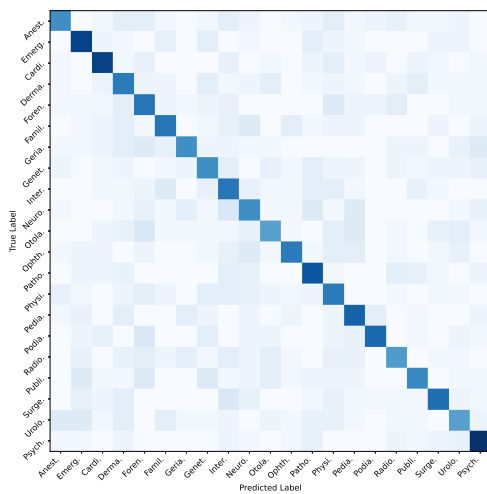
(b) Random Forest Manual dataset



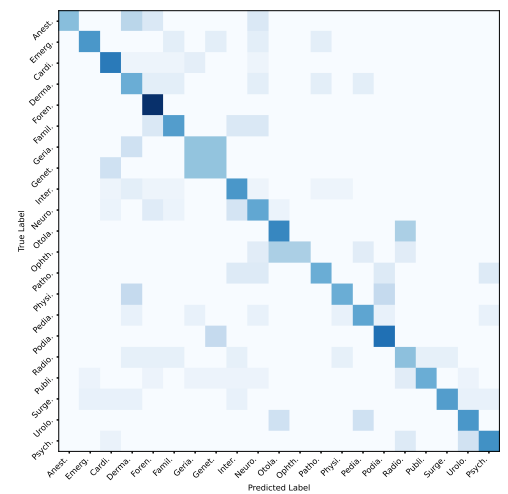
(c) PubMedBERT automatic testset



(d) PubMedBERT Manual dataset



(e) GPT-4 automatic testset



(f) GPT-4 Manual dataset

Figure 4-3.: Confusion Matrices of medical subfields classifiers on the automatic testset and manual dataset.

## 4.4. Discussion

In this chapter, we investigated methods for classification of medical documents into medical subfields they address. This task is an important personalization feature for medical search engines as it allows expert users to retrieve documents relevant to their fields of specialization. The chapter addresses three research questions:

- How can we classify a medical document into its medical subfield?
- Is the automatically collected dataset effective for this classification task?
- Can we replace fine-tuned models with zero-shot learning via LLMs?

In response to the first and the third research question, we demonstrated the efficacy of BERT-based models in classifying medical content based on its subfield. We also found that fine-tuned BERT models consistently achieve significantly better results compared to zero-shot GPT-4, especially when a significant amount of labeled training data is available. These findings align with other studies where fine-tuning or traditional machine learning methods outperformed zero-shot classification based on LLMs [181, 216], confirming the importance of domain-specific training for achieving high accuracy.

There is potential to improve the LLMs using different techniques such as few-shot learning [297], encoding examples within the query [221], or using supervised fine-tuning [58]. Implementing these techniques could enhance the performance of LLMs like GPT-4, making them more feasible for specialized tasks in the biomedical domain. However, an open-source LLM is needed to evaluate the effectiveness of these techniques.

With regard to the second question, our dual approach of using both manual and automatic datasets allowed for a comprehensive evaluation of classifier performance, ensuring robustness and generalizability. We found that the accuracy of annotation of our automatically annotated data set was comparable to that of the manual one, when automatically assessed by GPT-4 LLM. However, the overall classification performance remains modest, indicating that better training data could significantly enhance the accuracy and reliability of the classification models. Expanding and refining the training datasets, particularly with more high-quality and multi-label annotations, would likely improve the effectiveness of these classification methods.

One issue that arose from our datasets was the multiplicity of medical subfields a document can belong to. Due to the limited number of instances with multiple labels in the automatic dataset, we used single labels per instance, although the classification task based on the manual dataset is supposed to be with multiple labels. This limitation highlights the need for larger, high-quality multi-label datasets to fully explore the potential of multi-label classification in medical document retrieval. The automatic dataset does not guarantee that all documents can't fit in other labels. However, using the prediction's confidence values



could solve this issue by assigning subfields the highest two confidence scores subfields to the document or having a confidence threshold for classes.

The classification of medical research papers into specific subfields has several potential applications and is not limited to personalized information retrieval or enhanced search engine relevance. In clinical settings, decision support systems can leverage classified medical literature to provide evidence-based recommendations tailored to specific medical fields, improving patient care outcomes. Moreover, it could support researchers in quickly identifying pertinent literature in their field, aiding in literature reviews and the identification of research gaps. Educators can curate specialized content for teaching purposes, enhancing the quality of medical education. However, all these need to be evaluated from a user's perspective to prove the potential and scope.

Therefore, future research that builds on methods outlined here is necessary. It should focus on expanding multi-label classification capabilities, reflecting the task's interconnection nature. It should also integrate these models with a user profile to evaluate the effectiveness of this personalization method with actual field experts using a user study.

## 4.5. Summary

In this chapter the methods for personalizing literature search engines for medical experts is continued by developing and evaluating three document classification models: Random Forest Classifier, PubMedBERT and GPT-4 LLM on the task of classifying medical documents from MEDLINE repository into medical subfields they address. Section 4.1 outlines the relevance of this task for personalizing medical search engines for professionals and explains the task in the context of document classification. Due to the limited feasibility of manual annotation, an automatic dataset was compiled and subsequently assessed against a manually created dataset. Section 4.2 details the dataset compilation, evaluation, and the overall methodological setup. The classifier evaluation results are presented in Section 4.3 and further discussed in Section 4.4. The fine-tuned PubMedBERT model outperforms other models in our evaluations. However, the results, while promising, remain modest and suggest significant potential for improvement if trained on a higher-quality annotated dataset rather than the automatically collected one used in this study. The results also indicate that zero-shot LLMs are not yet capable of replacing fine-tuned models on the document classification task.



# 5. Bio-Concepts for Enhanced Precision in Medical Search

This chapter explores the benefit of bio-concepts to the relevance of search results for medical literature search. Bio-concepts are key biological and medical terms, such as genes, proteins, diseases, etc., which can be extracted from both medical documents and user queries. We explore the viability of incorporating bio-concepts into the relevance-matching process, guided by the following research questions:

- Can the explicit integration of bio-concepts into retrieval systems significantly enhance performance?
- Which integration method yields the best results?
- How do BERT-based methods compare with traditional approaches in handling bio-concepts?
- Can a synergistic approach that combines both BERT-based methods and bio-concepts offer superior retrieval outcomes?

We address these questions using the data from the TREC Precision Medicine tracks of 2017 to 2019 [208, 207, 209] and NDCG@10 [279] as our primary evaluation metric.

## 5.1. Introduction

Bio-concepts in medical literature encompass key biological and medical terms, entities, and relationships, including genes, diseases, symptoms, diagnostic procedures, treatments, cells, pathways, chemicals, anatomical structures, microorganisms, epidemiological terms, genetic variants, and biomarkers, which collectively form the foundation of biomedical research and clinical practice. The integration of bio-concepts into medical search engines represents a significant advancement over traditional keyword-based search algorithms.

Historically, keyword-based search algorithms like BM25 and its derivatives were central to retrieving articles through textual matching. These algorithms match the query search terms by the user with the document's content, retrieving results that contain those keywords. BM25 and similar algorithms can effectively handle term frequency saturation and

incorporate document length normalization, which makes it a reliable foundation for search engines such as PubMed [69, 210].

While these keyword-based algorithms are effective to some extent, they also have significant limitations in matching complex medical information, as they often fail to capture the context and semantic meaning of terms [179, 254]. For example, a search for “heart attack” might not retrieve documents that use only the term “myocardial infarction” despite them referring to the same medical condition. This shift has catalyzed investigations into alternative methods that more aptly address the nuances of biomedical information retrieval.

One technique that has shown effectiveness in search engines for health consumers is the integration of UMLS (Unified Medical Language System). UMLS helps with standardizing and disambiguating terms by structured vocabulary and ontology [6], showing improvement in understanding the context and relationships between medical terms [52, 33]. Further development in the same direction suggested the idea of bio-concepts for integrating domain-specific knowledge into information retrieval models [139, 146].

The extraction of bio-concepts does not only refer to annotating the terms with their type (medication, gene, etc) but also maps the term to its unique identifiers such as mapping “heart attack” to “<https://meshb.nlm.nih.gov/record/ui?ui=D009203>” which would allow retrieving several education materials based on this identifier. Similarly, genes such as “BRAF” would be mapped to the unique identifier “<https://www.ncbi.nlm.nih.gov/gene/673>”.

Wei et al. [282] highlight this trend, pointing to the growing importance of such integration for improving search outcomes in the biomedical field. The PubTator API has been instrumental in providing crucial annotations of biomedical literature with entities such as genes, diseases, and chemicals, thereby enriching content understanding and retrieval accuracy [280]. A similar system was proposed by Bravo et al. [29] aimed at identifying connections between bio-concepts, with a particular focus on genes and their associated diseases. A more specific use case is also presented by Lee et al. [138] by utilizing the bio-concepts to search and extract genomic variant information from biomedical literature.

The advanced models of extracting bio-concepts from the medical text allow enriching interactive IR in techniques proven the effectiveness in other domains like Query Expansion [185, 151], Highlighting Bio-Concepts [191, 289], or On-Demand Definitions [213].

**Query Expansion:** Expanding queries with bio-concepts, users can perform more precise exploratory searches. For instance, a query for “diabetes” can automatically include related bio-concepts like “insulin resistance” and “glucose metabolism”, providing a broader yet focused retrieval of relevant documents. Mustar et al. [185] and Lin et al. [151] demonstrated how bio-concepts could act as an additional signal in query processing systems. This method can potentially refine the context of queries, leading to a more focused retrieval of biomedical literature, directly aligning with the aims of our research.

**Highlighting Bio-Concepts:** During results screening, highlighting bio-concepts can enable faster relevance judgment. This helps users quickly identify key terms and concepts within the retrieved documents, which is highly needed as medical experts usually work in tight schedules. There could be two different techniques, like providing an interactive word cloud of all bio-concepts in the document abstract, allowing the users to quickly judge the relevance or refine the search query by adding the selected bio-concept. The other technique is coloring the bio-concept with color codes in the retrieved documents abstracts; this would allow users to identify key sentences, allowing for faster judgment.

**On-Demand Definitions:** Presenting definitions of bio-concepts on demand using unique identifiers can assist users when encountering unfamiliar terms, thus improving their understanding and the overall search experience. When the user clicks on one of the colored concepts in the abstract, the user will be able to see the definition and relevant educational material explaining the concept.

Leveraging bio-concepts, medical search engines can provide more accurate, contextually relevant, and user-friendly search experiences, ultimately improving the retrieval process and supporting better decision-making in the medical field. Despite the promising advancements, evaluating the integration of bio-concepts in medical IR from both system and user perspectives remains essential. All available studies consider the evaluation of the extraction quality for bio-concepts. However, the evaluation of an IR system integrating Bio-concept is still missing.

In this chapter, we explore the integration of bio-concepts — specifically genes, diseases, and chemicals — into retrieval systems and examine their potential to improve search precision and accuracy in the biomedical domain.

## 5.2. Materials and Methods

### 5.2.1. Dataset

We use the datasets published by TREC PM 2017, 2018, and 2019 research initiatives [208, 207, 209]. TREC PM aimed to advance the state of the art in biomedical information retrieval by exploring novel techniques and methods to improve the retrieval of pertinent medical information.

Medline collection [172] is used as the primary corpus for IR experiments. This collection is accessible via the PubMed<sup>1</sup> search engine. It is an extensive biomedical literature repository, crucial for academic research and clinical decision-making. It encompasses around 30

---

<sup>1</sup><https://pubmed.ncbi.nlm.nih.gov/>

million references to articles in life sciences with an emphasis on bio-medicine. Using the Medline collection not only offers a vast source of biomedical knowledge, but also presents a distinctive chance to tackle crucial issues in information retrieval leading to progress in medical research and enhancing access to healthcare information.

In the topics list, each topic comprises of three main fields: disease, gene, and demographic data, as shown in Table 5-1. Our method involved creating queries using these fields by combining the gene with the disease in a simple manner, omitting the demographic data <sup>2</sup>.

Table 5-1.: Example topics from the TREC PM 2017, 2018, and 2019.

Disease: lung cancer Gene: ROS1 Demographic: 71-year-old female
Disease: head and neck squamous cell carcinoma Gene: CDKN2A Demographic: 64-year-old male
Disease: colon cancer Gene: MLH1 methylation suppression (microsatellite instability) Demographic: 68-year-old male
Disease: melanoma Gene: BRAF (V600E), CDKN2A Deletion Demographic: 45-year-old female
Disease: mucosal melanoma Gene: KIT (L576P), KIT amplification Demographic: 62-year-old female

The relevance judgment process in the TREC PM 2017-2019 of document/topic pairs involved expert assessors systematically evaluating the retrieved documents and categorizing them into three distinct levels: “not relevant”, “partially relevant”, and “definitely relevant” [208]. These assessments were made based on the degree to which each document aligned with the provided topics and their associated information needs. “Definitely relevant” is assigned if the article focuses on PM, specifically addresses the exact disease, and examines the same gene precisely. “Partially relevant” is largely equivalent to “definitely relevant” but with the exception that disease can also be more general and gene can also be missing a variant or different variant. In case the article is neither definitely, nor partially relevant, it is assessed as ‘not relevant’.

<sup>2</sup>Demographic data was deemed relevant for clinical trial documents, but not for scientific papers

Table 5-2.: Example of bio-concepts could be found in an abstract.

Name	Type	Identifier
Melanoma	Disease	MESH:D008545
BRAF	Gene	673
V600	Variant	mVar:p  <i>Allele</i>  V 600;VariantGroup:0; CorrespondingGene:673; RS#:121913227;CorrespondingSpecies:9606

### 5.2.2. Documents Annotation with Bio-concepts

The initial phase of our experiment involved a comprehensive annotation of the Medline dataset with bio-concepts. Several studies discussed the idea of extracting concepts from medical text [263, 32, 196]. We facilitate the annotation process using PubTator Central. This tool is known for its accuracy and efficiency in extracting various entities, including genes, diseases, and chemicals, from text [280]. Our focus was primarily on the title and abstract sections of each article, wherein we enhanced these sections with bio-concepts metadata. This enhancement was crucial as it enabled us to leverage the rich bio-concepts data within these articles.

At this point, we have structured the document with three fields, title, abstract, and a list of bio-concepts, as shown in Figure 5-1. Each bio-concept has a type (Gene, Disease, Variant, Drug, or Cell Line), name (the bio-concept mentioned in the text), and unique identifier (such as MeSH terms for diseases<sup>3</sup> or NCBI Gene ID<sup>4</sup> for genes), which would be the same for terms referring to the same bio-concepts such as ‘heart attack’ and ‘myocardial infarction’. Therefore, when expanding the queries with bio-concepts in Section 5.2.3, we are expanding with the concepts identifier. Table 5-2 shows an example of three different bio-concepts that could be found in an article. (MeSH or Medical Subject Headings is the National Library of Medicine controlled vocabulary thesaurus used for indexing articles for PubMed).

The abstracts contain on average 8.5 bio-concepts each: 2.7 diseases, 2.1 species, 2.3 chemicals, and 1.3 genes. The effectiveness of PubTator in recognizing these biological entities has been well-documented [282]. Our dataset is therefore not only valuable for its bio-concepts information, but also aligned with established best practices in biomedical literature annotation [15].

<sup>3</sup><https://meshb.nlm.nih.gov/>

<sup>4</sup><https://www.ncbi.nlm.nih.gov/gene>

### Impact of BRAF mutation and BRAF inhibition on melanoma brain metastases.

Gummadi Tulasi & Dudek Arkadiusz Z

2015-09-14

PMID : 25426645

publication\_types :Journal Article , Research Support, N.I.H., Extramural .

✓ The impact of BRAF mutations in metastatic melanoma on the incidence of brain metastases and melanoma prognosis and the effect of BRAF inhibitors on the incidence of brain metastases has not been defined.

The impact of BRAF mutations in metastatic melanoma on the incidence of brain metastases and melanoma prognosis and the effect of BRAF inhibitors on the incidence of brain metastases has not been defined. Therefore, a retrospective analysis of patients with metastatic melanoma treated at three institutions was carried out to examine the impact of BRAF mutations and a BRAF inhibitor, vemurafenib, on the incidence of brain metastases. A retrospective review of 436 records revealed no difference in the incidence of brain metastases between patients with BRAF-mutated tumors versus those without (incidence rate ratio=1.11, 95% confidence interval: 0.80-1.53; P=0.53). A lower incidence of brain metastases was observed in patients with BRAF-mutated tumors who took vemurafenib before the development of brain metastases versus those who did not (incidence rate ratio=0.51, 95% confidence interval: 0.30-0.86; P=0.009). Although treatment with vemurafenib led to improvement in extracranial disease control, it did not significantly affect progression of existing intracranial disease and survival in these patients (P=0.7). Although our previous preclinical data have indicated that penetration of vemurafenib into the brain is limited, our retrospective analysis showed that there was a lower incidence of brain metastases in patients with BRAF-mutated tumors who took vemurafenib before the diagnosis of brain metastases.

Figure 5-1.: Example of abstract with bio-concepts.

### 5.2.3. Experimental Setup

To investigate the impact of bio-concepts on search engines, we test two retrieval approaches, Sparse retrieval using BM25 [210] and Hybrid retrieval (Sparse and Dense) by including a BERT-based model for re-ranking BM25 retrieved documents [8].

#### Sparse Retrieval

Sparse retrieval in IR-refers to methods that use explicit term matching between queries and documents, such as the TF-IDF, BM25, or BM25F algorithms. These methods produce a sparse representation of text in which each document is represented by a vector in a high-dimensional space, with the majority of the dimensions being zero. Sparse retrieval is computationally efficient and interpretable because it is based on the presence or absence of specific terms. It may, however, miss nuanced semantic relationships because it does not capture deeper context or meanings that extend beyond exact term matches [186].

We use BM25F as the baseline retrieval method, which helps us assess bio-concepts' impact on search engines and includes an evaluation of query processing techniques, allowing us to provide insights into their effectiveness in enhancing biomedical literature retrieval.

This study includes five retrieval models, each utilizing bio-concepts to enhance the retrieval process:



**Query Expanding Model:** This model involves bio-concept extraction from the query itself and expanding the query with extracted bio-concepts. The expanded query terms and the original query are then used for retrieval through the BM25F algorithm with equal weights across all fields. Query text matches the title and abstract sections, and query bio-concepts are used to match the bio-concepts extracted using PubTator API. This model is similar to the bio-concepts integration in Thalia [243] with the difference in the bio-concepts used.

**Weighted Fields Model:** Similar to the Query Expanding Model, the ranking in this model is based on bio-concepts, abstract, and title. However, a key distinction lies in the field weights assigned in that the fields are assigned with different importance weights (Section 5.2.5).

**Reranking Model:** After the initial retrieval of the top K documents using the baseline BM25F, this model re-ranks these documents using BM25 based on the relevance of bio-concepts extracted from the query. The aim is to reorder the retrieved documents to prioritize those highly relevant to the query bio-concepts.

**Filtering Model:** This model focuses on filtering the retrieved documents based on the presence of all bio-concepts from the query. Only articles containing all bio-concepts are considered, discarding those lacking any of the query's bio-concepts.

**Partial Filtering Model:** In contrast to the strictness of the *Filtering Model*, this model filters the retrieved documents based on the presence of a threshold percentage of the query's bio-concepts.

## Hybrid Retrieval

Hybrid retrieval is an information retrieval method combining the benefits of sparse and dense retrieval methods. This strategy combines the efficiency and interpretability of sparse retrieval with the deep semantic understanding of dense retrieval (via neural network models) [149]. Hybrid retrieval systems frequently use sparse retrieval for initial document filtering, then apply dense retrieval techniques for more nuanced reranking or analysis, as shown in Figure 5-2. This fusion provides a balanced solution for improving retrieval performance in terms of both relevance and computational efficiency [87, 307]. Several tools are developed to enable easy implementation of this retrieval architecture, e.g. Pyserini [150].

To perform this part of our experiment, we use the BM25F model (baseline) to retrieve the top K documents as initial retrieval, as in the re-ranking model above. We optimize the K value using cross-validation (Section 5.2.5)

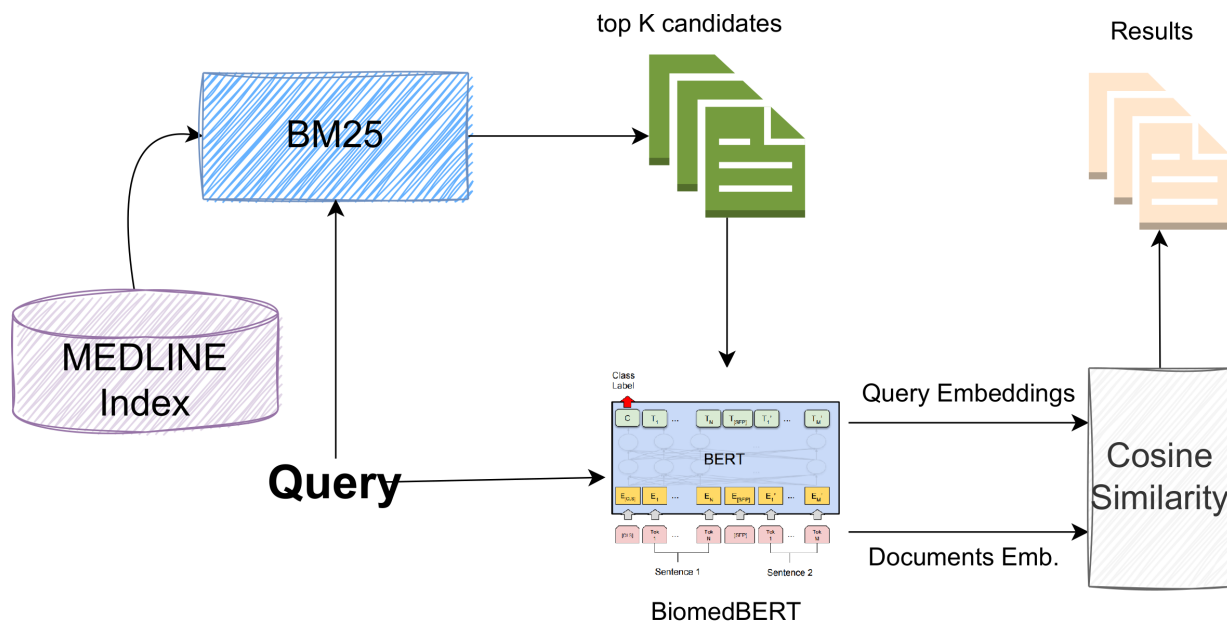


Figure 5-2.: Architecture of our Hybrid retrieval, illustrating the connection between BM25 that retrieves the candidate documents and BERT-based models.

As a dense retrieval component, we add BERT embeddings [55], which are calculated for the query and documents for ranking the documents based on cosine similarity. We use the BiomedBERT [86] model for calculating the embeddings. BiomedBERT is a domain-specific language model that has been pretrained for biomedical natural language processing (NLP) tasks. The model is pretrained from scratch using abstracts from PubMed. This model achieves state-of-the-art performance on many biomedical NLP tasks and currently holds the top score on the Biomedical Language Understanding and Reasoning Benchmark. This hybrid retrieval method reportedly improves the recall of sparse retrieval models and captures semantic relevance information [88].

We experiment with two retrieval models with and without bio-concepts:

**BiomedBERT:** This model is based on BiomedBERT embeddings without bio-concepts. Neither the documents, nor the queries contain any bio-concepts. The main goal of this model is to serve as a baseline for hybrid retrieval, against which models with bio-concepts can be compared. The query is as explained above for BM25F sparse baseline. The query and the top K relevant document are passed to the BiomedBERT model for extracting embeddings. As a model input to calculate the document embeddings, we concatenate the document title and abstract into a text sequence  $[[CLS], Title, [SEP], Abstract]$ . In this sequence, the  $[CLS]$  token is a special marker placed at the beginning of the input to capture the overall representation of the entire sequence, while the  $[SEP]$  token is used to separate the different input parts, helping the model distinguish between these parts of the document [55].

**Expanded BiomedBERT:** This model extracts bio-concepts from the query, and bio-concepts from  $K$  documents initially retrieved from the BM25F baseline. As a model input to calculate the BiomedBERT embeddings, the bio-concepts are concatenated at the end of the query text  $[[CLS], Query, [SEP]bio - concepts]$ . Similarly, to calculate the documents' embeddings, the bio-concepts concatenated at the end of the document text sequence  $[[CLS], Title, [SEP], Abstract, [SEP], bio - concepts]$ .

#### 5.2.4. Evaluation Metric

To evaluate the effectiveness of our models, we implemented the “Normalized Discounted Cumulative Gain at 10” (NDCG@10) as our core evaluation metric [279]. This choice was driven by the NDCG@10's ability to quantify the quality of the top ten results generated by our retrieval systems for every query, where the discounting element considers the fact that some users might stop before reaching rank ten. Central to the metric's appeal is its capacity to acknowledge varying levels of relevance among documents within the TREC PM collections. Irrelevant documents receive a score of 0, while partially relevant and definitely relevant documents receive scores of 1 and 2, respectively. This scoring schema is crucial, as it underscores the metric's sensitivity to document order, prioritizing the retrieval of the most pertinent documents at the highest ranks. By adopting NDCG@10, we ensure a nuanced evaluation that respects the inherent variance in document relevance, reinforcing the metric's suitability for our assessment needs.

#### 5.2.5. Parameter tuning

Several of our methods require the tuning of one or a few global parameters. For this purpose, we use cross-validation [229]: We divide each TREC PM collection into 5 folds to determine the optimum parameter value(s) for the four training folds, which we then apply to the fifth (testing) fold. This way, we calculate the NDCG@10 score for each fold and then compute the overall performance as the average NDCG@10 score across all folds. Also, the parameter values quoted in the following are the means over the five training runs.

For the *field weighting* in BM25F, it turned out that the optimum weight of the bio-concepts field is 0.7, while the abstract and title fields each receive a weight of 1 (across all folds).

The *reranking* variant of sparse retrieval was tested with different values between 10 and 1000 documents to optimise the  $K$  value. Based on the result shown in Figure 5-3, we select  $K = 40$  as it constantly shows the best performance over the three TREC PM collections.

For the *partial filtering* approach, we evaluated different subset sizes to optimize the subset size (the percentage of present bio-concepts in the document from the query bio-

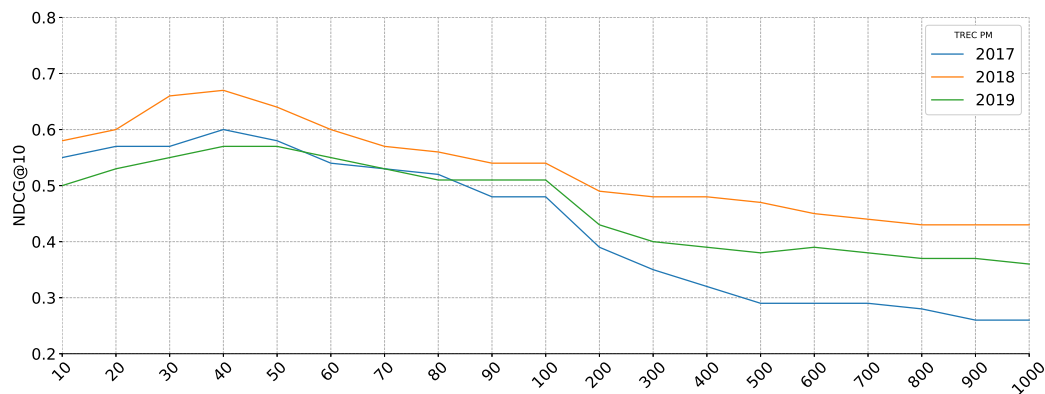


Figure 5-3.: Sparse retrieval reranking performance over different K values.

concepts). Figure 5-4 shows that this parameter has a strong influence on the results: strict filtering (100%) performs worst, while partial filtering with 50% achieves best results.

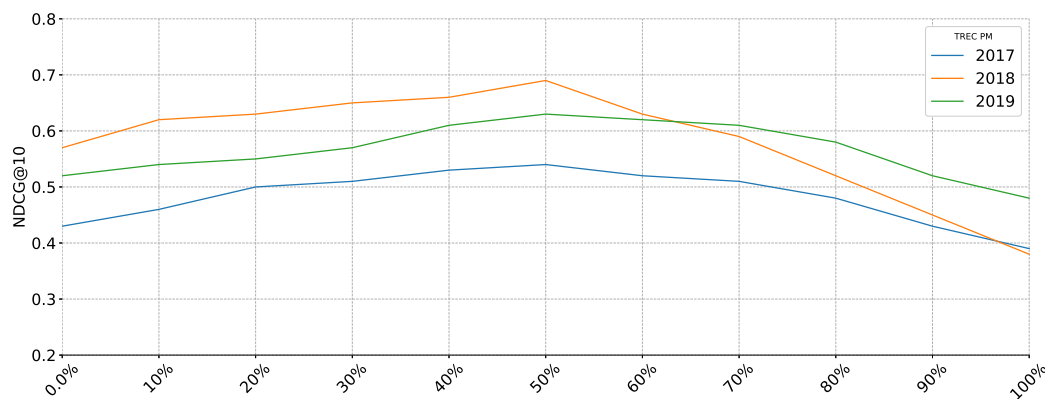


Figure 5-4.: Performance of the different filter strictness values of the document's bio-concepts subset percentage from all bio-concepts in the query.

With *Hybrid retrieval*, we performed experiments to choose the size  $K$  of the initial retrieval set, which is then reranked with the dense retrieval method. Optimizing the  $K$  value using the same method as in reranking for sparse retrieval, we received an optimum value of  $K = 50$  for this model.

### 5.3. Results

Table 5-3 outlines the performance of the retrieval models described in Section 5.2.3 across the TREC PM collections, assessed via the NDCG@10 metric, while a deeper analysis is

presented below. We only give the effect sizes of the improvements over the baselines and refrain from performing statistical tests on a re-used data set, as well as from reporting percent improvement over the baseline as suggested in [75].

Table 5-3.: Models NDCG@10 performance and effect sizes over TREC PM datasets.

Exp./TREC PM	2017	2018	2019
BM25F	0.43	0.57	0.52
Expanded	0.42 (-0.01)	0.55 (-0.02)	0.52 (+0.00)
Weighted	0.49 (+0.06)	0.62 (+0.05)	0.59 (+0.07)
Reranked	0.60 (+0.17)	0.67 (+0.10)	0.57 (+0.05)
Filter	0.39 (-0.04)	0.38 (-0.19)	0.48 (-0.04)
Partial Filter	0.54 (+0.11)	<b>0.69 (+0.12)</b>	<b>0.63 (+0.11)</b>
BiomedBERT	0.55	0.63	0.56
BiomedBERT+ bio-concepts	<b>0.56 (+0.01)</b>	0.67 (+0.04)	0.61 (+0.05)

### 5.3.1. Sparse Retrieval

The results of the sparse methods vary between positive and negative effect sizes compared to the BM25F baseline.

The Query Expanding Model, with bio-concepts incorporated as an additional field, exhibits comparable to slightly lower performance compared with the BM25F model. This might be due to the fact that bio-concepts are a different kind of terms, thus the *tf·idf* weights might not be fully compatible with those of ordinary text terms.

This problem is addressed by the Weighted Model, where the bio-concepts field got a weight of 0.7 only (vs. 1.0 for the text fields). Here we get a consistent performance improvement over the baseline, showing that integrating bio-concepts can lead to a performance improvement of a biomedical IR system.

This is also confirmed by the positive effect of the Re-Ranking Model, with positive effect sizes between 0.05 and 0.17. However, the K value (number of retrieved documents in the initial retrieval) needs to be carefully chosen because it can affect the performance significantly, as shown in Figure 5-3. The high effect size for 2017 correlates with the fact that in this case, the Baseline quality is much worse than those of the two other years. The same statement holds for the Expanded and the Weighted Model. We can only assume that the 2017 queries are somewhat different from those of the following years. Although this is not reflected in the query statistics shown in Table 7-1, the behavior of the partial filtering method for 2017 (see below) might be an explanation.

The Filter Model, which strictly only considers articles containing all bio-concepts, per-

Table 5-4.: TREC PM collections query statistics. Number of terms per query, number of bio-concepts per query, number of partially relevant documents per query, and number of fully relevant documents per query.

Year	#terms	#bio-concepts	#partially relevant docs	#fully relevant docs
2017	4.00 (Std=1.05)	3.83 (Std=0.91)	61.76	67.40
2018	3.76 (Std=1.40)	3.18 (Std=1.06)	42.92	68.84
2019	4.17 (Std=1.46)	3.65 (Std=1.07)	74.32	64.27

forms consistently less than the baseline. Analysing the performance per topic more closely, as illustrated in Figure 5-5, revealed that the strict bio-concept retrieval criteria led to some topics not retrieving any documents. However, where retrieval did occur, the filtering model consistently outperformed other strategies. The graphical representation in Figure 5-5 provides a summary of the effectiveness results for all retrieval methods.

