TUM SCHOOL OF COMPUTATION,
INFORMATION AND TECHNOLOGY

TECHNICAL UNIVERSITY OF MUNICH

Master's Thesis in Robotics, Cognition, Intelligence

Evaluating Adapter-based Knowledge-enhanced Language Models in the Biomedical Domain

Alexander Fichtl

TUM SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY

TECHNICAL UNIVERSITY OF MUNICH

Master's Thesis in Robotics, Cognition, Intelligence

Evaluating Adapter-based Knowledge-enhanced Language Models in the Biomedical Domain

Evaluation adapterbasierter Methoden zur Injektion von strukturellem Wissen in biomedizinische Sprachmodelle

Author:	Alexander Fichtl
Supervisor:	Prof. Dr. Florian Matthes
Advisor:	Juraj Vladika, M.Sc.
Submission Date:	October 15th, 2023

I confirm that this master's thesis in robotics, cognition, intelligence is my own work and I have documented all sources and material used.

Munich, October 15th, 2023 Alexander Fichtl

Acknowledgments

I would like to express my sincere gratitude to all the people who have supported me throughout the writing of my master's thesis. First and foremost, I would like to thank my advisor, Juraj Vladika, for his invaluable guidance, feedback, and encouragement. He has been a great mentor and a source of inspiration for me. I have learned a lot from him and I am grateful to have had the opportunity to work with him. I would also like to thank my supervisor, Prof. Florian Matthes, for his support and the opportunity to conduct my master's thesis at the chair for Software Engineering for Business Information Systems (sebis). He has been very helpful and generous with his resources. The same applies to my contacts at OntoChem, especially Dr. Claudia Bobach, who never tired of helping me and answering my never-ending questions.

Finally, I would like to thank my family, friends, and especially my girlfriend, Hannah, for their support and encouragement. They have shared their opinions, experiences, and knowledge with me. They have made this journey more meaningful and memorable.

I dedicate this thesis to all of them.

Abstract

In the rapidly evolving field of biomedical natural language processing (BioNLP), knowledge-enhanced language models (KELMs) have emerged as promising tools to bridge the gap between large-scale language models and domain-specific knowledge. KELMs can achieve higher factual accuracy and mitigate hallucinations by leveraging the potential of knowledge graphs (KGs). This work delves into the evaluation of such models within the biomedical domain. It aims to guide future research to ensure that BioNLP evolves in a manner most beneficial to the healthcare system. The thesis is structured in three parts, the main results of which are as follows **Literature Review:** First, a novel systematic literature review on adapter-based approaches to knowledge-enhancement gives an overview of the methodologies in the field. We explore the strengths and potential shortcomings of related work. We find that both open-domain and closed-domain approaches have been frequently explored along with a multitude of downstream tasks. We discovered that the biomedical domain has been the most frequently explored domain-specific field in the last four years, signifying our work's relevance. With the literature review, we provide an extensive resource for researchers exploring adapter-based KELMs. **Model Experiments:** In the second part of the thesis, we leverage the insights gained from the literature to design a pipeline for model experiments. Our industry partner, OntoChem, provided us with access to a novel knowledge graph for this section. We propose two sets of sub-graphs, "Onto20Rel" and "OntoType20Rel", with which we inject renowned biomedical large language models (LLMs). Our methodology leads to solid results on several downstream tasks, predominately "HoC", "PubMedBERT", and "BioASQ7b". For the BioASQ7b task, we report the best performance of all related works. We give a detailed interpretation of the results and report valuable insights. **Research Survey:** The final part of the thesis comprises a survey addressed to medical students and professionals. The survey sheds light on the implications of BioNLP in practice and sets the stage for further research in this domain. In particular, we find a growing adoption of NLP technology in the medical community. The survey discovers several anticipated and recognized applications of LLMs, particularly to reduce the time clinicians have to spend on monotonous tasks. At the same time, participants shared significant concerns, primarily surrounding data security and the sanctity of the fiduciary doctor-patient relationship. The survey results underscore the balance that needs to be struck between innovation and trust as we venture further into the realm of BioNLP in healthcare.

Keywords: Natural Language Processing (NLP), Pre-trained language models, Knowledge Graphs, Domain Knowledge, Knowledge Enhancement, Adapters, Biomedicine

Kurzfassung

In dem sich rasant entwickelnden Bereich der biomedizinischen natürlichen Sprachverarbeitung haben sich wissensgestärkte Sprachmodelle (sogenannte "KELMs") als vielversprechende Werkzeuge herausgestellt, um die Lücke zwischen großen Sprachmodellen und domänenspezifischem Wissen zu schließen. KELMs nutzen das Potential von Wissensgraphen um eine höhere faktische Genauigkeit zu erreichen und Halluzinationen von Sprachmodellen zu reduzieren. Diese Masterarbeit setzt sich mit der Evaluation solcher Modelle im biomedizinischen Bereich auseinander. Die Arbeit zielt darauf ab richtungsweisend für zukünftige Forschung zu sein. Die Arbeit ist in drei Teile gegliedert, deren Hauptergebnisse wie folgt sind: **Literaturanalyse:** Zunächst gibt eine neuartige systematische Literaturanalyse über adapterbasierte Ansätze zur Wissensverstärkung einen detaillierten Überblick über die existierenden Methoden und Ansätze in diesem Feld. Wir untersuchen die Stärken und Schwächen von verschiedenen Ansätzen in der Literatur. Wir stellen fest, dass sowohl Ansätze in offenen Dömanen, sowie domänenspezifische Ansätze häufig auftreten. Des Weiteren, wird eine Vielzahl an downstream-tasks in der Literatur erforscht. Wir haben festgestellt, dass der biomedizinische Bereich in den letzten vier Jahren das am häufigsten untersuchte domänenspezifische Feld war, was die Relevanz unserer Arbeit unterstreicht. Mit der Literaturanalyse bieten wir eine neue, umfangreiche Ressource für Forscher, die adapterbasierte KELMs erkunden. **Modell-Experimente:** Im zweiten Teil der Arbeit nutzen wir die gemachten Erkenntnisse aus der Literatur, um eine Pipeline für Sprachmodellexperimente zu entwerfen. Für diesen Abschnitt der Arbeit gab uns unser Industriepartner OntoChem Zugang zu einem neuartigen Wissensgraphen. Über unsere Methologie haben wir zwei Sets an Teilgraphen entwickelt: Onto20Rel und OntoType20Rel. Mit diesen konnten wir eine Anzahl an renommierten biomedizinische großen Sprachmodelle mit Wissen injizieren. Unsere Methodik führt zu soliden Ergebnissen bei mehreren downstream-tasks, insbesondere HoC, PubMedBERT und BioASQ7b. Im Fall von BioASQ7b haben wir die beste Modellperformanz unter allen verwandten Arbeiten erreicht. **Forschungsumfrage:** Der letzte Teil der Thesis besteht aus einer Umfrage, welche an Medizinstudent*innen und medizinisches Fachpersonal gerichtet ist. Die Umfrage beleuchtet die Auswirkungen von biomedizinischer natürlicher Sprachverarbeitung in der Praxis und legt den Grundstein für weitere Forschung in diesem Bereich. Insbesondere stellen wir eine wachsende Nutzung von NLP-Technologie in der Biomedizin fest. Die Umfrage legt verschiedene mögliche Anwendungen von Sprachmodellen offen. Teilnehmer*innen sehen großes Potential dabei die Zeit zu reduzieren, welche Ärzt*innen für monotone Aufgaben aufwenden. Gleichzeitig teilten die Teilnehmer*innen erhebliche Bedenken, besonders bei Datensicherheit und dem treuhändeischen Arzt-Patienten-Verhältniss. Die Umfrageergebnisse unterstreichen ein Gleichgewicht, welches zwischen Innovation und Vertrauen gefunden werden muss, um das Gesundheitswesen mithilfe von Sprachmodellen verbessern zu können.

Contents

1. Introduction

Time and health are two of our most valuable resources. Being short on either leads to strain and stress in our lives. We rely on medical professionals for healthcare; they rely on the healthcare system to enable them to work under good conditions. However, medical professionals worldwide are notoriously overworked and short on time, often due to the bureaucracy they have to deal with every working day. For example, a membership survey of the Marburger-Bund (2022) found that 60 percent of doctors in Germany spend three hours or more per day on documentation and administrative tasks. A nationwide study in the US (Overhagea & McCallie, 2020) led to similar results where, on average, physicians spend approximately 3.3 hours on electronic health records (EHRs), an average of 16 minutes and 14 seconds per encounter. While EHRs are important, they take away from the time spent on actual patient care and are one of the major factors physicians dislike about their profession (Overhagea & McCallie, 2020).

It is, therefore, an important task to relieve medical professionals of as much documentation work and administrative duties as possible so we can reduce stress and strain for patients and physicians alike. Recent advancements in the development of large language models (LLMs) could prove to be a significant aid in this task, especially if they are tuned for use in medicine and bio-medicine. For instance, specialized biomedical LLMs can help automate tasks such as analyzing medical records or generating reports based on patient data, allowing healthcare professionals to work more efficiently and effectively. Biomedical LLMs can also help improve communication between healthcare providers and patients, for example, by writing drafts to answer patient questions and mail inquiries. Overall, the progress of large language models has the potential to revolutionize how we approach healthcare, making it more personalized, efficient, and tailored to individual needs.

Of course, the use of language models does not come without risks. LLMs, in general, tend to lack knowledge awareness and are prone to so-called "hallucinations", which can lead to inaccurate predictions and make the models unusable. This issue is especially precarious in a high-risk domain such as healthcare. One possible solution to counteract hallucinations and improve the reliability of LLMs is knowledge enhancement: By leveraging expert knowledge from knowledge graphs (KGs), structured knowledge can be injected into LLMs. Such knowledge-enhanced language models (KELMs) are a promising approach to more factual accuracy and less hallucinations (Colon-Hernandez, Havasi, Alonso, et al., 2021; Wei, Wang, Zhang, et al., 2021). This thesis aims to research KELMs and their application in the biomedical domain.

1.1. Context

The starting point for this thesis was the opportunity to experiment with novel data from OntoChem, an industry partner of the chair of Software Engineering for Business Information Systems (sebis) at the Technical University of Munich (TUM). OntoChem provided us with access to vast amounts of structured biomedical knowledge in the form of entity-relation-entity triplets. We noticed that most research on KELMs in the biomedical domain relies on the publicly available "Unified Medical Language System" (UMLS) ontology (Bodenreider, 2004), suggesting that the field could profit from research that is leveraging different and more modern sources. During further investigation of the topic, we noticed two additional gaps in the field of biomedical KELM: A lack of a comprehensive study on adapter-based approaches and the non-existence of a survey that addresses medical professionals directly and involves them in the research in the field. This came as a surprise to us since the area of biomedical LLMs is subject to an ever-rising amount of research. Yet, it seems to miss a connection to the practicality of the results in clinics and medical centers. Eventually, our investigations and initial literature review led to three research questions (RQs), which will be stated in the following. The thesis is connected to the VeriSci research project at sebis. VeriSci is part of the Software Campus Framework sponsored by the Federal Ministry of Education and Research (BMBF).

1.2. Research Questions

Deducted from the context and initial literature review mentioned above, this thesis will focus on the research and methods that will answer three (RQs):

RQ1 What adapter-based approaches to knowledge-enhancement exist, and how do they compare to each other?

RQ2 Can we improve existing approaches with new methods and data from a private ontology?

RQ3 Is the research on biomedical KELMs relevant to medical professionals, and what factors hinder or support the deployment of the technology in practice?

1.3. Outline

This work is structured into five chapters. After the introduction in chapter 1, chapter 2 addresses past and ongoing research in biomedical NLP with a focus on KELMs. The third chapter explains the research questions and methods of the thesis. It describes the settings of the literature review, the model experiments, and the research survey addressed to medical professionals. Chapter 4 presents and discusses the previous chapter's results and answers the thesis's research questions. Finally, chapter 5 concludes the thesis by giving a concise summary and addressing possible shortcomings and further research opportunities. All supplementary material is provided in the appendix.

2. Background and Related Work

Few fields of research have been more active over the last decade than (bio)medicine and natural language processing (Cimini, Gabrielli, & Labini, 2014; Nwagwu, 2022). Since this thesis is nested right in between those two fields, it comes as no surprise that there is a significant amount of relevant literature. This chapter introduces the reader to past and ongoing research on biomedical NLP and KELMs and explains domain-specific terminology.

This chapter aims to lay the groundwork for understanding the methods and findings discussed in the later chapters of this work. For the systematic literature review on adapter-based approaches to knowledge enhancement, see chapter 3 and 4.

2.1. Natural Language Processing

The capacity to achieve fluency in a language, allowing comprehension and processing of spoken and written words, has historically been viewed as a distinctly human trait. Yet, in the past few decades, machines have made significant progress in this domain through NLP, a subdomain of computer science and linguistics that falls under the umbrella of artificial intelligence. Most individuals unknowingly interact with NLP systems daily. Prominent voice assistants rely on NLP to offer, for example, speech-to-text or translation services on smartphones or through home speakers. The same applies to the auto-correct functions and the customer service chatbots we commonly talk to on websites (Jurafsky & Martin, 2023). In the following, we will give an overview of NLP, highlighting the fields that are most relevant to the understanding of this work.