As the Filter Model struggled when no documents that contained all bio-concepts could be found, we explored an additional, more flexible approach, the Partial Filtering Model. This model achieved the best performance and effect size of 0.11 over the BM25F baseline and achieved equal or better performance in 75% of the topics. As can be seen from Figure 5-5a, even the Partial Filtering Method fails for some topics: For the 2017 collection, 7 topics (23%) return an NDCG@10 score of 0, which might be the reason why the re-ranking method performs better here. In contrast, for the 2018 collection only four topics (8%), and in 2019 two of the topics (5%) returned a zero score. This suggests that allowing users to filter concepts, especially with partial matching, yields a notably more effective approach to enhancing retrieval. Also, in an interactive setting, users can adjust the threshold percentage of concepts to be matched, thus avoiding the zero result problem.

### 5.3.2. Hybrid Retrieval

Several studies have demonstrated that language models perform at a state-of-the-art level in NLP tasks [176, 38]. Our results show that BiomedBERT outperforms the BM25F baseline model with an average effect size of 0.07, and encoding bio-concepts into the model further improved its performance, adding 0.03 to the effect size. Simple appending the bio-concepts to the BiomedBERT model achieves comparable results with the best-performing model.

When comparing the performance per topic as shown in Figure 5-5, we found that 'BiomedBERT+ bio-concepts' not only improved the overall performance but also the result of 65% of the topics. This suggests that encoding bio-concepts in the retrieval process would improve it regardless of the retrieval method.

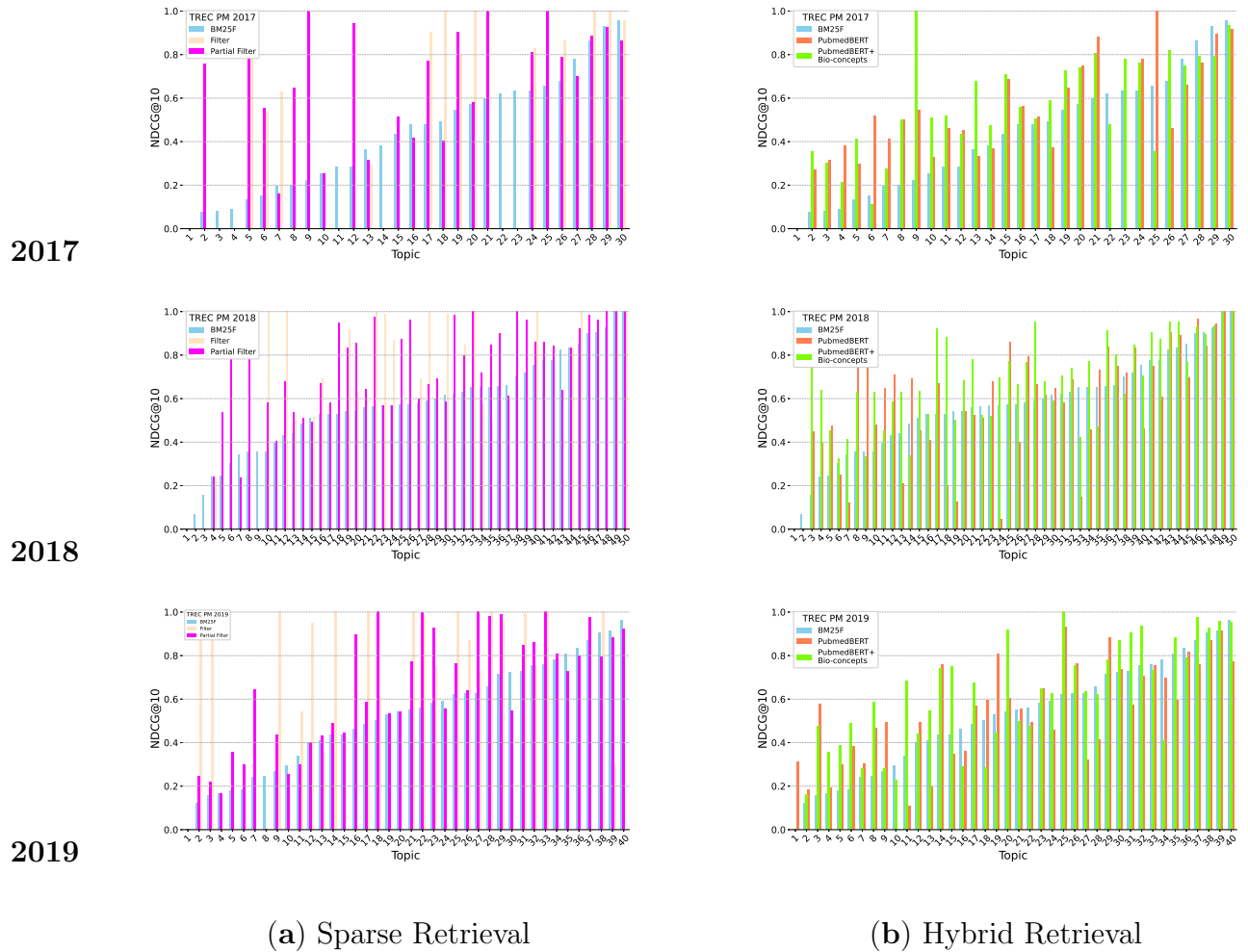


Figure 5-5.: BM25F model, Filtering models, and PubmedBERT NDCG@10 score per topic over TREC PM 2017, 2018, and 2019 collections sorted wrt BM25F.

## 5.4. Discussion

This chapter addressed three research question to investigate the value of bio-concepts for information retrieval from biomedical research literature:

- Can the explicit integration of bio-concepts into retrieval systems significantly enhance performance?
- Which integration method yields the best results?
- How do BERT-based methods compare with traditional approaches in handling bio-concepts?
- Can a synergistic approach that combines both BERT-based methods and bio-concepts

offer superior retrieval outcomes?

In response to our first research question, we can report that the findings of this study shed light on the nuanced role of bio-concepts in shaping the efficacy of retrieval models within the domain of biomedical research literature. The comparison and evaluation of diverse retrieval strategies have unearthed essential insights into their performance and practical implications. The popular biomedical search engine PubMed has adopted a learning-to-rank refinement of BM25F [69]. Therefore, there is a high probability that the improvements observed in our experiment will also carry over to the PubMed setting.

The observed lack of improvement in NDCG@10 scores for the Query Expanding model underscores a critical insight. Despite its theoretical potential to enhance retrieval by incorporating bio-concepts, the practical application did not yield a noticeable improvement in search result quality. This outcome suggests that simply expanding queries or reordering documents based on bio-concepts may not effectively increase the relevance and precision of retrieved articles. This limitation is particularly evident with the short queries typical of our TREC PM dataset, indicating that the effectiveness of partial filtering models may vary significantly with query length. (Most end-user queries in interactive retrieval are also short.) Consequently, the selection of partial filtering strictness should be meticulously tailored to accommodate collections featuring longer queries.

Commenting on the second research question, the Partial Filtering experiment marked a pivotal advancement, demonstrating that employing a lenient criterion for bio-concept matching—where only a subset of the query’s bio-concepts needs to be present in retrieved documents—significantly enhanced retrieval performance. This approach yielded a notable improvement compared to the baseline model, underscoring the efficacy of partial concept matching in the retrieval process. Recognizing that not every relevant article will contain all bio-concepts from a query, this experiment confirms the practical value and success of adopting a more flexible matching strategy.

Addressing the third and fourth research questions, our experiments with integrating bio-concepts into a BERT-based language model show that BERT also benefits from the integration of bio-concepts, demonstrating the necessity of bio-concepts for all search strategies. The results reveal that the hybrid model BiomedBERT does not outperform the sparse Partial Filtering model when bio-concepts are integrated into the search, but rather yields comparable results. However, the sparse approach has the advantage of higher transparency for users, as they see both the query terms and the bio-concepts that make a document appear in the top ranks - which helps users reformulate their queries. In contrast, the BERT reranking is more difficult to understand and thus less helpful in interactive retrieval.

A further notable performance was achieved by the Reranking model, although the results vary across the data sets. Re-ranking has potential benefits in scenarios with limited or no user interaction. This is particularly relevant in situations like the initial retrieval model



or when users lack the time to engage with filtered concepts, which is usually the case in clinical routines. Our results demonstrate that in these scenarios, re-ranking can still deliver good retrieval results and, depending on the data set, can even outperform Partial Filtering and BiomedBERT.

A recent study by Gupta et al. [89] combines n-grams and word embedding with lexical resources and uses a fine-tuned weighting method for improving retrieval. The method is applied to two collections from the TREC Clinical Decision Support tracks (a predecessor of the PM track collections used in our research), showing substantial improvements over other methods. It would be interesting to have a direct comparison of both methods; on the other hand, Gupta et al's approach lacks the transparency of our method (as illustrated in Figure 5-1), since it does not refer explicitly to biomedical concepts.

In the context of clinical application, the promising nature of our models necessitates a deeper investigation into their practical deployment in healthcare settings, where accuracy and the ability to adapt are paramount. These qualities are essential for supporting patient care and informed decision-making processes. Our research primarily concentrates on the technical efficiency of retrieval models; however, integrating considerations regarding the user experience and how professionals interpret the outcomes of these bio-concept-enhanced systems is critical. Consequently, conducting user studies or implementing feedback mechanisms to gauge the interaction with, and perceived value of, these systems by biomedical professionals will significantly contribute to advancing this area of research.

Our experiments advance the understanding of bio-concept integration in biomedical literature retrieval. It reveals gaps in current methodologies and highlights the potential of flexible, precision-driven approaches. However, the implications of this work extend beyond improving literature retrieval accuracy. Our findings can accelerate the research cycle, from hypothesis generation to clinical application, by facilitating more efficient access to relevant biomedical information. This can significantly impact areas such as drug discovery and personalized medicine.

Future research should investigate applying LLM techniques to enhance the scalability and accuracy of biomedical literature retrieval systems. It should also focus on refining these models and exploring their applicability in clinical practice, potentially revolutionizing how healthcare professionals access and utilize biomedical literature.

## 5.5. Summary

This chapter explores the integration of bio-concepts: genes, diseases, and chemicals, into retrieval systems and examines their potential to improve search precision and accuracy of the information retrieval from biomedical literature. Section 5.1 describes previous work

and explains why integrating bio-concepts into medical IR systems is beneficial. Section 5.2 describes the dataset and its annotation as well as our general methodology. Specifically, it addresses how bio-concepts can be integrated into sparse and dense retrieval models and the method of evaluating the bio-concepts contribution presented, including partial fitting as the best integration method applied in WisPerMed search engine for the user study discussed in Chapter 7. The results for each model are presented in Section 5.3 demonstrating that both sparse and hybrid retrieval strategies benefit from integrating bio-concepts, and that the performance of these retrieval strategies is comparable with significant improvements on partial filtering compared to BM25F results. These results are discussed with reference to the research questions in Section 5.4.

# 6. Level of Evidence Classification for Improving Medical Search Engine Credibility

This chapter address search result credibility as a further dimension of search result relevance for expert medical search engines. In medical research, the ability to retrieve information of high credibility is critical for informed clinical decisions. The field of Evidence-Based Medicine has already developed a framework for the assessment of medical publications' in terms of the strength of their empirical evidence base, also called Level of Evidence (LoE). Although an important factor in medical publication assessment, LoE is not usually reported along with the publication, necessitating an automatic classification method to accomplish this task. This chapter investigates a model for predicting a publication's LoE and examines the effectiveness of integrating this model into IR systems. Specifically, we address the following research questions:

- How can we train LoE classifiers?
- Are LoE classifiers trained on a specific medical subfield general enough for the whole field of medicine?
- What is the effect of LoE filtering on retrieval results?

## 6.1. Introduction

Credibility in the context of IR refers to the perceived trustworthiness and reliability of the information retrieved [197]. In medical search engines, credibility includes several factors, like the source's authority, the evidence supporting the information, and the author's credentials [217]. The quality and credibility of health information has been a challenge not only to health consumers [11] but also has been a major concern to health experts [64, 306]. These factors are crucial for medical professionals who rely on accurate and evidence-based information to make informed decisions.

Evidence based medicine (EBM) addresses this problem directly by integrating the best available external clinical evidence from systematic research into clinical practice [205, 145,

266, 245]. For this purpose, EBM critically relies on biomedical literature reviews to provide healthcare professionals with synthesized research findings, crucial for identifying effective healthcare interventions and understanding risk factors for diseases [43, 25]. EBM plays a critical role in ensuring that patient care decisions are informed by robust, current, and applicable scientific evidence, bridging the gap between research discoveries and clinical implementation.[233, 57]

Research evidence from systematic biomedical literature reviews is also used to inform further integral components of clinical practice like clinical guideline development and plays an important role in patient care counselling, where it enables healthcare providers to offer evidence-based advice and treatment options, as well as enhance patient trust and compliance with treatment protocols [140, 1].

A further element of clinical practice that crucially relies on the availability of systematic literature reviews is precision medicine, which aims to customize healthcare by tailoring medical decisions, treatments, practices or products to the individual patient. This approach heavily relies on integrating individual genetic information, environmental factors, and lifestyle into the clinical decision-making process. For this purpose, systematic reviews that aggregate and synthesise vast amounts of research data from diverse sources are integrated with patient records and genetic studies. They support the identification of effective treatments and the understanding of how individual differences affect disease processes and treatment outcomes. Precision medicine benefits from systematic reviews to ensure treatments are based on the best available evidence and tailored to individual patient needs.

To address this substantial need for a solid evidence base with high credibility, a framework for evaluating the credibility of medical documents [105] has already been developed within the field of EBM. It includes hierarchies of evidence, with clinical guidelines and systematic reviews, followed by RCTs, cohort studies, case reports, experts' opinions, and animal studies in the end, as shown in Figure 6-1. It also evaluates the used methods, expected biases, and other factors [54].

The EBM pyramid was further specified in the LoE framework, which categorizes medical research papers into 7 main distinct levels based on the strength and reliability of evidence reported [211, 54, 269]. This helps identify trustworthy information for clinical decision-making [211, 54]. This stratification, exemplified by the OCEBM<sup>1</sup> framework [105], ranges from highly rigorous and reliable systematic reviews of randomized controlled trials (Level 1a) to case studies with limited evidential value (Level 4) [26, 85]. Within this framework, each level holds unique significance, representing a specific study design and methodology [26]. The hierarchy includes the following Levels of Evidence (LoEs):

- **Level 1a: Systematic Reviews of Randomized Controlled Trials (RCTs).**

At the apex of the LoE pyramid are systematic reviews and meta-analyses of well-

---

<sup>1</sup>Oxford Centre for Evidence-Based Medicine <https://www.cebm.net/>

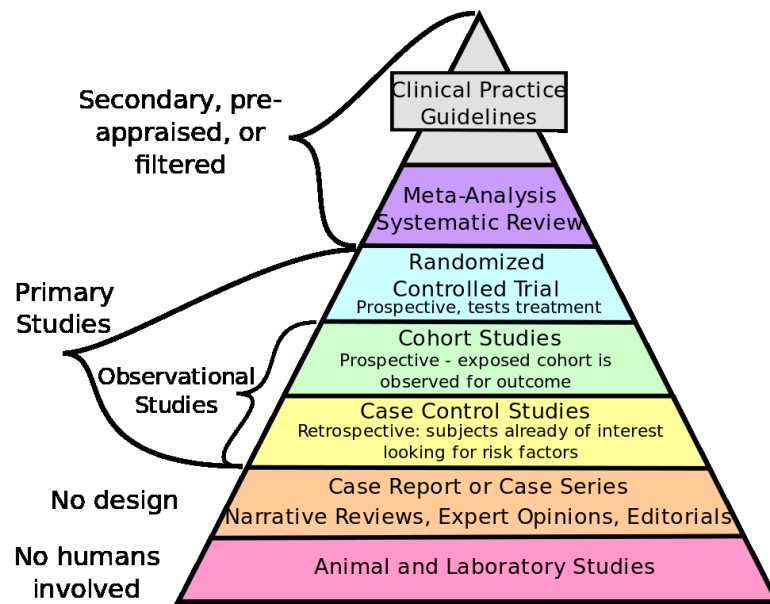


Figure 6-1.: Evidence-based Medicine Pyramid

conducted RCTs. Renowned for their comprehensive analysis of rigorous research, these reviews yield the most authoritative evidence.

- **Level 1b: Individual Randomized Controlled Trials (RCTs).** This level features individual RCTs that contribute crucial insights into causal relationships by evaluating interventions in controlled settings.
- **Level 2a: Systematic Reviews of Cohort Studies.** Systematic reviews of cohort studies provide valuable evidence regarding associations between interventions and outcomes in real-world settings.
- **Level 2b: Individual Cohort Studies.** Individual cohort studies at this level offer meaningful evidence about interventions' effects within specific populations.
- **Level 3a: Systematic Reviews of Case-Control Studies.** Systematic reviews of case-control studies extend insight into the associations between interventions and outcomes, offering a broader perspective.
- **Level 3b: Individual Case-Control Studies.** Individual case-control studies contribute evidence by exploring the relationships between interventions and outcomes within well-defined contexts.
- **Level 4: Case Series.** At this level, case series provide preliminary evidence about interventions' effects, although they are limited by their susceptibility to biases and confounding factors.

Although LoE is a crucial parameter for assessing a medical publication's significance,

it is often not explicitly stated in publications, creating a problem for medical information retrieval (IR), where the aim is to retrieve significant medical publications or their content. Recent advancements in EBM have emphasized the role of automation in enhancing the classification and credibility assessment of medical literature.

One proposal to address this was introduced by Marshall et al. [165, 166], by the RobotReviewer system, which automates the process of evaluating the risk of bias in RCTs by providing quality supporting text for bias assessment. This system provides great support in the process of preparing systematic reviews and meta-analyses of RCTs. The evaluation of the results indicated that RobotReviewer could match the performance of human reviewers in assessing the risk of bias [166, 167], which is confirmed by several subsequent studies [241, 103, 9].

A further contribution from Hartling et al. [94] highlighted the potential of such automation technologies to refine the quality and efficiency of systematic review, particularly in evaluating RCTs. These studies underscore the evolving role of machine learning and natural language processing in automating the identification and risk of bias assessment in RCTs and clinical trials, which are crucial components of the LoE framework.

One of the main limitations of these contributions is the requirement to download the complete document/medical publication for analysis, as well as the focus on only RCTs and clinical trials, which excludes the majority of publications [199]. This suggests a significant challenge for integration into medical search engines, given that not all medical publications are open-access to generate automatic access. Typically, medical search engines rely on publicly available information, such as titles and abstracts of publications. Therefore, it is important to develop a tool to extract LoE for all types of publications, not just RCTs and clinical trials, using publicly available information to facilitate easy integration into search engines.

This work addresses this problem and proposes an automatic approach to identifying and prioritizing significant works in medical research. First, we develop a classification method for automatically assigning LoE to medical publications, then we use the identified LoE as a search filter in an IR setting. We demonstrate on TREC PM 2017–2019 [208] collections that using LoE as a filter when retrieving medical papers leads to improved retrieval results, and that the gain is highest for highly evidential medical papers.

## 6.2. LoE Classifier

We view the problem of assigning LoE to medical publications as a classification task and explain in this section the training and the evaluation of the LoE classifier.

### 6.2.1. Data

We use a dataset derived from the Oncology Guidelines of the German Association of Scientific Medical Societies<sup>2</sup>. This dataset is unique in that it explicitly mentions the LoE of various medical publications as per the OCEBM framework. It includes 2816 publication - LoE pairs, extracted from unstructured PDFs.

The Oncology Guidelines mention publications as citations, which include the authors names, publication year, and publication title. This information is not sufficient for automatic LoE classification, which additionally requires some of the methodology, interventions, and clinical outcomes. This information can only be found in publication abstracts or full text. Therefore, we leverage the PubMed API<sup>3</sup> to enrich the initial dataset with abstracts and PubMed IDs.

The average word count in the abstracts is 263 (Std=97), slightly above the typical range for medical articles [7]. The prevalence of longer abstracts can be attributed to the frequent use of structured abstract formats within the medical literature [93]. Notably, we observe a positive correlation between the abstract length and the LoE classification as shown in Figure 6-1: publications with higher evidence levels tend to have longer abstracts (e.g. LoE 1a with a mean of 325 words (Std=163) than those with lower levels (LoE 3b and 4 with a mean of 233 words (Std=71)).

Table 6-1.: Average word count in abstracts per level.

Level	Word count
1a	325 (Std=163)
1b	264 (Std=74)
2a	274 (Std=93)
2b	250 (Std=59)
3a	258 (Std=75)
3b	233 (Std=71)
4	233 (Std=84)
All	263 (Std=97)

We split this data into a training dataset containing 1690 instances (60%) and a validation and testing dataset containing 563 instances (20%) each, ensuring a stratified representation across all classes.

<sup>2</sup>Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften, <https://www.awmf.org/>

<sup>3</sup><https://pubmed.ncbi.nlm.nih.gov/>

### 6.2.2. Experimental Setup

For the task of LoE classification, we focus on fine-tuning PubMedBERT [86]. PubMedBERT is a natural choice for this domain-specific classification task as it is a transformer-based model pre-trained using abstracts sourced directly from PubMed. Its efficacy has been well-established: It currently holds the top score on the Biomedical Language Understanding and Reasoning Benchmark, it excels in accurately interpreting the unique terminologies and context of biomedical texts, and it is proficient in handling the complexities of biomedical literature. The model is fine-tuned using the training set and hyperparameters are optimized using the validation set.

We develop the following classifiers:

**Random Forest (RF)** RF serves as our baseline. It is trained on the training set for multi-class classification. We use TF-IDF vectorization and chi-squared feature selection, and K-Fold cross-validation using the validation dataset, evaluating its performance with the macro-F1 score.

**Multi-Class-PubMedBERT** This classifier is directly fine-tuned on the training set to classify texts into specific LoE classes, with the macro-F1 score as the evaluation matrix.

**Reg-PubMedBERT** This is a regression approach, which assigns numeric values to LoE classes. PubMedBERT is fine-tuned to predict these values, by mapping different LoEs (1a, 1b, 2a, 2b, 3a, 3b, 4) to their respective numeric values (0, 1, 2, 3, 4, 5, 6). We used root-mean-square error (RMSE) for evaluation. To align the model's predictions with the original LoE classes and facilitate comparison with other classifiers using the F1 matrix, we mapped the predicted value to the nearest integer value and then used the same map to get predictions back to their corresponding LoE classes.

**Multi-Label-PubMedBERT** This classifier incorporates the multi-label classification approach, i.e. we transform the LoE categorization into a set of binary labels. Each label corresponds to a specific LoE class, effectively converting the problem into a multi-label classification task. This version enabled PubMedBERT to predict multiple labels simultaneously, accommodating the scenario where only one of the labels should be true while others are false. By modelling the LoE classification as a multi-label task, we aimed to capture potential overlap between LoE classes and assess the model's capacity to handle such nuances by looking at the prediction list that might contain multiple levels of evidence. For proper evaluation, we assigned the highest confidence value when multiple positive predictions.



**Ensemble Majority Vote** Ensemble methods are a well-established technique in classification that capitalizes on the strengths of diverse classifiers to enhance prediction accuracy and generalization [201]. We employed an Ensemble Majority Vote strategy, combining the strengths of the three PubMedBERT models (Multi-Class, Reg, and Multi-Label). This approach used majority voting to aggregate predictions from each model, enhancing the overall classification accuracy and robustness [309, 50].

### 6.2.3. Classifier Evaluation

We evaluate our LoE document classifiers using Macro F1 score, RMSE, and Confusion matrices.

#### Individual Classifiers Performance

Table 6-2 summarizes the performance of each classifier on the test dataset.

Table 6-2.: Level of Evidence Classifiers Performance on our test set. Macro F1 Score.

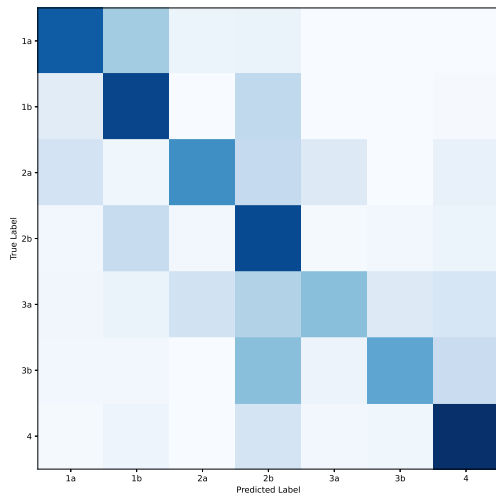
Model	F1 score	RMSE
Random Forest (RF)	0.59	1.30
Multi-Class-PubMedBERT	0.78	0.90
Reg-PubMedBERT	0.74	0.69
Multi-Label-PubMedBERT	0.79	0.90*
Majority voting	<b>0.83</b>	<b>0.65</b>

\* By considering the label of the highest confidence score as predicted class

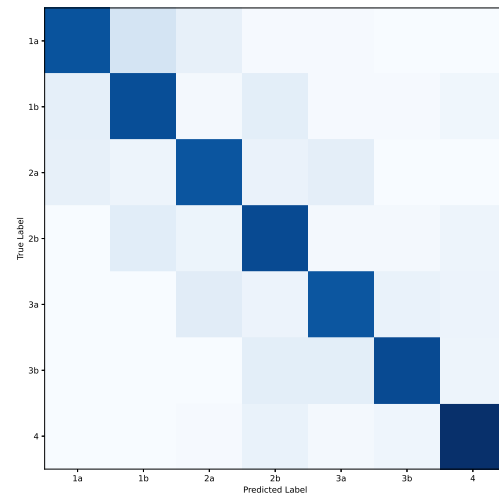
**RF Baseline** The RF model’s performance with a macro-F1 score of 0.59 and an RMSE of 1.30 did not surpass the deep learning models’ results. Nevertheless, the RF model shows robustness in effectively handling the challenges of multi-class LoE classification, as shown in the confusion matrix in Figure 6-2a.

**Multi-Class-PubMedBERT** scored 0.78 in F1 (+0.19 compared with baseline) and 0.90 in RMSE, showing effectiveness in multi-class categorization. However, after we analysed misclassification, we found that the model has some difficulties distinguishing closely related LoE classes as presented by the confusion matrix in Figure 6-2b.

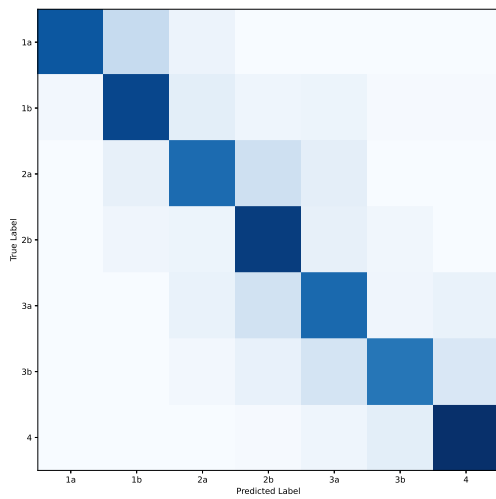
**Reg-PubMedBERT** exhibited strengths in capturing the ordered nature of LoE with an F1 score of 0.74 and the second-best RMSE of 0.69, indicating proficiency in differentiat-



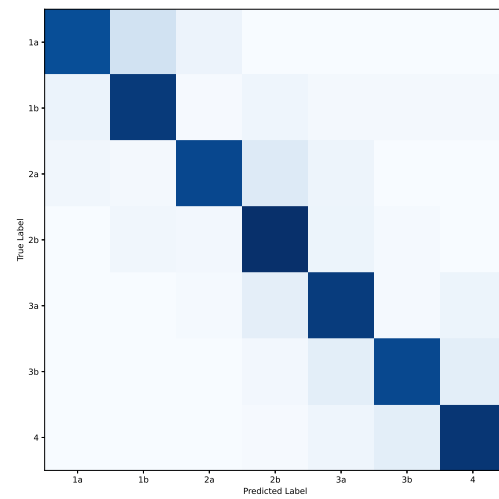
(a) Random Forest as a baseline



(b) Multi-Class-PubMedBERT



(c) Reg-PubMedBERT



(d) Majority voting

Figure 6-2.: Confusion Matrices using the test set per model.

ing between levels. This makes misclassified documents closer to the true labels, which is reflected in the smaller RMSE, as shown in the confusion matrix in Figure 6-2c.

**Multi-Label-PubMedBERT** performed best among individual classifiers with an F1 score of 0.79, adeptly handling documents with multiple LoE categories. A closer qualitative examination of this model’s performance revealed that some documents were assigned into

multiple LoE classes. This is a well known phenomenon, which was explored in the work of Murad et al. [183] bringing into question the clear demarcation between the evidence levels of the EBM pyramid. Instead, a nuanced perspective on LoEs has been proposed to align with the flexibility of multi-label classification as demonstrated by Multi-Label-PubMedBERT.

### Ensemble Majority Vote Performance

The Ensemble Majority Vote method combines the predictions of all three PubMedBERT models and demonstrates the best performance. It scores highest in F1 (0.83) and achieves an RMSE of 0.65, indicating its effectiveness in accurately categorizing medical literature by LoE, as shown in the confusion matrix in Figure 6-2d. This result emphasizes the significant role of collaborative intelligence in enhancing classification outcomes.

### Statistical Significance Analysis

We performed a statistical significance analysis on our machine learning models using a paired t-test. After applying Bonferroni correction ( $\alpha = 0.05/10$ ), we found that all deep learning models significantly outperformed the Random Forest baseline, indicating their effectiveness in LoE classification. However, no significant performance differences were observed among the deep learning models themselves, highlighting their comparable efficacy in evidence-based classification.

### Identifying Significant Terms

We utilized the LIME (Local Interpretable Model-Agnostic Explanations) explainer [206] to identify key terms influencing our model's predictions for different Levels of Evidence categories. This method provides insights by aggregating term scores, helping us to determine significant terms for each LoE level as shown in Figure 6-3. Such an approach enhanced the interpretability and transparency of our model, highlighting LoE-specific terms in the analyzed documents.

As shown in table 6-3, we analysed the top 10 contributing terms across the LoE levels in the test set. The results highlighted that our model was able to identify discriminating terms for each class. Moreover, we discovered common terms shared across multiple levels, such as “systematic review” in 1a (systematic reviews of RCTs), 2a (systematic reviews of cohort studies), and 3a (systematic reviews of case-control studies), and “RCT” in 1a and 1b (individual RCTs). Additionally, some less expected terms, like “risk” in 2a, 2b (individual cohort studies), 3a, and 3b (individual case-control studies), and “accuracy study” in 1a, 2a, and 3a (pertaining to Diagnostic Test Accuracy studies), emerged as significant classifiers.



**Text with highlighted words**

background saliva fundament oral health well mani factor impair saliva secret aduers effect prescrib medic auto immun diseas exampl sjogren syndrom radiotherapi head neck cancer sever studi suggest posit effect acupunctur oral dryness method pubm web scienc electron search refer list includ studi relev review manual search studi met inclus criteria systemat evalu two review assess includ studi confirm elig assess risk bia result ten random control tria investig effect acupunctur includ five tria compar acupunctur sham placebo acupunctur four tria compar acupunctur oral hygien usual care onli one clinic tria use oral care session control group for includ studi qualiti main outcom assess low although public suggest posit effect acupunctur either salivari flow rate subject dri mouth feel studi inconclus potenti effect acupunctur conclud insuffici evid avail conclud whether acupunctur evid base treatment option xerostomia hyposaliv further well design larger doubl blind tria requir determin potenti benefit acupunctur sampl size calcul perform initi studi

Figure 6-3.: Classification Explanation for a Document using LIME.

Interestingly, a specific therapy (“acupuncture”) only occurs among the terms of level 4, possibly indicating the lack of stronger evidence for this method.

The analysis of the LIME results suggests that the classifier is applicable across all fields of medicine, demonstrating its broad applicability and not being confined to specific areas. This finding also implies that the dataset used for training is robust and well-suited for developing LoE classifiers that can be effectively integrated into document classification systems.

### 6.2.4. Classifier Discussion

In this section, we report a method for classifying medical documents according to the empirical LoE, a framework already practiced in evidence-based-medicine. To this end, we addressed the first research question:

- How can we train LoE classifiers?

We have effectively demonstrated the automated application of the LoE framework for improving the retrieval of relevant medical publications. Our approach, leveraging fine-tuned PubMedBERT models, has proven adept at classifying medical publications based on their LoE with a high degree of accuracy (macro F1 = 0.83). This advancement addresses a significant gap in existing literature, where previous studies have largely focused on specific evidence levels, particularly RCTs and their systematic reviews. The higher transparency of our approach gives users full control over the LoE of the documents returned. Moreover, the method investigated here could be directly integrated into the existing PubMed search engine, by simply adding estimated LoE as an additional document attribute that can be referred to in the query.

Table 6-3.: Significant Terms in the Level of Evidence Classifier.