2.1.1. Overview

To begin with, we will summarize the most important developments in NLP methodologies. For further reading on the mentioned methods and developments, please refer to Dan Jurafsky's and James H. Martin's work on Speech and Language Processing (Jurafsky & Martin, 2023). They provide free access to their comprehensive and regularly updated book on all forms of NLP.

Traditional Methods Before the dawn of deep learning and its profound impact on the realm of NLP, various traditional techniques formed the bedrock of language-based algorithms and applications. These methodologies heavily relied on linguistic rules, pattern matching, statistical modeling, and symbolic approaches. With a mix of traditional methods, the automatic question-answerer "BASEBALL" proposed by Green, Wolf, Chomsky, and Laughery

(1961) was already able to read a baseball-related question from a punched card, process it, and print out an answer. Another famous system, ELIZA, often recognized as the first chatbot, simulated conversation by pattern matching and substitution methodology (Weizenbaum, 1966). Its most renowned variation, "DOCTOR", emulated a Rogerian psychotherapist, formulating its responses based on the user's input. While rule and pattern matching can be very precise for well-defined tasks, corresponding methods can be brittle and limited in handling the complexities and ambiguities of natural language. Each new linguistic phenomenon or exception could require the addition of new rules, making the system intricate and cumbersome. Moving away from hand-coded rules, statistical models started playing a significant role in NLP during the late 20th century. For example, Markov models and chains have been widely used for part-of-speech tagging, named entity recognition, and speech recognition (Almutiri & Nadeem, 2022). Their strength lies in their capability to consider context through state transitions, enabling them to effectively model sequential data like text or speech.

Modern Neural Networks and Transformers With the start of the deep learning boom, which was driven by the increasing affordability and accessibility of computing power and data, Recurrent Neural Networks (RNNs) and Long Short Term Memory (LTSM) emerged as leading techniques in NLP for a limited period of time. However, since the invention of the transformer by Vaswani, Shazeer, Parmar, et al. (2017), transformer-based language models are considered the state-of-the-art (SOTA) base architecture of NLP systems. These modes are usually pre-trained to give them a general understanding of natural language before "fine-tuning" them for a specific downstream purpose.[1] Devlin, Chang, Lee, and Toutanova (2019) presented BERT (Bidirectional Encoder Representations from Transformers), a then novel method of pre-training language representations bidirectionally that set new standards. Their pre-training approach, so-called "masked language modeling" (MLM), was to mask 15 percent of all input words of a sentence, feed the sequence to a bidirectional transformer, and finally predict the previously masked words. BERT became outstandingly successful and dominated most downstream tasks and applications in NLP over the more traditional machine learning text classification approaches (Gonz'alez-Carvajal & Garrido-Merch'an, 2020). Subsequently, a whole family of BERT-based models has been created since, some of which will be used in the experimental part of this thesis (see chapter 3).

Growth of Large Language Models After the described invention-driven leaps in NLP, the general trend for LLMs (and deep learning models in general) became to grow in size rapidly. For instance, while at its original publication, the largest version of BERT had 340 million parameters, more recent language models like GPT-3 (Brown, Mann, Ryder, et al., 2020) or PaLM (Singhal, Azizi, Tu, et al., 2022) already have 175 and 540 billion parameters, respectively. While this growth comes with significant performance boosts, experimenting with LLMs of this size is only possible for large tech companies that have access to vast

[1] From this point onward, any reference to LLMs inherently implies *pre-trained* large language models, unless otherwise specified.

amounts of computing power. As a consequence, researchers and data scientists with fewer resources now depend on parameter-efficient models and training algorithms.

2.1.2. Biomedical Natural Language Processing

Biomedical natural language processing (BioNLP) is an interdisciplinary field that combines NLP and biomedical science. It involves using computational techniques to analyze and interpret large volumes of biomedical text data such as electronic health records (EHRs), scientific literature, and clinical notes (Kalyan, Rajasekharan, & Sangeetha, 2022; B. Wang, Xie, Pei, et al., 2021). BioNLP has been used for various down-stream tasks such as information extraction (IE), text classification (TC), named entity recognition (NER), relation extraction (RE), and question answering (QA) (Gu, Tinn, Cheng, et al., 2020). The abundance of biomedical text data coupled with advances in NLP is resulting in novel BioNLP applications. Like in other NLP domains, BioNLP relied on rule-based systems and statistical models. However, nowadays, BioNLP is equally reliant on the availability of domain-specific LLMs that are trained on massive amounts of data (B. Wang, Xie, Pei, et al., 2021). An excellent reference for the change of the BioNLP field over the years is given by a survey on clinical natural language processing in the United Kingdom covering the period from 2007 to 2022 (Wu, Wang, Wu, et al., 2022).

Some of the challenges in BioNLP include the complexity of biomedical language, which often contains technical terms and jargon not commonly used in everyday language. Also, biomedical text data is often unstructured and noisy, making it difficult to extract meaningful information from it. Other challenges include the lack of standardization in biomedical language and the need for domain-specific language models. In the following, specialized biomedical LLMs will be discussed in preparation for their use in this thesis.

Biomedical Large Language Models

It was long assumed that for domain-specific applications, the open-domain pre-training of standard BERT models should be combined with fine-tuning (or further pre-training) on in-domain data ("mixed approach") for maximum performance. However, (Gu, Tinn, Cheng, et al., 2020) challenged this approach and showed that "domain-specific pre-training from scratch substantially outperforms continual pre-training of generic language models, thus demonstrating that the prevailing assumption in support of mixed-domain pre-training is not always applicable". They used the biomedical domain as a running example for their study and proposed "PubMedBERT", a model based on the original BERT architecture (Devlin, Chang, Lee, & Toutanova, 2019), but solely pre-trained on papers from PubMed[2], a free resource containing more than 36 million citations and abstracts of biomedical literature. PubMedBERT set new standards in biomedical NLP and consistently outperformed mixed approaches like BioBERT (Lee, Yoon, Kim, et al., 2019) or SciBERT (Beltagy, Lo, & Cohan, 2019).

[2]https://pubmed.ncbi.nlm.nih.gov/

Just like other LLMs, biomedical LLMs have been growing in size rapidly. For example, the SOTAs on many biomedical benchmarks (see following section) is constantly improved by ever larger models like BioGPT. Notably, one descendant of the BERT family of LLMs seems to be still competing (e.g., on the PubMedBERT benchmark on "PapersWithCode" [3] with much larger and more modern biomedical LLMs: BioLinkBERT. BioLinkBERT builds upon the LinkBERT (Yasunaga, Leskovec, & Liang, 2022) architecture, a self-supervised pre-training approach that is itself based on BERT (Devlin, Chang, Lee, & Toutanova, 2019). BioLinkBERT was created by incorporating citation links from academic papers in the pre-training process, enriching the model's understanding of knowledge dependencies that span a variety of papers (Yasunaga, Leskovec, & Liang, 2022). Because of its parameter/performance ratio and the limited computing resources available for this thesis, we decided to include BioLinkBERT in our methodology for the experimental section of our work (see chapter 3.

Benchmarks and State-of-the-art

There are various ways to assess the capability of BioNLP models. Apart from evaluating models on single specific downstream tasks, there exist two widely used comprehensive benchmarks combining a multitude of downstream tasks: The "Biomedical Language Understanding Evaluation" (BLUE) benchmark and the "Biomedical Language Understanding and Reasoning Benchmark" (BLURB) created at Microsoft Research (Gu, Tinn, Cheng, et al., 2020). Both benchmarks consist of a set of classic NLP tasks, such as Named Entity Recognition NER, Relation Extraction RE, and Question Answering QA, but set in the biomedical domain. In this work, we will run experiments on four of the tasks included in BLURB, namely HoC (Baker, Silins, Guo, et al., 2015), PubMedQA (Jin, Dhingra, Liu, et al., 2019), BioASQ7b (Nentidis, Bougiatiotis, Krithara, & Paliouras, 2020), and MedNLI (Romanov & Shivade, 2018). Reasons for this choice will be given in chapter 3.

For the HoC (Hallmarks of Cancer) task, "ten characteristics (i.e., hallmarks) of normal cells required for malignant growth have been proposed that provide an organizing principle to simplify the diversity of the biological processes leading to cancer" (Baker, Silins, Guo, et al., 2015). The HoC dataset contains cancer-related PubMed abstracts annotated according to the evidence they provide for such cancer hallmarks and requires models to solve the corresponding multi-label classification task. Examples of hallmarks are "Sustaining proliferative signaling (PS)", "Evading growth suppressors (GS)", or "Resisting cell death (CD)".

The two question-answering datasets, PubMedQA and BioASQ7b, require models to answer research questions extracted from PubMedQA abstracts and handcrafted by medical professionals, respectively. The answer choices in the case of PubMedQA are "yes", "no", or "maybe", where the source abstracts are given as context. An example question from the dataset is, "Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?" (Jin, Dhingra, Liu, et al., 2019). BioASQ7b is a binary classification task with

[3] https://paperswithcode.com/sota/question-answering-on-pubmedqa

questions created and curated by medical experts and answer choices "yes" and "no". A question example is "Is Baloxavir effective for influenza?" with an example context being "Baloxavir marboxil [...] is an oral cap-dependent endonuclease inhibitor that has been developed by Roche and Shionogi [...]".

Finally, MedNLI is an expert annotated dataset for Natural Language Inference (NLI) in the clinical domain. Romanov and Shivade (2018) explain NLI as "the task of determining whether a given hypothesis can be inferred from a given premise". For example, given the premise "She was not able to speak, but appeared to comprehend well" and hypothesis "Patient had aphasia" (a language disorder caused by damage in a specific area of the brain), the model has to decide between an "entailment", "contradiction", or "neutral" predicted label (here: entailment).

Table 2.1 shows the current SOTA for the four chosen tasks.

Task/SOTA	Score	Model
HoC	87.3 (MicroF1)	NCI_BERT (Peng, Yan, & Lu, 2019)
PubMedQA	81.8 (Acc)	Med-PaLM 2 (Singhal, Tu, Gottweis, et al., 2023)
BioASQ7b	94.8 (Acc)	BioLinkBERT-large (Yasunaga, Leskovec, & Liang, 2022)
MedNLI	86.6 (Acc)	SciFive (Phan, Anibal, Tran, et al., 2021)

Table 2.1.: SOTA for select tasks from the BLURB benchmark according to the best of our knowledge and PapersWithCode (https://paperswithcode.com/) leaderboards. The metrics are indicated with the scores and correspond to the metrics used in BLURB

2.2. Knowledge Enhanced Language Models

The original motivation for developing KELMs was to introduce structured knowledge to the unstructured nature of LMs to increase knowledge awareness and reduce hallucinations (Wei, Wang, Zhang, et al., 2021). With the steady progress of LLMs over the last years, there arose a multitude of KELM approaches and models that could leverage the expert knowledge in knowledge graphs (KGs). The following will explain the concept of KGs, and existing literature on KELMs will be presented.

2.2.1. Knowledge Graphs

KGs have seen a rising prominence in NLP research over the past decade (Schneider, Schopf, Vladika, et al., 2022) and are an essential element of most approaches to KELMs. Hogan, Blomqvist, Cochez, et al. (2020) define a KG as "a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities". Ji, Pan, Cambria, et al. (2020) published a comprehensive survey on KGs and, following existing literature, defined the concept of a KG

as "$\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{F}\}$, where $\mathcal{E}, \mathcal{R}$ and $\mathcal{F}$ are sets of entities, relations and facts, respectively. A fact is denoted as a triple $(h, r, t) \in \mathcal{F}$". In our methodology, we follow the documentation of Irmer, Bobach, and Böhme (2019) for practical reasons related to our methodologies and denote these triples as (s, r, o), where s is a subject, r is a relation, and o is an object. Both s and o are entities that come from an entity set E, while relations come from a relation set R. Depending on the source and purpose of a KG, entities, and relations can take on various shapes. For example, a relation can take the shape of a single word like "inhibits", a short phrase like "relates to ", or a compound term including, for example, chemical or medical categories such as "[protein] relates to [disease]" or "[substance] induces [physiology]". A textual connection is vital because it serves as a link between the graph structure and natural language, simplifying the integration of information from KGs into language models and the associated learning processes. Examples of popular KGs that are important for this work are UMLS (Bodenreider, 2004), DBpedia (Auer, Bizer, Kobilarov, et al., 2007a), and ConceptNet (Speer, Chin, & Havasi, 2017). They will be discussed in the later chapters of this thesis.

2.2.2. Approaches to Knowledge Enhancement

At the time of writing, there are several studies that try to give an overview of KELMs and classify different approaches. Colon-Hernandez, Havasi, Alonso, et al. (2021) "examine a variety of approaches to integrate structured knowledge into current language models and determine challenges, and possible opportunities to leverage both structured and unstructured information sources". They review existing literature in three categories: Input-centered strategies (1) center around altering either the structure of the input or the selected data, which is fed into the base LLMs. Architecture-focused approaches (2), on the other hand, "[involve] either adding additional layers that integrate knowledge in some way with the contextual representations or modifying existing layers to manipulate things such as attention mechanisms. [...] These approaches utilize adapter-like mechanisms to be able to inject information into the models" (Colon-Hernandez, Havasi, Alonso, et al., 2021).