1a		1b		2a		2b	
term	score	term	score	term	score	term	score
accur predict	2.11	achiev complet	1.92	cohort studi	1.30	cohort studi	1.62
accur stage	1.85	achiev patient	1.91	accuraci detect	1.14	accrual	1.42
accuraci respect	1.72	activ control	1.58	systemat review	1.09	acquisit	1.14
rect	1.42	activ intervent	1.56	meta analysi	1.02	accept	1.11
meta analysi	1.31	activ surveil	1.25	exposur	0.98	access	1.08
systemat review	1.30	rect	1.21	longitudin	0.95	accru	1.01
accuraci studi	1.17	control set	1.12	access	0.74	longitudin	0.89
accuraci clinic	1.16	acut delay	0.98	accur stage	0.73	risk	0.61
achiev	1.15	acut	0.79	accuraci studi	0.64	administr	0.21
activ treatment	1.02	adjuv	0.71	risk	0.59	affect patient	0.14

3a		3b		4	
term	score	term	score	term	score
systemat review	1.24	case control	1.60	small sampl	1.69
epidemiolog	1.21	case definit	1.41	preliminari evid	1.32
case definit	1.17	exposur	1.02	exploratori research	0.99
abnorm	1.12	risk	0.49	uncontrol studi	0.98
exposur	1.11	advers reaction	0.31	acupunctur treatment	0.68
absent	0.98	affect patient	0.30	patient characterist	0.60
accuraci respect	0.88	age	0.29	acupunctur effect	0.51
accuraci studi	0.71	age diagnosi	0.23	analysi reveal	0.22
risk	0.64	advers effect	0.19	analysi identifi	0.22
accur stage	0.51	affect surviv	0.10	affect	0.13

Although our experiment shows the potential of using LoE in Medline, one limitation that needs to be considered is the potential bias from using the oncology guideline dataset for training the classifiers. Medline collection contains publications where LoE can't be applied, such as bioinformatics. To apply it in real-world applications, we could introduce a new class, "others", where the model confidence score is below a threshold or when multiple positive labels are in the multi-label classifier.

### 6.3. Levels of Evidence as a filter in medical IR

In this experiment, we investigate the benefit of LoE classification for the IR of medical publications using TREC Precision Medicine (PM) datasets from 2017 to 2019 [208, 207, 209].

### 6.3.1. Data

We used the TREC PM dataset in evaluating the effectiveness of LoE in IR as it was used to evaluate the effectiveness of bio-concepts (Section 5.2.1). It is important to mention that the criteria for relevance did not refer to the LoE of the documents.

We categorize each abstract in the Medline collection into its respective LoE category using our ensemble classifier. Figure 6-4 shows the distribution of LoE classes in Medline data. Most frequent are Level 4 documents (41% of the collection), which require the smallest empirical basis. The highest LoE 1a and 1b each represent only 7% of the documents. This distribution highlights the predominance of lower-evidence articles in medical literature and underscores the importance of our approach in focusing on evidence quality in IR.

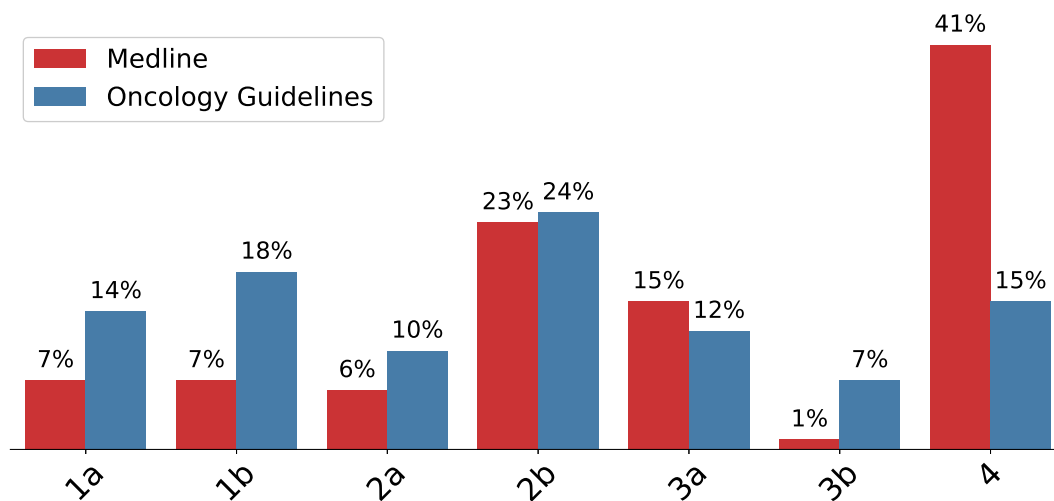


Figure 6-4.: The distribution of LoE Classes in the Medline Dataset and Oncology Guidelines (Classifier Dataset).

### 6.3.2. Experimental Setup

Our experiment utilizes the BM25 retrieval method applied to documents of all LoE classes ('All') as a baseline for our IR process [210]. The impact of LoE classification is tested by filtering the documents based on their LoE as follows:

- *LoE3+*: LoE categories 3b to 1a, i.e. case-control studies or higher LoE.
- *LoE2+*: LoE categories 2b to 1a, i.e. cohort studies or higher LoE.
- *LoE1*: LoE categories 1a and 1b, i.e. RCTs only.

The performance of each model was assessed using infNDCG, R-Prec, and P@10 matrices, as these are the official matrices used to report on the datasets. Also, we report the “Normalized discounted cumulative gain @10” (NDCG@10) matrix, which considers the position of relevant documents, giving higher weight to documents appearing earlier in the search results. We chose it as LoE is a user-centric application, making the early results more likely to be seen and used. As we are re-using a test collection, performing statistical tests here would contradict statistical testing theory [75]. Instead, we give the effect sizes, which indicate substantial improvements over the baseline.

### 6.3.3. Results

As shown Table 6-4 using LoE to filter out document set to be searched improves the retrieval effectiveness as measured by NDCG@10 score. The retrieval of RCT documents with highest LoEs is the most successful. Moreover, there is a clear trend in improving NDCG when the minimum LoE is increased. For all three collections, the strictest filter (LoE1) outperformed all other methods, with substantial NDCG improvements (0.08 ... 0.11) over the baseline. Moreover, as shown in Table 6-5, our LoE1 model improved the performance of the baseline on all matrices. It also outperformed each of the best-reported runs on infNDCG matrix and provided comparable results on R-Prec<sup>4</sup>. In addition, the retrieval quality of our method is accompanied with the guarantee of returning only documents of the highest evidence.

Table 6-4.: Models’ NDCG@10 performance on TREC PM datasets

Exp./Year	2017	2018	2019
<i>All</i>	0.46	0.59	0.54
<i>LoE3+</i>	0.48 (0.02)	0.60 (0.01)	0.57 (0.03)
<i>LoE2+</i>	0.49 (0.03)	0.64 (0.05)	0.58 (0.04)
<i>LoE1</i>	<b>0.54</b> (0.08)	<b>0.69</b> (0.10)	<b>0.65</b> (0.11)

Numbers in parentheses show the effect size when comparing with the baseline “All”.

<sup>4</sup>Note that these are pessimistic estimates, as unjudged documents only retrieved by our method are treated as irrelevant

Table 6-5.: Models' InfNDCG/R-Prec/P@10 performance on TREC PM datasets.

Exp./Year	2017	2018	2019
<i>All</i>	0.43/0.27/0.52	0.50/0.32/0.58	0.47/0.30/0.57
<i>LoE3+</i>	0.45/0.28/0.54	0.52/0.34/0.60	0.50/0.31/0.58
<i>LoE2+</i>	0.47/0.28/0.54	0.55/0.36/0.61	0.52/0.31/0.61
<i>LoE1</i>	<b>0.52/0.30/0.55</b>	<b>0.57/0.38/0.61</b>	<b>0.58/0.34/0.61</b>
<i>Top run</i>	0.46/ <b>0.30/0.64</b>	0.56/0.37/ <b>0.71</b>	<b>0.58/0.36/0.65</b>

Best reported runs per matrix, meaning model performing best on P@10 not same as model performing best on infNDCG.

## 6.4. Discussion

In this chapter, we report a method for increasing credibility of medical search engines such as WisPerMed as discussed in Section 6.2.4, by classifying medical documents according to the empirical LoE, a framework already practiced in evidence-based-medicine, but to date not integrated into search engines. To this end, we addressed the following research questions:

- How can we train LoE classifiers?
- Are LoE classifiers trained on a specific medical subfield general enough for the whole field of medicine?
- What is the effect of LoE filtering on retrieval results?

In this chapter, we have developed an effective method for automating the estimation of LoE medical publications. This method is integrated into filtering retrieved documents based on the LoE. A key finding of our work is the effect of LoE filtering in directing attention towards the most reliable 14% of documents, while enhancing retrieval quality at the same time. This aspect is particularly crucial in the medical domain, where accessing accurate and high-quality information rapidly can make a pivotal difference in patient care and medical research. On the other hand, LoE2 or LoE3 papers may also be searched for in case there are no relevant answers at the top level of credibility, e.g. when the user is interested in more recent methods for which higher level studies are not available yet.

In our experiment, the LoE1 model outperformed the best-reported runs on the three datasets [208, 207, 209] in terms of infNDCG and provided comparable results in R-Prec matrix. This demonstrated the effectiveness of using LoE as a filter in medical IR, improving the relevance and reliability of retrieved documents. These improvements over integrating the LoE filter in the BM25 baseline suggest that these benefits could extend to the other stronger baselines.

In the user study detailed in Chapter 7 with medical professionals, we investigate if these



experimental enhancements also lead to measurable benefits in real-world clinical settings. By engaging real users in this research, we intend to validate the practical utility of our LoE integrated system in enhancing search experiences and outcomes.

## 6.5. Summary

This chapter addresses a key credibility challenge faced by current medical search engines for experts, which lies in identifying significant, evidence-backed medical publications. Although relevant and widely used in evidence-based medical practice, the Level of Evidence framework has not yet been fully automatised and tested for medical IR (Section 6.1). We introduce a classification model for tagging medical research abstracts with LoE levels and demonstrate that a vast number of medical publications without LoE tags can be successfully and fully automatically enriched with this crucial information (Section 6.2). Our retrieval results confirm that LoE is an effective filter that improves results in a fully automatic retrieval scenario (Section 6.3). As we discuss in Section 6.4, these results suggest that our LoE based approach to medical IR is a viable and robust tool to evidence-based medical practice, which can facilitate and improve medical decision-making, leading to better patient care.



# 7. User Evaluation of WisPerMed

This chapter evaluates WisPerMed, a search engine for medical literature which implements the methods for medical search engine optimization developed and described in the previous chapters. Having developed and evaluated the various building components that aim to improve medical information retrieval, we now have all the building blocks in place for an advanced medical search engine. The next step is to evaluate whether these methods truly enhance the efficiency and helpfulness of information retrieval for medical professionals. The aim is to evaluate WisPerMed from a medical expert user's perspective in an information search scenario that simulates a realistic information search at the point of care, where a clinician needs to research treatment options for a patient.

In this chapter, we focus on evaluating Levels of Evidence (LoE) and bio-concepts, excluding personalization aspects. Assessing personalization would require the creation and refinement of user profiles, a process demanding significant time and user engagement. The resource-intensive nature of this task does not justify the benefits, as the effectiveness of personalization in enhancing search engine performance is already well-documented across many domains. This makes its advantages less contentious compared to the relatively novel aspects of LoE and bio-concepts.

Therefore, the user evaluation in this chapter aims to answer two research questions:

- Does using WisPerMed improve the efficiency of search compared to PubMed, a standard search engine for medical literature?
- Do users find the credibility assessment by Levels of Evidence helpful for the search task?
- Do users find interactivity through highlighting bio-concepts helpful for the search task?

## 7.1. Introduction

As Chapters 5 and 6 demonstrated, modelling Level of Evidence (LoE) framework and the integration of bio-concepts improve the performance of information retrieval systems, when the task is to retrieve medical research papers. However, improvement is best seen as a

user-centric, task-focused category [53], and information retrieval research should focus on helping users understand the information they retrieve, rather than merely improving the retrieval process. This necessitates a user study that evaluates medical expert users' search behaviour in realistic settings and measures the efficiency and efficacy of their search.

Our retrieval results suggest that LoE and bio-concepts would improve search performance in real world situations, where medical experts need an efficient access to scientific publications from large repositories of biomedical literature such as MEDLINE. However, a search engine that successfully integrates LoE and bio-concepts into improved user interfaces to support decision-making is as yet unexplored.

The available medical search engines, most prominently PubMed, provide sophisticated similarity-based retrieval [177, 122], which involves ranking documents based on their similarity with the query terms using advanced IR algorithms such as vector space models and deep learning techniques to ensure relevant results. These developments in the field of medical information retrieval systems and other proposals [90, 80] focus on improving effectiveness and efficiency by considering not only the time but also the accuracy, relevance, and accessibility of search results. Advanced techniques such as semantic query expansion and tensor factorization are being integrated into search engines to better handle complex medical queries and incorporate domain-specific knowledge, thereby improving retrieval performance [275]. Other systems integrated technologies such as personalized search algorithms, which adapt to user profile preferences and historical search behaviour to provide more relevant results [170].

Despite the progress, however, there is still a lot of space for improvement with respect to addressing practitioners' search needs. Often medical professionals are left with extensive lists of publications, whose relevance and level of credibility and authority they need to establish [100, 47]. Such search is essentially inefficient, which can be problematic in urgent clinical scenarios, such as identifying the latest treatment options for diseases like multi-drug-resistant tuberculosis or assessing therapeutic interventions during public health emergencies [203, 160]. The information search in medical publication repositories can be both time-consuming and overwhelming, particularly for those requiring quick access to evidence-based medical information [222, 71] and state of the art search engines for biomedical literature still need substantial improvement before they can successfully address medical professionals' information search needs.

To address this gap, we incorporate Levels of Evidence (LoE) and bio-concepts into WisPerMed, our search engine for retrieving publications from medical paper repositories. WisPerMed allows users to filter search results based on LoE and bio-concepts, highlights bio-concepts in the text, and displays the LoE for each article, thereby enhancing the identification of relevant publications. Additionally, bio-concepts are presented in a word cloud format to assist users in quickly assessing document relevance.

To evaluate the efficiency and efficacy of search using WisPerMed, we conducted a user

study, in which we compare WisPerMed with PubMed, a standard search engine [188] for biomedical literature. The study involved 131 medical experts performing predefined search tasks on both platforms in a controlled, randomized order. This chapter presents the design and development of our search engine, outlines the methodology of our user study, and provides a comparative analysis of performance and user experience relative to PubMed [188].

## 7.2. Methodology

To evaluate the efficiency of the integration of LoE and bio-concepts in biomedical search engines, we designed the user study to gather quantitative and qualitative data to assess search efficiency with specific search engine features. For that, we first developed WisPerMed, a medical search engine that integrates LoE and bio-concepts. LoE enables users to filter the retrieved documents based on their evidence level, such as retrieving documents with LoE between 1a and 2b. Bio-concepts allow users to filter out documents that do not mention selected bio-concepts, present all bio-concepts in the text in a word cloud, and highlight bio-concepts within the text. We also developed a comparable interface without the WisPerMed features with the PubMed API for retrieving documents. This setup allows us to evaluate the feature's efficiency using a comparative, within-subjects design.

### 7.2.1. Implementation

In this section, we describe the implementation of key features in WisPerMed. We focused on integrating two main components: Level of Evidence (LoE) and bio-concepts, to enhance the search efficacy for medical professionals. To make sure that our search engine is fast in responding to user queries, we stored all LoE and bio-concepts for all Medline articles in a database. Therefore, once we receive a query from the user and retrieve the documents from PubMed API, we tag the documents with feature values and send them to the interface to display.

#### Integration of LoE

Our search engine automatically incorporates a classification model to assign LoE to biomedical publications. This classifier was proposed in Chapter 6 and trained on a labelled dataset of MEDLINE documents, utilizing a combination of text analysis and machine learning techniques. The classifier achieved an F1 score of 83% on a separate test set and kept the misclassified documents as close as possible to the true label, indicating strong performance in identifying the LoE. Based on the OXFORD framework [105], the model classifies literature abstracts into seven distinct levels of evidence (1a, 1b, 2a, 2b, 3a, 3b, 4). Using this

classifier, we enabled users to filter search results by their LoE, allowing them to identify the most authoritative and relevant publications quickly.

As shown in Figure 7-1, we decided to design the LoE element to the retrieved documents based on a concept most users are aware of as Nutri-Score [99]. This design aims to display the LoE level using a grading system users are familiar with, without the need to review external resources outside the search engine.



Figure 7-1.: Level of Evidence.

### Integration of Bio-concepts

Integrating bio-concepts involved developing an NLP pipeline to identify and extract genes, diseases, chemicals, and others from the text of publications. Bio-concept extraction was facilitated using PubTator API, which has been discussed in several research papers [280, 281]. Moreover, integrating bio-concepts into IR and its effectiveness from a system perspective is addressed in Chapter 5. From a search engine interface perspective, these bio-concepts were then highlighted in different colours within search results, and a word cloud representation of the most prevalent bio-concepts in a document was generated to aid in quick relevance assessment, as shown in Figure 7-2. The interface also allowed users to filter results based on the bio-concepts by selecting them from the word cloud or the filters under the search bar.

### 7.2.2. User Study Setup

The study setup was cleared by our institution's ethics committee under approval number "2311ISFS0138". The study was conducted in a controlled environment, ensuring that each participant used both the developed two search interfaces:

1. PubMed: a baseline interface of PubMed with basic functionalities such as a search bar and two filters (publication type and publication year). Using the PubMed API, we retrieved the documents and presented the top 10 results to users as shown in Figure 7-3a.
2. WisPerMed: this interface is similar to the PubMed baseline interface, enriched with LoE and bio-concepts filters. Additionally, we show the LoE indicator for each doc-

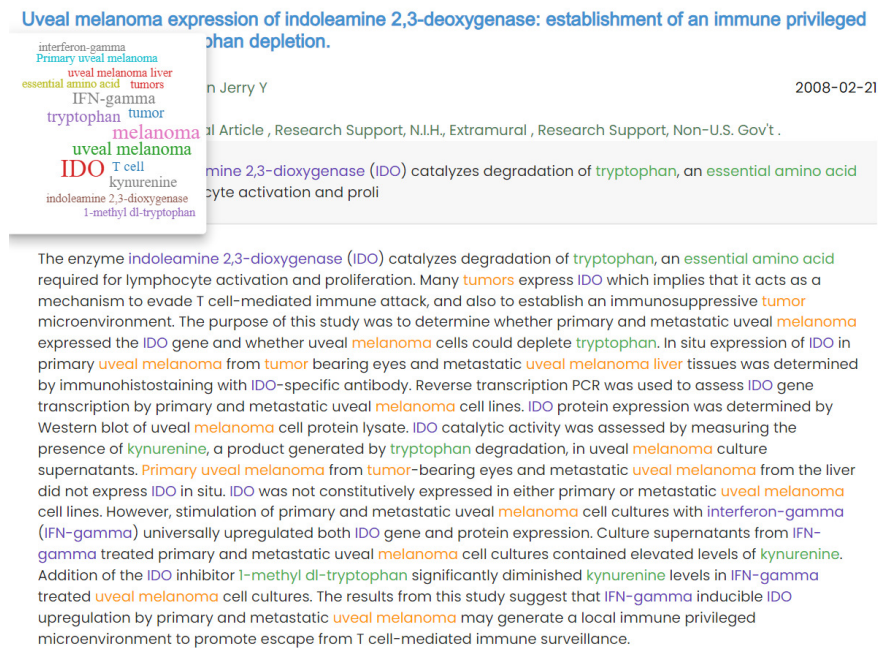


Figure 7-2.: Screenshot from WisPerMed on a selected article with highlighted bio-concepts.

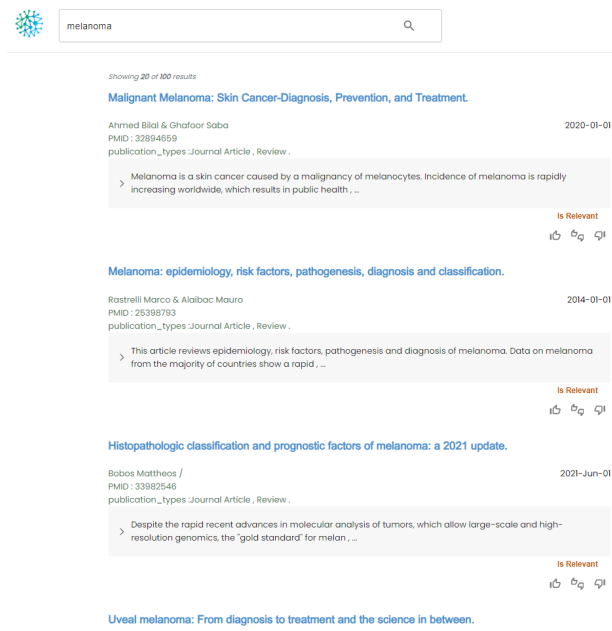
ument in the retrieved list. Bio-concepts are highlighted in the text, and interactive bio-concepts are provided for each article as shown in Figure 7-3b.

We decided to use the PubMed baseline interface and not the actual interface to control any learning effects from using an interface they are familiar with. Since each user performs a search task on each search engine, we kept the search engines' order randomized to mitigate order bias and learning effects. By taking these measures into the user study setup, we aimed to reduce the effect of biases and confounding factors.

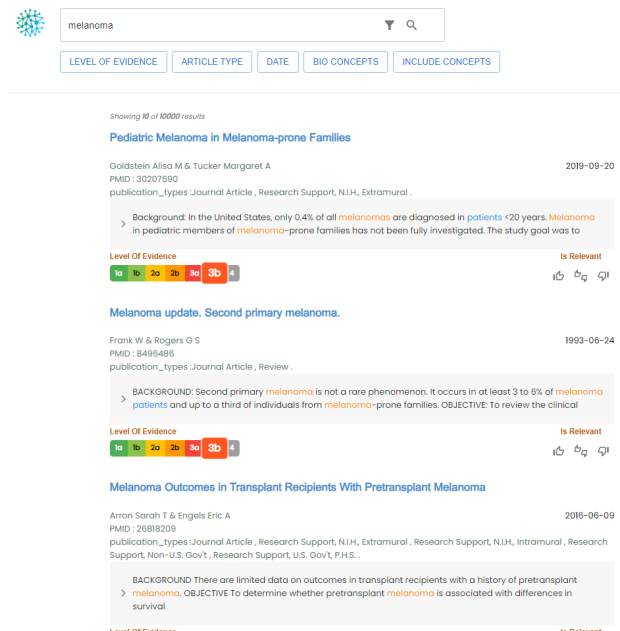
## Participants

The target participants in the user study were medical experts. We recruited a total of 131 participants, comprising medical students and practising doctors, to capture a broad spectrum of experiences and perspectives within the medical field. This approach acknowledges the wide variance in experiences and knowledge from the academic training phase to real-world clinical practice. All participants volunteered without any compensation.

Participants were recruited from three German institutions. The responses were categorized into two professional groups: 79 (60%) medical students and 52 (40%) practising doctors, as shown in Figure 7-4a. Of these participants, 68 (51%) started the search task using the PubMed search engine, while 63 (49%) began with WisPerMed, as shown in Figure 7-4b. Specifically, 30 medical doctors started with PubMed for the first search task,



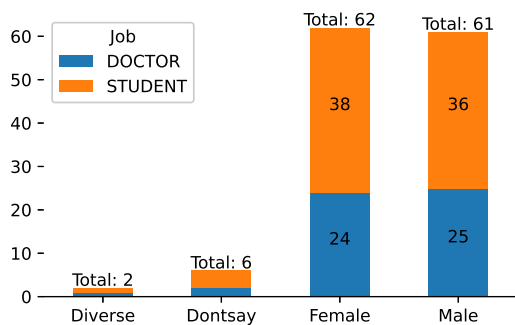
(a) PubMed



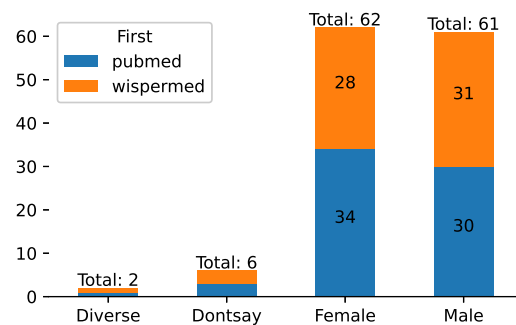
(b) WisPerMed

Figure 7-3.: Screenshots from PubMed and WisPerMed interfaces of the retrieved list of documents.

while 22 doctors started with WisPerMed. Similarly, 38 students began with PubMed, and 41 students started with WisPerMed.



(a) Participants' professional categories.



(b) First search engine.

Figure 7-4.: Participants' gender distribution over professional categories and first search engine to start the search task with.

The gender distribution of participants was as follows: 62 female, 61 male, 2 diverse, and 6 undisclosed, as shown in Figure 7-4. The age distribution had a mean of 26.4 years (Std=4.6, min=18, and max=40).

As part of the demographic information collected, we assessed the participants' profi-



ciency in English, given that the study was conducted in Germany. All participants had English language skills ranging from upper intermediate (B2) to professional English (C1). We also inquired about their experience levels: years of study for students and years of professional experience for doctors. As shown in Figure 7-5a, students were between the first and seventh years of study, with a mean of 4.2 years. Doctors had between one and ten years of professional experience, with a mean of 3.8 years. This mix was intended to capture a diverse range of experiences and perspectives within the medical field.

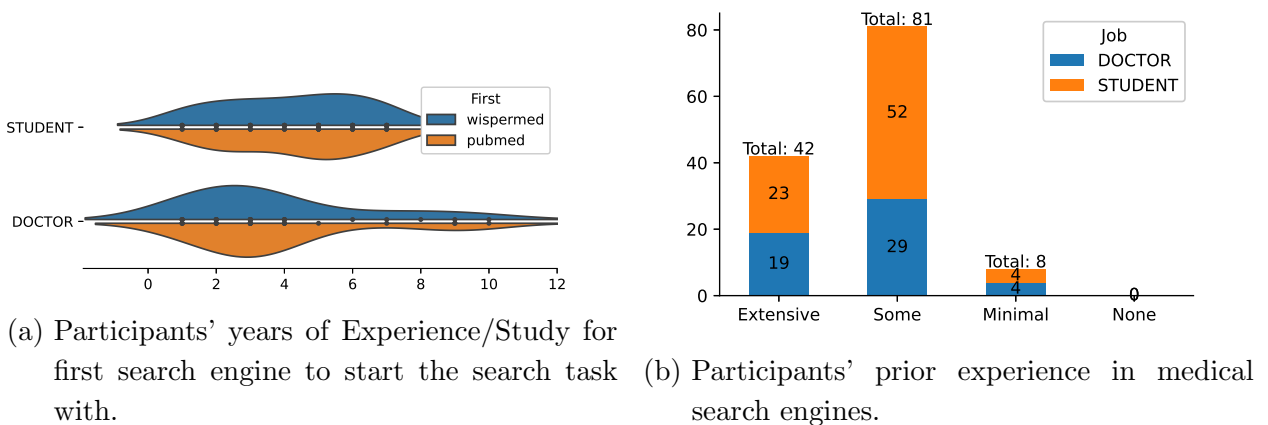


Figure 7-5.: Overview of Participants' Experience and Background in Medical Search Engines

The final demographic question explored the participants' previous experience with medical search engines, such as PubMed. Participants were asked to rate their experience on a scale from "Extensive experience" to "No experience." As shown in Figure 7-5b, over 93% of participants had experience in the top two categories. Only 8 participants reported having minimal search experience, and none reported having no experience at all.

## Scenario

Since each user needed to try every search engine, we had to design two search scenarios (Patient cases) of comparable complexity. These tasks were developed based on everyday scenarios encountered by medical professionals, which could require the retrieval of information on specific diseases, treatments, and recent research findings. The Scenario was presented in German, as the participants usually deal with patients in German. This is the English translation of the scenario and patient cases.

"Imagine you are a dermatologist interested in the field of dermatological oncology. You are a person who values both the accuracy of information and the efficiency of finding relevant information.

Imagine you have received two patients suffering from a certain type of skin cancer. You need to find the best treatment options for each of the two patients. You will use two different search engines to perform this task. Each search engine will be used for one patient.

For each task, you will be asked to find 10 relevant articles for the treatment to decide which treatment strategy is the best for this patient.”

Patient descriptions (assigned in random order):

“The patient has received a first-time diagnosis of metastatic melanoma with brain metastasis.”

“The patient was diagnosed with Stage III basal cell carcinoma.”

## Procedure

Due to the busy schedules of medical practitioners, conducting a training session on the WisPerMed search engine was not an option. Therefore, participants saw the search engine for the first time during the user study.

The user study was conducted in person. Participants were greeted and sat down in front of a provided laptop with the study opened in a browser window. They first received a digital information letter and gave informed consent. They then completed the demographic questions (occupation, years of experience, age, gender, search experience, language level) and proceeded to a description of their task: finding ten relevant articles for a given medical condition (see Tasks). Upon confirmation, they started with the first search engine (order randomized) and the first medical condition. The description of the medical condition was displayed below the search engine in the browser window throughout the whole search session. After marking relevant articles and clicking on 'finish the search', they were confronted with the second scenario and second search engine. After completing the second search task, they filled in the post-task questionnaire about LoE, bio-concepts highlighting, and word cloud, and were thanked in a debriefing note.

### 7.2.3. Data Collection

Data collection focused on quantitative and qualitative data such as time taken to complete each search task, number of queries executed, responses to specific questions on a Likert scale, and an open-ended section for detailed feedback. These measures are designed to align with our research goals of evaluating the efficiency of integrating LoE and bio-concepts into biomedical search engines. This aims to improve search performance and meet the information needs of medical professionals.

The collected data can be categorised into two categories:

### **Quantitative Measures**

Quantitative measures were used to assess the efficiency of the WisPerMed compared to the baseline PubMed search engine. During the search tasks, we automatically recorded the submitted queries and the time participants took to complete each search engine's task. We compared the performance of WisPerMed and the baseline PubMed by analysing the time taken to complete tasks and the number of queries executed. We used a non-parametric statistical test (Wilcoxon Signed-Rank with Bonferroni correction at  $\alpha = 0.01$ ) to evaluate the significance of the differences observed between the two search engines. This comparison helps us assess the efficiency improvements provided by WisPerMed's integrated features, namely LoE and bio-concepts.

After completing both search tasks, participants answered six questions on a Likert scale with five levels ranging ("very unlikely", "unlikely", "neutral", "likely", and "very likely") except for the first question, which is a yes/no question. The questions were:

1. Were you aware of the LoE concept before?
2. How helpful did you find the LoE information to complete the task?
3. How helpful was the colouring of the bio-concepts for your search?
4. To what extent did it help you to judge the article's relevance?
5. How helpful was the word cloud for your search?
6. To what extent did it help you to judge the article's relevance?

These questions aim to assess the helpfulness of integrating the LoE and bio-concepts features in medical search engines. To our knowledge, no previous assessment of these features has been done, although extracting bio-concepts is a well-studied topic, and coloring has been used in other experimental search engines [280, 263, 32, 196]. Therefore, we asked participants two questions per interface feature (LoE, coloring bio-concepts, and word cloud).

The results of these quantitative questions were analysed using descriptive statistics (means and the distribution of responses). These measures help evaluate the perceived helpfulness of the integrated features, thereby addressing our goal of improving search performance and facilitating evidence-based decision-making.

### **Qualitative Analysis**

At the end of the quantitative questions, we provided four optional open-ended questions ("What did you like?", "What did you not like?", "What did you miss?", and "Recommen-

dations?") to provide feedback on the WisPerMed search engine. This feedback would allow us to identify comment themes and suggestions for improving the search engine.

## 7.3. Results

### 7.3.1. Quantitative Analysis

In this task, we evaluate the efficiency of the WisPerMed search engine in completing search tasks in terms of time and number of queries aspects compared with the PubMed search engine. The aim is to highlight the efficiency of integrating LoE and bio-concepts in medical search engines from the users' perspective.

Table 7-1.: Comparison of Search Task Efficiency for WisPerMed and PubMed: This table shows the time and number of queries needed to complete tasks, broken down by which search engine participants saw first.

(a) Number of queries per task.				(b) Completion time per task.			
Search Engine	WisPerMed	PubMed	Wilcoxon test	Search Engine	WisPerMed	PubMed	Wilcoxon test
Seen First	Mean (Std)	Mean (Std)	P value	Seen First	Mean (Std)	Mean (Std)	P value
All	4.39 (1.35)	5.52 (1.40)	7.6e-13	All	611 (244)	752 (267)	4.5e-9
WisPerMed	4.55 (1.37)	5.22 (1.19)	9.8e-3	WisPerMed	671 (270)	697 (234)	0.9999
PubMed	4.25 (1.32)	5.80 (1.51)	1.9e-9	PubMed	556 (205)	802 (286)	4.5e-12

#### Number of Queries to Complete Task

We automatically captured the queries submitted by both search engines. We found that WisPerMed required fewer queries to complete the search task, with a mean of 4.39 (Std=1.35) queries compared with 5.52 (Std=1.40) queries on PubMed, as shown in Figure 7-6c. The results are significant using Wilcoxon test at  $\alpha = 0.01$ , with a large effect size of 0.822.

When WisPerMed was shown to participants first, similar results were observed, with a mean of 4.55 (Std=1.37) on WisPerMed and 5.22 (Std=1.19) on PubMed, as shown in Figure 7-6b, with an effect size of 0.522. The results are significant using Wilcoxon test at  $\alpha = 0.01$ .