For both (1) and (2), there are various ways to incorporate information from a KG $\mathcal{G}$ into the LLM Θ_0: A prevalent method involves creating a dataset by transforming the facts $\mathcal{F}$ into a set of words or sentence-like structure. Then $\Phi_{\mathcal{G}}$ can be learned through classical MLM as training objective $\mathcal{L}_{\mathcal{G}}$. Accordingly to Meng, Liu, Clark, et al. (2021), implementing $\Phi_{\mathcal{G}}$ through entity prediction is widely popular, with other methods including relation classification (R. Wang, Tang, Duan, et al., 2020), entity linking (Peters, Neumann, Logan, et al., 2019), and next sentence prediction (Goodwin & Demner-Fushman, 2020).

Finally, output-focused approaches (3) work by changing either the output structure or the losses used in the base transformer model. However, here, the authors mention only the methodology of SemBERT (Zhang, Wu, Hai, et al., 2019) to fall under this category. SemBERT enhances the output of a BERT model with entity embeddings. In their survey, Colon-Hernandez, Havasi, Alonso, et al. (2021) also mention hybrid frameworks that utilize

a mixture of the three categories. One of their main conclusions is that there are "still opportunities at exploiting adapter-based injections" (situated in the architecture-focused category), which further increased our interest in following an adapter-based approach.

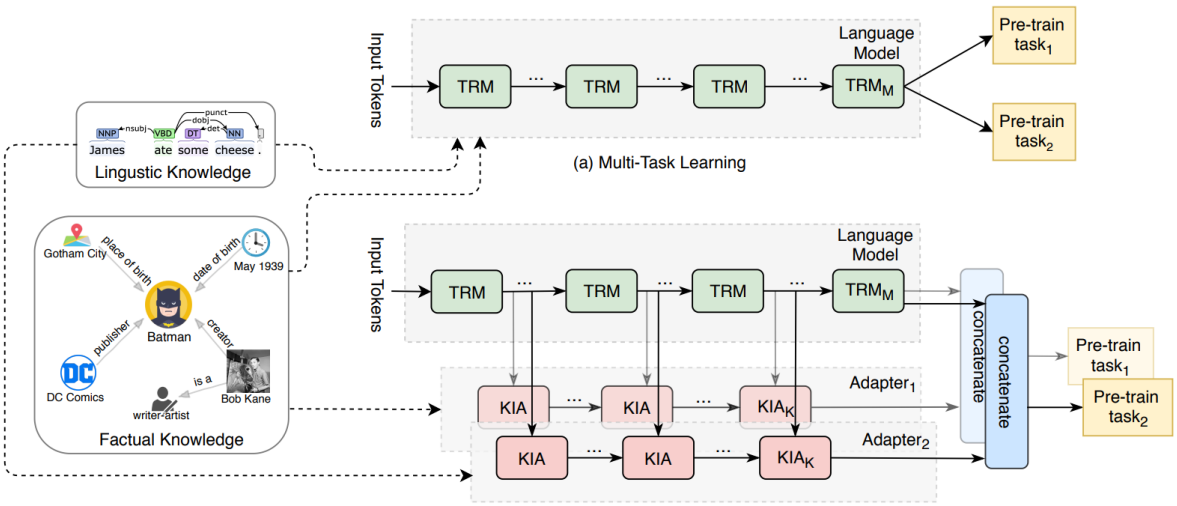

Figure 2.1.: An example of (a) knowledge enhancement in a classical "fine-tuning" approach and (b) adapter-based knowledge enhancement as applied by R. Wang, Tang, Duan, et al. (2020). Image from R. Wang, Tang, Duan, et al. (2020)

The second study by Wei, Wang, Zhang, et al. (2021) was published just a few months after the survey by Colon-Hernandez, Havasi, Alonso, et al. (2021). They review a large number of studies on KELMs and classify them using three taxonomies: (1) knowledge sources, (2) knowledge granularity, and (3) application areas. Within (1), they explored the integration of knowledge from knowledge sources such as linguistic knowledge, encyclopedia knowledge, and commonsense and domain-specific knowledge. They introduce representative methods for every source and discuss the corresponding knowledge the methods exploit. The second taxonomy (2) acknowledges the common approach to using KGs as a source of knowledge. In their study, Wei, Wang, Zhang, et al. (2021) group models by the granularity of knowledge they incorporate from the KGs. Levels of granularity are text-based knowledge, entity knowledge, relation triples, and KG sub-graphs. Notably, one of the significant works of literature by Meng, Liu, Clark, et al. (2021) that we used in our work draws on the utilization of graph partitioning and KG sub-graphs. In chapter 3, we explain the methods in detail. Lastly, with the third taxonomy (3), the authors discuss in detail how knowledge enhancement can improve natural language generation and understanding. They also review popular benchmarks that can be used for task performance evaluation of KELMs (Wei, Wang, Zhang, et al., 2021).

These two field studies by Colon-Hernandez, Havasi, Alonso, et al. (2021) and Wei, Wang, Zhang, et al. (2021) on the variations and classification of KELM approaches were our starting

point for the discussion of KELMs and initially proved to be very valuable. However, although they address some adapter-based studies like K-Adapter (R. Wang, Tang, Duan, et al., 2020), they did not mention most other adapter-based KELMs, and notably none from the biomedical domain. For some of the missing papers, like (Q. Lu, Dou, & Nguyen, 2021), this is likely due to an overlap of writing and publishing. Nevertheless, this lack of coverage led to our decision to conduct a systematic literature search on the topic as part of this thesis.

A third study on KELMs by Yang, Chen, Li, et al. (2023) was published while the writing of this thesis was already well underway. For completeness, we will still briefly give an overview here: The survey categorizes existing literature by "knowledge type" and by the point of time where the injection takes place: before, during, or after training, roughly translating to the input, architecture, and output focused approach categories form Colon-Hernandez, Havasi, Alonso, et al. (2021). Utilizing adapters can, therefore, be considered a "during training" injection. The paper strongly focuses on the advancement of LLMs of the size of the GPT series and can be viewed as the most contemporary survey. However, although they mention three works that use adapters for knowledge enhancement, Yang, Chen, Li, et al. (2023) still do not provide a comprehensive study of adapter-based approaches.

2.3. Adapters

In the following, an overview of adapters for LLMs and their individual functionalities and applications will be given. This section is meant to establish a conceptual understanding of adapter-based approaches to KELMs.

2.3.1. Overview

Adapters (Bapna & Firat, 2019; Houlsby, Giurgiu, Jastrzebski, et al., 2019; Pfeiffer, Kamath, Rücklé, et al., 2020) are a simple yet effective approach to mitigate issues that arise when training parameter-heavy transformer-based LLMs for downstream tasks. Broadly speaking, adapters are small bottleneck feed-forward layers inserted within each layer of a LLM. The small amount of additional parameters allows for the injection of new data or knowledge without having to fine-tune the whole model. This feat is usually accomplished by freezing the layers of the base model with its millions or billions of parameters while only updating the adapter weights (e.g., through entity prediction). Due to the lightweight nature of adapters, this approach allows for short training times with relatively low computing resource requirements. This way, adapters can be used for quick and cheap downstream-task fine-tuning in a traditional language modeling sense or enhancement of knowledge by injecting domain knowledge.

Because it is possible to train adapters individually, they can be used for multi-task training by specializing one adapter for each task or multi-domain knowledge injection by specializing adapters (individually or in sets) to different domains. Leveraging adapters in LLMs also

has positive "side effects": Adapters can avoid catastrophic forgetting by introducing new task-specific parameters (Houlsby, Giurgiu, Jastrzebski, et al., 2019; Pfeiffer, Kamath, Rücklé, et al., 2020) and, in transfer learning, adapters have even been shown to improved stability and adversarial robustness for various downstream tasks (Han, Pang, & Wu, 2021).

The specifics of how and where adapters are added to a LLM depend on the adapter type. In the following, the most common adapter types will be described.

2.3.2. Adapter Types

Houlsby Adapter The Houlsby Adapter (Houlsby, Giurgiu, Jastrzebski, et al., 2019) was the first adapter to be used for transfer learning in NLP. The idea was based on adapter modules initially introduced by Rebuffi, Bilen, and Vedaldi (2017) in the computer vision domain. The two main principles stayed the same: Adapters require a relatively small number of parameters when compared to the base model and a near-identity initialization. These principles "ensure that the total model size grows relatively slowly when more [transfer] tasks are added. A near-identity initialization is required for stable training of the adapted model" (Houlsby, Giurgiu, Jastrzebski, et al., 2019). The optimal architecture of the Houlsby Adapter was determined by meticulous experimenting and tuning; the result can be seen in figure 2.2. In a classical transformer structure (Vaswani, Shazeer, Parmar, et al., 2017), they add the adapter module once after the multi-headed attention and once after the two feed-forward layers. The modules project the d-dimensional layer features of the base model into a smaller dimension, m, then apply a non-linearity (like ReLU) and project back to d dimensions. The configuration also hosts a skip-connection, and the output of each sub-layer is forwarded to a layer normalization (Ba, Kiros, & Hinton, 2016). Including biases, $2md + d + m$ parameters are added per layer, accounting to only 0.5 to 8 percent of the parameters of the original BERT model used by the authors when setting $m << d$ (Houlsby, Giurgiu, Jastrzebski, et al., 2019).

Bapna and Firat Adapter In contrast to Houlsby, Giurgiu, Jastrzebski, et al. (2019), Bapna and Firat (2019) only introduce one adapter module in each transformer layer: They keep the adapters after the multi-headed attention (so-called "top" adapters) while dropping the adapters after the feed-forward layers (so-called "bottom" adapters) of the transformer (refer to figure 2.2 for better understanding of the component positions). Moreover, while (Houlsby, Giurgiu, Jastrzebski, et al., 2019) re-train layer normalization parameters for every domain, they "simplify this formulation by leaving the parameters frozen, and introducing new layer normalization parameters for every task, essentially mimicking the structure of the transformer feed-forward layer" (Bapna & Firat, 2019).

Pfeiffer Adapter and AdapterFusion The approaches of Bapna and Firat (2019) and Houlsby, Giurgiu, Jastrzebski, et al. (2019) did not allow for information sharing between tasks. Pfeiffer, Kamath, Rücklé, et al. (2020) introduce Adapter Fusion, a two-stage algorithm that addresses the sharing of information encapsulated in adapters that were trained on different tasks. In the first stage, they train the adapters in either single-task or multi-task setup for a total of N

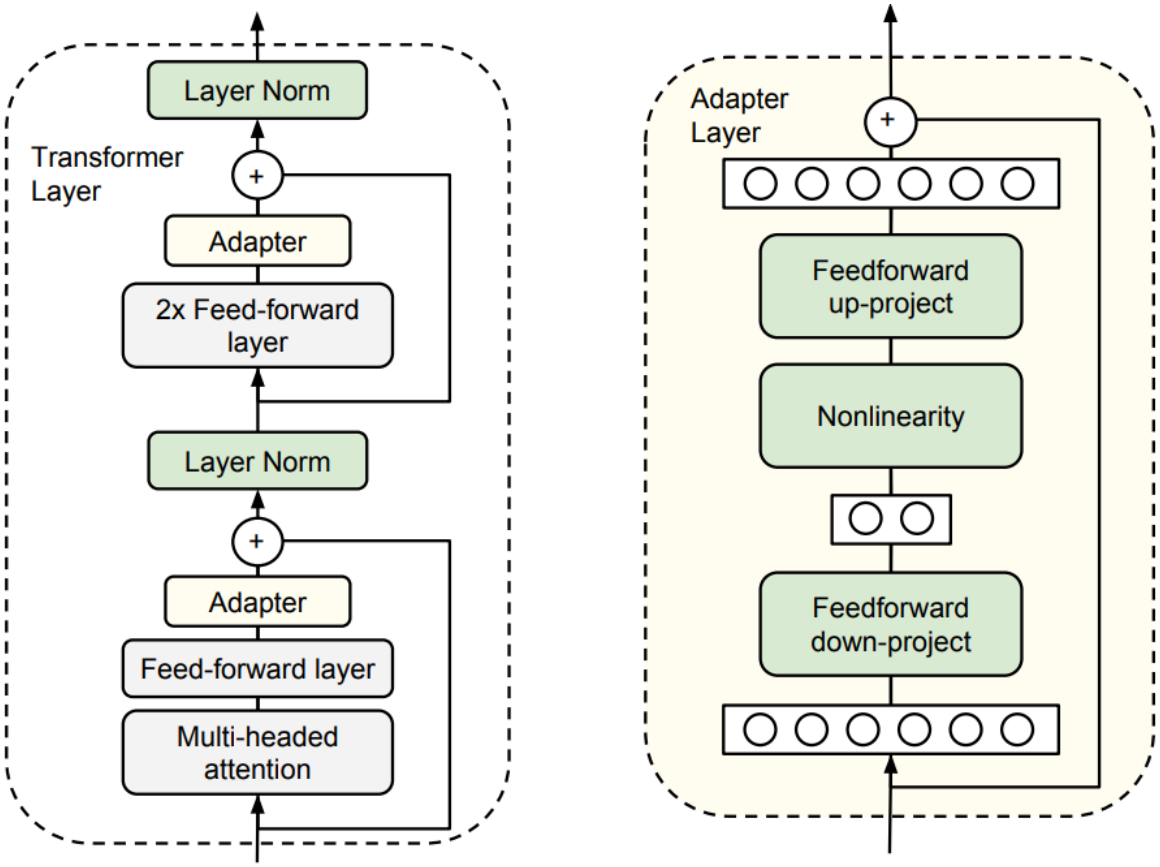

Figure 2.2.: Location of the adapter module in a transformer layer (left) and architecture of the Houlsby Adapter (right). All green layers are trained on the fine-tuning data, including the adapter itself, the layer normalization parameters, and the final classification layer, which is not shown in the figure. Image with permission from Houlsby, Giurgiu, Jastrzebski, et al. (2019)

tasks similar to the Houlsby Adapter, but only keeping the top adapters, similar to the Bapna and Firat Adapter. As a second step, they combine the set of N adapters with AdapterFusion: They fix the parameters Θ and all adapters Φ, and finally introduce parameters Ψ that learn to combine the N task adapters for the given target task (Pfeiffer, Kamath, Rücklé, et al., 2020):

$$\Psi_m \leftarrow \underset{\Psi}{\operatorname{argmin}} L_m\left(D_m; \Theta, \Phi_1, \ldots, \Phi_N, \Psi\right)$$

Here, Ψ_m are the learned AdapterFusion parameters for task m. In the process, "the training dataset of m is used twice: once for training the adapters Φ_m and again for training Fusion parameters Ψ_m, which learn to compose the information stored in the N task adapters" (Pfeiffer, Kamath, Rücklé, et al., 2020). With their approach of separating knowledge extraction and knowledge composition, they further improve the ability of adapters to avoid catastrophic forgetting and interference between tasks and training instabilities.

K-Adapter R. Wang, Tang, Duan, et al. (2020) follow a substantially different approach where the adapters work as "outside plug-ins". In their work, an adapter model consists of K adapter layers (hence the name) that contain N transformer layers as well as two projection layers. Similar to the approaches above, a skip connection is added, but it is applied across the two projection layers here. The adapter layers are plugged in among varying transformer layers of the pre-trained model. The authors explain that they "concatenate the output hidden feature of the transformer layer in the pre-trained model and the output feature of the former adapter layer, as the input feature of the current adapter layer. For each knowledge-specific adapter, we concatenate the last hidden features of the pre-trained model and adapter as the final output feature of this adapter model" (R. Wang, Tang, Duan, et al., 2020). K-Adapter is also the first work where factual knowledge injection is explicitly mentioned as an application, apart from the typical parameter-efficient fine-tuning on downstream tasks.

There exist adapter architectures that are different from the four adapter types mentioned here (like the "Parallel Adapter" (J. He, Zhou, Ma, et al., 2021) or the adapter architecture by Stickland and Murray (2019)). However, as the upcoming comprehensive literature survey will show, these architectures are either unique to specific papers or have not found broader applications in the field of KELMs. Either way, these approaches are out of the scope of this thesis and will not be discussed here.

3. Methodology

This chapter details the methodology we employed for the systematic literature review, the model experiments, and the research survey. There will be one section dedicated to each of these three thesis components. The methodology is designed to answer the three research questions described in the introduction. The results will be presented in chapter 4.

3.1. Systematic Literature Review

To begin with, we will describe the search strategy of the systematic literature review on adapter-based approaches to KELMs. We largely followed the procedure of Kitchenham, Pearl Brereton, Budgen, et al. (2009) for systematic literature reviews in software engineering. First, the inclusion and exclusion criteria will be stated. Then, the methodology of the data collection and the subsequent data analysis will be explained.

3.1.1. Inclusion and exclusion criteria

The search strategy for the systematic literature review of this thesis included literature that fulfilled the following inclusion criteria:

- Peer-reviewed articles from the ACM[1], ACL[2], and IEEE Xplore[3]
- Articles which fulfilled the search string ("adapter" OR "adapter-based") AND ("language model" OR "nlp" OR "natural language processing") AND ("injection" OR "knowledge")
- Articles published after February 2, 2019 (publication of the Houlsby Adapter, the first LLM adapter)
- Articles which address the topic of adapter-based knowledge-enhanced language models

We also included some articles not found on the mentioned databases if they fulfilled the remaining criteria and were significantly relevant to our work.

Articles on the following topics were excluded:

- Articles published before February 2, 2019

[1] https://dl.acm.org/
[2] https://aclanthology.org/
[3] https://ieeexplore.ieee.org/Xplore/home.jsp

- Duplicate versions of the same article (when multiple versions of an article were found in different journals, only the most recent version was included)
- Articles where Adapters were used for NLP, but for use-cases other than knowledge-enhancement (such as few-shot learning or model debiasing)
- Articles written in a language other than English

3.1.2. Data collection

The data extracted from each included document were:

- Source (journal or publication platform)
- Full reference
- Main topic area
- Facts of interest such as adapter architecture, domain, and downstream tasks within the papers
- A short summary of the study, including the main research questions and the answers

3.1.3. Data analysis

The collected data was tabulated to show:

- Source and publication dates of the studies
- Adapter architectures used in the papers
- Distribution of papers across domains (highlighting the biomedical domain)
- Distribution of papers across downstream tasks
- Results on biomedical NLP benchmarks (if relevant)

In chapter 4, the analysis results are interpreted and presented together with interesting and valuable qualitative findings from the papers.

3.2. Model Experiments

In this section, we describe our approach to leveraging data from OntoChem's SciWalker platform together with adapters to improve existing approaches to biomedical KELMs. In order to find the best settings for the experiments, information from the literature review on adapter-based methods to KELMs was required. Therefore, the experiment setting was largely designed after the initial results of the literature survey were present (especially the results addressing the biomedical domain).

The first step to determine the best setting for our experiments was to recreate the most promising works on adapter-based biomedical KELMs that we found during the literature survey and run some initial experiments. We did so for KEBLM ("Knowledge-Enhanced Biomedical Language Models" (Lai, Zhai, & Ji, 2023)) and MoP ("Mixture of Partitions" (Meng, Liu, Clark, et al., 2021)) since they openly shared their code and brought forth very recent and successful frameworks. Unfortunately, we could not recreate all of the results of KEBLM because we did not have access to enough computing power. The MoP framework, on the other hand, proved to be more accessible and low-resource friendly while enabling the training of models with almost equivalent performance (depending on the task). For this reason, we decided that the experiment setting in this thesis should be based on the MoP framework, an overview of which is given in figure 3.1. However, we altered the data source and processing and changed the used models and the training approach to a certain extent. More details on this will follow in the upcoming sections. Please refer to the results of the systematic literature survey in section 4.1 for more information on the MoP and KEBLM framework.

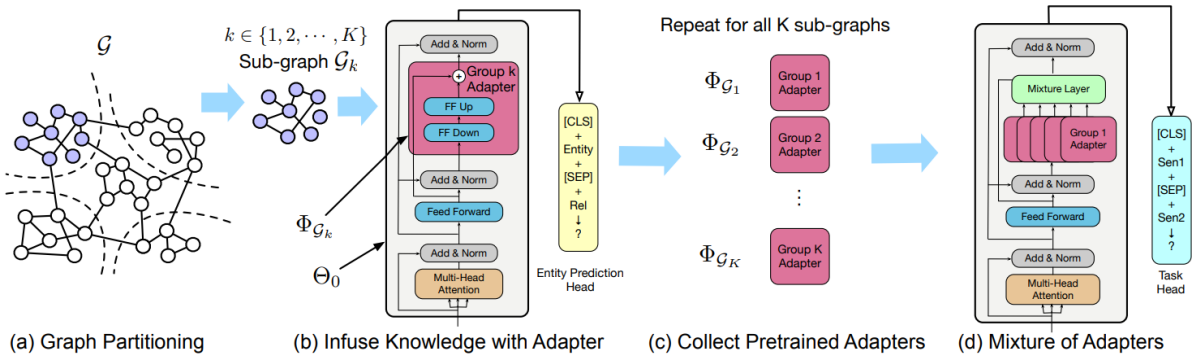

Figure 3.1.: Overview of the MoP processing pipeline from graph partitioning to the mixture of adapters. Image with permission from Meng, Liu, Clark, et al. (2021).

3.2.1. Data Pre-processing

For the model experiments, we have to process two separate data sources: the data from OntoChem that we use for the knowledge enhancement and the fine-tuning data from the BLURB benchmark tasks. The following describes the necessary pre-processing steps for the two data sources.

OntoChem Data

To begin with, we had to tackle the challenge of processing the biomedical data triplets provided by our industry partner, OntoChem. This thesis's data is publicly available on the SciWalker platform hosted by OntoChem. More precisely, the data is made available

through the "FactFinder" functionality of SciWalker[4], where data triplets can be downloaded based on custom settings. Figure 3.2 illustrates how the FactFinder can be used to find entity-relation-entity triplets.

OntoChem extracted all of the data that we use from MedLine[5], a bibliographic database from the US National Library of Medicine's (NLM), as described by Irmer, Bobach, and Böhme (2019). MedLine contains more than 30 million references to journal articles focusing on chemistry and bio-medicine. Further information on OntoChem's extraction process and original data sources can be found in Appendix A. Thanks to the collaboration between OntoChem and sebis, we were given access to a single file containing the, at the time, newest extraction of triplets available to OntoChem. This significantly sped up the data collection process since we did not have to download sets of data triplets individually from the FactFinder.

Triplet Structure In figure 3.2, examples of data triplets can be seen. A triplet consists of 3 parts: An entity subject, an entity object, and the relation between them. The relations provided by OntoChem are unique to the type of entities that the relation connects, so there can be several types of "induces" relations. For example, there is one possible relation for a "substance" as a subject and a "disease" as an object ([substance] induces [disease]), and another one for a "physiology" as a subject and a "disease" as an object ([physiology] induces [disease]). In contrast to other works, like MoP or KEBLM, we decided to leverage this entity-type information for our knowledge injections. Following Meng, Liu, Clark, et al. (2021), we determined the 20 most frequent relations in the data. However, we did so twice: Once with the relations including the entity types and once with the "fused" relations where no difference is made between entity types. We named the two approaches "Onto20Rel" for the fused version and "OntoType20Rel" for the version including the entity types. As a consequence of the more fine-grained representation of the relations by the OntoType20Rel approach, it contains a smaller number of total triplets, which can be seen in table 3.1. The total number of unique entities in the Onto20Rel and OntoType20Rel set is **159,833**.

Sub-graph Construction Finally, we needed to construct knowledge sub-graphs via graph partitioning to effectively handle the computational complexity associated with massive biomedical KGs. The partitioning strategy had to solve the NP-complete "balanced graph partition" problem at scale in order to "(1) maximize the number of resulting knowledge triples to retain as much factual knowledge as possible [and] (2) balance nodes over partitions to reduce the overall parameters across different entity prediction heads" (Meng, Liu, Clark, et al., 2021). We follow the suggestion of MoP to use the METIS algorithm (Auer, Bizer, Kobilarov, et al., 2007b) for the partitioning process since it can solve the balanced graph partitioning efficiently even for KGs of the size of OntoChem.

[4]https://sciwalker.com/analytics/factfinder
[5]https://www.nlm.nih.gov/medline/index.html

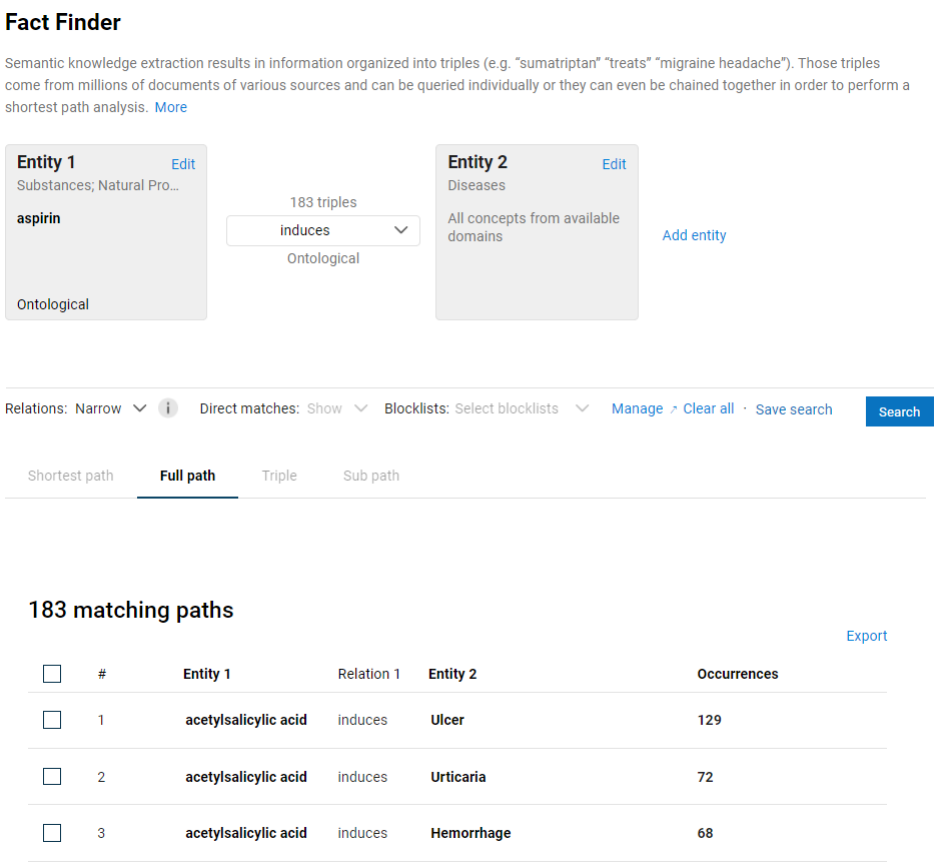

Figure 3.2.: Screenshot of the FactFinder functionality on the SciWalker platform. Here, the entity "aspirin" and the relation "induces" are preset. The results list triplets with the preferred name for aspirin (Acetylsalicyclic acid). All of the data can be accessed for further processing through the "Export" button.