#### Time to Complete Task

We also automatically captured the time it takes to complete the search task, from loading the search engine's homepage to moving to the next tasks or feedback page. As shown in

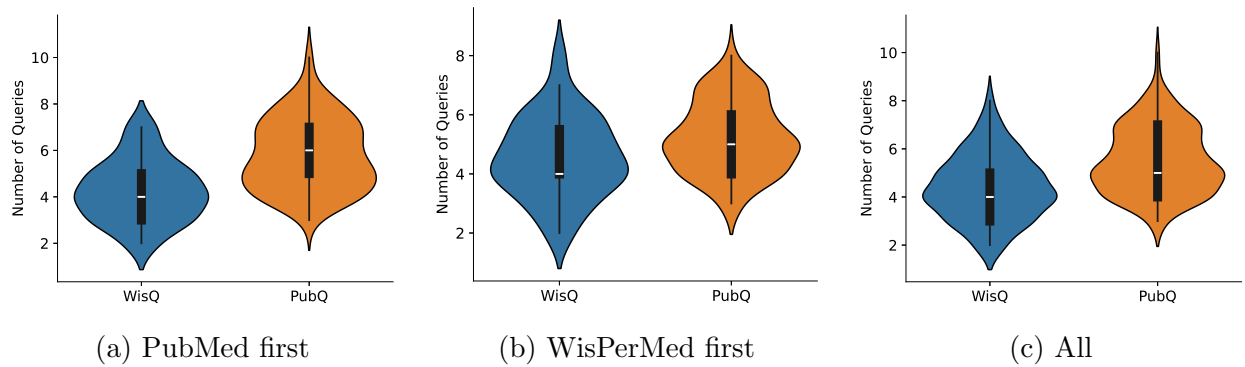


Figure 7-6.: Violin distribution for the number of queries required to complete the search task per search engine in case PubMed is shown first for participants, WisPerMed is shown first for participants, or all together.

Figure 7-7c, it took less time to complete the search task on WisPerMed, 611 (Std=244) seconds, and 752 (Std=267) seconds on PubMed. This leads to a large effect size of 0.551 and significant differences using the Wilcoxon test at  $\alpha = 0.01$ .

However, in the case where WisPerMed was shown first, we found that WisPerMed required only less time by a mean of 671 (Std=270) seconds and 697 (Std=234) seconds for PubMed, as shown in Figure 7-7b, with a small effect size of 0.103. The Wilcoxon test at  $\alpha = 0.01$  showed that this difference is not significant.

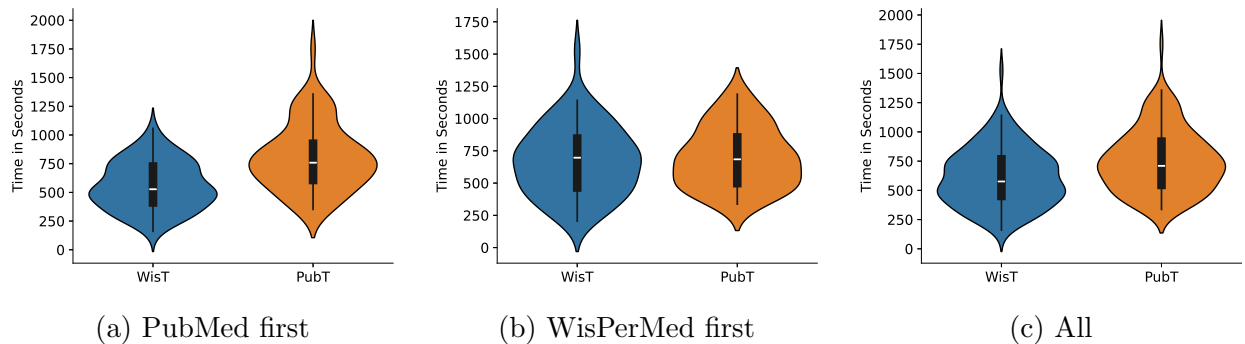


Figure 7-7.: Violin plot distribution for spent time to complete the search task per search engine in case PubMed is shown first for participants, WisPerMed is shown first for participants, or all together. Time is measured in seconds.

### Helpfulness Evaluation

We evaluated the participant's feedback on the helpfulness of the interface aspects (LoE, bio-concepts, and word cloud). Table 7-2 provides a descriptive analysis of the questions.

Table 7-2.: Participant's quantitative feedback questions and results on LoE, bio-concepts highlight, and word cloud.

No.	Question	Yes	No
1	Were you aware of the LoE concept before?	117	14
No.	Question	Mean	Std
2	How helpful did you find the LoE information to complete the task?	3.5	1.4
3	How helpful was the colouring of the bio-concepts for your search?	3.9	0.9
4	To what extent did it help you to judge the article's relevance?	3.5	0.9
5	How helpful was the word cloud for your search?	3.8	1.0
6	To what extent did it help you to judge the article's relevance?	3.6	1.0

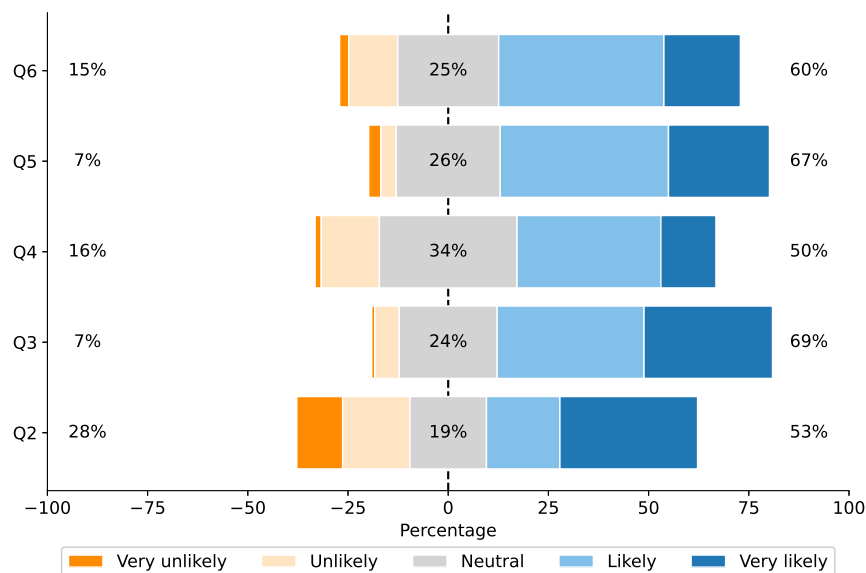


Figure 7-8.: Participants feedback responses distribution (Question 2 - Question 6).

**Levels of Evidence (LoE)** Most participants were aware of the LoE concept before the study: 89% were already familiar with it, and only 11% reported being unfamiliar with it. The helpfulness of displaying the LoE per article was rated at a mean of 3.5 (Std=1.4) on a 5-point Likert scale, with 28% negative reactions. Figure 7-8 shows the distribution of responses across the five Likert scale items.

**Bio-concepts** Most participants found the coloring of bio-concepts helpful in the search task, with a mean of 3.9 (Std=0.9) on a 5-point Likert scale and only 7% negative reactions. This indicates the positive impact this aspect had on the participants. Moreover, the other question aimed to find out its impact on judging articles relevance, with an average answer of 3.5 (Std=0.9) and 16% negative responses. Figure 7-8 shows the distribution of the

responses across the five Likert scale items.

**Word Cloud** The word cloud feature of our WisPerMed interface utilizes bio-concepts (which represent the core topic of the article’s abstract in a weighted format) was considered moderately helpful. The mean helpfulness rating is 3.8 out of 5 (Std=1.0), with only 7% negative reactions. Moreover, as shown in Figure 7-8, “about 60%” of participants found word clouds useful for relevance judgment, rating their usefulness at 3 or higher with a mean of 3.6 (Std=1.0) with 15% negative reactions.

### 7.3.2. Qualitative Feedback

Table 7-3 presents the common themes of users’ qualitative feedback that emerged from participants’ open answers. The users’ feedback highlighted the simplicity and the clear visualization of the different search engine components. This included interactive components such as word cloud and biomedical concepts coloring. They also received the LoE design, which is similar to the Nutri score, which can be seen in most of the food packages in Germany.

Table 7-3.: Participant’s most important feedback per question.

What did you like?	What did you not like?	What did you miss?	Recommendations?
Coloring the biological concepts	Highlighting species in abstracts	Reminder of LoE levels definition	Show abstracts summary line
LoE as Nutri-score	Query suggestions needs improvement	Citations count	Display the word cloud next to abstract
Word Cloud was well-received	Incomplete authors list	Search for article does not contain a bio-concept	Improve relevance of query suggestions
Intuitive and user-friendly interface	Need to open bio-concepts category to include topic	Coloring concepts in title	Incorporate more educational materials for bio-concepts

Although most participants liked the idea of coloring biomedical concepts, which can be shown from the distribution of the responses in Figure 7-8, but some users did not like the idea of coloring one type of biomedical concept, namely species. This can be understood as the search tasks mainly addressing oncological topics. Therefore, all species are seen as “human,” which makes it sound useless. However, this can be seen as useful in other medical fields where different species are of interest. Other users focused on the improvement of query

suggestions, this can be seen as potential future work for evaluating it as it is not part of the scope of this user study.

The user's feedback also mentioned useful suggestions for improving the search engine, such as small features like citation count, searching for articles that do not mention a bio-concept, showing the full list of authors, coloring the titles with bio-concepts like in abstracts, and showing LoE levels definitions.

Other suggested features need more investigation, like summarizing the abstract with one or two statements, allowing the word cloud to be displayed next to the abstract, or incorporating educational material on bio-concepts that would be more useful to junior students. The recommendation for integrating educational material was already developed feature but disabled as shown in Figure 7-9. This feature is integrated as a popup visible once the user clicks on a colored bio-concept in the abstract (Figure 7-2). The information provided using the concept identifiers discussed in Section 5.2.2.

Overall, participants appreciated WisPerMed's intuitive design, although some noted some missing features as a drawback. The feedback emphasized the simplicity and clear visualization of the various search engine components. Furthermore, feedback proposed potential improvements in the search engine, such as including a one-line summary using LLMs for the complete abstract before expanding the full text.

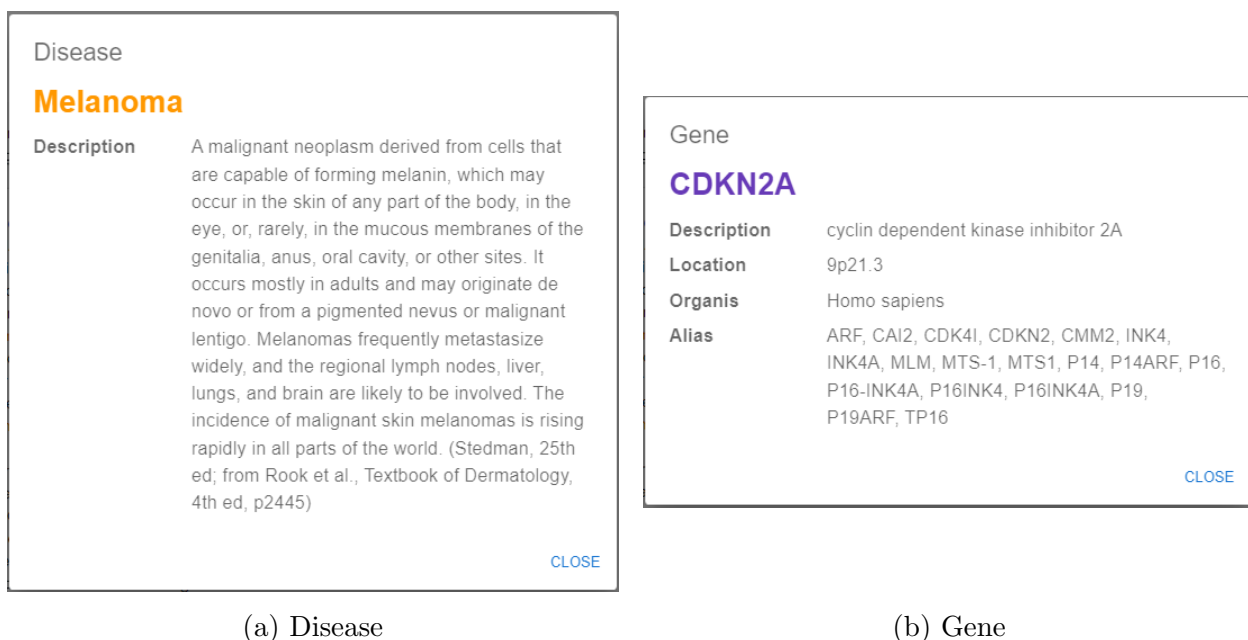


Figure 7-9.: WisPerMed educational material popups. Example on Melanoma disease and CDKN2A Gene.



## 7.4. Discussion

The user evaluation in this chapter addressed two research questions:

- Does using WisPerMed improve the efficiency of search compared to PubMed, a standard search engine for medical literature?
- Do users find the credibility assessment by Levels of Evidence helpful for the search task?
- Do users find interactivity through highlighting bio-concepts helpful for the search task?

In response to the first question, the integration of LoE and bio-concepts into the WisPerMed search engine confirmed the system-based evaluation and demonstrated significant improvements in search efficiency by reducing number of queries and time required to complete the search tasks.

Regarding the second question, the results suggest high helpfulness of WisPerMed to medical expert users. Our results give a strong indication that the LoE feature enabled users to quickly identify the most authoritative and relevant publications by categorizing publications based on the robustness of their empirical evidence. This finding aligns with the system-based evaluation, which suggested that a structured approach to evidence quality can facilitate the information retrieval process for medical professionals.

Similarly, the use of bio-concepts enhanced the semantic precision of searches by coloring entities. This enabled users to assess the results more effectively, which contributed to reducing search time. Moreover, the positive feedback on color-highlighting bio-concepts and the word cloud feature support the effectiveness of integrating text analysis tools in biomedical literature searches.

An interesting observation is that the effect size of the final search engine is much better than the major two parts, which are LoE (Chapter 6) and bio-concepts (Chapter 5) offline experiments. This difference can be justified as the offline and online metrics measure different aspects: while NDCG focuses on single-query quality, online metrics evaluate entire sessions, which capture user interaction aspects that enhance search performance. In the online setting, users can refine queries and utilize filters to overcome the limitations observed in offline experiments, such as failing in certain queries as in figure 5-5. Additionally, the visualization of LoE and bio-concepts helps users quickly identify relevant documents and adjust their queries when necessary, leading to the improved effect size in the final search engine.

WisPerMed with its integrated capabilities has direct implications for clinical and research settings. Reducing the time and number of queries to access high-quality and relevant studies will not only support medical professionals in their ongoing educational and clinical

efforts, but also contribute to the broader goal of improving patient outcomes through faster access to critical medical knowledge.

Despite the promising results, there are some limitations. First, we were comparing a new search engine with one that was familiar to most participants (93% of the participants had experience with medical search engines) – so the results are a lower bound of the real quality of WisPerMed. Although the sample size was acceptable, it was limited to participants from German medical institutions, which may not fully represent medical professionals worldwide. Moreover, there are also inherited limitations in WisPerMed, such as the reliance on accurately annotating documents with the LoE and bio-concepts. Misclassifications could potentially lead to misinformation or ignored relevant studies; on the other hand, retrieval is hardly ever perfect, and the loss due to misclassification is more than outweighed by the observed gain in terms of retrieval quality and search time.

Future research could investigate the comparative framework to include more biomedical search engines with more diverse demographic users. Investigating the scalability of the proposed features and their applicability in other specialized fields of medicine could further validate the utility of LoE and bio-concepts in broader contexts. Moreover, investigating the integration of LoE in self-RAG architecture [12] could improve the biomedical LLMs-generated responses not only with relevant information but also more credible. Finally, based on participants' feedback, potential improvements include refining the user interface to provide more intuitive navigation and incorporating features such as automated summary generation for quick insights into the research articles.

## 7.5. Summary

This chapter describes the user study aimed to evaluate the contribution of Level of Evidence (LoE) models and bio-concepts to search efficiency and efficacy in a realistic scenario, in which medical experts search in biomedical publication repositories to find best treatment options for their patients. LoE and bio-concepts were implemented into our search engine WisPerMed, which was then presented to the medical expert users along with PubMed and compared on the task of retrieving medical information needed at the point of care. Section 7.1 describes the rationale behind the user study. The study's setup and methods are described in Section 7.2. The results are presented in Section 7.3, showing that users find the credibility assessment by LoE and bio-concept-based interactive elements helpful in the search task. In addition, the results revealed that LoE and the interactive features using bio-concepts reduce the time and number of queries needed to complete the search task compared to PubMed. As the discussion of the results in Section 7.4 states, despite limitations, WisPerMed's integrated capabilities make it a viable and useful tool for biomedical literature search, which can contribute to the broader goal of improving patient outcomes

through faster access to relevant specialized knowledge.



# 8. Retrieval Augmented Generation

This chapter explores how conversational search engines can be used for searching for information in medical research publications. Conversational search engines are systems based on Large Language Models (LLMs) that enable users to interact and retrieve information through natural language dialogue. Their main challenge in searching through biomedical literature is accurately interpreting complex medical queries and delivering precise, relevant, and evidence-based responses from vast and specialized datasets.

The chapter investigates the integration of Levels of Evidence (LoE) into LLMs and examines the incorporation of WisPerMed within Retrieval-Augmented Generation (RAG) - a framework that combines retrieval of relevant documents or information from large datasets with the generative capabilities of LLMs to improve the relevance and the quality of outputs. The primary objective is to enhance the performance of conversational search engines driven by LLMs. The chapter covers the fine-tuning process of LLMs, the integration of LoE, and the evaluation of WisPerMed's role in Llama-based LLMs. Specifically, the chapter seeks to address the following research questions:

- What is the impact of fine-tuning LLMs for medical search tasks?
- How does the integration of LoE influence the performance of LLMs?
- Can we use WisPerMed for retrieval in the RAG setup?

## 8.1. Introduction

Recent developments in NLP in general and in LLMs in particular have revolutionized several fields, including IR and search engines. LLMs have transformed search engines from the traditional search, where users write a query and have to navigate through the retrieved document, to more interactive and user-friendly platforms known as conversational search engines. Unlike traditional search engines, conversational search engines –presented by tools like ChatGPT– can generate direct, contextually relevant answers to user queries, significantly simplifying the process of accessing information in complex domains such as medicine [310, 121].

This shift to conversational search engines led to a significant development in how infor-

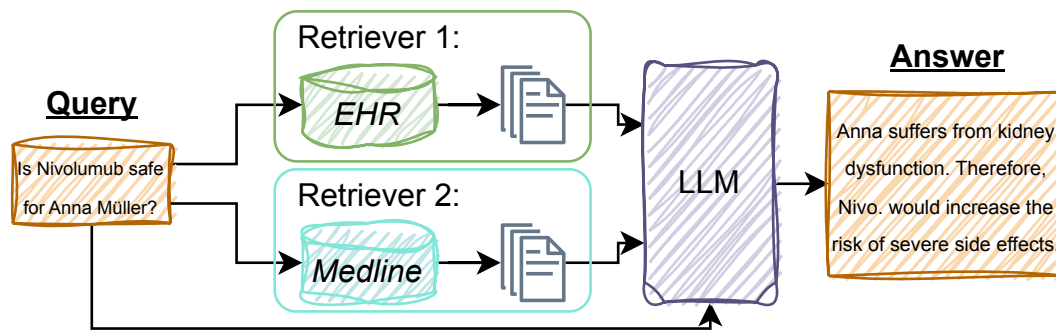


Figure 8-1.: Example of RAG-based architecture.

mation is retrieved and consumed. Traditional search engines operate on presenting a list of documents to users ranked by relevance. The responsibility to sift through these results to extract the necessary information and piece together a coherent answer from multiple sources lies with the user. This process, although effective, can be time-consuming in complex and interconnected fields like medicine, where the accuracy and credibility of information are essential [310].

On the other hand, conversational search engines offer a more streamlined approach by allowing users to interact with results and information using NLP queries. This allows it to process complex questions and generate synthesized responses that directly address the user's needs.

This new approach to accessing information raises a crucial question regarding the LLM's efficiency in biomedical information generation tasks. While LLM's ability to generate coherent and contextually appropriate responses is impressive, the accuracy and reliability of these responses in the highly specialized and critical field of medicine need a thorough evaluation. LLMs, despite their sophistication, are not immune to errors. They can produce convincing but factually incorrect information, a phenomenon known as "hallucination" [242]. In the context of healthcare, where misinformation can have serious, even life-threatening consequences, this issue is of particular concern [264].

One promising method to mitigate the risk of hallucinations in LLMs is the use of RAG. RAG is a technique that enhances the performance of LLMs by combining the strengths of both retrieval-based and generation-based approaches [242, 295]. In a RAG model, the system first retrieves relevant information from a large corpus of text based on the user's query (Similar to WisPerMed). This retrieved information is then used to guide the LLM in generating a response. By grounding the generation process in actual data, RAG reduces the likelihood of hallucinations and increases the accuracy and relevance of the output. Figure 8-1 shows an illustration of RAG-based architecture.

RAG is particularly useful in the biomedical domain because it allows LLMs to generate responses that are not only contextually appropriate but also supported by reliable sources.

This method ensures that the answers provided are based on existing medical literature or data, which is crucial for maintaining the integrity of information in healthcare. However, this opens the question: how does RAG affect the accuracy of generated answers in biomedical search contexts?

Building on the RAG framework, self-reflection is another advanced technique to enhance the reliability of LLM outputs [12]. Self-reflection is an extension of RAG, where the model not only retrieves information and generates a response but also evaluates its own output iteratively. This process involves the model critiquing its initial response, checking for inconsistencies or inaccuracies, and refining its answer accordingly. By incorporating self-reflection, LLMs can further improve the quality and coherence of their responses, making them more reliable for use in sensitive fields like medicine. This also raises the question: How much effect does self-reflection have over RAG in the medical domain?

Self-reflection allows the model to engage in a form of introspection, where it assesses the logical consistency and factual correctness of its generated text. This iterative refinement process helps minimize errors and ensure that the final output is well-grounded in evidence. In the context of biomedical information retrieval, this approach can significantly enhance the model's ability to provide accurate and trustworthy information, thereby mitigating the risks associated with LLM-generated content in healthcare.

While the advanced techniques, RAG and self-reflection, have proven effective in many fields in terms of improving accuracy and trustworthiness, the medical field has its own evidence-based trustworthiness standards as defined in the Level of Evidence Framework. Incorporating this framework into the medical publication search has proven effective both from the system's perspective (Chapter 6) and the user perspective (Chapter 7). This motivates the investigation of the impact of LoE in LLMs on the accuracy and reliability of their outputs in the biomedical context. By embedding LoE into LLMs, we can potentially enhance the credibility of the generated responses and show that LoE can potentially improve user trust.

## 8.2. Materials and Methods

This section outlines the material and methods employed in this experiment to investigate and evaluate the application of LoE in LLMs.

### 8.2.1. Dataset

We use the "PubMedQA" dataset [115] for our experiment, which is designed to answer biomedical research questions using short answers (yes, no, or maybe) or long open-ended

answers. The dataset was collected from PubMed abstracts, which makes it suitable for evaluating the integration of medical LLMs with RAG. PubMedQA has 1k expert-annotated instances, 61k unlabelled, and 211k automatically generated instances. Figure 8-2 presents an instance example from the dataset. The dataset is extended with the article ID on PubMed to make it easier to extract extra information such as authors, journals, MeSH terms, and others. For our experiment, we split the 1k manually annotated instances into 70% for fine-tuning and 30% for testing and excluded the short answers as we were interested only in the open-ended answers.

**Question:** Is cD30 expression a novel prognostic indicator in extranodal natural killer/T-cell lymphoma, nasal type?

**Context:**  
(**BACKGROUND**) Extranodal natural killer/T-cell lymphoma, nasal type (ENKTL), is an aggressive type of lymphoma whose standard treatment and validated prognostic model have not yet been defined.  
(**METHODS**) CD30 expression was detected using immunohistochemistry in 96 ENKTL patients, and the data were used to evaluate its relationship with clinical features, treatment response and prognosis.

**Long Answer:**  
Our results showed that expression of CD30 was not related to response to treatment but was an independent prognostic factor for both OS and PFS in ENKTL, nasal type, which suggests a role for CD30 in the pathogenesis of this disease and may support the incorporation of anti-CD30-targeted therapy into the treatment paradigm for ENKTL.

**Answer:** Yes

Figure 8-2.: An instance [147] of PubMedQA dataset.

### 8.2.2. Experimental setup

This experiment aims to explore the impact of integrating LoE into the input prompt on an LLM's response. We augmented the dataset by the LoE level as well as some descriptions on all levels. To tag instances with corresponding LoE, we used the WisPerMed search engine (Chapter 7), which indexes all PubMed articles to retrieve the LoE using the article



abstracts.

Analyzing the distribution of LoE in the PubMedQA dataset and comparing it with Medline, we found both datasets have a similar distribution over classes, as can be seen in Figure 8-3. This shows that the dataset is representative of Medline in terms of the LoE distribution.

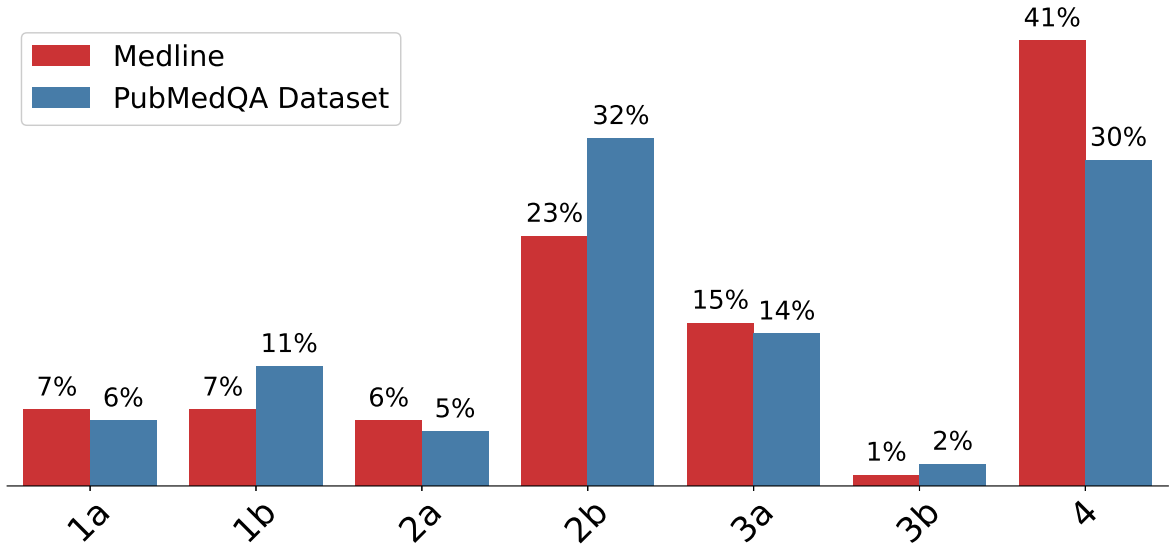


Figure 8-3.: The distribution of LoE Classes in the Medline Dataset and PubMedQA Dataset.

The performance of each model was assessed using ROUGE scores and cosine similarity. ROUGE scores measure the overlap of n-grams between the generated text and reference text. Although effective, they do not capture the factual correctness of alignment with the context’s evidence level. Therefore, we included semantic similarity levels between the generated text and provided answers using PubMed-BERT-based cosine similarity (Section 2.3.2).

Several studies investigate LLM’s sensitivity to prompt formats, which could influence results. Therefore, to minimize this effect, we used the same formats for all experiments. Figure 8-4 shows our fine-tuning prompt, and Figure 8-5 shows the evaluation prompt used for all models.

### Zero-Shot:

In this part of the experiment, we do not use the fine-tuning dataset and evaluate directly on the evaluation set. This aims to show the effect of fine-tuning.

```
Question: {question}
Context: {context}
Levels of Evidence: {LoE value}

Answer: {answer}
```

Figure 8-4.: Fine-tuning prompt. For the baseline model, the LoE information was excluded.

### Baseline:

We fine-tuned the LLM on the original dataset without including the LoE information. It is trained on all the documents without any exclusions. This model works as a baseline reference for evaluating the LoE-based LLM models.

### Baseline + LoE Models:

To measure the impact of integrating LoE into the LLMs, we first use the dataset with LoE tags. Then, we fine-tune the model based on the following setups to systematically test each LoE level:

- LoE4+: This model is simply fine-tuned on documents from all categories similarly to the baseline, but with LoE information appended to the prompt.
- LoE3+: This model excludes documents with LoE 4 (observation articles and case series), i.e. it includes articles with LoE categories 3b to 1a (case-control studies or higher LoE).
- LoE2+: This model included LoE categories 2b to 1a, i.e. cohort studies or higher LoE.
- LoE1: This model included LoE categories 1a and 1b, i.e. RCTs only.

### RAG zero-shot:

We append the documents retrieved by WisPerMed (Chapter 7) as context to the LLM model with zero-shot learning. This allows the model to use contextually relevant information to generate responses aiming to improve the LLM performance without being fine-tuned on similar documents.

Each question below is accompanied by contextual information tagged with its specific Level of Evidence. As you formulate answers, please ensure that they are informed by and reflect the level of evidence provided.

**Example: (Question:)** Does a diet rich in antioxidants and low in saturated fats reduce the risk of Alzheimer’s disease?

**(Context:)** Level of Evidence 1b. Research from controlled trials suggests that diets rich in antioxidants and low in saturated fats may help reduce the risk of cognitive decline and dementia.

**(Answer with Source:)** Yes, a diet rich in antioxidants and low in saturated fats can reduce the risk of Alzheimer’s disease. Such diets promote brain health by minimizing inflammation and oxidative stress, which are critical factors contributing to cognitive decline and the onset of Alzheimer’s disease. Antioxidants help neutralize free radicals, while healthy fats support overall brain function and structure. Therefore, maintaining a diet with these characteristics can significantly mitigate the risk factors associated with Alzheimer’s disease. (Source: Alzheimer’s Association, 2021).

**Levels of Evidence (LoE) Map:**

LoE 1a: Systematic reviews of randomized controlled trials.

LoE 1b: Individual Randomized Controlled Trials.

LoE 2a: Systematic Reviews of Cohort Studies. LoE 2b: Individual Cohort Studies.

LoE 3a: Systematic Reviews of Case-Control Studies.

LoE 3b: Individual Case-Control Studies.

LoE 4: Case Series.

Use the following pieces of context given below, each tagged with their Level of Evidence (LoE). Give attention to the LoE as it indicates the strength of the evidence behind the information.

**Context:** {context}

**Question:** {question}

Figure 8-5.: Evaluation prompt. For the baseline model, the LoE information was dropped.

**RAG baseline:**

After fine-tuning the baseline on PubMedQA dataset, we used it in RAG architecture with appending retrieved documents from WisPerMed (Chapter 7) as context. This allowed us to measure the effect size of LLM models fine-tuning in RAG setup.

**RAG baseline + LoE:**

To evaluate the LoE in the RAG architecture, we extend the prompts used in “RAG baseline” model with LoE information. This included all LoE levels from 1a ... 4.

## 8.3. Results

Our results are divided into two sections: system evaluation (Section 8.3.1), which shows the performance of the models, and application (Section 8.3.2), which presents the user interface of the LLM-based search engine.

### 8.3.1. System Evaluation

We performed the experiments on two small open-source LLMs: “Llama-2-7b” [262] and “Llama-3-8B” [4]. The results are summarized in Table 8-1 and Table 8-2.

The RAG-based experiments were performed on “Llama-2-7b” only as it aims to evaluate the effect of RAG rather than LLMs itself. The RAG based experiment results are summarized in table 8-3.

Model	ROUGE Score	Semantic Similarity
Zero-shot	11.12	56.29
Baseline	22.40	90.20
LoE 4+	28.69	95.98
LoE 3+	<b>28.95</b>	<b>96.84</b>
LoE 2+	28.60	95.97
LoE 1	28.04	94.76

Table 8-1.: Experiment performance on “Llama-2-7b” LLM.

**Fine-tuning impact** Fine-tuning the baseline models results in a significant improvement in both ROUGE scores and semantic similarity metrics. For “Llama-2-7b” (Table 8-1), fine-tuning increased the ROUGE score from 11.12 in the zero-shot setting to 22.40, and semantic

Model	ROUGE Score	Semantic Similarity
Zero-shot	12.24	83.87
Baseline	26.40	91.04
LoE 4+	<b>33.12</b>	94.11
LoE 3+	32.92	<b>94.96</b>
LoE 2+	31.26	94.86
LoE 1	32.75	93.80

Table 8-2.: Experiment performance on “Llama-3-8B” LLM.

Model	ROUGE Score	Semantic Similarity
zero-shot without RAG	11.12	56.29
RAG zero-shot	18.23	86.61
Baseline without RAG	22.40	90.20
RAG baseline	23.20	93.25
RAG baseline + LoE	<b>23.61</b>	<b>93.85</b>

Table 8-3.: RAG-based experiments performance on “Llama-2-7b” LLM.

similarity improved from 56.29 to 90.20. Similarly, for “Llama-3-8b” (Table 8-2), fine-tuning led to an increase in the ROUGE score from 12.24 to 26.40, and in semantic similarity from 83.87 to 91.04. These results confirm the importance of task-specific fine-tuning to enhance model performance, consistent with findings reported in Chapter 4, where fine-tuned BERT and Random Forest models outperformed zero-shot GPT-4.