The resulting sub-graphs are transformed into training data, which is split into three separate files:

- Entity2Id: Maps all unique entities to their corresponding IDs
- Relation2Id: Maps the twenty most frequent relations to their corresponding IDs
- Train2Id: The main training file, where the data triplets are saved with their entity and relation IDs

BLURB Data

In chapter 2, we described four of the BLURB benchmark downstream tasks with example data, namely HoC (Baker, Silins, Guo, et al., 2015), PubMedQA (Jin, Dhingra, Liu, et al., 2019), BioASQ7b (Nentidis, Bougiatiotis, Krithara, & Paliouras, 2020), and MedNLI (Romanov &

Onto20Rel	#Triplets	OntoType20Rel	#Triplets
relates to	708,076	[protein] relates to [disease]	295,841
induces	502,512	[substance] induces [physiology]	282,721
modulates	326,534	[food] contains [compound]	269,211
treats	225,279	[substance] treats [disease]	247,348
inhibits	219,720	[biomarker] of [disease]	205,604
is analyzed by	195,291	[substance] is analyzed by [method]	130,275
produces	173,979	[plant] produces [compound]	102,270
increases activity of	148,673	[protein] induces [physiology]	85,411
contains	133,241	[compound] increases activity of [protein]	85,196
increases	110,803	[compound] decreases activity of [protein]	72,311
detects	93,373	[substance] inhibits [physiology]	68,728
decreases activity of	85,425	[protein] is a [biomarker]	65,558
prevents	82,574	[anatomy] produces [protein]	64,206
increases expression of	80,771	[substance] prevents [disease]	60,260
expresses	62,142	[protein] induces [disease]	59,577
attenuates	54,865	[substance] modulates [protein]	54,533
decreases expression of	51,152	[protein] is analyzed by [method]	54,250
binds to	49,206	[method] treats [disease]	35,768
is a	47,435	[method] detects [physiology]	33,504
affects expression of	37,399	[protein] modulates [physiology]	24,332
Total	**3,388,450**		**2,296,904**

Table 3.1.: Comparison of triplet numbers for the twenty most frequent relations for the Onto20Rel set and the OntoType20Rel set

Shivade, 2018). Computing resource limitations restrict us from incorporating additional tasks: fine-tuning all of our model and KG combinations on the complete BLURB benchmark was not feasible. We chose the four tasks named above for their popularity and best comparability with existing research (they had a significant overlap with the tasks explored by PubMedBERT, LinkBERT, MoP, KEBLM, and DAKI). Moreover, while we can not cover the full BLURB benchmark, the chosen task spectrum still gives a diverse representation of biomedical NLP.

We access the openly available datasets from the website of the BLURB benchmark [6]. The data requires minimal pre-processing and can be loaded with the Hugging Face transformers and datasets library (Wolf, Debut, Sanh, et al., 2020). Depending on the downstream task,

[6] https://microsoft.github.io/BLURB/

the data is encoded in JSON or TSV format, and we processed it directly with standard data loaders.

3.2.2. Model Training

The most important (and time-consuming) phase of the model experiments is the model training. In our approach, the training is divided into two parts: The adapter knowledge injection and the full model fine-tuning. In the following, we first describe the training environment and then the two training phases.

Training Environment

To facilitate our experiments, we established a training environment on Google Colab[7]. Google Colab offers cost-effective access to GPUs like T4s and V100s, which are essential for training LLMs. Moreover, Google Colab's integration with Jupyter Notebooks makes it easy to streamline deep learning experiments and code documentation. As our base models, we chose the base versions of PubMedBERT (Gu, Tinn, Cheng, et al., 2020) and BioLinkBERT (Yasunaga, Leskovec, & Liang, 2022), which we already described in the literature overview in chapter 2. We made this choice because of the high performance of these models when factoring in their relatively small amounts of parameters and because of their popularity in the BioNLP domain. In our experiments, we loaded all base models from Hugging Face (Wolf, Debut, Sanh, et al., 2020) and chose PyTorch (Paszke, Gross, Massa, et al., 2019) as our deep learning framework.

Knowledge Injection

The knowledge injection is the most crucial aspect of our methodology. It is what transforms our vanilla LLMs into KELMs. After partitioning the knowledge graph into subgraphs, we leverage the adapterhub library (Pfeiffer, Rücklé, Poth, et al., 2020) to infuse Pfeiffer adapters (Pfeiffer, Kamath, Rücklé, et al., 2020) with knowledge from each of the 20 sub-graphs. The used training objective is entity prediction, where entities in KG triplets are masked and then predicted by the LLMs (see KELM section in chapter 2). As already discussed in chapter 2, this approach is flexible and efficient, as it doesn't necessitate fine-tuning the parameters of the underlying LLMs and can have additional benefits like mitigating the catastrophic forgetting issue.

Fine Tuning

Task-specific fine-tuning is carried out for the four chosen BLURB benchmark downstream tasks. We aligned our hyperparameters with the settings of Meng, Liu, Clark, et al. (2021) and Gu, Tinn, Cheng, et al. (2020): We employ Adam [30] alongside the typical slanted triangular learning rate schedule, with a warm-up for the initial 10 percent of steps and a

[7] https://colab.research.google.com/

cool-down for the subsequent 90 percent, and set the dropout probability at 0.1. Furthermore, we followed Meng, Liu, Clark, et al. (2021) and Pfeiffer, Kamath, Rücklé, et al. (2020) by introducing mixture layers and AdapterFusion to route valuable knowledge from the adapters to downstream tasks automatically. Given the random initialization of the task-specific model and dropout, outcomes can fluctuate based on different random seeds, particularly for the small PubMedQA and BioASQ7b datasets. For a more accurate representation, we follow Gu, Tinn, Cheng, et al. (2020) and Meng, Liu, Clark, et al. (2021) and present average results from ten iterations for BioASQ7b and PubMedQA, five iterations HoC, and three for MedNLI.

In an ideal scenario, each model on each dataset should undergo distinct hyperparameter adjustments. However, considering the extensive computational load to evaluate every possible combination and the necessity to average over varied runs, this was not feasible with our limited computing resources. Notably, Meng, Liu, Clark, et al. (2021) "observe that the development performance is not very sensitive to hyperparameter selection, as long as they are in a ballpark range". Therefore, we only slightly adjusted the learning rate and batch size to support our limited computing resources while maintaining as much performance as possible (using the development set). All specific values will be provided in Appendix A for reproducibility.

3.2.3. Model Evaluation

In evaluating our models, we adhere to best practices established in the BLURB benchmark. For the HoC task, we utilize the micro F1 metric, while for all other tasks, accuracy serves as the primary evaluation metric. To contribute to the existing body of knowledge, we compare our findings with several key works, including MoP (Meng, Liu, Clark, et al., 2021), KEBLM (Lai, Zhai, & Ji, 2023), DAKI (Q. Lu, Dou, & Nguyen, 2021), PubMedBERT (Gu, Tinn, Cheng, et al., 2020), and BioLinkBERT (Yasunaga, Leskovec, & Liang, 2022). This comparative analysis provides insights into what kinds of advancements adapters-based KELMs can bring to the field compared to each other and the baseline models. For our best models, we also use qualitative probing to gain further insights into the inner workings of our methods. To do so, we directly compare select predictions on the test set between enhanced models and their base-version.

3.3. Research Survey

Finally, this section will describe the methodology used for the third part of this work: The research survey addressed to medical professionals. Scheetz, Rothschild, McGuinness, et al. (2021) find that "few studies have examined clinician perceptions of new AI technologies on healthcare provision and the clinical workforce". Medical professionals have hands-on experience with the challenges and requirements of medical practice. Their insights are invaluable in understanding where the application of NLP could be most beneficial and where it might face challenges. By knowing what medical professionals prioritize or are

concerned about, developers and researchers can tailor biomedical NLP solutions to better fit the actual needs of the medical community. They can also highlight potential pitfalls or risks associated with integrating NLP into medical practice, ensuring that developers are aware of and can mitigate these risks. Last but not least, acceptance by the end-users is vital for any technology, especially in the critical field of medicine. Understanding and addressing their concerns can lead to faster and broader adoption of NLP solutions (Scheetz, Rothschild, McGuinness, et al., 2021).

3.3.1. Survey design

While AI and NLP experts understand the technology, medical professionals understand patient care. This survey was designed to bridge the gap between these two, fostering a comprehensive understanding that benefits both sides. We primarily based our study on a survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology, and radiation oncology at the University of Melbourne (Scheetz, Rothschild, McGuinness, et al., 2021). We did so in the hopes of making our findings in the field of biomedical NLP directly comparable to their findings in the area of general biomedical AI (with a focus on computer vision). The survey consisted of 20 short questions and was designed to take approximately 10 minutes to complete. A brief introduction was provided within the survey for medical professionals unfamiliar with LLMs and their applications in the (bio)medical field. The survey was hosted on Google Forms[8].

Given that medical professionals are notoriously stressed and short on time, we already expected a low return rate before the start of the survey. Therefore, in contrast to Scheetz, Rothschild, McGuinness, et al. (2021), who had better connections to the medical domain and more resources, we opened the study to medical professionals of all disciplines and medical students. We did so, hoping to reach a sufficiently high participant count and diversity that allows for meaningful insights. Participants from the related AI study were all members of Australian and New Zealand Royal Colleges. In our survey, we contacted participants primarily by reaching out directly to universities, hospitals, and medical centers in Munich and the surrounding countryside, as well as personal contacts of the authors and advisor Juraj Vladika (including some international contacts from related research). For questions about the area of expertise and participants' experiences, students were asked to refer to their desired (future) area of expertise and focus of study.

Participants were prompted to give online electronic informed consent before survey commencement. The survey questions, data, and the document containing the introduction to biomedical NLP, which was included in the survey, are provided under Appendix C.

[8]https://www.google.com/forms/about/

3.3.2. Data analysis methods

We conducted a complete-case analysis where incomplete surveys were excluded from the process. The data was analyzed and plotted using the Python matplotlib (J. D. Hunter, 2007) and pandas (pandas development team, 2020) libraries. While Scheetz, Rothschild, McGuinness, et al. (2021) compared findings between their three target groups (ophthalmologists, radiologists, radiation oncologists, and dermatologists), we differentiated whether a participant was already working or studying medicine. While we would have liked to additionally compare the results alongside the profession, location of practice, and years of experience, the limited number of participants would have restricted the quality of insights gained from this analysis. We compared our findings to the results of Scheetz, Rothschild, McGuinness, et al. (2021). The data supporting the findings of this survey is given in Appendix C.

3.3.3. Preliminary status of the survey

In spite of our larger target group, reaching medical professionals proved to be tremendously difficult. The main hurdles were the very stressful schedules of doctors and nurses and their subsequent reluctance to make time for an unpaid survey, and the bureaucracy and general difficulty connected with reaching medical professionals at a large scale. The formal process to conduct the survey at the TUM "Klinikum rechts der Isar" (MRI) is still ongoing at the time of writing. Therefore, we have decided to give the survey in this thesis a preliminary status in preparation for the survey at the MRI. We will use the results to gather initial insights and experience. The survey will then be expanded to the MRI at sebis when the formal process is completed and the survey is approved by the Ethical Committee of the university hospital.

4. Results

Our methodology to answer the three research questions of this thesis was described in chapter 3. This chapter will present the results. Due to their importance for this chapter, we repeat the RQs here once more:

RQ1 What adapter-based approaches to knowledge-enhancement exist, and how do they compare to each other?

RQ2 Can we improve existing approaches with new methods and data from a private ontology?

RQ3 Is the research on biomedical KELMs relevant to medical professionals, and what factors hinder or support the deployment of the technology in practice?

The results of the literature review will answer RQ 1. The literature review and model experiments will be used to answer RQ 2. Finally, RQ 3 will be answered by the survey targeted to medical students and professionals.

4.1. Literature Review

This section will present the results of the systematic literature review on adapter-based knowledge-enhancement . The results will be interpreted in the context of the experimental setting and RQ 1 of the thesis.

4.1.1. Overview

Figure 4.1 shows the source distribution for all included papers. Fifty-nine papers were found by applying the search string as a command on the ACL, ACM, and IEEE search engines. Due to their importance for the thesis, we included three additional papers from other sources. These papers were found through online search and paper references during the general research process. In summary, after the abstract screening, 25 articles met all inclusion criteria (and no exclusion criteria). After the full paper screening, 22 papers remained to form the final paper pool of the survey. While 22 total included papers seems like a small absolute number, it is necessary to consider the small niche and harsh inclusion criteria for the survey. Also, adapters are a very recent invention in the scope of NLP research, and all papers were published in 2020 at the latest.