**Impact of LoE** The integration of LoE information as context for the LLMs showed varying effects on performance. For the “Llama-2-7b” model, the inclusion of unfiltered LoE data (LoE 4+) led to improvements in both ROUGE score and semantic similarity, with values rising to 28.69 (+6.29) and 95.98 (+5.78), respectively. However, further filtering by LoE produced mixed results. The LoE 3+ configuration slightly outperformed the LoE 4+ with a ROUGE score of 28.95 and semantic similarity of 96.84, suggesting some benefit from excluding lower LoE data. Nonetheless, further filtering (LoE 2+ and LoE 1) led to decreased performance, indicating that while the inclusion of LoE information is beneficial, excessive filtering might exclude useful contextual data.

The Llama-3-8B model showed a similar pattern. The LoE 4+ configuration achieved the highest ROUGE score (33.12), while the LoE 3+, LoE 2+, and LoE 1 configurations slightly underperformed in comparison, with ROUGE scores of 32.92, 31.26, and 32.75, respectively. Interestingly, the LoE 3+ configuration produced the highest semantic similarity score (94.96), despite a slight decrease in the ROUGE score (32.92). This suggests that for the more advanced “Llama-3-8b” model, the benefits of LoE integration are more nuanced,

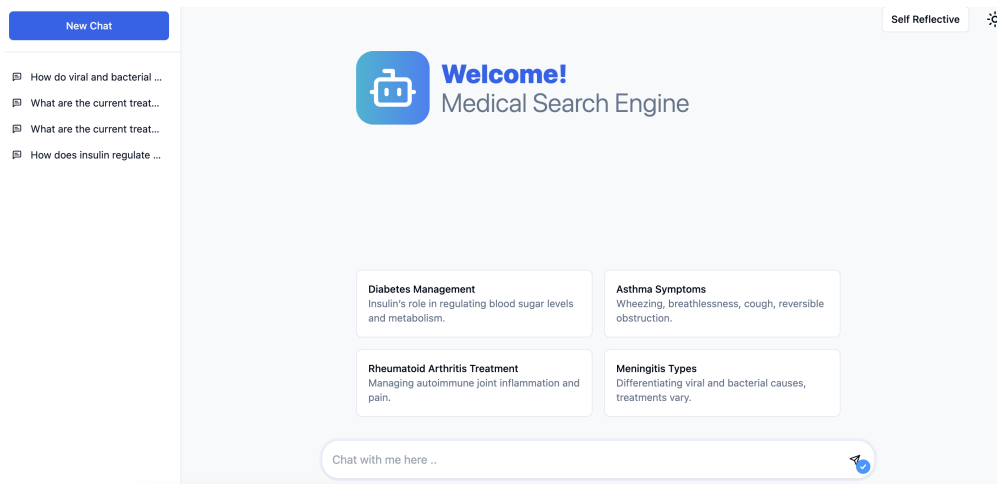
potentially depending on the balance between data filtering and the complexity of the model.

**Comparison of Llama models** Between the two models, “Llama-3-8b” demonstrated substantially higher ROUGE scores across all configurations compared to “Llama-2-7b”. For instance, the ROUGE score for the LoE 4+ configuration was 33.12 for “Llama-3-8b”, compared to 28.69 for “Llama-2-7b”. However, “Llama-3-8b” exhibited lower semantic similarity scores than “Llama-2-7b”, even in the best-performing LoE configurations. This suggests that while “Llama-3-8b” may have a better capacity for retrieving and summarizing relevant information, “Llama-2-7b” was more consistent in maintaining semantic coherence. This performance disparity highlights the trade-offs between different models, with “Llama-3-8b” leveraging its advanced architecture for certain tasks but not necessarily outperforming in all aspects [4].

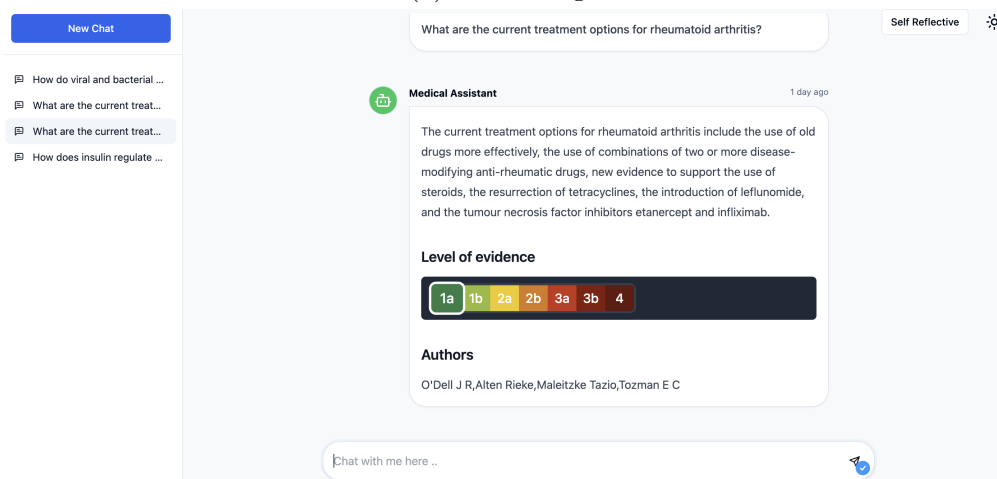
**Impact of RAG** As shown in Table 8-3, “Llama-2-7b” without fine-tuning when extended with relevant documents automatically via WisPerMed search engine showed significant improvements in both ROUGE score from 11.12 to 18.23 and semantic similarity from 56.29 to 86.61. These results confirm that, indeed, RAG improves the quality of generated responses in LLMs [41, 144]. Moreover, the fine-tuning of LLMs in RAG setup still improves the results to 23.40 and 93.20 on ROUGE score and semantic similarity, respectively. Comparing the fine-tuned model with and without RAG, we see that RAG improved the ROUGE score with 0.80 and semantic similarity of 3.05. Although these differences may appear minimal, they could be a point of contention for further investigation, as addressing them might improve the factuality that these metrics do not currently account for [144]. Finally, LoE improved the RAG performance with a minimal difference, but it underperformed the LoE 4+, which is the same setup without RAG.

### 8.3.2. Application: Medical LLMs Search Engine

Figure 8-6 shows an implementation of an LLM-based conversational search engine with a user-friendly interface that allows users to interact with evidence-based LLMs when searching for medical information. The system incorporates user interface elements similar to those found in other AI chatbots to maintain user familiarity and ensure intuitive interaction. We include the LoE for the article that generated the response as an indicator for the evidence base of the information and allow the user to switch between the different models. This system can offer more reliable and accessible support to healthcare professionals, which potentially saves more time during their busy schedules. Moreover, it can be considered a starting point for developing an LLM powered search engine for medical experts, which could allow for more customization based on their needs. However, this system needs to be



(a) Home Page



(b) Response

Figure 8-6.: LLM-based search engine with LoE levels for generated responses.

evaluated from a user perspective before integrating it into clinical settings<sup>1</sup>.

## 8.4. Discussion

In this chapter, we investigated the impact of integrating LoE into LLMs within the medical domain. This aim is to enhance the accuracy and reliability of generated responses in biomedical information retrieval tasks. The study explored both direct fine-tuning LLMs with annotation and the application of RAG to improve LLM outputs. Three research questions were addressed:

<sup>1</sup>LLM search engine source code can be found under: <https://github.com/samehfrihat/Medical-Conversational-LLM>

- What is the impact of fine-tuning LLMs for medical search tasks?
- How does the integration of LoE influence the performance of LLMs?
- Can we use WisPerMed for retrieval in the RAG setup?

In response to the first research question, the results demonstrated that fine-tuning LLMs significantly improves their performance in generating appropriate and semantically coherent responses. This is supported by the increases in both ROUGE scores and semantic similarity metrics. Fine-tuning enhances the model’s ability to produce text that closely aligns with reference answers. This aligns with prior research, which has established that task-specific fine-tuning is crucial for optimizing LLM performance, particularly in specialized domains such as medicine [274].

Regarding the second research question, the integration of LoE into LLMs yielded better results, suggesting that while the inclusion of LoE information generally enhances response accuracy, the degree of filtering based on LoE can have varying effects. Models fine-tuned with LoE 4+ information performed well, indicating that a broader inclusion of evidence types supports the generation of accurate responses. However, further filtering to higher LoE levels (e.g., LoE 3+ and above) led to mixed outcomes, with some performance gains in semantic similarity but not necessarily in ROUGE scores. This suggests a potential trade-off between the specificity of the evidence and the richness of contextual information available to the model.

Our analysis, which included two LLM models, highlighted the importance of LLM selection in determining performance outcomes. The “Llama-3-8b” model generally outperformed the “Llama-2-7b” in terms of ROUGE scores, suggesting superior capabilities in information retrieval and summarization. However, the “Llama-2-7b” model exhibited slightly higher semantic similarity scores, indicating better consistency in maintaining the semantic integrity of the information. These findings underscore the importance of selecting an appropriate model based on the specific requirements of the task at hand, particularly in balancing retrieval performance with the preservation of semantic coherence.

Finally, in answer to the third research question, the use of RAG techniques to augment LLM outputs with IR-based information demonstrated the potential improvement of the models. IR systems also retrieve relevant information but usually need cognitive power from users to judge the relevance and extract the needed information, which LLM models are excellent at. This would allow users to benefit from the power of IR systems and language models. This also has the potential of reducing the risk of “hallucinations” [237]—a critical concern in biomedical applications where accuracy is paramount.

This work can be considered a starting point for more investigation in the field. Several future research directions could be considered:

- **Human and Expert-Based Evaluation:** While this study focused on system-level



performance metrics, future work should include an evaluation by medical experts to assess the reliability, usability, and ethical implications of using LLM-based systems in clinical settings.

- **Impact of Citation on User Trust:** Our proposed work included the citation of generated responses, which potentially could lead to overtrust [134]. This needs further investigation to understand the influence on user decision-making and trust. Qualitative methods such as surveys and think-aloud studies could provide valuable insights.
- **Prompt Sensitivity and Engineering:** Several studies investigated the prompts sensitivity in general LLMs [225]. This could be more relevant in the medical domain, given the diversity of query formats used by medical experts. This requires a systematic evaluation and development of a robust prompt engineering framework to enhance model performance and user satisfaction.
- **Further Refinement of LoE Integration:** Although the LoE integration showed significant improvement, the study did not investigate further sophisticated methods for integrating LLMs. Research on investigating further methods could balance between contextual richness and specificity (The reliance on certain types of information for generation).
- **Integration of Self-RAG:** Self-RAG is built on the RAG framework by incorporating an additional layer of iterative self-assessment and refinement, where the model critiques and improves its generated responses based on the retrieved evidence. This could potentially improve accuracy, reliability, user trust, and user acceptance [12].

## 8.5. Summary

This chapter investigates the potential of conversational search engines for biomedical research literature retrieval (Section 8.1). To this end the chapter presents the impact of integrating WisPerMed with Large Language Models (LLMs) in a Retrieval Augmented Generation (RAG) setup. It explores the inclusion of Levels of Evidence (LoE) information retrieved by WisPerMed into LLMs and examines its potential to improve the answers generated from biomedical literature. Section 8.2 describes the dataset and the experimental setup. Specifically, two versions of Llama models are evaluated to show the impact of fine-tuning on LLMs. The results for each model are presented in Section 8.3 demonstrating that LLMs benefits from fine-tuning, expanding the context with LoE and the retrieved list from WisPerMed search engine. These results are discussed in Section 8.4.



## 9. Discussion and Conclusion

Our research reported in this work aimed to improve medical literature search engines by transitioning from traditional bag-of-word approaches to those considering semantic relationships and multidimensional relevance. Traditional IR systems often overlook the contextual factors surrounding medical searches. However, these factors have been shown to enhance retrieval processes for medical search engines aimed at health consumers [157, 311]. There has been limited research on the potential improvements when focusing on health experts [311]. Therefore, this research focuses on investigating contextual aspects and evaluating their impact on medical IR systems for experts and how they can be used to improve the retrieval process. To that end, we set out to (1) investigate different contextual aspects important to medical experts, (2) develop methods to integrate these aspects in IR systems, (3) build a search engine that supports medical experts with their needs for literature search, and (4) evaluate the search engine from system and user perspectives.

In Chapter 3, we analyzed the method to estimate document difficulty aspects (readability and technicality) from medical literature, allowing the development of personalized search engines to better align retrieved results with users' knowledge. Our models outperformed existing methods, which were not designed specifically for this purpose. Similarly, in Chapter 4, we developed automatic methods to predict medical subfields from medical literature. These methods, the first of its kind for this task, would allow personalized search engines to consider users' fields of expertise in retrieval.

In Chapter 5, we utilized PubTator [280] to extract the bio-concepts (genes, chemicals, diseases) present in medical documents and user queries, allowing the IR system to retrieve results not only based on the terms but also on the overlapping bio-concepts. We investigated several integration techniques using TREC PM collections, which showed the effectiveness of considering bio-concepts in medical IR. We also utilized bio-concepts to develop interactive elements in Chapter 7, such as presenting all bio-concepts of each document in an interactive Word cloud, which allows users to refine the search results in exploratory search format and faster document screening. Concepts in the abstracts are highlighted with color codes, allowing for faster relevance judgment. We also provided users with a feature to show bio-concept definitions on demand.

In Chapter 6, we allowed search engines to consider the evidence base of medical literature by integrating the level of evidence framework. We not only developed and investigated

effective techniques to predict the level but also evaluated the effectiveness of the integration from a system perspective using TREC PM 2017–2019 collections. This method improves relevance and significance at the same time.

These aspects and elements are then integrated into the WisPerMed search engine as presented in Chapter 7, which allows us to compare it with the PubMed search engine using predefined medical cases prepared by medical experts. The results showed that WisPerMed requires less time and fewer queries to complete search tasks and that WisPerMed achieves high user satisfaction, validating the suggested improvements from the system evaluations.

The recent developments in the field of Large Language Models (LLMs) and conversational search engines motivated the investigation of integrating LoE into LLMs and RAG frameworks in Chapter 8. The main goal of this experiment was to measure the impact of integrating WisPerMed and LoE into LLMs (using an RAG setup) on the relevance of generated answers. The results showed that LoE and WisPerMed hold the potential to improve the syntactic and semantic quality of generated responses.

This research provides new insights into how specific aspects can be effectively leveraged to enhance search performance, highlighting the value of an integrated approach that combines multiple contextual aspects. The results inform the theoretical framework by demonstrating the practical benefits of integrating contextual and personalized aspects into search engines. They support the notion that a multidimensional approach to relevance can enhance the efficiency and effectiveness of search systems by considering not only semantic relevance to the user query but also other related aspects of the search context.

A detailed discussion of how each of the factors we investigated improves the IR process and the search in biomedical repositories is presented and discussed in detail in each of the chapters above. In the remainder of this chapter we discuss our findings with respect to our general research questions (Section 9.1) and examine how the identified aspects improve medical literature retrieval for healthcare professionals. In Section 9.2, we explore the practical implications of these findings for the development and implementation of medical search engines. In Section 9.3, we provide an overview of the general limitations and present some directions for future work in Section 9.4 that could build on the contributions of this work. Finally, we summarize the key contributions in in Section 9.5.

## 9.1. Discussion of Research Questions

Our overarching goal to improve the expert users' search in biomedical research publication databases involved addressing several general research questions (cf. Section 1.2). In this section we discuss our findings relative to each of these research questions.

### 9.1.1. Personalization

First, we explored how medical documents can be tagged for document difficulty level and medical subfield they belong to. The rationale behind these experiments was to enhance the documents with meta-information that could be used to match with information in potential user profiles, thus enabling personalized search for biomedical literature that takes into account a user's language difficulty preference and the desired medical field of specialization. To this end, we asked the following research question:

**RQ1: Can we categorize medical research publications into their difficulty levels and medical subfields?**

The study demonstrated that categorizing medical research publications based on difficulty level is feasible with our annotated dataset using domain experts. The collected dataset can be considered a solid foundation for investigating different methods of estimating these aspects. The use of machine learning models to predict document difficulty and classify medical subfields shows a strong baseline for these tasks. Although the medical subfield classifier's performance is modest, it can be improved with high-quality training data to replace the automatic dataset used. These personalization features estimation would support personalized search engines, enhancing the relevance of retrieved results.

With the recent advance of LLMs, it was important to establish how a zero-shot classification approach via LLM compares to substantially more resource-intensive classification via fine-tuned models. We investigated this on the task of classifying documents into medical subfields in order to answer the following research question:

**RQ2: Can LLMs classify medical literature into its corresponding medical subfield?**

Our results indicate that fine-tuned models still offer superior performance in classifying medical literature into specific subfields compared to LLMs. However, zero-shot learning via LLMs is considered a good option when training data is limited or unavailable.

### 9.1.2. Bio-concepts

Our next research question concerned including bio-concepts like genes, diseases, and chemicals in the search for medical literature. Previous work has shown that bio-concepts have the potential to make the search more precise and the results more relevant to the users, as they allow for document retrieval based on semantic relevance. It was, however, not well understood, to what extent bio-concepts benefit both the retrieval process and the user's experience in medical publication search. Our research questions addressed this gap as follows:

**RQ3: Does query expansion with bio-concepts improve the relevance of retrieved medical publications?**

Query expansion using bio-concepts significantly improves retrieval quality. This was evidenced by evaluating sparse and dense retrieval methods with different integrating techniques using TREC PM collections. Bio-concepts incorporate domain-specific entities such as genes, diseases, and chemicals as a semantic relevance dimension, leading to more precise and contextually appropriate results. This approach improves the accuracy of search outcomes and aligns the results more closely with the complex needs of medical professionals.

**RQ4: How do bio-concepts affect user search experience when included as interactive features, such as abstract highlights, word clouds, or query expansions?**

The inclusion of bio-concepts as interactive features positively impacted the user search experience. This was reflected in the expert's feedback on the responses to the open-ended questions. Medical professionals' reactions on the quantitative scales appreciated the visual and interactive elements, such as abstract highlights and word clouds, which allowed for quicker identification of key information and more efficient query refinement. These features were particularly useful in helping users navigate complex biomedical literature, making the search process more intuitive and user-friendly.

**9.1.3. Levels of Evidence**

The credibility and trustworthiness of the search results is a key dimension in the multi-dimensional space that defines the relevance of biomedical publication search results. The evidence-based medical practice builds on well established credibility standards defined in frameworks, such as Level of Evidence (LoE) Framework. Yet, the evidence base of medical publications has to date not been considered in medical publication search. We explored the benefits of integration of existing professional guidelines into the search process by addressing the following research questions:

**RQ5: Can Level of Evidence be identified in medical research papers?**

Our results confirmed that LoE can be effectively identified in medical research papers using machine learning techniques. This task can be seen as a classification task or as a regression task to reduce the distance between misclassified labels, both providing good performance using the collected dataset from the German oncology guidelines. Moreover, the use of methods for explaining predictions allowed for further trust in the predicted labels.

**RQ6: Does the inclusion of Levels of Evidence as a filter improve the efficiency and effectiveness of medical publication retrieval?**

Including the level of evidence as a filter in IR showed significant improvements in the efficiency and effectiveness of the retrieval process. This allows to retrieve not only relevant publications but also prioritize significant work and high-quality research. The evaluation using the TREC PM collections showed that LoE filtering outperformed the best-published runs, underscoring the value of this feature in enhancing retrieval outcomes.

**RQ7: How useful is the indication of a publications Level of Evidence to medical professionals when searching for relevant research publications?**

The results of the user study indicated that LoE was found to be highly useful for the medical professionals. The integration of LoE as design similar to the Nutri-score in food packaging made it intuitive and self-explanatory, also allowed the user to prioritize their reading and decision-making processes. The LoE information and the bio-concepts presented together not only improved the accuracy of the results, but reduced the time and effort required by the users to refine queries.

**RQ8: What is the impact of Levels of Evidence in a retrieval augmented generation (RAG) setting?**

Our research explored the application of LoE in LLMs and RAG architecture, finding that it enhances the quality of retrieved responses on syntactic (ROUGE score) and semantic (PubMedBERT embeddings cosine similarity) levels. The integration provided evidence-backed outputs, thus improving the reliability and trust of AI-generated advice or summaries. However, further investigation is needed to fully understand the potential and limitations of this approach.

## 9.2. Implications for Practice

The findings of this research have several practical implications for the development and enhancement of medical search engines for medical experts:

### Customization and Personalization

A search engine should incorporate detailed user profiles that consider the user's field of expertise, language proficiency, and search behavior. This personalization approach ensures

that healthcare professionals receive search results that are most relevant to their specific needs and knowledge level. This relevant information is then evaluated not only by topicality relevance but also by considering other concepts, such as whether a document about melanoma might be relevant for a radiologist but not a dermatologist.

### **Integration of Credibility Metrics**

Our experiments showed that incorporating metrics such as the level of evidence into the ranking algorithm enhanced not only the credibility and trustworthiness of retrieved documents but also the effectiveness of the results. This is crucial for medical professionals who rely on high-quality, evidence-based information for clinical decision-making. Moreover, it is also clear that medical professionals need to control the retrieved level of evidence based on the performed task, such as when performing systematic reviews for case-controlled studies, where less evident results are targeted.

### **Interactive Elements**

The inclusion of interactive features such as bio-concept coloring, word cloud, educational definitions on demand, and query expansion tools allows users to engage more deeply with search results. These features help users to quickly identify relevant information, refine their searches, and make faster judgments about document relevance. This contributes to the field of exploratory search when conducting literature review tasks for the medical domain when a user is less sure of the exact search query and keeps updating it.

### **Efficiency and Usability**

Integrating personalization and context aspects in search engines can significantly enhance the efficiency of the search process for medical professionals by reducing search time and the number of queries required. This is particularly important given the time constraints and high stakes involved in medical decision-making.

## **9.3. Limitations**

The concept of contextualization in the biomedical publication search for medical experts is complex. Despite the promising results we presented in this research, several general limitations were identified in this research:

- **Technical Constraints:** Our WisPerMed search engine aims to serve in clinical settings and is based on several machine learning models and NLP techniques, which



while effective, are dependent on the quality and quantity of the training data. Continuous updates and retraining with new data are necessary to maintain and improve performance.

- **Data Limitations:** The extraction and analysis were based on publicly available abstracts and titles, which may not capture the full depth of the medical documents in some cases. Access to full-text articles could further improve the accuracy of the contextual aspects. However, this would face several privacy challenges, as most medical articles are not published in an open-access format.
- **Scope of Evaluation:** Our evaluations were conducted from both a system perspective using the TREC PM collections, which consistently showed improvements, and from a user perspective through a study involving a significant number of medical experts, yielding considerable results. However, to generalize the findings, a more diverse set of evaluations is necessary. The current evaluation from the user study focused primarily on real-world clinical scenarios within the dermatological field. Thus, further evaluations across other medical fields are required to ensure the broader applicability of the results.

## 9.4. Future Work

The field of medical search engines has received considerable attention since the early days of domain-specific IR systems. However, little attention has been given to supporting medical experts compared with health consumers in their information search. This makes it an area of huge potential for development. Similarly, the findings of this thesis open several avenues for future research and development:

- **Result Summarization:** Medical experts often need to quickly gather the latest information on specific aspects of a medical case, such as diagnosis, medication, and treatment. Currently, this process is time-consuming, requiring manual review of numerous documents. To streamline this, future work will focus on developing an aspect-based summarization feature in the WisPerMed search engine. This feature will allow users to select specific aspects of interest, and the system will generate concise summaries from the retrieved documents, color-coded by aspect, and linked back to the original text (allowing the user to navigate to the document mentioning the selected sentence). This line of research will be pursued by Research Project 23 of WisPerMed RTG, which will utilize LLMs to enhance search efficiency, enabling physicians to rapidly extract relevant information and improve clinical decision-making. This initiative will significantly save time and effort in medical information retrieval.
- **Integration with Clinical Systems:** The clinical system stores important clinical

aspects which, if integrated, could help IR system as contextual aspects. For example, searching for clinical trials is highly dependent on the inclusion and exclusion criteria, which depend on patients' age, gender, medical history, and other features. Therefore, integrating the WisPerMed search engine with electronic health records (EHR) and other clinical decision support systems could provide seamless access to relevant literature within the practitioners' workflow. Therefore, WisPerMed will be integrated into the clinical dashboard systems based on the development of research project 9 of the WisPerMed RTG. This would potentially save more time for the users, resulting in higher efficiency.

- **Learning to Rank Integration:** To improve the retrieval effectiveness of the WisPerMed search engine, future work direction will focus on integrating Learning to Rank (LTR) techniques. LTR will optimize the ranking of search results by incorporating contextual aspects such as bio-concepts, personalization features (readability, technicality), and medical subfields. By leveraging the automatically collected user interaction data and feedback, the system will continuously learn and adapt to the specific needs and preferences of individual users, enhancing personalization.
- **Broader User Studies:** To gain more comprehensive insights into the effectiveness of the WisPerMed search engine, future work will focus on conducting larger and more diverse user studies. This will involve engaging participants from various medical specialties and geographical locations to evaluate the system's performance across different contexts and user needs. The studies will assess usability attributes such as learnability, efficiency, memorability, errors, and satisfaction using both subjective methods (surveys, cognitive walkthroughs) and objective methods (observational interviews, log file recording, session recordings). This initiative, part of research project 20 of the WisPerMed RTG, aims to uncover any remaining weaknesses and enhance the system's overall usability and effectiveness in diverse clinical settings.
- **Advanced Interaction Techniques:** Exploring more sophisticated interactive features, such as adaptive learning systems and real-time user feedback mechanisms, could further enhance user experience and retrieval performance.

## 9.5. Conclusion

In conclusion, this thesis demonstrates that integrating contextual and personalized aspects into medical search engines significantly enhances their efficiency and effectiveness for healthcare professionals. By addressing the unique needs of medical experts through advanced natural language processing and machine learning techniques, WisPerMed provides a robust solution for retrieving high-quality medical information.

The implementation of features such as the level of evidence, document readability and technicality, and bio-concepts has shown to improve both the relevance and usability of search results. User studies confirmed that these enhancements lead to faster and more accurate information retrieval, ultimately supporting better clinical decision-making and patient care.

The implications for practice suggest that future medical search engines should continue to develop and integrate these advanced features, ensuring that they are tailored to the specific needs of healthcare professionals. Furthermore, ongoing research and development in this area will likely yield even more sophisticated tools and techniques, further improving the landscape of medical information retrieval.

Overall, this research contributes significantly to the field of medical information retrieval, providing a foundation for future innovations and improvements that can benefit healthcare professionals and, by extension, patient outcomes.



# A. Appendix

## A.1. Complete Version of The Annotation Tool

Medical experts used this annotation tool to annotate medical abstracts with document difficulty aspects [3](#) and medical subfields [4](#). The source code can be found on GitHub <https://github.com/samehfrihat/TechnicalityLevelAnnotationTool>.

# Willkommommen Beim WisPerMed Annotationstool!

Künstliche Intelligenz Und Entscheidungsunterstützung In Der Medizin

## Los Geht's!

*Vielen dank, dass Sie uns bei unserer Forschung zur Verbesserung von KI in der Medizin unterstützen!  
Sie sind nur einen Schritt davon entfernt den Unterschied zu machen!*

*Sind Sie im Bereich der Medizin tätig und würden gerne mehr erfahren oder uns unterstützen? Dann kontaktieren Sie uns einfach via [E-mail](#)*

*Um mit der Annotation starten zu können beantworten Sie bitte die folgenden Fragen zu Ihren Sprachkenntnissen und stimmen der Daten- und Cookienutzung zu. Alle gesammelten Daten werden unter einem automatisch angelegten und komplett anonymisierten Nutzer gespeichert.*

Wie gut sprechen Sie Deutsch?: \*  ▼

Wie gut sprechen Sie Englisch?: \*  ▼

In welcher Kategorie ordnen Sie sich ein? \*  ▼

Ich bin mit der Nutzung der Daten zu Forschungs- und Publikationszwecken einverstanden.  \*

Ich bin damit einverstanden, dass die Webseite funktionale Cookies speichert  \*

**STARTEN**

# Willkommen Beim WisPerMed Annotationstool

Hilf Uns Künstliche Intelligenz Und Entscheidungsunterstützung In Der  
Medizin Zu Verbessern

Sie Haben Bisher 0 Abstracts Annotiert  
Bitte Annotieren Sie Mindestens 2 Abstracts

**ANNOTATION STARTEN!**

Vielen Dank, Dass Sie Uns Bei Unserer Forschung Unterstützen!





Bitte lesen Sie die Abstracts sorgfältig durch und bewerten Sie die jeweilige Formalität und Lesbarkeit auf einer Skala zwischen 0 und 100 Prozent.

Topic_ID	630f64c3952929d86fe02ac9
Titel	[Risk factors for the oral use of antibiotics and treatment incidence of weaners in Switzerland]
Englisches Abstract	<p>In the present study, risk factors for the use of oral antibiotics in weaned piglets were collected on 112 pig farms by a personal questionnaire. The most common indication for an antibiotic group therapy was diarrhoea, and the most frequently used antibiotic was Colistin. On average, 27.33 daily doses in the control farms and 387.21 daily doses in the problem farms per 1000 weaners were administered on a given day. The significant risk factors in the multivariate model were poor hygiene in the water supply of suckling piglets, less than two doses of prestarter feed daily, lack of an all-in-and-all-out production system in weaners, no herd book performance data analysis, and less than two of the legally prescribed veterinary visits per year. Furthermore, the treatment incidence of weaners for oral antibiotics was calculated on the basis of the drug inventory. This study provides evidence that the use of oral antibiotics in weaners can be reduced by interventions in hygiene and management.</p>
	<p><b>Grad der Formalität - Englisches Abstract:50% *</b></p> <p>0 <input type="range" value="50"/> 100</p> <p><b>Grad der Lesbarkeit - Englisches Abstract:50% *</b></p> <p>0 <input type="range" value="50"/> 100</p>
Deutsches Abstract	<p>Publisher: In der vorliegenden Arbeit wurden auf 112 Schweinezuchtbetrieben Risikofaktoren für den Verbrauch von oralen Antibiotika bei Absetzferkeln anhand einer persönlichen Befragung erhoben. Die häufigste Indikation für eine antibiotische Gruppentherapie war Durchfall und das am meisten verwendete Antibiotikum Colistin. Im Durchschnitt wurden bei den Kontrollbetrieben 27.33 Tagesdosen und bei den Problembetrieben 387.21 Tagesdosen pro 1000 Absetzferkel an einem Tag verabreicht. Als signifikante Risikofaktoren im multivariaten Modell wurden mangelnde Tränkehygiene im Abferkelstall, keine oder weniger als zweimal tägliche Prästarterfüttergabe, kontinuierliche Bestossung des Absetzstalls, keine Herdebuch-Leistungsdatenauswertung und weniger als zwei der gesetzlich vorgeschriebenen Tierarzneimittelbesuche (TAM-Besuche) pro Jahr durch den Bestandestierarzt festgestellt. Ferner wurde anhand der Arzneimittelinventarlisten auf den Betrieben die Tierbehandlungsinzidenz der Absetzferkel für oral verabreichte Antibiotika berechnet. Dieses Ergebnis deutet darauf hin, dass der orale Antibiotikaverbrauch im Absetzstall durch Interventionen im Hygiene- und Managementbereich reduziert werden kann.</p>
	<p><b>Grad der Formalität - Deutsches Abstract:50% *</b></p> <p>0 <input type="range" value="50"/> 100</p> <p><b>Grad der Lesbarkeit - Deutsches Abstract:50% *</b></p> <p>0 <input type="range" value="50"/> 100</p>
Themen-bereich *	<input type="text" value="Select an option..."/>

\* erforderliches Feld

ABSENDEN



## **A.2. Complete Version of the User Study of Chapter 7**



## Willkommen zu unserer Studie zur medizinischen Literatursuche!

Vielen Dank, dass Sie an unserer Studie teilnehmen, die darauf abzielt, Suchmaschinen für medizinisches Fachpersonal zu verbessern.

Diese Studie ist Teil des WisPerMed RTG. Wir sind ein interdisziplinäres Team aus Informatikern, Psychologen und Ärzten. Unser Ziel ist es, die Personalisierung von medizinischem Wissen und die datenbasierte Entscheidungsunterstützung am Point of Care voranzutreiben.

Ihre Teilnahme ist vollständig freiwillig und Sie können jederzeit ohne Verpflichtungen von der Studie zurücktreten. Mit Ihrer Teilnahme erklären Sie sich einverstanden, dass die gesammelten Daten in einem Forschungspapier präsentiert werden. Wir sammeln keine sensiblen Daten, die Sie identifizieren können. Ihre Privatsphäre ist geschützt. Im Forschungspapier werden ausschließlich die aggregierten Ergebnisse aller Teilnehmer veröffentlicht.

Während dieser Studie werden Ihnen zwei spezifische Aufgaben zur medizinischen Literatursuche gestellt, und die Studie wird voraussichtlich etwa 15 Minuten dauern. Ihre Beiträge werden erheblich zu unseren laufenden Forschungsbemühungen beitragen.

Bei Fragen zur Studie zögern Sie bitte nicht, Kontakt aufzunehmen mit:

Sameh Frihat

[Sameh.Frihat@uni-due.de](mailto:Sameh.Frihat@uni-due.de)

Duisburg-Essen Universität

- Ich habe die oben genannten Bedingungen gelesen.
- Ich bin 18 Jahre oder älter.

LOS GEHT'S!

Weitere Informationen zu WisPerMed finden Sie unter: [WisPerMed.com](http://WisPerMed.com)

## Demografische Daten

Beruf/Rolle  
Arzt/Ärztin ▼

Berufserfahrung in Jahren \*

Geschlecht ▼

Alter \*

Englischkenntnisse ▼

Deutschkenntnisse ▼

Vorherige Erfahrung mit medizinischen Suchmaschinen wie PubMed: \*

Keine  Gering  Einige  Umfassend

FORTSETZEN


## Kontext:

Stellen Sie sich vor, Sie sind Dermatologe und interessieren sich für das Gebiet der dermatologischen Onkologie. Sie sind eine Person, die sowohl Wert auf die Genauigkeit der Informationen als auch auf die Effizienz beim Finden relevanter Informationen legt.

## Aufgabe:

Stellen Sie sich vor, Ihnen wurden zwei Patienten zugeteilt, die jeweils an einer bestimmten Art von Hautkrebs leiden. Sie müssen die besten Behandlungsoptionen für jeden der beiden Patienten finden. Sie werden dazu zwei verschiedene Suchmaschinen verwenden. Jede Suchmaschine wird für einen Patienten verwendet.


Für jede Aufgabe sollen Sie 10 relevante Artikel zur Behandlung finden, um zu entscheiden, welche Behandlungsstrategie für diesen Patienten die beste ist. Jeder Artikel sollte ein hohes Evidenzniveau haben. Wenn Sie keine 10 Artikel finden können, können Sie jederzeit zum nächsten Punkt übergehen.

Wenn Sie einen relevanten Artikel finden, markieren Sie diesen bitte als relevant mit dem Daumen-hoch-Button: .

Die Suchmaschine wird die Anzahl der Anfragen und die für die Aufgabe aufgewendete Zeit notieren.

Die Suchmaschine wurde, wie auch andere frei verfügbare Suchmaschinen wie PubMed, in Englischer Sprache entwickelt.

STARTEN



## Patient 1:

Der Patient wurde erstmals mit metastasierendem Melanom mit Hirnmetastasen diagnostiziert.

Bitte versuchen Sie, 10 relevante Artikel zu möglichen Behandlungen mit dieser Suchmaschine zu finden.

Anzahl relevanter Artikel: 0

NÄCHSTER PATIENT



melanoma



- LEVEL OF EVIDENCE
- ARTICLE TYPE
- DATE
- BIO CONCEPTS
- INCLUDE CONCEPTS

Showing 10 of 10000 results

## Pediatric Melanoma in Melanoma-prone Families

Goldstein Alisa M & Tucker Margaret A

2019-09-20

PMID : 30207590

publication\_types :Journal Article , Research Support, N.I.H., Extramural .

> Background: In the United States, only 0.4% of all melanomas are diagnosed in patients <20 years. Melanoma in pediatric members of melanoma-prone families has not been fully investigated. The study goal was to

Level Of Evidence

Is Relevant



## Melanoma update. Second primary melanoma.

Frank W & Rogers G S

1993-06-24

PMID : 8496486

publication\_types :Journal Article , Review .

> BACKGROUND: Second primary melanoma is not a rare phenomenon. It occurs in at least 3 to 6% of melanoma patients and up to a third of individuals from melanoma-prone families. OBJECTIVE: To review the clinical

### Patient 1:

Der Patient wurde erstmals mit metastasierendem Melanom mit Hirnmetastasen diagnostiziert.

Bitte versuchen Sie, 10 relevante Artikel zu möglichen Behandlungen mit dieser Suchmaschine zu finden.

Anzahl relevanter Artikel: 0

NÄCHSTER PATIENT



**WisPerMed**

Wissens- und datenbasierte  
Personalisierung von Medizin  
am Point of Care



## Patient 2:

Der Patient wurde mit Basalzellkarzinom im Stadium III diagnostiziert.

Bitte versuchen Sie, 10 relevante Artikel zu möglichen Behandlungen mit dieser Suchmaschine zu finden.

Anzahl relevanter Artikel: 0

NÄCHSTER



melanoma



Showing 20 of 100 results

## Malignant Melanoma: Skin Cancer-Diagnosis, Prevention, and Treatment.

Ahmed Bilal & Ghafoor Saba

2020-01-01

PMID : 32894659

publication\_types :Journal Article , Review .

- > Melanoma is a skin cancer caused by a malignancy of melanocytes. Incidence of melanoma is rapidly increasing worldwide, which results in public health , ...

Is Relevant



## Melanoma: epidemiology, risk factors, pathogenesis, diagnosis and classification.

Rastrelli Marco & Alaibac Mauro

2014-01-01

PMID : 25398793

publication\_types :Journal Article , Review .

- > This article reviews epidemiology, risk factors, pathogenesis and diagnosis of melanoma. Data on melanoma from the majority of countries show a rapid , ...

### Patient 2:

Der Patient wurde mit Basalzellkarzinom im Stadium III diagnostiziert.

Bitte versuchen Sie, 10 relevante Artikel zu möglichen Behandlungen mit dieser Suchmaschine zu finden.

Anzahl relevanter Artikel: 0

NÄCHSTER





Vielen Dank, dass Sie an unserer Studie teilgenommen haben. Hier ist der letzte Schritt. Bitte geben Sie Feedback zu Ihrer Erfahrung mit der WisPerMed-Suchmaschine. Eine der Suchmaschinen zeigte zusätzliche Informationen zu den Artikeln an, z. B. das Evidenzniveau und hervorgehobene Konzepte:

## Melanoma genetics.

Read Jazlyn & Hayward Nicholas K

2016-10-12

PMID : 26337759

publication\_types :Journal Article , Research Support, Non-U.S. Gov't , Review .

> Approximately 10% of melanoma cases report a relative affected with melanoma, and a positive family history is associated with an increased risk of de

Level Of Evidence



Is Relevant



### 1. Level of Evidence (LoE)



Waren Sie sich vorher des LoE-Konzepts bewusst?  Ja  Nein\*

Wie hilfreich fanden Sie die LoE-Informationen, um die Aufgabe zu erledigen?  Überhaupt nicht hilfreich      Sehr hilfreich

### 2. Bio Concepts

Approximately 10% of melanoma cases report a relative affected with melanoma, and a positive family history is associated with an increased risk of developing melanoma. Although the majority of genetic alterations associated with melanoma development are somatic, the underlying presence of heritable melanoma risk genes is an important component of disease occurrence. Susceptibility for some families is due to mutation in one of the known high penetrance melanoma predisposition genes: CDKN2A, CDK4, BAP1, POT1, ACD, TERF2IP and TERT. However, despite such mutations being implicated in a combined total of approximately 50% of familial melanoma cases, the underlying genetic basis is unexplained for the remainder of high-density melanoma families. Aside from the possibility of extremely rare mutations in a few additional high penetrance genes yet to be discovered, this suggests a likely polygenic component to susceptibility, and a unique level of personal melanoma risk influenced by multiple low-risk alleles and genetic modifiers. In addition to conferring a risk of cutaneous melanoma, some melanoma predisposition genes have been linked to other cancers, with cancer clustering observed in melanoma families at rates greater than expected by chance. The most extensively documented association is between CDKN2A germ line mutations and pancreatic cancer, and a cancer

Wie hilfreich war die Farbgebung der bio-medizinischen Konzepte für Ihre Suche?  Überhaupt nicht hilfreich      Sehr hilfreich

Inwiefern hat es Ihnen geholfen, die Relevanz des Artikels zu beurteilen?  Überhaupt nicht hilfreich      Sehr hilfreich

### 3. Word Cloud



Wie hilfreich war die Wortwolke für Ihre Suche?  Überhaupt nicht hilfreich      Sehr hilfreich

Inwiefern hat es Ihnen geholfen, die Relevanz des Artikels zu beurteilen?  Überhaupt nicht hilfreich      Sehr hilfreich

### 4. Bitte geben Sie abschließend ein Feedback zu Ihrer Gesamterfahrung mit der WisPerMed-Suchmaschine.

Was hat Ihnen gefallen?

Was hat Ihnen nicht gefallen?

Was haben Sie vermisst?

Empfehlungen?

**EINREICHEN**

### **A.3. WisPerMed Search Result Page**

The search engine source code can be found on GitHub <https://github.com/samehfrihat/WisPerMedSearchEngine>.



Melanoma



LEVEL OF EVIDENCE ARTICLE TYPE DATE BIO CONCEPTS INCLUDE CONCEPTS

BRAF V600E

Showing 10 of 10000 results

### Prognostic Role of BRAFV600E Cellular Localization in Melanoma.

Abd Elmageed Zakaria Y & Kandil Emad 2019-07-15
PMID : 29369798
publication\_types : Journal Article , Research Support, N.I.H., Extramural , Research Support, Non-U.S. Gov't .

BACKGROUND: Approximately half of cutaneous melanoma tissues harbor BRAFV600E mutations, resulting in a constitutive activation of the mitogen-activated protein kinase (MAPK) pathway. Nuclear-cytoplasmic

Level Of Evidence Is Relevant
1a 1b 2a 2b 3a 3b 4

### In-depth genomic data analyses revealed complex transcriptional and epigenetic dysregulations of BRAFV600E in melanoma

Guo Xingyi & Zhao Zhongming 2016-01-14
PMID : 25890285
publication\_types : Journal Article , Research Support, N.I.H., Extramural , Research Support, Non-U.S. Gov't .

Background The recurrent BRAF driver mutation V600E (BRAFV600E) is currently one of the most clinically relevant mutations in melanoma. However, the genome-wide transcriptional and epigenetic dysregulations

Level Of Evidence Is Relevant
1a 1b 2a 2b 3a 3b 4

### Detection of BRAF p.V600E Mutations in Melanoma by Immunohistochemistry Has a Good Interobserver Reproducibility.

Marin Cristi & Emile Jean-François 2014-02-18
PMID : 23651150
publication\_types : Journal Article , Research Support, Non-U.S. Gov't .

CONTEXT: Assessment of BRAF p.V600E mutational status has become necessary for treatment of patients with metastatic melanoma. Detection of p.V600E mutation by immunohistochemistry was recently reported in

Level Of Evidence Is Relevant
1a 1b 2a 2b 3a 3b 4

### BRAFV600E Mutant Allele Frequency (MAF) Influences Melanoma Clinicopathologic Characteristics

Soria Xavier & Marti Rosa M 2015-03-03
PMID : 34680222
publication\_types : Journal Article .

Simple Summary The mutational load of BRAFV600E in melanomas has been described as a possible prognostic biomarker but there is no information about the mutant allele frequency (MAF) variability of BRAFV600E

Level Of Evidence Is Relevant
1a 1b 2a 2b 3a 3b 4

### Quantitative cell-free circulating BRAFV600E mutation analysis by use of droplet digital PCR in the follow-up of patients with melanoma being treated with BRAF inhibitors.

Sanmamed Miguel F & González Alvaro 2015-03-03
PMID : 25411185
publication\_types : Journal Article .

BACKGROUND: Around 50% of cutaneous melanomas harbor the BRAF (V600E) mutation and can be treated with BRAF inhibitors. DNA carrying this mutation can be released into circulation as cell-free BRAF (V600E)

Level Of Evidence Is Relevant
1a 1b 2a 2b 3a 3b 4

### BRAF V600E mutational load as a prognosis biomarker in malignant melanoma

Sevilla Arrate & Alonso Santos 2020-06-22
PMID : 32168325
publication\_types : Journal Article , Research Support, Non-U.S. Gov't .

Analyzing the mutational load of driver mutations in melanoma could provide valuable information regarding its progression. We aimed at analyzing the heterogeneity of mutational load of BRAF V600E in biopsies

Level Of Evidence Is Relevant
1a 1b 2a 2b 3a 3b 4

### Prospective immunohistochemical analysis of BRAF V600E mutation in melanoma.

Thiel Alexandra & Ristimäki Ari 2015-03-20
PMID : 25442222
publication\_types : Journal Article , Research Support, Non-U.S. Gov't .

The v-rat murine sarcoma viral oncogene homolog B1 (BRAF) V600E mutation is the most common activating genetic alteration of this oncogene and a predictive marker for the therapeutic use of BRAF inhibitors in

Level Of Evidence Is Relevant
1a 1b 2a 2b 3a 3b 4

### BRAFV600E Expression Is Homogenous and Associated with Nonrecurrent Disease and Better Survival in Primary Melanoma.

Naimy Soraya & Rahbek Ojerdum Lise Mette 2015-03-03
PMID : 36657398
publication\_types : Journal Article .

BACKGROUND: Superficial spreading melanomas (SSMs) are the most common type of melanoma and cause the majority of skin cancer deaths. More than 50% of cases harbor a mutation in the BRAF gene that activates

Level Of Evidence Is Relevant
1a 1b 2a 2b 3a 3b 4

### BRAFV600E protein expression and outcome from BRAF inhibitor treatment in BRAFV600E metastatic melanoma

Wilmott J S & Long G V 2013-04-23
PMID : 23403819
publication\_types : Journal Article , Research Support, Non-U.S. Gov't .

Background: To examine the association between level and patterns of baseline intra-tumoural BRAFV600E protein expression and clinical outcome of BRAFV600E melanoma patients treated with selective BRAF

Level Of Evidence Is Relevant
1a 1b 2a 2b 3a 3b 4

### The BRAF(V600E) causes widespread alterations in gene methylation in the genome of melanoma cells.

Hou Peng & Xing Mingzhao 2012-08-30
PMID : 22189819
publication\_types : Journal Article , Research Support, N.I.H., Extramural .

Although BRAF (V600E) is well known to play an important role in the tumorigenesis of melanoma, its molecular mechanism, particularly the epigenetic aspect has been incompletely understood. Here, we

Level Of Evidence Is Relevant
1a 1b 2a 2b 3a 3b 4



skin cancer



READABILITY LEVEL OF EVIDENCE ARTICLE TYPE DATE BIO CONCEPTS

INCLUDE CONCEPTS

X BRAF X BRAFV600E X MELANOMA

Showing 10 of 10000 results

### Immunohistochemistry as an accurate tool for evaluating BRAF-V600E mutation in 130 samples of papillary thyroid cancer.

Abd Elmaged Zakaria Y & Kandil Emad 2017-07-19
PMID : 2709446
publication\_types : Comparative Study , Journal Article .

BACKGROUND: BRAFV600E mutation has been investigated by immunohistochemistry and has shown high sensitivity and specificity. We aim to investigate the accuracy of immunohistochemistry versus molecular testing

Level Of Evidence: 1a 1b 2a 2b 3a 3b 4 (4 selected)
Readability: hard
Is Relevant: Yes

### Inhibiting BRAF Oncogene-Mediated Radioresistance Effectively Radiosensitizes BRAFV600E Mutant Thyroid Cancer Cells by Constraining DNA Double-strand Break Repair

Robb Ryan & Williams Terence M 2020-09-14
PMID : 31097454
publication\_types : Journal Article , Research Support, N.I.H., Extramural , Research Support, Non-U.S. Gov't .

Purpose: Activating BRAF mutations, most commonly BRAFV600E, are a major oncogenic driver of many cancers. We explored whether BRAFV600E promotes radiation resistance and whether selectively targeting

Level Of Evidence: 1a 1b 2a 2b 3a 3b 4 (2b selected)
Readability: hard
Is Relevant: Yes

### Evaluation of the expression levels of BRAFV600E mRNA in primary tumors of thyroid cancer using an ultrasensitive mutation assay

Tran Tien Viet & Ho Tho Huu 2021-01-22
PMID : 32357891
publication\_types : Evaluation Study , Journal Article .

Background The BRAFV600E gene encodes for the mutant BRAFV600E protein, which triggers downstream oncogenic signaling in thyroid cancer. Since most currently available methods have focused on detecting

Level Of Evidence: 1a 1b 2a 2b 3a 3b 4 (2b selected)
Readability: hard
Is Relevant: Yes

### Clinicopathologic Features and Prognosis of BRAF Mutated Colorectal Cancer Patients

Guan Wen-Long & Wang Feng-Hua 2017-07-19
PMID : 33330032
publication\_types : Journal Article .

Background BRAFV600E mutation is associated with poor prognosis of colorectal cancer (CRC) patients, but the comparison of clinic-pathologic features between V600E and non-V600E mutation was not well-known in

Level Of Evidence: 1a 1b 2a 2b 3a 3b 4 (2b selected)
Readability: hard
Is Relevant: Yes

### Preliminary Study on the Identification of BRAFV600E Mutation in Colorectal Cancer by Near-Infrared Spectroscopy.

Duan Jiale & Wang Yaolan 2017-07-19
PMID : 33376356
publication\_types : Journal Article .

Introduction: In metastatic colorectal cancer (mCRC), the B-type Raf kinase (BRAF) V600E mutation is a molecular biomarker of poor prognosis and is of great importance to drug target. Currently, the commonly

Level Of Evidence: 1a 1b 2a 2b 3a 3b 4 (2b selected)
Readability: hard
Is Relevant: Yes

### BRAF V600E mutations in right-side colon cancer: Heterogeneity detected by liquid biopsy.

Ueda Koji & Yoshida Hiroshi 2022-06-21
PMID : 35172933
publication\_types : Journal Article , Observational Study .

INTRODUCTION: The prognosis for metastatic colorectal cancer patients (mCRC) with the BRAFV600E mutation is poor. BRAFV600E mutation frequency is reportedly low among Asians; however, the frequency of the

Level Of Evidence: 1a 1b 2a 2b 3a 3b 4 (2b selected)
Readability: expert
Is Relevant: Yes

### The Effect of BRAF V600E Mutation on Lymph Node Involvement in Papillary Thyroid Cancer.

Sahin Samet & Baglan Tolga 2017-07-19
PMID : 33778379
publication\_types : Journal Article .

Objectives: Papillary thyroid cancer (PTC) is the most common well-differentiated thyroid cancer. Lymph node (LN) metastasis is frequently seen in PTC. The effect of BRAFV600E mutation on PTC-associated LN

Level Of Evidence: 1a 1b 2a 2b 3a 3b 4 (2b selected)
Readability: hard
Is Relevant: Yes

### THE RELATIONSHIP OF BRAFV600E MUTATION STATUS TO FDG PET/CT AVIDITY IN THYROID CANCER: A REVIEW AND META-ANALYSIS.

Santhanam Prasanna & Ladenson Paul W 2018-07-31
PMID : 29144823
publication\_types : Journal Article , Meta-Analysis , Review , Systematic Review .

OBJECTIVE: Papillary thyroid cancer (PTC) harboring a BRAFV600E gene mutation has been shown to exhibit aggressive tumor behavior and carries higher risks of recurrence and disease-specific death. In this

Level Of Evidence: 1a 1b 2a 2b 3a 3b 4 (2a selected)
Readability: hard
Is Relevant: Yes

### Genome-wide alterations in gene methylation by the BRAF V600E mutation in papillary thyroid cancer cells

Hou Peng & Xing Minghao 2012-03-27
PMID : 21937738
publication\_types : Journal Article , Research Support, N.I.H., Extramural .

The BRAF V600E mutation plays an important role in the tumorigenesis of papillary thyroid cancer (PTC). To explore an epigenetic mechanism involved in this process, we performed a genome-wide DNA methylation

Level Of Evidence: 1a 1b 2a 2b 3a 3b 4 (4 selected)
Readability: medium
Is Relevant: Yes

### Clinical Significance of BRAF Non-V600E Mutations in Colorectal Cancer: A Retrospective Study of Two Institutions.

Shimada Yoshitumi & Wakai Toshitumi 2019-04-26
PMID : 30463788
publication\_types : Journal Article , Research Support, Non-U.S. Gov't .

BACKGROUND: Recent advances in next-generation sequencing have enabled the detection of BRAF V600E mutations as well as BRAF non-V600E mutations in a single assay. The present work aimed to describe the

Level Of Evidence: 1a 1b 2a 2b 3a 3b 4 (3a selected)
Readability: hard
Is Relevant: Yes

# Bibliography

- [1] ABUL-HUSN, Noura S. ; KENNY, Eimear E.: Personalized medicine and the power of electronic health records. In: *Cell* 177 (2019), Nr. 1, S. 58–69
- [2] AGOSTI, Maristella ; SMEATON, Alan: *Information retrieval and hypertext*. Springer Science & Business Media, 2012
- [3] AGYAPONG, VIO ; KIRrane, R ; BANGARU, R: Medical confidentiality versus disclosure: Ethical and legal dilemmas. In: *Journal of Forensic and Legal Medicine* 16 (2009), Nr. 2, S. 93–96
- [4] AI@META: Llama 3 Model Card. (2024)
- [5] ALLOT, Alexis ; LEE, Kyubum ; CHEN, Qingyu ; LUO, Ling ; LU, Zhiyong: LitSuggest: a web-based system for literature recommendation and curation using machine learning. In: *Nucleic acids research* 49 (2021), Nr. W1, S. W352–W358
- [6] AMOS, Liz ; ANDERSON, David ; BRODY, Stacy ; RIPPLE, Anna ; HUMPHREYS, Betsy L.: UMLS users and uses: a current overview. In: *Journal of the American Medical Informatics Association* 27 (2020), Nr. 10, S. 1606–1611
- [7] ANDRADE, Chittaranjan: How to write a good abstract for a scientific paper or conference presentation. In: *Indian journal of psychiatry* 53 (2011), Nr. 2, S. 172
- [8] ARABZADEH, Negar ; YAN, Xinyi ; CLARKE, Charles L.: Predicting efficiency/effectiveness trade-offs for Dense vs. Sparse retrieval strategy selection. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, S. 2862–2866
- [9] ARNO, Anneliese ; THOMAS, James ; WALLACE, Byron ; MARSHALL, Iain J. ; MCKENZIE, Joanne E. ; ELLIOTT, Julian H.: Accuracy and Efficiency of Machine Learning–Assisted Risk-of-Bias Assessments in “Real-World” Systematic Reviews: A Noninferiority Randomized Controlled Trial. In: *Annals of Internal Medicine* 175 (2022), Nr. 7, S. 1001–1009
- [10] ARNOTT SMITH, Catherine ; PATRICK, Tim ; RHYNER, Paula ; SWAIN, Deborah ; DAVOLIO, Leonard ; MORRIS, Ted: Problems with the distribution of your health and medical information. In: *Proceedings of the American Society for Information Science and Technology* 44 (2007), Nr. 1, S. 1–6

- 
- [11] ARORA, Neeraj K. ; HESSE, Bradford W. ; RIMER, Barbara K. ; VISWANATH, Kasisomayajula ; CLAYMAN, Marla L. ; CROYLE, Robert T.: Frustrated and confused: the American public rates its cancer-related information-seeking experiences. In: *Journal of general internal medicine* 23 (2008), S. 223–228
- [12] ASAI, Akari ; WU, Zeqiu ; WANG, Yizhong ; SIL, Avirup ; HAJISHIRZI, Hannaneh: Self-rag: Learning to retrieve, generate, and critique through self-reflection. In: *arXiv preprint arXiv:2310.11511* (2023)
- [13] ATCHERSON, Samuel R. ; DELAUNE, Ashley E. ; HADDEN, Kristie ; ZRAICK, Richard I. ; KELLY-CAMPBELL, Rebecca J. ; MINAYA, Carlos P.: A computer-based readability analysis of consumer materials on the American Speech-Language-Hearing Association website. In: *Contemporary Issues in Communication Science and Disorders* 41 (2014), Nr. Spring, S. 12–23
- [14] AZEEZ, Nureni A. ; VAN DER VYVER, Charles: Security and privacy issues in e-health cloud-based system: A comprehensive content analysis. In: *Egyptian Informatics Journal* 20 (2019), Nr. 2, S. 97–108
- [15] BADA, Michael ; HUNTER, Lawrence: Desiderata for ontologies to be used in semantic annotation of biomedical documents. In: *Journal of Biomedical Informatics* 44 (2011), Nr. 1, S. 94–101
- [16] BAEZA-YATES, Ricardo ; RIBEIRO-NETO, Berthier [u. a.]: *Modern information retrieval*. ACM press New York, 1999
- [17] BAJPAI, Nidhi ; ARORA, Deepak: An estimation of user preferences for search engine results and its usage patterns. In: *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications: Proceedings of ICACNI 2016, Volume 2* Springer, 2018, S. 255–264
- [18] BASCH, Corey H. ; FERA, Joseph ; GARCIA, Phillip: Readability of influenza information online: implications for consumer health. In: *American Journal of Infection Control* 47 (2019), Nr. 11, S. 1298–1301
- [19] BECKER, Shirley A.: A study of web usability for older adults seeking online health resources. In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 11 (2004), Nr. 4, S. 387–406
- [20] BEHRENS, Heike: How difficult are complex verbs? Evidence from German, Dutch and English. (1998)
- [21] BERGE, Geir T. ; GRANMO, Ole-Christoffer ; TVEIT, Tor O. ; GOODWIN, Morten ; JIAO, Lei ; MATHEUSSEN, Bernt V.: Using the Tsetlin machine to learn human-interpretable rules for high-accuracy text categorization with medical applications. In: *IEEE Access* 7 (2019), S. 115134–115146

- [22] BERNERS-LEE, Tim ; CAILLIAU, Robert ; GROFF, Jean-François ; POLLERMANN, Bernd: World-Wide Web: the information universe. In: *Internet Research* 2 (1992), Nr. 1, S. 52–58
- [23] BHAVNANI, Suresh K.: Domain-specific search strategies for the effective retrieval of healthcare and shopping information. In: *CHI'02 extended abstracts on human factors in computing systems*, 2002, S. 610–611
- [24] BODENREIDER, Olivier: The unified medical language system (UMLS): integrating biomedical terminology. In: *Nucleic acids research* 32 (2004), Nr. suppl\_1, S. D267–D270
- [25] BOLUYT, N ; SCHOLTEN, RJ ; OFFRINGA, M: The practice of systematic reviews. XI. The Cochrane Library. In: *Nederlands tijdschrift voor geneeskunde* 147 (2003), Nr. 52, S. 2572–2577
- [26] BORAWSKI, Kristy M. ; NORRIS, Regina D. ; FESPERMAN, Susan F. ; VIEWEG, Johannes ; PREMINGER, Glenn M. ; DAHM, Philipp: Levels of evidence in the urological literature. In: *The Journal of urology* 178 (2007), Nr. 4, S. 1429–1433
- [27] BORLUND, Pia: The concept of relevance in IR. In: *Journal of the American Society for information Science and Technology* 54 (2003), Nr. 10, S. 913–925
- [28] BOZDAG, Engin: Bias in algorithmic filtering and personalization. In: *Ethics and information technology* 15 (2013), S. 209–227
- [29] BRAVO, Àlex ; PIÑERO, Janet ; QUERALT-ROSINACH, Núria ; RAUTSCHKA, Michael ; FURLONG, Laura I.: Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. In: *BMC bioinformatics* 16 (2015), S. 1–17
- [30] BROWN, David: A review of the PubMed PICO tool: using evidence-based practice in health education. In: *Health promotion practice* 21 (2020), Nr. 4, S. 496–498
- [31] CAMBAZOGLU, Berkant B. ; BAEZA-YATES, Ricardo: Scalability challenges in web search engines. In: *Advanced topics in information retrieval*. Springer, 2011, S. 27–50
- [32] CAMPOS, David ; MATOS, Sérgio ; OLIVEIRA, José Luís: A modular framework for biomedical concept recognition. In: *BMC bioinformatics* 14 (2013), S. 1–21
- [33] CAN, Aysu B. ; BAYKAL, Nazife: MedicoPort: A medical search engine for all. In: *Computer methods and programs in biomedicine* 86 (2007), Nr. 1, S. 73–86
- [34] CAO, Zhe ; QIN, Tao ; LIU, Tie-Yan ; TSAI, Ming-Feng ; LI, Hang: Learning to rank: from pairwise approach to listwise approach. In: *Proceedings of the 24th international conference on Machine learning*, 2007, S. 129–136
- [35] CASTRO, Victor M. ; DLOGACH, Dmitriy ; FINAN, Sean ; YU, Sheng ; CAN, Anil ; ABD-

- EL-BARR, Muhammad ; GAINER, Vivian ; SHADICK, Nancy A. ; MURPHY, Shawn ; CAI, Tianxi [u. a.]: Large-scale identification of patients with cerebral aneurysms using natural language processing. In: *Neurology* 88 (2017), Nr. 2, S. 164–168
- [36] CERI, Stefano ; BOZZON, Alessandro ; BRAMBILLA, Marco ; DELLA VALLE, Emanuele ; FRATERNALI, Piero ; QUARTERONI, Silvia ; CERI, Stefano ; BOZZON, Alessandro ; BRAMBILLA, Marco ; DELLA VALLE, Emanuele [u. a.]: An introduction to information retrieval. In: *Web information retrieval* (2013), S. 3–11
- [37] CHALL, JS. *Readability: An appraisal of research and application. Bureau of Educational Research Monographs, No. 34. Columbus.* 1958
- [38] CHANG, Yupeng ; WANG, Xu ; WANG, Jindong ; WU, Yuan ; YANG, Linyi ; ZHU, Kaijie ; CHEN, Hao ; YI, Xiaoyuan ; WANG, Cunxiang ; WANG, Yidong [u. a.]: A survey on evaluation of large language models. In: *ACM Transactions on Intelligent Systems and Technology* (2023)
- [39] CHARIF, Ali B. ; HASSANI, Kasra ; WONG, Sabrina T. ; ZOMAHOUN, Hervé Tchala V. ; FORTIN, Martin ; FREITAS, Adriana ; KATZ, Alan ; KENDALL, Claire E. ; LIDDY, Clare ; NICHOLSON, Kathryn [u. a.]: Assessment of scalability of evidence-based innovations in community-based primary health care: a cross-sectional study. In: *Canadian Medical Association Open Access Journal* 6 (2018), Nr. 4, S. E520–E527
- [40] CHEN, Jia ; LIU, Yiqun ; MAO, Jiaxin ; ZHANG, Fan ; SAKAI, Tetsuya ; MA, Weizhi ; ZHANG, Min ; MA, Shaoping: Incorporating query reformulating behavior into web search evaluation. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, S. 171–180
- [41] CHEN, Jiawei ; LIN, Hongyu ; HAN, Xianpei ; SUN, Le: Benchmarking large language models in retrieval-augmented generation. In: *Proceedings of the AAAI Conference on Artificial Intelligence* Bd. 38, 2024, S. 17754–17762
- [42] CHEN, Ye ; ZHOU, Ke ; LIU, Yiqun ; ZHANG, Min ; MA, Shaoping: Meta-evaluation of online and offline web search evaluation metrics. In: *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, 2017, S. 15–24
- [43] CHISHOLM, Orin ; SHARRY, Patrick ; PHILLIPS, Lawrence: Multi-criteria decision analysis for benefit-risk analysis by national regulatory authorities. In: *Frontiers in medicine* 8 (2022), S. 820335
- [44] COLLINS-THOMPSON, Kevyn ; BENNETT, Paul N. ; WHITE, Ryen W. ; DE LA CHICA, Sebastian ; SONTAG, David: Personalizing web search results by reading level. In: *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, S. 403–412



- [45] COLLINS-THOMPSON, Kevyn ; CALLAN, Jamie: Predicting reading difficulty with statistical language models. In: *Journal of the American Society for Information Science and Technology* 56 (2005), Nr. 13, S. 1448–1462
- [46] DA COSTA, Marcelo José M.: THE MULTIDISCIPLINARY LOOK AT THE ONCOLOGICAL PATIENT: PALLIATIVE CARE, SPIRITUALITY AND BIOETHICS. In: *Health and Society* 3 (2023), Nr. 02, S. 277–315
- [47] CULLEN, Rowena J.: In search of evidence: family practitioners' use of the Internet for clinical information. In: *Journal of the Medical Library Association* 90 (2002), Nr. 4, S. 370
- [48] DAEI, Azra ; SOLEYMANI, Mohammad R. ; ASHRAFI-RIZI, Hasan ; ZARGHAM-BOROUJENI, Ali ; KELISHADI, Roya: Clinical information seeking behavior of physicians: A systematic review. In: *International journal of medical informatics* 139 (2020), S. 104144
- [49] DALE, Edgar ; CHALL, Jeanne S.: A formula for predicting readability: Instructions. In: *Educational research bulletin* (1948), S. 37–54
- [50] DANG, Huong ; LEE, Kahyun ; HENRY, Sam ; UZUNER, Özlem: Ensemble BERT for Classifying Medication-mentioning Tweets. In: *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. Barcelona, Spain : Association for Computational Linguistics, Dezember 2020, S. 37–41
- [51] DEEPSET-AI ; WINDOWS PHONE CENTRAL (Hrsg.). *state-of-the-art German BERT model trained from scratch*
- [52] DENECKE, Karl: Semantic structuring of and information extraction from medical documents using the UMLS. In: *Methods of Information in Medicine* 47 (2008), Nr. 05, S. 425–434
- [53] VAN DER VEGT, Anton ; ZUCCON, Guido ; KOOPMAN, Bevan: Do better search engines really equate to better clinical decisions? If not, why not? In: *Journal of the Association for Information Science and Technology* 72 (2021), Nr. 2, S. 141–155
- [54] DESAI, Vishal S. ; CAMP, Christopher L. ; KRYCH, Aaron J.: What is the hierarchy of clinical evidence? In: *Basic Methods Handbook for Clinical Orthopaedic Research: A Practical Guide and Case Based Research Approach* (2019), S. 11–22
- [55] DEVLIN, Jacob ; CHANG, Ming-Wei ; LEE, Kenton ; TOUTANOVA, Kristina: Bert: Pre-training of deep bidirectional transformers for language understanding. In: *arXiv preprint arXiv:1810.04805* (2018)
- [56] DIAS, Amanda G. ; MILIOS, Evangelos E. ; DE OLIVEIRA, Maria Cristina F.: Trivir: A visualization system to support document retrieval with high recall. In: *Proceedings of the ACM Symposium on Document Engineering 2019*, 2019, S. 1–10