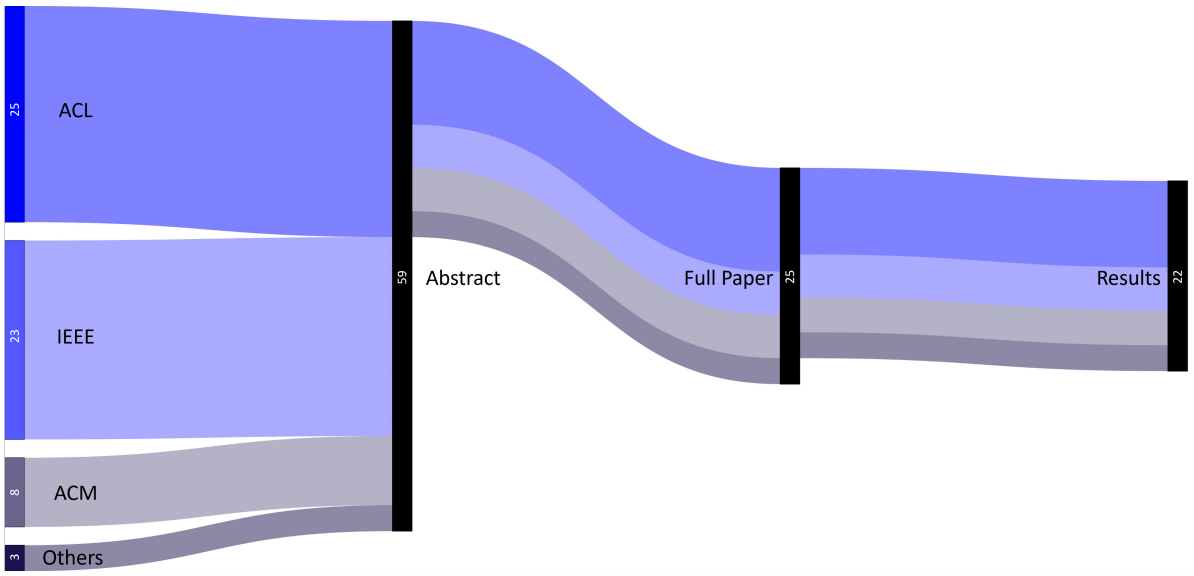

Figure 4.1.: Visualization of the literature sources and the selection process

Table 4.1 gives an overview of all papers included in the survey. The columns depict the following information: **paper:** The authors of the paper, together with the publication date; **adapter type:** The type or method of the adapter used were found in different journals, only the most recent version was included); **scope:** Domain scope that indicates whether the approach is open or closed domain; **coverage:** Domain coverage that indicates whether a closed domain approach is specific to a single domain or multiple domains; **biomed:** Indicates whether the paper explicitly addresses the biomedical field; **task:** Lists abbreviations for the downstream tasks covered within the paper; **nickname:** A nickname or shorthand reference for the paper or model framework (if given by the authors).

4.1.2. Data Analysis

Next, we will take a closer look at the paper-specific data collected from the survey. This section starts with a quantitative analysis showcasing and interpreting quantitative distributions. Afterward, we report significant qualitative insights from the papers.

Quantitative Analysis

Year-wise Distribution To begin with, we assess how many papers were published each year to get a sense of the trend and growth in the area (4.2).

There has been a noticeable increase in publications on adapter-based approaches to knowledge-enhanced language models in recent years, especially from 2022 onward. This trend suggests growing interest and research activity in the domain.

paper	adapter type	scope	coverage	biomed	task	nickname
Zou et al. (2022)	K-Adapter	open	/	no	RC	K-MBAN
Moon et al. (2021)	Houlsby	open	/	no	MT	/
Yu et al. (2023)	Unique	open	/	no	SL	CSBERT
Quian et al. (2022)	Unique	open	/	no	SR	/
Li et al. (2023)	Houlsby	closed	multi	no	SF	/
Li et al. (2023)	Unique	open	/	no	SC	CKGA
Nguyen et al.(2023)	Pfeiffer	open	/	no	SA	/
Lai et al. (2023)	Pfeiffer	closed	single	yes	QA, NLI, EL	KEBLM
Guo et al. (2022)	Unique	open	/	no	NER	/
Chronopoulou et al. (2023)	Bapna and Firat	closed	both	no	LM	Adapter-Soup
Wold et al. (2022)	Houlsby	open	/	no	LAMA	/
Chronopoulou et al. (2022)	Unique	closed	multi	no	LM	/
Hung et al. (2022)	Pfeiffer	closed	multi	no	TOD	DS-TOD
Emelin et al. (2022)	Houlsby	closed	multi	no	TOD	/
Xu et al. (2022)	Bapna and Firat	open	/	no	KGD	KnowBERT
Kær et al. (2021)	Pfeiffer	closed	multi	yes	NER, STC	mDAPT
Lu et al. (2021)	K-Adapter	closed	single	yes	NLI	DAKI
Majewska et al. (2021)	Pfeiffer	open	/	no	EE	/
Lauscher et al. (2020)	Houlsby	open	/	no	GLUE	/
Meng et al. (2021)	Pfeiffer	closed	single	yes	BLURB	MoP
Wang et al. (2020)	K-Adapter	open	/	no	RCL, ET, QA	K-Adapter
Xie et al. (2022)	Pfeiffer	closed	single	yes	ES	KeBioSum

Table 4.1.: Overview of the results for the literature survey, including all papers and their references. The task acronyms are explained in the glossary at the end of the thesis. The dotted lines separate the database sources: First come the IEEE papers, then ACM, ACL, and finally, the papers from other sources.

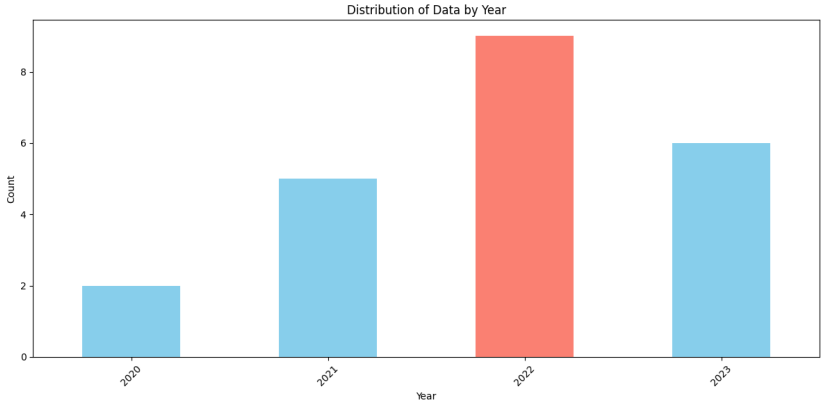

Figure 4.2.: Year-wise distribution of publications

Adapter Type Distribution Next, we evaluate the popularity and variety of adapter types used across the papers (4.3).

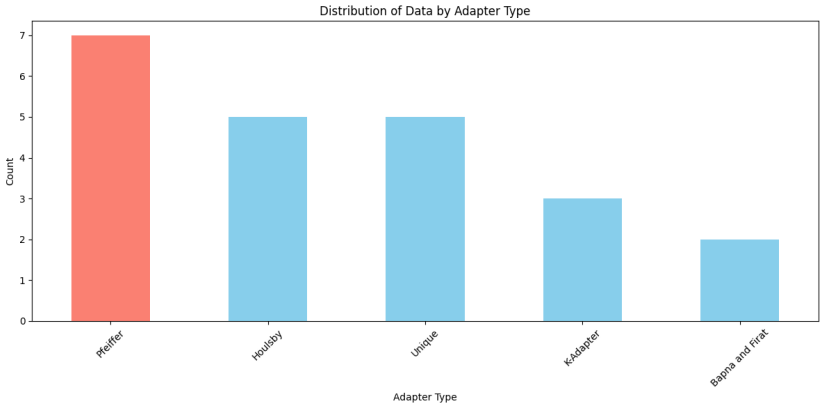

Figure 4.3.: Distribution of adapter types being used in the articles

The "Pfeiffer" adapter type stands out as the most common, which suggests that the architecture is the most popular methodology in the field. This popularity is likely not only an achievement of the performance of the adapter but also of the well-established adapter-hub platform, which, although offering other options, uses adapters with the Pfeiffer configuration by default. This finding showcases a need and trend to build custom adapters well-suited to individual tasks. In the upcoming years, we will likely see many novel adapter architectures. The "K-Adapter" and "Bapna and Firat" adapters are the least frequently mentioned architectures, suggesting that these approaches are less well-established. Overall, a variety of adapter types are present, indicating a diverse range of methodologies being explored.

Domain Analysis Third, we analyze the distribution of papers across the domain scope and coverage to understand domain-specific preferences in the literature.

The first plot in figure 4.4 shows that the open-domain scope is the most popular, with a significant number of papers exploring adapter-based approaches within the open domain. The popularity likely is caused by the interest in creating LLMs with a common-sense understanding or "world knowledge". Researchers see this understanding or knowledge as the next step towards general artificial intelligence (Davis, 2021; Marcus, 2019).

As illustrated by the second plot in figure 4.4, the single- and multi-domain approaches are split almost evenly within the closed-domain papers. There is most likely no further insight in this distribution.

Finally, the third plot addresses the coverage of the biomedical domain. In absolute numbers, only five papers focus on the biomedical domain, but relative to other parts, the biomedical field is by far the most prominent of all domain-specific approaches. However, the survey is biased toward the biomedical domain: We included two papers from sources

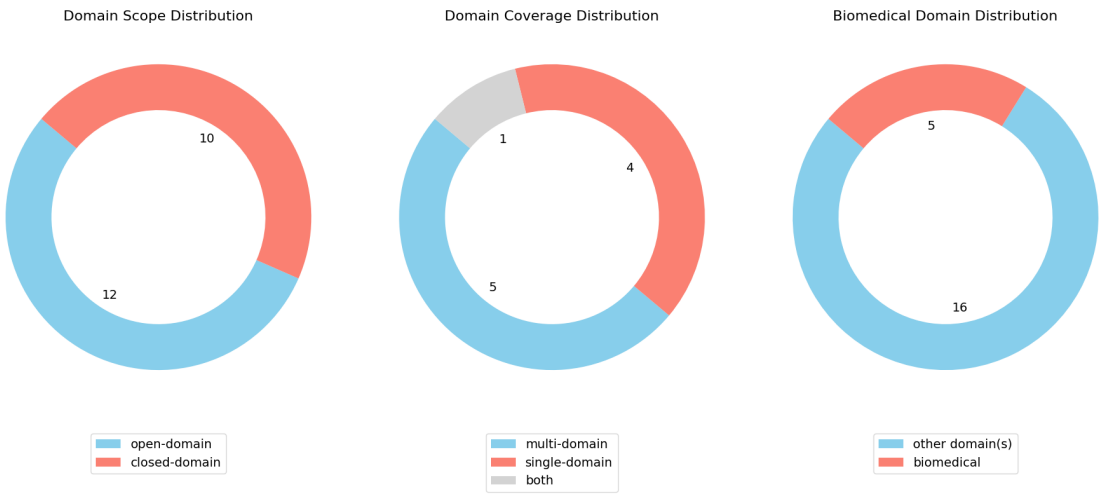

Figure 4.4.: Distribution of domain scope, coverage, and the biomedical domain

other than the three main platforms after we found them while researching biomedical NLP specifically. Nevertheless, with three articles dedicated to the biomedical domain without bias, it still is the most frequent domain. The popularity likely comes down to (bio)medicine historically being one of the most active research domains on its own (Cimini, Gabrielli, & Labini, 2014).

Task Distribution A highly diverse range of tasks is being explored throughout the papers, which signifies the versatility and potential of adapter-based approaches across different natural language processing tasks. However, combined with the limited number of papers in the survey, the approach-versatility prevents further meaningful quantitative analysis. Still, tasks such as Reading Comprehension (RC), Named Entity Recognition (NER), and Question Answering (QA) appear to be popular areas of focus in the literature. Figure 4.5 provides a word cloud of all keywords in the downstream tasks as a visualization, showing that there is also a focus on tasks with a dialogue or sentiment component.

In summary, we report the following quantitative findings:

- Adapter-based KELMs are a new development in NLP, but there has been fast-growing interest in adapter-based approaches in recent years.
- Various adapter methodologies are being explored, with some well-established ones and many novel approaches.
- Research predominantly focuses on open-domain tasks. For closed-domain approaches, the biomedical domain is the most popular.

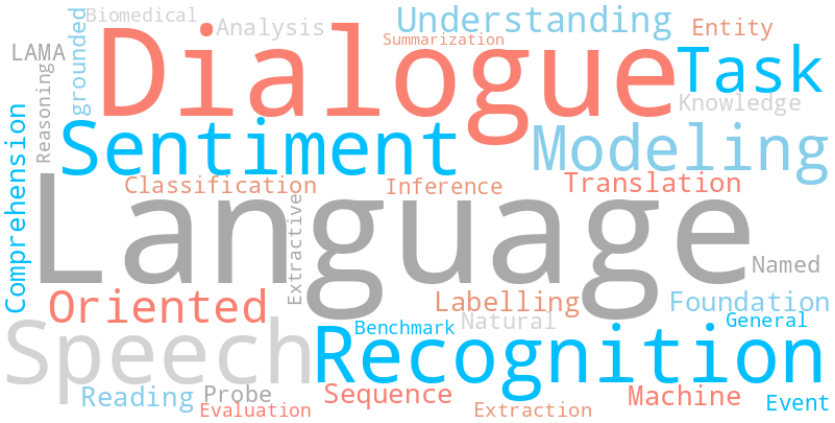

Figure 4.5.: Wordcloud of keywords in the task distribution

- A large diversity of downstream tasks is being addressed in the literature.