- [57] DJULBEGOVIC, Benjamin ; GUYATT, Gordon H.: Progress in evidence-based medicine: a quarter century on. In: *The lancet* 390 (2017), Nr. 10092, S. 415–423
- [58] DONG, Guanting ; YUAN, Hongyi ; LU, Keming ; LI, Chengpeng ; XUE, Mingfeng ; LIU, Dayiheng ; WANG, Wei ; YUAN, Zheng ; ZHOU, Chang ; ZHOU, Jingren: How abilities in large language models are affected by supervised fine-tuning data composition. In: *arXiv preprint arXiv:2310.05492* (2023)
- [59] DUBAY, William H.: The principles of readability. In: *Online Submission* (2004)
- [60] EKSTRAND, Michael D. ; DAS, Anubrata ; BURKE, Robin ; DIAZ, Fernando [u. a.]: Fairness in information access systems. In: *Foundations and Trends® in Information Retrieval* 16 (2022), Nr. 1-2, S. 1–177
- [61] EL-ANSARI, Anas ; BENI-HSSANE, Abderrahim ; SAADI, Mostafa: An improved modeling method for profile-based personalized search. In: *Proceedings of the 3rd international conference on networking, information systems & security*, 2020, S. 1–6
- [62] ELLIOTT, John O. ; SHNEKER, Bassel F.: A health literacy assessment of the epilepsy.com website. In: *Seizure* 18 (2009), Nr. 6, S. 434–439
- [63] ENTIN, Eileen B.: Relationships of measures of interest, prior knowledge, and readability to comprehension of expository passages. (1981)
- [64] EYSENBACH, Gunther ; KÖHLER, Christian: How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. In: *Bmj* 324 (2002), Nr. 7337, S. 573–577
- [65] FAGGIOLI, Guglielmo ; DIETZ, Laura ; CLARKE, Charles L. ; DEMARTINI, Gianluca ; HAGEN, Matthias ; HAUFF, Claudia ; KANDO, Noriko ; KANOULAS, Evangelos ; POTTHAST, Martin ; STEIN, Benno [u. a.]: Perspectives on large language models for relevance judgment. In: *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, 2023, S. 39–50
- [66] FAJARDO, Michael A. ; WEIR, Kristie R. ; BONNER, Carissa ; GNJIDIC, Danijela ; JANSEN, Jesse: Availability and readability of patient education materials for de-prescribing: an environmental scan. In: *British journal of clinical pharmacology* 85 (2019), Nr. 7, S. 1396–1406
- [67] FAN, Yixing ; GUO, Jiafeng ; MA, Xinyu ; ZHANG, Ruqing ; LAN, Yanyan ; CHENG, Xueqi: A linguistic study on relevance modeling in information retrieval. In: *Proceedings of the Web Conference 2021*, 2021, S. 1053–1064
- [68] FERRANTE, Marco ; FERRO, Nicola ; FUHR, Norbert: Towards meaningful statements in IR evaluation: Mapping evaluation measures to interval scales. In: *IEEE Access* 9 (2021), S. 136182–136216

- [69] FIORINI, Nicolas ; CANESE, Kathi ; STARCHENKO, Grisha ; KIREEV, Evgeny ; KIM, Won ; MILLER, Vadim ; OSIPOV, Maxim ; KHOLODOV, Michael ; ISMAGILOV, Rafis ; MOHAN, Sunil [u. a.]: Best match: new relevance search for PubMed. In: *PLoS biology* 16 (2018), Nr. 8, S. e2005343
- [70] FITZSIMMONS, Paul R. ; MICHAEL, BD ; HULLEY, Joane L. ; SCOTT, G O.: A readability assessment of online Parkinson's disease information. In: *The journal of the Royal College of Physicians of Edinburgh* 40 (2010), Nr. 4, S. 292–296
- [71] FORD, Jennifer ; KORJONEN, Helena: Information needs of public health practitioners: a review of the literature. In: *Health Information & Libraries Journal* 29 (2012), Nr. 4, S. 260–273
- [72] FREJ, Jibril ; CHEVALLET, Jean-Pierre ; SCHWAB, Didier: Knowledge based transformer model for information retrieval. In: *Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020)* Bd. 2621, 2020
- [73] FRIHAT, Sameh ; BECKMANN, Catharina L. ; HARTMANN, Eva M. ; FUHR, Norbert: Document Difficulty Aspects for Medical Practitioners: Enhancing Information Retrieval in Personalized Search Engines. In: *Applied Sciences* 13 (2023), Nr. 19, S. 10612
- [74] FUHR, Norbert: A probability ranking principle for interactive information retrieval. In: *Information Retrieval* 11 (2008), S. 251–265
- [75] FUHR, Norbert: Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. In: *SIGIR Forum* 51 (2017), Nr. 3, S. 32–41
- [76] FUHR, Norbert ; GIACHANOU, Anastasia ; GREFENSTETTE, Gregory ; GUREVYCH, Iryna ; HANSELOWSKI, Andreas ; JARVELIN, Kalervo ; JONES, Rosie ; LIU, YiquN ; MOTHE, Josiane ; NEJDL, Wolfgang [u. a.]: An information nutritional label for online documents. In: *ACM SIGIR Forum* Bd. 51 ACM New York, NY, USA, 2018, S. 46–66
- [77] FUNG, Adrian Chi H. ; LEE, Mathew Hon L. ; LEUNG, Ling ; CHAN, Ivy Hau Y. ; KENNETH, Wong: Internet Health Resources on Nocturnal Enuresis—A Readability, Quality and Accuracy Analysis. In: *European Journal of Pediatric Surgery* (2023)
- [78] GARCÍA, Marcos Antonio M. ; RODRÍGUEZ, Roberto P. ; RIFÓN, Luis A.: Leveraging Wikipedia knowledge to classify multilingual biomedical documents. In: *Artificial intelligence in medicine* 88 (2018), S. 37–57
- [79] GARCIA-GATHRIGHT, Jean ; HOSEY, Christine ; THOMAS, Brian S. ; CARTERETTE, Ben ; DIAZ, Fernando: Mixed methods for evaluating user satisfaction. In: *Proceedings of the 12th ACM conference on recommender systems*, 2018, S. 541–542
- [80] GOBEILL, Julien ; CAUCHETEUR, Déborah ; MICHEL, Pierre-André ; MOTTIN, Luc ; PASCHE, Emilie ; RUCH, Patrick: SIB Literature Services: RESTful customizable

- search engines in biomedical literature, enriched with automatically mapped biomedical concepts. In: *Nucleic acids research* 48 (2020), Nr. W1, S. W12–W16
- [81] GOEURIOT, Lorraine ; JONES, Gareth J. ; KELLY, Liadh ; LEVELING, Johannes ; HANBURY, Allan ; MÜLLER, Henning ; SALANTERÄ, Sanna ; SUOMINEN, Hanna ; ZUCCON, Guido: ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. In: *CLEF 2013 online working notes* (2013)
- [82] GRABER, Mark A. ; ROLLER, Cathy M. ; KAEBLE, Betsy [u. a.]: Readability levels of patient education material on the World Wide Web. In: *Journal of Family Practice* 48 (1999), S. 58–61
- [83] GRANITZER, Michael ; VOIGT, Stefan ; FATHIMA, Noor A. ; GOLASOWSKI, Martin ; GUETL, Christian ; HECKING, Tobias ; HENDRIKSEN, Gijs ; HIEMSTRA, Djoerd ; MARTINOVIČ, Jan ; MITROVIĆ, Jelena [u. a.]: Impact and development of an Open Web Index for open web search. In: *Journal of the Association for Information Science and Technology* 75 (2024), Nr. 5, S. 512–520
- [84] GRIMM, Angela ; HÜBNER, Julia: Nonword repetition by bilingual learners of German: the role of language-specific complexity. In: *Bilingualism and Specific Language Impairment, Bi-SLI* 201 (2017)
- [85] GROUP, Evidence-Based Medicine W. ; GUYATT, Gordon ; RENNIE, Drummond [u. a.]: *Users' guides to the medical literature: a manual for evidence-based clinical practice*. AMA Press, 2002
- [86] GU, Yu ; TINN, Robert ; CHENG, Hao ; LUCAS, Michael ; USUYAMA, Naoto ; LIU, Xiaodong ; NAUMANN, Tristan ; GAO, Jianfeng ; POON, Hoifung. *Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing*. 2020
- [87] GUO, Jiafeng ; CAI, Yinqiong ; FAN, Yixing ; SUN, Fei ; ZHANG, Ruqing ; CHENG, Xueqi: Semantic models for the first-stage retrieval: A comprehensive review. In: *ACM Transactions on Information Systems (TOIS)* 40 (2022), Nr. 4, S. 1–42
- [88] GUO, Tonglei ; GUO, Jiafeng ; FAN, Yixing ; LAN, Yanyan ; XU, Jun ; CHENG, Xueqi: A comparison between term-based and embedding-based methods for initial retrieval. In: *Information Retrieval: 24th China Conference, CCIR 2018, Guilin, China, September 27–29, 2018, Proceedings 24* Springer, 2018, S. 28–40
- [89] GUPTA, Supriya ; SHARAFF, Aakanksha ; NAGWANI, Naresh K.: Query based biomedical document retrieval for clinical information access with the semantic similarity. In: *Multim. Tools Appl.* 83 (2024), Nr. 18, S. 55305–55317
- [90] HANAUER, David A. ; MEI, Qiaozhu ; LAW, James ; KHANNA, Ritu ; ZHENG, Kai: Supporting information retrieval from electronic health records: a report of University

- of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). In: *Journal of biomedical informatics* 55 (2015), S. 290–300
- [91] HARDENBOL, Alec X. ; KNOLS, Bram ; LOUWS, Mathijs ; MEULENDIJK, Michiel ; ASKARI, Marjan: Usability aspects of medication-related decision support systems in the outpatient setting: a systematic literature review. In: *Health informatics journal* 26 (2020), Nr. 1, S. 72–87
- [92] HARPER, David J. ; KELLY, Diane: Contextual relevance feedback. In: *Proceedings of the 1st international Conference on information interaction in Context*, 2006, S. 129–137
- [93] HARTLEY, James: Current findings from research on structured abstracts. In: *Journal of the Medical Library Association* 92 (2004), Nr. 3, S. 368
- [94] HARTLING, Lisa ; GATES, Allison: Friend or Foe? The Role of Robots in Systematic Reviews. In: *Annals of Internal Medicine* 175 (2022), Nr. 7, S. 1045–1046
- [95] HASHAVIT, Anat ; WANG, Hongning ; LIN, Raz ; STERN, Tamar ; KRAUS, Sarit: Understanding and mitigating bias in online health search. In: *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2021, S. 265–274
- [96] HÄTTY, Anna ; SCHLECHTWEG, Dominik ; DORNA, Michael ; IM WALDE, Sabine S.: Predicting degrees of technicality in automatic terminology extraction. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, S. 2883–2889
- [97] HE, Jiyin ; QVARFORDT, Pernilla ; HALVEY, Martin ; GOLOVCHINSKY, Gene: Beyond actions: Exploring the discovery of tactics from user logs. In: *Information Processing & Management* 52 (2016), Nr. 6, S. 1200–1226
- [98] HEDMAN, Amy S.: Using the SMOG formula to revise a health-related document. In: *American Journal of Health Education* 39 (2008), Nr. 1, S. 61–64
- [99] HERCBERG, Serge ; TOUVIER, Mathilde ; SALAS-SALVADO, Jordi. *The nutri-score nutrition label*. 2021
- [100] HERRETT, Emily ; GALLAGHER, Arlene M. ; BHASKARAN, Krishnan ; FORBES, Harriet ; MATHUR, Rohini ; VAN STAA, Tjeerd ; SMEETH, Liam: Data resource profile: clinical practice research datalink (CPRD). In: *International journal of epidemiology* 44 (2015), Nr. 3, S. 827–836
- [101] HERSH, William R.: *Information retrieval: a health and biomedical perspective*. Bd. 3. Springer, 2009

- [102] HERSKOVIC, Jorge R. ; TANAKA, Len Y. ; HERSH, William ; BERNSTAM, Elmer V.: A day in the life of PubMed: analysis of a typical day's query log. In: *Journal of the American Medical Informatics Association* 14 (2007), Nr. 2, S. 212–220
- [103] HIRT, Julian ; MEICHLINGER, Jasmin ; SCHUMACHER, Petra ; MUELLER, Gerhard: Agreement in Risk of Bias Assessment Between RobotReviewer and Human Reviewers: An Evaluation Study on Randomised Controlled Trials in Nursing-Related Cochrane Reviews. In: *Journal of Nursing Scholarship* 53 (2021), Nr. 2, S. 246–254
- [104] HOCKETT, Charles F.: A course in modern linguistics. (1958)
- [105] HOWICK, Jeremy: Oxford Centre for Evidence-Based Medicine: Levels of Evidence. In: <https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009> (2009)
- [106] HSIEH-YEE, Ingrid: Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. In: *Journal of the american society for information science* 44 (1993), Nr. 3, S. 161–174
- [107] HU, Jian ; WANG, Gang ; LOCHOVSKY, Fred ; SUN, Jian-tao ; CHEN, Zheng: Understanding user's query intent with wikipedia. In: *Proceedings of the 18th international conference on World wide web*, 2009, S. 471–480
- [108] HU, Ze ; ZHANG, Zhan ; YANG, Haiqin ; CHEN, Qing ; ZHU, Rong ; ZUO, Decheng: Predicting the quality of online health expert question-answering services with temporal features in a deep learning framework. In: *Neurocomputing* 275 (2018), S. 2769–2782
- [109] INGWERSEN, Peter ; JÄRVELIN, Kalervo ; BELKIN, Nick ; LARSEN, Birger ; WHITE, Ryan W.: Proceedings of the ACM SIGIR 2005 Workshop on Information Retrieval in Context (IRiX). (2005)
- [110] ISLAMAJ DOGAN, Rezarta ; MURRAY, G C. ; NÉVÉOL, Aurélie ; LU, Zhiyong: Understanding PubMed® user search behavior through log analysis. In: *Database* 2009 (2009), S. bap018
- [111] JABER, Areej ; MARTÍNEZ, Paloma: Disambiguating Clinical Abbreviations using Pre-trained Word Embeddings. In: *HEALTHINF*, 2021, S. 501–508
- [112] JAMEEL, Shoaib ; QIAN, Xiaojun: An unsupervised technical readability ranking model by building a conceptual terrain in LSI. In: *2012 Eighth International Conference on Semantics, Knowledge and Grids* IEEE, 2012, S. 39–46
- [113] JANSEN, Bernard J. ; BOOTH, Danielle L. ; SPINK, Amanda: Determining the user intent of web search engine queries. In: *Proceedings of the 16th international conference on World Wide Web*, 2007, S. 1149–1150

- [114] JIANG, Jiepu ; HASSAN AWADALLAH, Ahmed ; SHI, Xiaolin ; WHITE, Ryan W.: Understanding and predicting graded search satisfaction. In: *Proceedings of the eighth ACM international conference on web search and data mining*, 2015, S. 57–66
- [115] JIN, Qiao ; DHINGRA, Bhuwan ; LIU, Zhengping ; COHEN, William W. ; LU, Xinghua: Pubmedqa: A dataset for biomedical research question answering. In: *arXiv preprint arXiv:1909.06146* (2019)
- [116] JU, Michelle R. ; KARALIS, John D. ; BLACKWELL, James-Michael ; MANSOUR, John C. ; POLANCO, Patricio M. ; AUGUSTINE, Mathew ; YOPP, Adam C. ; ZEH, Herbert J. ; WANG, Sam C. ; POREMBKA, Matthew R.: Inaccurate clinical stage is common for gastric adenocarcinoma and is associated with undertreatment and worse outcomes. In: *Annals of Surgical Oncology* 28 (2021), S. 2831–2843
- [117] KAINGADE, Rasika M. ; TIRMARE, Hemant A.: Personalization of Web Search based on privacy protected and auto-constructed user profile. In: *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) IEEE*, 2015, S. 818–823
- [118] KANE, Lorna ; CARTHY, Joe ; DUNNION, John: Readability applied to information retrieval. In: *European Conference on Information Retrieval* Springer, 2006, S. 523–526
- [119] KARASNEH, Reema ; AL-MISTAREHI, Abdel-Hameed ; AL-AZZAM, Sayer ; ABUHAMMAD, Sawsan ; MUFLIH, Suhaib M. ; HAWAMDEH, Sahar ; ALZOUBI, Karem H.: Physicians’ knowledge, perceptions, and attitudes related to patient confidentiality and data sharing. In: *International Journal of General Medicine* (2021), S. 721–731
- [120] KELLY, Diane [u. a.]: Methods for evaluating interactive information retrieval systems with users. In: *Foundations and Trends® in Information Retrieval* 3 (2009), Nr. 1–2, S. 1–224
- [121] KELLY, Dominique ; CHEN, Yimin ; CORNWELL, Sarah E. ; DELELLIS, Nicole S. ; MAYHEW, Alex ; ONAOLAPO, Sodiq ; RUBIN, Victoria L.: Bing chat: the future of search engines? In: *Proceedings of the Association for Information Science and Technology* 60 (2023), Nr. 1, S. 1007–1009
- [122] KHATTAB, Omar ; ZAHARIA, Matei: Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020, S. 39–48
- [123] KIM, Hyeoneui ; GORYACHEV, Sergey ; ROSEMBLAT, Graciela ; BROWNE, Allen ; KESSELMAN, Alla ; ZENG-TREITLER, Qing: Beyond surface characteristics: a new health text-specific readability measurement. In: *AMIA Annual Symposium Proceedings* Bd. 2007 American Medical Informatics Association, 2007, S. 418

- [124] KIM, Jeongkyun ; SO, Seongeun ; LEE, Hee-Jin ; PARK, Jong C. ; KIM, Jung-jae ; LEE, Hyunju: DigSee: disease gene search engine with evidence sentences (version cancer). In: *Nucleic acids research* 41 (2013), Nr. W1, S. W510–W517
- [125] KIM, Jin Y. ; COLLINS-THOMPSON, Kevyn ; BENNETT, Paul N. ; DUMAIS, Susan T.: Characterizing web content, user interests, and search behavior by reading level and topic. In: *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012, S. 213–222
- [126] KIM, Jin Y. ; CROFT, W B.: A field relevance model for structured document retrieval. In: *European Conference on Information Retrieval* Springer, 2012, S. 97–108
- [127] KLARE, George R.: The formative years. In: *Readability: Its past, present, and future* (1988), S. 14–34
- [128] KLATT, Edward C. ; KLATT, Carolyn A.: How much is too much reading for medical students? Assigned reading and reading rates at one medical school. In: *Academic Medicine* 86 (2011), Nr. 9, S. 1079–1083
- [129] KOCABALLI, Ahmet B. ; BERKOVSKY, Shlomo ; QUIROZ, Juan C. ; LARANJO, Liliana ; TONG, Huong L. ; REZAZADEGAN, Dana ; BRIATORE, Agustina ; COIERA, Enrico: The personalization of conversational agents in health care: systematic review. In: *Journal of medical Internet research* 21 (2019), Nr. 11, S. e15360
- [130] KOLUKISA, Burak ; DEDETURK, Bilge K. ; DEDETURK, Beyhan A. ; GULSEN, Abdulkadir ; BAKAL, Gokhan: A Comparative Analysis on Medical Article Classification Using Text Mining & Machine Learning Algorithms. In: *2021 6th International Conference on Computer Science and Engineering (UBMK)* IEEE, 2021, S. 360–365
- [131] KOO, Terry K. ; LI, Mae Y.: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. In: *Journal of chiropractic medicine* 15 (2016), Nr. 2, S. 155–163
- [132] KRITIKOU, Yiouli ; DEMESTICHAS, Panagiotis ; ADAMOPOULOU, Evgenia ; DEMESTICHAS, Konstantinos ; THEOLOGOU, Michael ; PARADIA, Maria: User Profile Modeling in the context of web-based learning management systems. In: *Journal of Network and Computer Applications* 31 (2008), Nr. 4, S. 603–627
- [133] KRUSCHWITZ, Udo ; HULL, Charlie [u. a.]: Searching the enterprise. In: *Foundations and Trends® in Information Retrieval* 11 (2017), Nr. 1, S. 1–142
- [134] KÜPER, Alisa ; KRÄMER, Nicole: Psychological Traits and Appropriate Reliance: Factors Shaping Trust in AI. In: *International Journal of Human–Computer Interaction* 0 (2024), Nr. 0, S. 1–17
- [135] LAHAV, Dan ; FALCON, Jon S. ; KUEHL, Bailey ; JOHNSON, Sophie ; PARASA, Sra-vanthi ; SHOMRON, Noam ; CHAU, Duen H. ; YANG, Diyi ; HORVITZ, Eric ; WELD,



- Daniel S. [u. a.]: A search engine for discovery of scientific challenges and directions. In: *Proceedings of the AAAI Conference on Artificial Intelligence* Bd. 36, 2022, S. 11982–11990
- [136] LASHKARI, Arash H. ; MAHDAVI, Fereshteh ; GHOMI, Vahid: A boolean model in information retrieval for search engines. In: *2009 International Conference on Information Management and Engineering* IEEE, 2009, S. 385–389
- [137] LEE, Hee-Jin ; DANG, Tien C. ; LEE, Hyunju ; PARK, Jong C.: OncoSearch: cancer gene search engine with literature evidence. In: *Nucleic acids research* 42 (2014), Nr. W1, S. W416–W421
- [138] LEE, Kyubum ; WEI, Chih-Hsuan ; LU, Zhiyong: Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature. In: *Briefings in bioinformatics* 22 (2021), Nr. 3, S. bbaa142
- [139] LEE, Sunwon ; KIM, Donghyeon ; LEE, Kyubum ; CHOI, Jaehoon ; KIM, Seongsoon ; JEON, Minji ; LIM, Sangrak ; CHOI, Donghee ; KIM, Sunkyu ; TAN, Aik-Choon [u. a.]: BEST: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. In: *PloS one* 11 (2016), Nr. 10, S. e0164680
- [140] LEEFLANG, Mariska M. ; DEEKS, Jonathan J. ; GATSONIS, Constantine ; BOSSUYT, Patrick M.: Systematic reviews of diagnostic test accuracy. In: *Annals of internal medicine* 149 (2008), Nr. 12, S. 889–897
- [141] LENSTRA, Noah ; ROBERTS, Joanne: Public libraries and health promotion partnerships: Needs and opportunities. In: *Evidence Based Library and Information Practice* 18 (2023), Nr. 1, S. 76–99
- [142] LEROY, Gondy ; XU, Jennifer ; CHUNG, Wingyan ; EGGERS, Shauna ; CHEN, Hsinchun: An end user evaluation of query formulation and results review tools in three medical meta-search engines. In: *International journal of medical informatics* 76 (2007), Nr. 11-12, S. 780–789
- [143] LETT, Elle ; ORJI, Whitney U. ; SEBRO, Ronnie: Declining racial and ethnic representation in clinical academic medicine: a longitudinal study of 16 US medical specialties. In: *PLoS One* 13 (2018), Nr. 11, S. e0207274
- [144] LEWIS, Patrick ; PEREZ, Ethan ; PIKTUS, Aleksandra ; PETRONI, Fabio ; KARPUKHIN, Vladimir ; GOYAL, Naman ; KÜTTLER, Heinrich ; LEWIS, Mike ; YIH, Wen-tau ; ROCKTÄSCHEL, Tim [u. a.]: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: *Advances in Neural Information Processing Systems* 33 (2020), S. 9459–9474
- [145] LEWIS, Ruth A. ; HUGHES, Dyfrig ; SUTTON, Alex J. ; WILKINSON, Clare: Quantitative evidence synthesis methods for the assessment of the effectiveness of treatment

- sequences for clinical and economic decision making: a review and taxonomy of simplifying assumptions. In: *Pharmacoeconomics* 39 (2021), Nr. 1, S. 25–61
- [146] LI, Peng-Hsuan ; CHEN, Ting-Fu ; YU, Jheng-Ying ; SHIH, Shang-Hung ; SU, Chan-Hung ; LIN, Yin-Hung ; TSAI, Huai-Kuang ; JUAN, Hsueh-Fen ; CHEN, Chien-Yu ; HUANG, Jia-Hsin: pubmedKB: an interactive web server for exploring biomedical entity relations in the biomedical literature. In: *Nucleic Acids Research* 50 (2022), Nr. W1, S. W616–W622
- [147] LI, Pengfei ; JIANG, Li ; ZHANG, Xinke ; LIU, Jun ; WANG, Hua: CD30 expression is a novel prognostic indicator in extranodal natural killer/T-cell lymphoma, nasal type. In: *BMC cancer* 14 (2014), S. 1–8
- [148] LI, Xiangsheng ; MAO, Jiaxin ; MA, Weizhi ; LIU, Yiqun ; ZHANG, Min ; MA, Shaoping ; WANG, Zhaowei ; HE, Xiuqiang: Topic-enhanced knowledge-aware retrieval model for diverse relevance estimation. In: *Proceedings of the Web Conference 2021*, 2021, S. 756–767
- [149] LI, Yizhi ; LIU, Zhenghao ; XIONG, Chenyan ; LIU, Zhiyuan: More robust dense retrieval with contrastive dual learning. In: *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, 2021, S. 287–296
- [150] LIN, Jimmy ; MA, Xueguang ; LIN, Sheng-Chieh ; YANG, Jheng-Hong ; PRADEEP, Ronak ; NOGUEIRA, Rodrigo: Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, S. 2356–2362
- [151] LIN, Sheng-Chieh ; YANG, Jheng-Hong ; LIN, Jimmy: Contextualized query embeddings for conversational search. In: *arXiv preprint arXiv:2104.08707* (2021)
- [152] LIU, Jingjing ; LIU, Chang ; BELKIN, Nicholas J.: Personalization in text information retrieval: A survey. In: *Journal of the Association for Information Science and Technology* 71 (2020), Nr. 3, S. 349–369
- [153] LIU, Jiqun: *A Behavioral Economics Approach to Interactive Information Retrieval: Understanding and Supporting Boundedly Rational Users*. Bd. 48. Springer Nature, 2023
- [154] LIU, Tie-Yan [u. a.]: Learning to rank for information retrieval. In: *Foundations and Trends® in Information Retrieval* 3 (2009), Nr. 3, S. 225–331
- [155] LIU, Xiaoyong ; CROFT, W B. ; OH, Paul ; HART, David: Automatic recognition of reading levels from user queries. In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, S. 548–549

- [156] LIU, Yanmeng ; JI, Meng ; LIN, Shannon S. ; ZHAO, Mengdan ; LYV, Ziqing: Combining readability formulas and machine learning for reader-oriented evaluation of online health resources. In: *IEEE Access* 9 (2021), S. 67610–67619
- [157] LOPES, Carla T.: Context-based health information retrieval. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, S. 845–845
- [158] LOPES, Carla T.: Health Information Retrieval–State of the art report. In: *arXiv preprint arXiv:2205.09083* (2022)
- [159] LYKKE, Marianne ; LARSEN, Birger ; LUND, Haakon ; INGWERSEN, Peter: Developing a test collection for the evaluation of integrated search. In: *Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings 32* Springer, 2010, S. 627–630
- [160] MACKWAY-JONES, K: Towards evidence based emergency medicine: best BETs from the Manchester Royal Infirmary. In: *Emergency medicine journal* 20 (2003), Nr. 4, S. 362–362
- [161] MALONE, Tara ; CLIFTON, Shari: Using focus groups to evaluate a multiyear consumer health outreach collaboration. In: *Journal of the Medical Library Association: JMLA* 109 (2021), Nr. 4, S. 575
- [162] MANNING, Christopher D. ; RAGHAVAN, Prabhakar ; SCHÜTZE, Hinrich: *Introduction to information retrieval*. Cambridge university press, 2008
- [163] MAO, Jiaxin ; LIU, Yiqun ; ZHOU, Ke ; NIE, Jian-Yun ; SONG, Jingtao ; ZHANG, Min ; MA, Shaoping ; SUN, Jiashen ; LUO, Hengliang: When does relevance mean usefulness and user satisfaction in web search? In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, S. 463–472
- [164] MARKOV, Ilya ; DE RIJKE, Maarten: What should we teach in information retrieval? In: *ACM SIGIR Forum* Bd. 52 ACM New York, NY, USA, 2019, S. 19–39
- [165] MARSHALL, Iain J. ; KUIPER, Joël ; WALLACE, Byron C.: Automating risk of bias assessment for clinical trials. In: *proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2014, S. 88–95
- [166] MARSHALL, Iain J. ; KUIPER, Joël ; WALLACE, Byron C.: RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. In: *Journal of the American Medical Informatics Association* 23 (2016), Nr. 1, S. 193–201
- [167] MARSHALL, Iain J. ; WALLACE, Byron C.: Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. In: *Systematic reviews* 8 (2019), S. 1–10

- [168] MARTÍNEZ GARCÍA, Laura ; ARÉVALO-RODRÍGUEZ, Ingrid ; SOLÀ, Ivan ; HAYNES, R. B. ; VANDVIK, Per O. ; ALONSO-COELLO, Pablo ; GROUP, Updating Guidelines W.: Strategies for monitoring and updating clinical practice guidelines: a systematic review. In: *Implementation Science* 7 (2012), S. 1–10
- [169] MC LAUGHLIN, G. H.: SMOG grading—a new readability formula. In: *Journal of reading* 12 (1969), Nr. 8, S. 639–646
- [170] MCCREADIE, Richard ; MACDONALD, Craig ; OUNIS, Iadh ; BRASSEY, Jon: A study of personalised medical literature search. In: *Information Access Evaluation. Multilinguality, Multimodality, and Interaction: 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings 5* Springer, 2014, S. 74–85
- [171] MCGOWAN, Jessie ; SAMPSON, Margaret ; SALZWEDEL, Douglas M. ; COGO, Elise ; FOERSTER, Vicki ; LEFEBVRE, Carol: PRESS peer review of electronic search strategies: 2015 guideline statement. In: *Journal of clinical epidemiology* 75 (2016), S. 40–46
- [172] NATIONAL LIBRARY OF MEDICINE (NLM),  
: *MEDLINE Overview*. 2024. – [Online; accessed 30-March-2024]
- [173] MEILLIER, Andrew ; PATEL, Shyam: Readability of healthcare literature for gastro-paresis and evaluation of medical terminology in reading difficulty. In: *Gastroenterology Research* 10 (2017), Nr. 1, S. 1
- [174] MENG, Shunmei ; FAN, Shaoyu ; LI, Qianmu ; WANG, Xinna ; ZHANG, Jing ; XU, Xiaolong ; QI, Lianyong ; BHUIYAN, Md Zakirul A.: Privacy-aware factorization-based hybrid recommendation method for healthcare services. In: *IEEE Transactions on Industrial Informatics* 18 (2022), Nr. 8, S. 5637–5647
- [175] MIKOLOV, Tomas ; SUTSKEVER, Ilya ; CHEN, Kai ; CORRADO, Greg S. ; DEAN, Jeff: Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems* 26 (2013)
- [176] MIN, Bonan ; ROSS, Hayley ; SULEM, Elior ; VEYSEH, Amir Pouran B. ; NGUYEN, Thien H. ; SAINZ, Oscar ; AGIRRE, Eneko ; HEINTZ, Ilana ; ROTH, Dan: Recent advances in natural language processing via large pre-trained language models: A survey. In: *ACM Computing Surveys* 56 (2023), Nr. 2, S. 1–40
- [177] MITRA, Bhaskar ; DIAZ, Fernando ; CRASWELL, Nick: Learning to match using local and distributed representations of text for web search. In: *Proceedings of the 26th international conference on world wide web*, 2017, S. 1291–1299
- [178] MORRIS, John H. ; SOMAN, Karthik ; AKBAS, Rabia E. ; ZHOU, Xiaoyuan ; SMITH, Brett ; MENG, Elaine C. ; HUANG, Conrad C. ; CERONO, Gabriel ; SCHENK, Gundolf ;