Qualitative Analysis

This section of the analysis highlights recurring themes and individual insights from the papers. Fully summarizing all articles was outside the scope of this survey. However, the qualitative insights still provide an overview of the specific areas of focus and technical details discussed in the papers. We address both open and closed-domain papers in this section but focus on the biomedical domain separately in the next section. This particular emphasis follows the theme of this thesis and is especially important for understanding the following experimental part of this work (section 4.2).

General Knowledge The quantitative analysis showed that open-domain approaches are more popular than their close-domain counterparts. Subsequently, there is also a large variety in the used frameworks, knowledge sources, and overall goals of the papers. Two commonly used KGs for general knowledge are ConceptNet (Speer, Chin, & Havasi, 2017) and DBpedia (Auer, Bizer, Kobilarov, et al., 2007a), which we already explored in chapter 4.1. Two example works that use these KGs are Wold (2022) and the CKGA ("knowledge graph-based adapter") by G. Lu, Yu, Yan, and Xue (2023). Wold (2022) train adapter modules on sub-graphs of ConceptNet to inject factual knowledge into LLMs. They evaluate their framework on the Concept-Net Split of the LAMA Probe (Petroni, Rocktäschel, Lewis, et al., 2019) and see increasing performance while only adding 2.1% of additional parameters to the original models. Wold (2022) also mention the biomedical domain explicitly as a possible domain-specific application of their work, but they stick to the open domain and do not evaluate their approach on any biomedical benchmarks.

CKGA (G. Lu, Yu, Yan, & Xue, 2023), on the other hand, tackle aspect-level sentiment classification by leveraging knowledge from DBpedia. They "link aspects to [DBpedia] end extract an aspect-related sub-graph. Then, a pre-trained language model and the knowledge

graph embedding are utilized to encode the common-sense knowledge of entities based on which the corresponding knowledge is extracted with a graph convolutional networks [sic]" (G. Lu, Yu, Yan, & Xue, 2023).

Linguistic Knowledge Instead of only including factual knowledge, some works additionally inject linguistic knowledge into adapters (Majewska, Vulić, Glavaš, et al., 2021; R. Wang, Tang, Duan, et al., 2020; Yu & Yang, 2023; Zou, Zhang, Song, et al., 2022). While LLMs already encode a range of syntactic and semantic properties of language, Majewska, Vulić, Glavaš, et al. (2021) explain that they "are still prone to fall back on superficial cues and simple heuristics to solve downstream tasks, rather than leverage deeper linguistic information". Their paper explores and uses the interplay between verb meaning and argument structure. They use the gained knowledge to enhance LLMs with Pfeiffer Adapters to improve English event extraction and machine translation in other languages. Another example is the work of Zou, Zhang, Song, et al. (2022) on machine reading comprehension (MRC). They proposed the K-MBAN model to integrate external knowledge, both linguistic and factual, into LLMs through K-Adapters.

Domain-specific Knowledge (nonbiomedical) Chronopoulou, Peters, and Dodge (2022) propose a parameter-efficient approach to domain-adaptation using adapters. They "represent domains as a hierarchical tree structure where each node in the tree is associated with a set of adapter weights". Their work focused on specializing adapters in website domains like booking.com and yelp.com. In another instance, Chronopoulou, Peters, Fraser, and Dodge (2023) propose "AdapterSoup". In this framework, they also use adapters for domain-specific tasks but use "an approach that performs weight-space averaging of adapters trained on different domains". AdapterSoup can be helpful in a variety of domain-specific approaches in low-resource settings, especially when only a small amount of data on a specific subdomain is obtainable and closely related adapters are available instead.

Other Insights R. He, Liu, Ye, et al. (2021) criticize that "existing work only focuses on the parameter-efficient aspect of adapter-based tuning while lacking further investigation on its effectiveness". They address this issue with their work and "show that adapter-based tuning better mitigates forgetting issues than fine-tuning since it yields representations with less deviation from those generated by the initial [pre-trained language model]" (R. He, Liu, Ye, et al., 2021). They found that adapter-based approaches outperform fine-tuning in low-resource and cross-lingual settings and are "more robust to overfitting and less sensitive to changes in learning rates" (R. He, Liu, Ye, et al., 2021).

Biomedical Adapter-based Works

We have already seen that the biomedical domain is the most prevalent among the closed-domain approaches to adapter-based KELMs. We have found the works of DAKI (Q. Lu, Dou, & Nguyen, 2021), MoP (Meng, Liu, Clark, et al., 2021), KeBioSum (Xie, Bishop, Tiwari,

& Ananiadou, 2022), and KEBLM (Lai, Zhai, & Ji, 2023), to be the most impactful. We will discuss them individually (by order of publication) in this section. We report the benchmark results of all four papers and compare them to our final results (where they overlap) in table 4.2 in section 4.2. For completeness, we refer to Kær Jørgensen, Hartmann, Dai, and Elliott (2021) for information on the m-DAPT framework, which addresses multi-lingual domain adaptation for biomedical LLMs

DAKI According to the results of our literature survey, DAKI ("Diverse Adapters for Knowledge Integration") was the first work to use adapters specifically for knowledge enhancement in the biomedical domain. Q. Lu, Dou, and Nguyen (2021) leverage data from the UMLS meta-thesaurus (Bodenreider, 2004) and UMLS Semantic Network groups concepts, but also from Wikipedia articles for diseases as proposed by Y. He, Zhu, Zhang, et al. (2020). Their architecture consists of three main components: the base LLM, the knowledge-specific adapters, and an adapter integration module. Following the K-Adapter architecture (R. Wang, Tang, Duan, et al., 2020), their integration module "aims to adaptively integrate the knowledge adapters by assigning them different importance weights, as opposed to the simple concatenation of the outputs of adapters" (Q. Lu, Dou, & Nguyen, 2021). They evaluate their final model's performance over three biomedical NLP tasks (QA, NLI, NER). Specifically, MEDIQA-2019 TRECQA-2017 MEDNLI BC5CDR NCBI. The main observations from the evaluation are that DAKI can improve the performance of BERT, RoBERTa, and ALBERT, as reflected in DAKI-BERT, DAKI-RoBERTa, and DAKI-ALBERT (Q. Lu, Dou, & Nguyen, 2021).

KeBioSum Xie, Bishop, Tiwari, and Ananiadou (2022) state that their work is "the first study exploring the inclusion of fine-grained biomedical knowledge with PLMs for biomedical extractive summarization". They are also the first to use adapters in this context. In particular, their KeBioSum framework first identifies medical knowledge in documents through PICO ("Population, Intervention, Comparison, and Outcome"), a structural medical knowledge representation method often used for creating literature search queries (Huang, Lin, & Demner-Fushman, 2006). The gained information is then injected into the LLMs Pfeiffer Adapters utilizing the adapter-hub plattform (Pfeiffer, Rücklé, Poth, et al., 2020). Similar to DAKI, the primary knowledge source stated by the authors is UMLS. Their proposed KeBioSum framework efficiently improves extractive summarization in the biomedical domain, providing further proof that pre-trained LLMs can be enhanced through knowledge injection with adapters. However, given that our own work does not include experiments on text summarization, we can not compare our results to the experimental results of KeBioSum.

MoP Next, we have the work of Meng, Liu, Clark, et al. (2021), which is essential for the experimental section of our thesis. They propose to use a "Mixture of Partitions" (MoP) enabled by AdapterFusion (Pfeiffer, Kamath, Rücklé, et al., 2020). They recognize that KGs like UMLS, which can be several gigabytes large (Bodenreider, 2004), are very expensive to train on in their entirety. This is especially true for language modeling objectives like entity prediction that require calculating a softmax over all entities (Meng, Liu, Clark, et al., 2021).

MoP tackles this problem through graph partitioning. In particular, they use the Auer, Bizer, Kobilarov, et al. (2007b) algorithm, which we have already mentioned in our methodology in chapter 3, to partition the graph into smaller, more manageable sub-graphs. A set of adapters (MoP uses Pfeiffer Adapters) can then be trained on these sub-graphs individually using the entity prediction objective. MoP then utilizes AdapterFusion to amalgamate the knowledge encapsulated in the adapters.

With their experiments, Meng, Liu, Clark, et al. (2021) show that using only a subset of triplets limited to the 20 most frequent relations from the KG works best for most downstream tasks on the BLURB benchmark. They reported significant performance gains in their final results and published their code on GitHub.

KEBLM KEBLM ("Knowledge-Enhanced Biomedical Language Models") was published during the work on this thesis. They were the first to acknowledge the heavy reliance of related work on the UMLS ontology. Subsequently, their framework's trademark is that it allows for the inclusion of a variety of knowledge types from multiple sources into biomedical LLMs. While they still use UMLS as a source, they also draw data from PubChem (Kim, Chen, Cheng, et al., 2020) and the MSI KG (Ruiz, Zitnik, & Leskovec, 2021). In contrast to DAKI, which also utilizes more than one source, KEBLM includes a knowledge consolidation phase after the knowledge injection, where they "teach the fusion layers to effectively combine knowledge from both the original PLM and newly acquired external knowledge by using a large collection of unannotated texts" (Lai, Zhai, & Ji, 2023). Like MoP, they use AdapterFusion and Pfeiffer Adapters and evaluate their framework on a selection of downstream tasks, including MedNLI and PubMedQA.

Notably, all existing approaches use the UMLS ontology as one of their primary knowledge sources. While KEBLM and DAKI try to diversify their data sources to some extent, this circumstance, nevertheless, justifies our approach to leverage data from OntoChem to contribute to the diversification of knowledge sources in the field.

4.1.3. Review Summary

In summary, the systematic literature review portrayed what adapter-based approaches to knowledge-enhancement exist and how they compare to each other. The findings underscore the nascent yet fast-growing interest in this domain. The diversity of adapter types, with the "Pfeiffer" adapter type being predominant, allows for various approaches to the overall model architectures. From a task perspective, the range of NLP tasks covered in the literature signifies the adaptability and potential of adapter-based approaches. Moreover, both the open-domain and closed-domain scope have garnered significant attention, reflecting the broader ambition of developing language models with generalized as well as specialized knowledge and understanding. Notably, the biomedical domain stands out as the most explored closed-domain area, underscoring the significance of accurate knowledge representation in this

intricate area.

In conclusion, adapter-based knowledge-enhancement represents a unique and flexible way to integrate vast, diverse knowledge into LLMs without expensive fine-tuning. With the in-depth analysis given in this chapter, we discovered what adapter-based approaches to knowledge-enhancement exist in the literature and how they compare to each other. Therefore, we have successfully answered RQ1 of this thesis and provided a novel and extensive resource for other researchers in the field.

4.2. Model Experiments

All model experiments were carried out according to our proposed methodology in chapter 3. In the following, we will report the benchmark results, qualitative probing, and give further information on the course of the experiments and an interpretation of the results. Unfortunately, not all data triplets provided to us by OntoChem were fully usable due to missing ID mappings. Several projects running at OntoChem in parallel slowed the communication on this issue, and only about two-thirds of the provided triplets could be used for the experimental part of this thesis.

Shortcomings of BioLinkBERT During our initial experiments with BioLinkBERT, our attempts at reproducing the original results of Yasunaga, Leskovec, and Liang (2022) could not live up to the strong performance reported by the authors. After a thorough investigation of the code and documentation provided by the authors, we realized that they seemingly employed a different evaluation process compared to the standard (Gu, Tinn, Cheng, et al., 2020) methodology: Instead of averaging over a set of runs for datasets, they only reported the results of a single run. Only for the one seed that was given in their documentation did their model reach the performance reported in their publication. For example: Yasunaga, Leskovec, and Liang (2022) reported an accuracy of 70.20 on PubMedQA, but when averaged over 10 runs (best practice), the performance drops to 56.76 ± 3.00. We attempted to contact the authors multiple times (via mail and LinkedIn) to discuss this issue, but we did not get a response to this day. We, therefore, had to assume that their reported results were possibly lucky runs. To account for this, we rerun their experiments with best-practice run-averaging for HoC, PubMedQA, BioASQ7b, and MEDNLI. For our evaluation and model comparison, we used these new results instead (see Table 4.2).

4.2.1. Experiment Results

We present our experiment results in direct comparison with the results of the public releases of SciBERT, BioBERT, PubMedBERT, BioLinkBERT (reproduced), MoP, and KEBLM. The results allow for a quantitative and qualitative assessment of our experiment methodologies in the competitive and active field of adapter-based KELMs. For reproducibility, all run seeds and the specific hyperparameters are listed in Appendix A.