- RIZK-JACKSON, Angela [u. a.]: The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. In: *Bioinformatics* 39 (2023), Nr. 2, S. btad080
- [179] MOSKOVITCH, Robert ; WANG, Fei ; PEI, Jian ; FRIEDMAN, Carol. *JASIST special issue on biomedical information retrieval*. 2017
- [180] MOWSHOWITZ, Abbe ; KAWAGUCHI, Akira: Assessing bias in search engines. In: *Information Processing & Management* 38 (2002), Nr. 1, S. 141–156
- [181] MU, Yida ; WU, Ben P. ; THORNE, William ; ROBINSON, Ambrose ; ALETRAS, Nikolaos ; SCARTON, Carolina ; BONTCHEVA, Kalina ; SONG, Xingyi: Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. In: *arXiv preprint arXiv:2305.14310* (2023)
- [182] MUHAMAD, Nor A. ; SELVARAJAH, Vinesha ; DHARMARATNE, Anuja ; INTHIRAN, Anushia ; DALI, Nor Soleha M. ; CHAIYAKUNAPRUK, Nathorn ; LAI, Nai M.: Online Searching as a Practice for Evidence-Based Medicine in the Neonatal Intensive Care Unit, University of Malaya Medical Center, Malaysia: Cross-sectional Study. In: *JMIR Formative Research* 6 (2022), Nr. 4, S. e30687
- [183] MURAD, M H. ; ASI, Noor ; ALSAWAS, Mouaz ; ALAHDAB, Fares: New evidence pyramid. In: *BMJ Evidence-Based Medicine* 21 (2016), Nr. 4, S. 125–127
- [184] MURRAY, Elizabeth ; LO, Bernard ; POLLACK, Lance ; DONELAN, Karen ; CATANIA, Joe ; LEE, Ken ; ZAPERT, Kinga ; TURNER, Rachel: The impact of health information on the Internet on health care and the physician-patient relationship: national US survey among 1.050 US physicians. In: *Journal of medical internet research* 5 (2003), Nr. 3, S. e17
- [185] MUSTAR, Agnès ; LAMPRIER, Sylvain ; PIWOWARSKI, Benjamin: On the study of transformers for query suggestion. In: *ACM Transactions on Information Systems (TOIS)* 40 (2021), Nr. 1, S. 1–27
- [186] NGUYEN, Thong ; MACAVANEY, Sean ; YATES, Andrew: A Unified Framework for Learned Sparse Retrieval. In: *European Conference on Information Retrieval* Springer, 2023, S. 101–116
- [187] NATIONAL LIBRARY OF MEDICINE (NLM),  
: *LitVar*. <https://www.ncbi.nlm.nih.gov/research/litvar2/>. 2024. – [Online; accessed 30-March-2024]
- [188] NATIONAL LIBRARY OF MEDICINE (NLM),  
: *PubMed*. <http://www.ncbi.nlm.nih.gov>. 2024. – [Online; accessed 30-March-2024]
- [189] O’SULLIVAN, Lydia ; SUKUMAR, Prasanth ; CROWLEY, Rachel ; MCAULIFFE, Eilish ; DORAN, Peter: Readability and understandability of clinical research patient infor-

- mation leaflets and consent forms in Ireland and the UK: a retrospective quantitative analysis. In: *BMJ open* 10 (2020), Nr. 9, S. e037994
- [190] OTT, Niels ; MEURERS, Detmar: Information retrieval for education: Making search engines language aware. In: *Themes in Science and Technology Education* 3 (2011), Nr. 1-2, S. 9–30
- [191] OZYEGEN, Ozan ; KABE, Devika ; CEVIK, Mucahit: Word-level text highlighting of medical texts for telehealth services. In: *Artificial Intelligence in Medicine* 127 (2022), S. 102284
- [192] PALOTTI, Joao ; GOEURIOT, Lorraine ; ZUCCON, Guido ; HANBURY, Allan: Ranking health web pages with relevance and understandability. In: *Proceedings of the 39th international ACM SIGIR conference on Research and development in information retrieval*, 2016, S. 965–968
- [193] PALOTTI, João ; HANBURY, Allan ; MÜLLER, Henning: Exploiting health related features to infer user expertise in the medical domain. In: *Web Search Click Data workshop at WSCM, New York City, NY, USA*, 2014
- [194] PALOTTI, Joao ; ZUCCON, Guido ; HANBURY, Allan: MM: a new framework for multidimensional evaluation of search engines. In: *Proceedings of the 27th ACM international conference on information and knowledge management*, 2018, S. 1699–1702
- [195] PALOTTI, Joao ; ZUCCON, Guido ; HANBURY, Allan: Consumer health search on the web: study of web page understandability and its integration in ranking algorithms. In: *Journal of medical Internet research* 21 (2019), Nr. 1, S. e10986
- [196] PAPANIKOLAOU, Nikolas ; PAVLOPOULOS, Georgios A. ; PAFILIS, Evangelos ; THEODOSIOU, Theodosios ; SCHNEIDER, Reinhard ; SATAGOPAM, Venkata P. ; OUZOUNIS, Christos A. ; ELIOPOULOS, Aristides G. ; PROMPONAS, Vasilis J. ; ILIOPOULOS, Ioannis: BioTextQuest+: a knowledge integration platform for literature mining and concept discovery. In: *Bioinformatics* 30 (2014), Nr. 22, S. 3249–3256
- [197] PASI, Gabriella ; VIVIANI, Marco: Information credibility in the social web: Contexts, approaches, and open issues. In: *arXiv preprint arXiv:2001.09473* (2020)
- [198] PATEL, Chirag R. ; CHERLA, Deepa V. ; SANGHVI, Saurin ; BAREDES, Soly ; ELOY, Jean A.: Readability assessment of online thyroid surgery patient education materials. In: *Head & neck* 35 (2013), Nr. 10, S. 1421–1425
- [199] PATRICK, Timothy B. ; DEMIRIS, George ; FOLK, Lillian C. ; MOXLEY, David E. ; MITCHELL, Joyce A. ; TAO, Donghua: Evidence-based retrieval in evidence-based medicine. In: *Journal of the Medical Library Association* 92 (2004), Nr. 2, S. 196
- [200] PENNINGTON, Jeffrey ; SOCHER, Richard ; MANNING, Christopher D.: Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical*

- methods in natural language processing (EMNLP)*, 2014, S. 1532–1543
- [201] POLIKAR, Robi: Ensemble learning. In: *Ensemble machine learning: Methods and applications* (2012), S. 1–34
- [202] POWELL, Lauren ; KRIVANEK, Taylor ; DESHPANDE, Sagar ; LANDIS, George: Assessing Readability of FDA-Required Labeling for Breast Implants. In: *Aesthetic Surgery Journal Open Forum* Bd. 5 Oxford University Press US, 2023, S. ojad027–009
- [203] RAINE, Todd ; THOMA, Brent ; CHAN, Teresa M. ; LIN, Michelle: FOAMS earch.net: A custom search engine for emergency medicine and critical care. In: *Emergency Medicine Australasia* 27 (2015), Nr. 4, S. 363–365
- [204] RASMUSSEN, Edie: Evaluation in information retrieval. In: *The MIR/MDL evaluation project white paper collection edition 3* (2003), S. 45–49
- [205] RATNANI, Iqbal: *Gap Between Practices of Evidence-based Medicine and Personalized Medicine: A Barrier in Learning*, Dissertation, 2021
- [206] RIBEIRO, Marco T. ; SINGH, Sameer ; GUESTRIN, Carlos: ” Why should i trust you?” Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, S. 1135–1144
- [207] ROBERTS, Kirk ; DEMNER-FUSHMAN, Dina ; VOORHEES, Ellen M. ; HERSH, William R. ; BEDRICK, Steven ; LAZAR, Alexander J.: Overview of the TREC 2018 precision medicine track. In: *The... text REtrieval conference: TREC. Text REtrieval Conference* NIH Public Access, 2018
- [208] ROBERTS, Kirk ; DEMNER-FUSHMAN, Dina ; VOORHEES, Ellen M. ; HERSH, William R. ; BEDRICK, Steven ; LAZAR, Alexander J. ; PANT, Shubham: Overview of the TREC 2017 precision medicine track. In: *The... text REtrieval conference: TREC. Text REtrieval Conference* Bd. 26 NIH Public Access, 2017
- [209] ROBERTS, Kirk ; DEMNER-FUSHMAN, Dina ; VOORHEES, Ellen M. ; HERSH, William R. ; BEDRICK, Steven ; LAZAR, Alexander J. ; PANT, Shubham ; MERIC-BERNSTAM, Funda: Overview of the TREC 2019 precision medicine track. In: *The... text REtrieval conference: TREC. Text REtrieval Conference* Bd. 1250 NIH Public Access, 2019
- [210] ROBERTSON, Stephen ; ZARAGOZA, Hugo [u. a.]: The probabilistic relevance framework: BM25 and beyond. In: *Foundations and Trends® in Information Retrieval* 3 (2009), Nr. 4, S. 333–389
- [211] ROSNER, Anthony L.: Evidence-based medicine: revisiting the pyramid of priorities. In: *Journal of Bodywork and Movement Therapies* 16 (2012), Nr. 1, S. 42–49
- [212] SANDERSON, Mark [u. a.]: Test collection based evaluation of information retrieval

- systems. In: *Foundations and Trends® in Information Retrieval* 4 (2010), Nr. 4, S. 247–375
- [213] SANTANA, Idaira R. ; MASON, Anne ; GUTACKER, Nils ; KASTERIDIS, Panagiotis ; SANTOS, Rita ; RICE, Nigel: Need, demand, supply in health care: working definitions, and their implications for defining access. In: *Health Economics, Policy and Law* 18 (2023), Nr. 1, S. 1–13
- [214] SANTOSH, KC ; BELAÏD, Abdel: Document information extraction and its evaluation based on client’s relevance. In: *2013 12th International Conference on Document Analysis and Recognition IEEE*, 2013, S. 35–39
- [215] SATHIAN, Brijesh ; SREEDHARAN, Jayadevan ; BABOO, Suresh N. ; SHARAN, Krishna ; ABHILASH, ES ; RAJESH, E: Relevance of sample size determination in medical research. In: *Nepal Journal of Epidemiology* 1 (2010), Nr. 1, S. 4–10
- [216] SAVELKA, Jaromir ; ASHLEY, Kevin D.: The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. In: *Frontiers in Artificial Intelligence* 6 (2023)
- [217] SBAFFI, Laura ; ROWLEY, Jennifer: Trust and credibility in web-based health information: a review and agenda for future research. In: *Journal of medical Internet research* 19 (2017), Nr. 6, S. e218
- [218] SCHAMBER, Linda: Relevance and information behavior. In: *Annual review of information science and technology (ARIST)* 29 (1994), S. 3–48
- [219] SCHAMBER, Linda ; EISENBERG, Michael: Relevance: The Search for a Definition. (1988)
- [220] SCHARDT, Connie ; ADAMS, Martha B. ; OWENS, Thomas ; KEITZ, Sheri ; FONTELO, Paul: Utilization of the PICO framework to improve searching PubMed for clinical questions. In: *BMC medical informatics and decision making* 7 (2007), S. 1–6
- [221] SCHLAG, Imanol ; SUKHBAATAR, Sainbayar ; CELIKYILMAZ, Asli ; YIH, Wen-tau ; WESTON, Jason ; SCHMIDHUBER, Jürgen ; LI, Xian: Large language model programs. In: *arXiv preprint arXiv:2305.05364* (2023)
- [222] SCHUERS, Matthieu ; GRIFFON, Nicolas ; KERDELHUE, Gaëtan ; FOUBERT, Quentin ; MERCIER, Alain ; DARMONI, Stéfan J: Behavior and attitudes of residents and general practitioners in searching for health information: from intention to practice. In: *International journal of medical informatics* 89 (2016), S. 9–14
- [223] SCHUTH, Anne ; HOFMANN, Katja ; RADLINSKI, Filip: Predicting search satisfaction metrics with interleaved comparisons. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015, S. 463–472



- [224] SCHWARZ, Julia ; MORRIS, Meredith: Augmenting web pages and search results to support credibility assessment. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, S. 1245–1254
- [225] SCLAR, Melanie ; CHOI, Yejin ; TSVETKOV, Yulia ; SUHR, Alane: Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In: *arXiv preprint arXiv:2310.11324* (2023)
- [226] SELVARAJ, Prabha ; BURUGARI, Vijay K. ; SUMATHI, D ; NAYAK, Rudra K. ; TRIPATHY, Ramamani: Ontology based recommendation system for domain specific seekers. In: *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* IEEE, 2019, S. 341–345
- [227] SHAFARI, Mohammad ; HU, Qing ; AFSARMANESH, Hamideh ; HUANG, Zhisheng ; TEN TEIJE, Annette ; VAN HARMELEN, Frank [u. a.]: A Task-based Comparison of Linguistic and Semantic Document Retrieval Methods in the Medical Domain. In: *EMSA-RMed@ ESWC*, 2016
- [228] SHANAVAS, Niloofer ; WANG, Hui ; LIN, Zhiwei ; HAWK, Glenn: Ontology-based enriched concept graphs for medical document classification. In: *Information Sciences* 525 (2020), S. 172–181
- [229] SHAO, Jun: Linear model selection by cross-validation. In: *Journal of the American statistical Association* 88 (1993), Nr. 422, S. 486–494
- [230] SHARIFPOUR, Romina ; WU, Mingfang ; ZHANG, Xiuzhen: Large-scale analysis of query logs to profile users for dataset search. In: *Journal of Documentation* 79 (2023), Nr. 1, S. 66–85
- [231] SHARMA, Himanshu ; JANSEN, Bernard J.: Automated evaluation of search engine performance via implicit user feedback. In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005, S. 649–650
- [232] SHENOI, S. ; LY, Vickie ; SONI, Sarvesh ; ROBERTS, Kirk: Developing a Search Engine for Precision Medicine. In: *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science 2020* (2020), S. 579–588
- [233] SHERIDAN, Desmond J. ; JULIAN, Desmond G.: Achievements and limitations of evidence-based medicine. In: *Journal of the American College of Cardiology* 68 (2016), Nr. 2, S. 204–213
- [234] SHERRIFF, Alicia ; ALLY, Hamza ; MAHOMED, Wasim ; RAE, Heather ; SCHARKE-NECHT, RORY ; SEALANYANE, Seipati ; JOUBERT, Gina: The understanding of medical abbreviations across different medical departments in a South African hospital setting. In: *Communication & Medicine (Equinox Publishing Group)* 14 (2017), Nr. 1

- [235] SHI, Lei ; SONG, Guangjia ; CHENG, Gang ; LIU, Xia: A user-based aggregation topic model for understanding user's preference and intention in social network. In: *Neurocomputing* 413 (2020), S. 1–13
- [236] SHOEMAKER, Sarah J. ; WOLF, Michael S. ; BRACH, Cindy: Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. In: *Patient education and counseling* 96 (2014), Nr. 3, S. 395–403
- [237] SHUSTER, Kurt ; POFF, Spencer ; CHEN, Moya ; KIELA, Douwe ; WESTON, Jason: Retrieval augmentation reduces hallucination in conversation. In: *arXiv preprint arXiv:2104.07567* (2021)
- [238] SI, Luo ; CALLAN, Jamie: A statistical model for scientific readability. In: *Proceedings of the tenth international conference on Information and knowledge management*, 2001, S. 574–576
- [239] SINGH, Vaibhav K. ; SINGH, Vinay K.: Vector space model: an information retrieval system. In: *Int. J. Adv. Engg. Res. Studies/IV/II/Jan.-March* 141 (2015), Nr. 143
- [240] VAN DER SLUIS, Frans ; VAN DEN BROEK, Egon L.: Using complexity measures in information retrieval. In: *Proceedings of the third symposium on information interaction in context*, 2010, S. 383–388
- [241] SOBOCZENSKI, Frank ; TRIKALINOS, Thomas A. ; KUIPER, Joël ; BIAS, Randolph G. ; WALLACE, Byron C. ; MARSHALL, Iain J.: Machine learning to help researchers evaluate biases in clinical trials: a prospective, randomized user study. In: *BMC Medical Informatics and Decision Making* 19 (2019), S. 1–12
- [242] SOONG, David ; SRIDHAR, Sriram ; SI, Han ; WAGNER, Jan-Samuel ; SÁ, Ana Caroline C. ; YU, Christina Y. ; KARAGOZ, Kubra ; GUAN, Meijian ; HAMADEH, Hisham ; HIGGS, Brandon W.: Improving accuracy of GPT-3/4 results on biomedical data using a retrieval-augmented language model. In: *arXiv preprint arXiv:2305.17116* (2023)
- [243] SOTO, Axel J. ; PRZYBYŁA, Piotr ; ANANIADOU, Sophia: Thalia: semantic search engine for biomedical abstracts. In: *Bioinformatics* 35 (2019), Nr. 10, S. 1799–1801
- [244] SPINK, Amanda ; YANG, Yin ; JANSEN, Jim ; NYKANEN, Pirrko ; LORENCE, Daniel P. ; OZMUTLU, Seda ; OZMUTLU, H C.: A study of medical and health queries to web search engines. In: *Health Information & Libraries Journal* 21 (2004), Nr. 1, S. 44–51
- [245] SPRING, BONNIE ; MARCHESE, SARA H. ; STEGLITZ, JEREMY: History and process of evidence-based practice in mental health. In: *Evidence-based practice in action: Bridging clinical science and intervention* (2019), S. 9–27
- [246] SQUIRES, Janet E. ; GRAHAM, Ian D. ; HUTCHINSON, Alison M. ; LINKLATER, Stefanie ; BREHAUT, Jamie C. ; CURRAN, Janet ; IVERS, Noah ; LAVIS, John N. ; MICHIE,

- Susan ; SALES, Anne E. [u. a.]: Understanding context in knowledge translation: a concept analysis study protocol. In: *Journal of advanced nursing* 71 (2015), Nr. 5, S. 1146–1155
- [247] SQUIRES, Janet E. ; GRAHAM, Ian D. ; HUTCHINSON, Alison M. ; MICHIE, Susan ; FRANCIS, Jill J. ; SALES, Anne ; BREHAUT, Jamie ; CURRAN, Janet ; IVERS, Noah ; LAVIS, John [u. a.]: Identifying the domains of context important to implementation science: a study protocol. In: *Implementation Science* 10 (2015), S. 1–9
- [248] STOSSEL, Lauren M. ; SEGAR, Nora ; GLIATTO, Peter ; FALLAR, Robert ; KARANI, Reena: Readability of patient education materials available at the point of care. In: *Journal of general internal medicine* 27 (2012), S. 1165–1170
- [249] STRAGE, Katya ; STACEY, Stephen ; MAUFFREY, Cyril ; PARRY, Joshua A.: The interobserver reliability of clinical relevance in medical research. In: *Injury* 54 (2023), S. S66–S68
- [250] SUI, Yujia ; ZHANG, Bin: Determinants of the perceived credibility of rebuttals concerning health misinformation. In: *International Journal of Environmental Research and Public Health* 18 (2021), Nr. 3, S. 1345
- [251] SUZIEDELYTE, Agne: How does searching for health information on the Internet affect individuals' demand for health care services? In: *Social science & medicine* 75 (2012), Nr. 10, S. 1828–1835
- [252] SWIRE-THOMPSON, Briony ; LAZER, David [u. a.]: Public health and online misinformation: challenges and recommendations. In: *Annu Rev Public Health* 41 (2020), Nr. 1, S. 433–451
- [253] SZMUDA, Tomasz ; ÖZDEMİR, Cathrine ; ALI, Shan ; SINGH, Akshita ; SYED, Mohammad T. ; SŁONIEWSKI, Paweł: Readability of online patient education material for the novel coronavirus disease (COVID-19): a cross-sectional health literacy study. In: *Public health* 185 (2020), S. 21–25
- [254] TAMINE, Lynda ; GOEURIOT, Lorraine: Semantic information retrieval on medical texts: Research challenges, survey, and open issues. In: *ACM Computing Surveys (CSUR)* 54 (2021), Nr. 7, S. 1–38
- [255] TAMINE-LECHANI, Lynda ; BOUGHANEM, Mohand ; DAOUD, Mariam: Evaluation of contextual information retrieval effectiveness: overview of issues and research. In: *Knowledge and Information Systems* 24 (2010), S. 1–34
- [256] TAN, Chenhao ; GABRILOVICH, Evgeniy ; PANG, Bo: To each his own: personalized content selection based on text comprehensibility. In: *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012, S. 233–242
- [257] TARANOVA, Anastasia ; BRASCHLER, Martin: Textual complexity as an indicator of

- document relevance. In: *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43* Springer, 2021, S. 410–417
- [258] TAWFIK, Andrew A. ; KOCHENDORFER, Karl M. ; SAPAROVA, Dinara ; AL GHENAIMI, Said ; MOORE, Joi L.: Using semantic search to reduce cognitive load in an electronic health record. In: *2011 IEEE 13th International Conference on e-Health Networking, Applications and Services IEEE*, 2011, S. 181–184
- [259] THIRUNAVUKARASU, Arun J. ; TING, Darren Shu J. ; ELANGOVAN, Kabilan ; GUTIERREZ, Laura ; TAN, Ting F. ; TING, Daniel Shu W.: Large language models in medicine. In: *Nature medicine* 29 (2023), Nr. 8, S. 1930–1940
- [260] THOMAS, Paul ; MOFFAT, Alistair ; BAILEY, Peter ; SCHOLER, Falk ; CRASWELL, Nick: Better effectiveness metrics for serps, cards, and rankings. In: *Proceedings of the 23rd australasian document computing symposium*, 2018, S. 1–8
- [261] TOMS, Elaine G. ; LATTE, Celeste: How consumers search for health information. In: *Health informatics journal* 13 (2007), Nr. 3, S. 223–235
- [262] TOUVRON, Hugo ; MARTIN, Louis ; STONE, Kevin ; ALBERT, Peter ; ALMAHAIRI, Amjad ; BABAEI, Yasmine ; BASHLYKOV, Nikolay ; BATRA, Soumya ; BHARGAVA, Prajjwal ; BHOSALE, Shruti [u. a.]: Llama 2: Open foundation and fine-tuned chat models. In: *arXiv preprint arXiv:2307.09288* (2023)
- [263] TSURUOKA, Yoshimasa ; MIWA, Makoto ; HAMAMOTO, Kaisei ; TSUJII, Jun'ichi ; ANANIADOU, Sophia: Discovering and visualizing indirect associations between biomedical concepts. In: *Bioinformatics* 27 (2011), Nr. 13, S. i111–i119
- [264] UMAPATHI, Logesh K. ; PAL, Ankit ; SANKARASUBBU, Malaikannan: Med-halt: Medical domain hallucination test for large language models. In: *arXiv preprint arXiv:2307.15343* (2023)
- [265] UPRETY, Sagar ; TIWARI, Prayag ; DEHDASHTI, Shahram ; FELL, Lauren ; SONG, Dawei ; BRUZA, Peter ; MELUCCI, Massimo: Quantum-like structure in multidimensional relevance judgements. In: *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42* Springer, 2020, S. 728–742
- [266] UTTLEY, Lesley ; INDAVE, Blanca I. ; HYDE, Chris ; WHITE, Valerie ; LOKUHETTY, Dilani ; CREE, Ian: Invited commentary—WHO Classification of Tumours: How should tumors be classified? Expert consensus, systematic reviews or both? In: *International journal of cancer* 146 (2020), Nr. 12, S. 3516
- [267] VAN DER VEGT, A. ; ZUCCON, G. ; KOOPMAN, B. ; DEACON, Anthony J.: Impact of a Search Engine on Clinical Decisions Under Time and System Effectiveness Constraints:

- Research Protocol. In: *JMIR Research Protocols* 8 (2019)
- [268] VELTRI, Lauren W. ; MILTON, Denái R ; DELGADO, Ruby ; SHAH, Nina ; PATEL, Krina ; NIETO, Yago ; KEBRIAELI, Partow ; POPAT, Uday R. ; PARMAR, Simrit ; ORAN, Betul [u. a.]: Outcome of autologous hematopoietic stem cell transplantation in refractory multiple myeloma. In: *Cancer* 123 (2017), Nr. 18, S. 3568–3575
- [269] VAN DE VLIET, Peter ; SPRENGER, Tobias ; KAMPERS, Linde F. ; MAKALOWSKI, Jennifer ; SCHIRRMACHER, Volker ; STÜCKER, Wilfried ; VAN GOOL, Stefaan W.: The Application of Evidence-Based Medicine in Individualized Medicine. In: *Biomedicines* 11 (2023), Nr. 7, S. 1793
- [270] VOGEL, Stephen N. ; SULTAN, Thomas R. ; TEN EYCK, Raymond P.: Cyanide poisoning. In: *Clinical toxicology* 18 (1981), Nr. 3, S. 367–383
- [271] VOORHEES, Ellen M. ; HARMAN, Donna K. [u. a.]: *TREC: Experiment and evaluation in information retrieval*. Bd. 63. Citeseer, 2005
- [272] VYDISWARAN, VG V. ; MEI, Qiaozhu ; HANAUER, David A. ; ZHENG, Kai: Mining consumer health vocabulary from community-generated text. In: *AMIA Annual Symposium Proceedings* Bd. 2014 American Medical Informatics Association, 2014, S. 1150
- [273] VYDISWARAN, VG V. ; REDDY, Manoj: Identifying peer experts in online health forums. In: *BMC medical informatics and decision making* 19 (2019), S. 41–49
- [274] WANG, Guangyu ; YANG, Guoxing ; DU, Zongxin ; FAN, Longjun ; LI, Xiaohu: ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation. In: *arXiv preprint arXiv:2306.09968* (2023)
- [275] WANG, Haolin ; ZHANG, Qingpeng ; YUAN, Jiahu: Semantically enhanced medical information retrieval system: a tensor factorization based approach. In: *Ieee Access* 5 (2017), S. 7584–7593
- [276] WANG, Jiajia ; HUANG, Jimmy X. ; TU, Xinhui ; WANG, Junmei ; HUANG, Angela J. ; LASKAR, Md Tahmid R. ; BHUIYAN, Amran: Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. In: *ACM Computing Surveys* 56 (2024), Nr. 7, S. 1–33
- [277] WANG, Lih-Wern ; MILLER, Michael J. ; SCHMITT, Michael R. ; WEN, Frances K.: Assessing readability formula differences with written health information materials: application, results, and recommendations. In: *Research in Social and Administrative Pharmacy* 9 (2013), Nr. 5, S. 503–516
- [278] WANG, Liupu ; WANG, Juexin ; WANG, Michael ; LI, Yong ; LIANG, Yanchun ; XU, Dong: Using Internet search engines to obtain medical information: a comparative study. In: *Journal of medical Internet research* 14 (2012), Nr. 3, S. e74

- [279] WANG, Yining ; WANG, Liwei ; LI, Yuanzhi ; HE, Di ; LIU, Tie-Yan: A theoretical analysis of NDCG type ranking measures. In: *Conference on learning theory* PMLR, 2013, S. 25–54
- [280] WEI, Chih-Hsuan ; ALLOT, Alexis ; LEAMAN, Robert ; LU, Zhiyong: PubTator central: automated concept annotation for biomedical full text articles. In: *Nucleic acids research* 47 (2019), Nr. W1, S. W587–W593
- [281] WEI, Chih-Hsuan ; KAO, Hung-Yu ; LU, Zhiyong: PubTator: a web-based text mining tool for assisting biocuration. In: *Nucleic acids research* 41 (2013), Nr. W1, S. W518–W522
- [282] WEI, Chih-Hsuan ; KAO, Hung-Yu ; LU, Zhiyong: Text mining tools for assisting literature curation. In: *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2014, S. 590–591
- [283] WHITE, Ryen W.: *Interactions with search systems*. Cambridge University Press, 2016
- [284] WHITE, Ryen W. ; DUMAIS, Susan T. ; TEEVAN, Jaime: Characterizing the influence of domain expertise on web search behavior. In: *Proceedings of the second ACM international conference on web search and data mining*, 2009, S. 132–141
- [285] WHITE, Ryen W. ; HORVITZ, Eric: Studies of the onset and persistence of medical concerns in search logs. In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2012, S. 265–274
- [286] WHITE, Ryen W. ; ROTH, Resa A.: *Exploratory search: Beyond the query-response paradigm*. Morgan & Claypool Publishers, 2009 ( 3)
- [287] WIENER, R C. ; WIENER-PLA, Regina: Literacy, pregnancy and potential oral health changes: The internet and readability levels. In: *Maternal and child health journal* 18 (2014), S. 657–662
- [288] WILLIS, Lawrence ; GOSAIN, Ankush: Readability of patient and family education materials on pediatric surgical association websites. In: *Pediatric Surgery International* 39 (2023), Nr. 1, S. 156
- [289] WINCHELL, Adam ; LAN, Andrew ; MOZER, Michael: Highlights as an early predictor of student comprehension and interests. In: *Cognitive Science* 44 (2020), Nr. 11, S. e12901
- [290] WOLFE, Jeremy M. ; EVANS, Karla K. ; DREW, Trafton ; AIZENMAN, Avigael ; JOSEPHS, Emilie: How do radiologists use the human search engine? In: *Radiation protection dosimetry* 169 (2016), Nr. 1-4, S. 24–31
- [291] WOOLF, Steven H. ; CHAN, Evelyn C. ; HARRIS, Russell ; SHERIDAN, Stacey L. ; BRADDOCK III, Clarence H. ; KAPLAN, Robert M. ; KRIST, Alex ; O’CONNOR,

- Annette M. ; TUNIS, Sean. *Promoting informed choice: transforming health care to dispense knowledge for decision making*. 2005
- [292] WORRALL, Amy P. ; CONNOLLY, Mary J. ; O'NEILL, Aine ; O'DOHERTY, Murray ; THORNTON, Kenneth P. ; McNALLY, Cora ; McCONKEY, Samuel J. ; DE BARRA, Eoghan: Readability of online COVID-19 health information: a comparison between four English speaking countries. In: *BMC Public Health* 20 (2020), Nr. 1, S. 1–12
- [293] WYNN, Thomas A. ; RAMALINGAM, Thirumalai R.: Mechanisms of fibrosis: therapeutic translation for fibrotic disease. In: *Nature medicine* 18 (2012), Nr. 7, S. 1028–1040
- [294] XIE, Iris: *Interactive information retrieval in digital environments*. IGI global, 2008
- [295] XIONG, Guangzhi ; JIN, Qiao ; LU, Zhiyong ; ZHANG, Aidong: Benchmarking retrieval-augmented generation for medicine. In: *arXiv preprint arXiv:2402.13178* (2024)
- [296] YAN, Xin ; SONG, Dawei ; LI, Xue: Concept-based document readability in domain specific information retrieval. In: *Proceedings of the 15th ACM international conference on Information and knowledge management*, 2006, S. 540–549
- [297] YAO, Bingsheng ; CHEN, Guiming ; ZOU, Ruishi ; LU, Yuxuan ; LI, Jiachen ; ZHANG, Shao ; LIU, Sijia ; HENDLER, James ; WANG, Dakuo: More Samples or More Prompt Inputs? Exploring Effective In-Context Sampling for LLM Few-Shot Prompt Engineering. In: *arXiv preprint arXiv:2311.09782* (2023)
- [298] YILMAZ, Emine ; VERMA, Manisha ; CRASWELL, Nick ; RADLINSKI, Filip ; BAILEY, Peter: Relevance and effort: An analysis of document utility. In: *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, 2014, S. 91–100
- [299] YILMAZ, Zeynep A. ; WANG, Shengjin ; YANG, Wei ; ZHANG, Haotian ; LIN, Jimmy: Applying BERT to document retrieval with birch. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, 2019, S. 19–24
- [300] YOON, Sung H.: Personalized web search using query based user profile. In: *Journal of the Korea Academia-Industrial Cooperation Society* 17 (2016), Nr. 2, S. 690–696
- [301] ZAMBORLINI, Veruska ; HU, Qing ; HUANG, Zhisheng ; DA SILVEIRA, Marcos ; PRUSKI, Cedric ; TEN TELJE, Annette ; VAN HARMELEN, Frank: Knowledge-driven paper retrieval to support updating of clinical guidelines: a use case on PubMed. In: *International Workshop on Process-oriented Information Systems in Healthcare* Springer, 2016, S. 71–89

- [302] ZENG-TREITLER, Qing ; GORYACHEV, Sergey ; TSE, Tony ; KESELMAN, Alla ; BOXWALA, Aziz: Estimating consumer familiarity with health terminology: a context-based approach. In: *Journal of the American Medical Informatics Association* 15 (2008), Nr. 3, S. 349–356
- [303] ZHAI, Chengxiang: Interactive information retrieval: Models, algorithms, and evaluation. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, S. 2444–2447
- [304] ZHANG, Fan ; LIU, Yiqun ; MAO, Jiaxin ; ZHANG, Min ; MA, Shaoping: User behavior modeling for web search evaluation. In: *AI Open* 1 (2020), S. 40–56
- [305] ZHANG, Fan ; MAO, Jiaxin ; LIU, Yiqun ; XIE, Xiaohui ; MA, Weizhi ; ZHANG, Min ; MA, Shaoping: Models versus satisfaction: Towards a better understanding of evaluation metrics. In: *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval*, 2020, S. 379–388
- [306] ZHANG, Yan ; SUN, Yalin ; XIE, Bo: Quality of health information for consumers on the web: a systematic review of indicators, criteria, tools, and evaluation results. In: *Journal of the Association for Information Science and Technology* 66 (2015), Nr. 10, S. 2071–2084
- [307] ZHANG, Yanzhao ; LONG, Dingkun ; XU, Guangwei ; XIE, Pengjun: HLATR: enhance multi-stage text retrieval with hybrid list aware transformer reranking. In: *arXiv preprint arXiv:2205.10569* (2022)
- [308] ZHENG, Kai ; MEI, Qiaozhu ; HANAUER, David A.: Collaborative search in electronic health records. In: *Journal of the American Medical Informatics Association* 18 (2011), Nr. 3, S. 282–291
- [309] ZHOU, Zhi-Hua ; ZHOU, Zhi-Hua: *Ensemble learning*. Springer, 2021
- [310] ZHU, Yutao ; YUAN, Huaying ; WANG, Shuting ; LIU, Jiongnan ; LIU, Wenhan ; DENG, Chenlong ; DOU, Zhicheng ; WEN, Ji-Rong: Large language models for information retrieval: A survey. In: *arXiv preprint arXiv:2308.07107* (2023)
- [311] ZUCCON, Guido ; KOOPMAN, Bevan: SIGIR 2018 Tutorial on Health Search (HS2018) A Full-day from Consumers to Clinicians. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, S. 1391–1394