Benchmark Results

↓ model | dataset →	HoC	PubMedQA	BioASQ7b	MedNLI
SciBERT-base	$80.52_{\pm 0.60}$	$57.38_{\pm 4.22}$	$75.93_{\pm 4.20}$	$81.19_{\pm 0.54}$
+ MoP	$81.79^{\dagger}_{\pm 0.66}$ ↑	$54.66_{\pm 3.10}$	$78.50^{\dagger}_{\pm 4.06}$ ↑	$81.20_{\pm 0.37}$
+ KEBLM	/	59.0	/	82.14
BioBERT-base	$81.41_{\pm 0.59}$	$60.24_{\pm 2.32}$	$77.50_{\pm 2.92}$	$82.42_{\pm 0.59}$
+ MoP	$82.53^{\dagger}_{\pm 1.08}$ ↑	$61.04_{\pm 4.81}$ ↑	$80.79^{\dagger}_{\pm 4.40}$ ↑	$82.93_{\pm 0.55}$ ↑
+ KEBLM	/	**68.00**	/	84.24
+ DAKI	/	/	/	83.41
PubMedBERT-base	$82.25_{\pm 0.46}$	$55.84_{\pm 1.78}$	$87.71_{\pm 4.25}$	$84.18_{\pm 0.19}$
+ MoP	$\mathbf{83.26}^{\dagger}_{\pm 0.32}$ ↑	$62.84^{\dagger}_{\pm 2.71}$ ↑	$90.64^{\dagger}_{\pm 2.43}$ ↑	$\mathbf{84.70}_{\pm 0.19}$ ↑
+ OntoType20Rel (ours)	$82.17_{\pm 0.62}$	$55.40_{\pm 5.57}$	$86.36_{\pm 3.07}$	$83.94_{\pm 0.63}$
+ Onto20Rel (ours)	$82.39_{\pm 0.65}$ ↑	$56.12_{\pm 2.91}$ ↑	$84.36_{\pm 4.73}$	$83.97_{\pm 0.59}$
BioLinkBERT-base (ours)	$82.21_{\pm 0.87}$	$56.76_{\pm 3.00}$	$91.29_{\pm 3.18}$	$84.1_{\pm 0.03}$
+MoP (ours)	$82.36_{\pm 0.57}$ ↑	$63.62^{\dagger}_{\pm 5.31}$ ↑	$91.50_{\pm 2.25}$ ↑	$83.78_{\pm 0.09}$
+OntoType20Rel (ours)	$82.37_{\pm 0.42}$ ↑	$60.46_{\pm 5.81}$ ↑	$\mathbf{92.14}_{\pm 2.30}$ ↑	$82.84_{\pm 0.34}$
+Onto20Rel (ours)	$82.24_{\pm 1.25}$ ↑	$63.28^{\dagger}_{\pm 4.46}$ ↑	$90.57_{\pm 3.14}$	$83.69_{\pm 0.55}$

Table 4.2.: Final results for the model experiments: The best results for every task are in bold. "↑" denotes that improvements are observed when compared to the base model. "†" denotes a statistically significant better result over the base model (T-test, $p < 0.05$). For all MoP metrics, we took the S20Rel metric for better comparability. Because of their unclarified status, we excluded all original BioLinkBERT results in favor of our reproduced results. DAKI and KEBLM did not evaluate on all of our used benchmarks and did not provide standard deviation or p-tests, but we still included their results for completeness.

Table 4.2 shows the results of our model evaluation on the four chosen tasks of the BLURB benchmark. In eight instances, we were able to improve the BioLinkBERT-base model, either with the pure MoP approach or our own approach with the OntoType20Rel KG. However, not all results led to a successful T-test. Moreover, in all instances apart from the BioASQ7b benchmark, KEBLM, and the original MoP approach stayed at the top of the leaderboard utilizing PubMedBERT and BioBERT.

An interesting result that we have to investigate further is the relatively worse performance of our approach on PubMedBERT compared to BioLinkBERT, even when factoring in the stronger base performance of BioLinkBERT. When the base models don't match, it is hard to distinguish whether performance gains or losses come from the difference in base models or the difference in the adapter-based approaches. Here, the base models of BioLinkBERT generally perform better than those of PubMedBERT or SciBERT over a variety of tasks.

Therefore, whenever we use BioLinkBERT, we cannot say how much of the performance gains come from the superiority of our approach versus the superiority of the base model.

Unfortunately, due to an oversight in logging, not all precision and recall values have been saved for the final runs. While the evaluation metrics required for the BLURB tasks in table 4.2 were most important, we would have liked to dive deeper into the run details. However, all run configurations and seeds were saved, enabling us to repeat the runs in the future when we have again access to the necessary computing power.

4.2.2. Qualitative Probing

We probed our best model, BioLinkBERT-base+OntoType20Rel, for qualitative insights on model prediction and compared it with the base version of the model, BioLinkBERT-base (both models fine-tuned on BioASQ7b). In the following, we give some excerpts from the two models' predictions on the test set and interpret them, in the hopes of gaining some insights on the inner workings of our method:

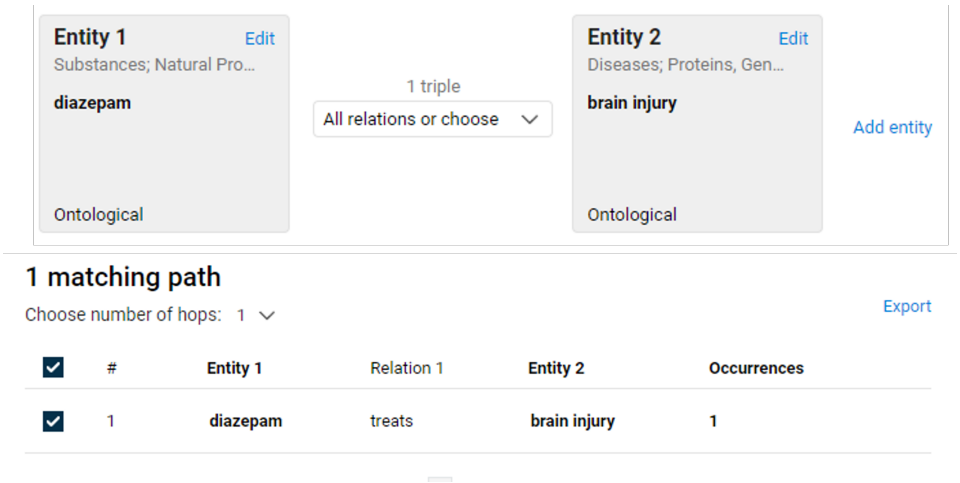

Figure 4.6.: Illustration of successful relation search for "diazepam" and "brain injury"

Example 1

Question: Can Diazepam be beneficial in the treatment of traumatic brain injury?

Context: The present experiment examined the effects of diazepam, a positive modulator at the GABA(A) receptor, on survival and cognitive performance in traumatically brain-injured animals.

Predictions: BioLinkBERT+OntoType20Rel: yes BioLinkBERT: no True Label: yes

Results: While BioLinkBERT-base without knowledge-enhancement answered the question incorrectly, our BioLinkBERT-base+OntoType20Rel model gave the correct answer. Diazepam (first marketed as Valium) is listed as an entity in the OntoChem KG (see Figure 4.6 where it has a direct relation to brain injuries ("diazepam [substance] treats [disease] brain injury"). It is likely that, thanks to the injection of this knowledge, the enhanced model was able to deduce the answer to the question, while the base model was not.

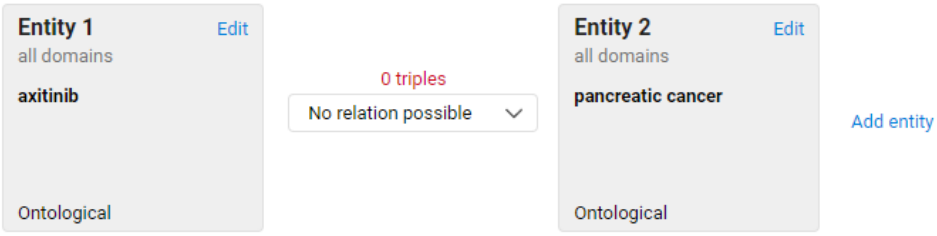

Figure 4.7.: Illustration of failed relation search for "axitinib" and "pancreatic cancer"

Example 2

Question: Does axitinib prolong the survival of pancreatic cancer patients?

Context: Axitinib/gemcitabine, while tolerated, did not provide survival benefit over gemcitabine alone in patients with advanced pancreatic cancer from Japan or other regions [...].

Predictions: BioLinkBERT+OntoType20Rel: no BioLinkBERT: yes True Label: no

Results: Here, the base version of BioLinkBERT incorrectly predicted that axitinib does prolong the survival of pancreatic cancer patients, while our BioLinkBERT-base+OntoType20Rel model gave the correct negative answer. This time, there is no relation between axitinib and any form of cancer listed in the OntoChem KG (see Figure 4.7. Therefore, our enhanced model might have been able to rely on its injected knowledge and deduce that there are no such connections between the entities in the question.

4.2.3. Limitations and possible deficiencies

MoP (Lai, Zhai, & Ji, 2023; Q. Lu, Dou, & Nguyen, 2021; Meng, Liu, Clark, et al., 2021) reported that the performance gains were due to the models learning factual knowledge and gaining knowledge awareness. However, the currently best overall performing models, like BioGPT, do not employ specific knowledge enhancement technology. Assuming that structured knowledge is necessary to solve BLURB tasks, and larger models did not explicitly learn from KGs in a structured way, this could mean that either the assumption is false

or massive models like BioGPT can learn factual knowledge awareness through training on unstructured text. This would mean that KELMs are not providing something entirely different for LM training and might only perform better thanks to the additional parameters and training data.

To investigate the influence and relevance of each component of an adapter-based approach, the authors of DAKI (Q. Lu, Dou, & Nguyen, 2021) performed an ablation study where they successively removed the adapter components from their work while keeping their knowledge controller. Essentially, "the results of the ablated versions demonstrate varying degrees of performance drop, indicating the necessity of each component" (Q. Lu, Dou, & Nguyen, 2021). Lai, Zhai, and Ji (2023) also provide an ablation study coming to similar results. While this removes some doubts, further investigations regarding the added parameter count are necessary. We will leave this to future research.

4.2.4. Discussion

Finally, we can answer research question 2: Can we improve existing approaches with new methods and data from a private ontology? The systematic literature review has shown strong results from various approaches to adapter-based knowledge enhancement in the biomedical domain across existing literature. While the SOTA on tasks where very large language models, like BioGPT, compete, was out of reach for low-resource settings like ours, our approach utilizing data from OntoChem was able to stand its own against other adapter-based parameter efficient works like MoP, KEBLM or DAKI.

In summary, we can answer the research question with a cautious yes: In the case of BioAQS7b, we were able to improve existing approaches with data from OntoChem as a private Ontology. In several other instances, our method approximately matched the performance of others. This finding diversifies the pool of viable sources for KGs for future research. It makes researchers less reliant on UMLS (Bodenreider, 2004), especially while the SciWalker platform and its FactFinder tool remain free of charge. We also solidify adapter-based knowledge enhancement as a viable and performant tool for NLP with our experiments in question answering (PubMedQA, BioASQ7b), semantic classification (HoC) and, NLI (MedNLI) in closed domain settings.

However, some questions remain open: Why did our approach not work as well for PubMedBERT? How would the results look with the full KG from OntoChem that might be better connected? A possible explanation to explore for the first question could be that the injected knowledge from OntoChem resonates better with the information encapsulated in the linked pre-training data from BioLinkBERT than with the unlinked data used by PubMedQA. To answer the second question, we plan on running new experiments in the future with the full KGs from OntoChem.

4.3. Survey of Medical Professionals and Students

In the following, we will present the results of the research survey. We follow Scheetz, Rothschild, McGuinness, et al. (2021) in presentation style and compare our findings to theirs where possible and appropriate. We will discuss the survey questions within their respective thematic groups. If there are no significant insights to be gained from a question, we will skip it to keep the results concise with a high informative value. However, all survey results are included in Appendix C. The survey is meant as a preliminary study and will be expanded on in future research at the TUM "Klinikum rechts der Isar" (MRI).

4.3.1. Participant Background

The survey achieved a total of 34 participants over two months, 33 of whom submitted usable questionnaires. Almost all clinicians were reached through personal contact within Munich and the surrounding countryside. Visiting medical centers and reaching out to other researchers in the biomedical domain proved less fruitful, although exact numbers can't be determined due to the anonymous nature of the survey.

As already mentioned in chapter 3, we decided to open the survey to medical students and medical professionals in a larger sense. The occupation distribution looks as follows:

- "Ongoing studies in the field of (bio)medicine": 25 participants
- "Working as a doctor, (bio)medical specialist, teaching or research staff": 5 participants
- "Ongoing training in (bio)medicine" (including residency): 2 participants
- "Working as a pharmacist": 1 participant

Therefore, we have a total count of 25 students and eight medical professionals. Figure 4.8 shows the (desired) location of practice of all participants. Similar to the study by Scheetz, Rothschild, McGuinness, et al. (2021), the metropolitan area was predominant.

4.3.2. Current Knowledge and Use of NLP

Exactly a third of the participants (n=11; 33.3%) considered their knowledge of LLMs to be average compared to their colleagues. Overall, only a small number believed they had an excellent (n=5, 15.2%) or very poor (n=4; 12.1%) understanding (See figure 4.9). Interestingly, almost 30% (n=10) reported that they feel they have a worse understanding of the technology than their colleagues, suggesting that many participants could feel left behind with the rapid developments around them. Responses were similar across students and professionals.

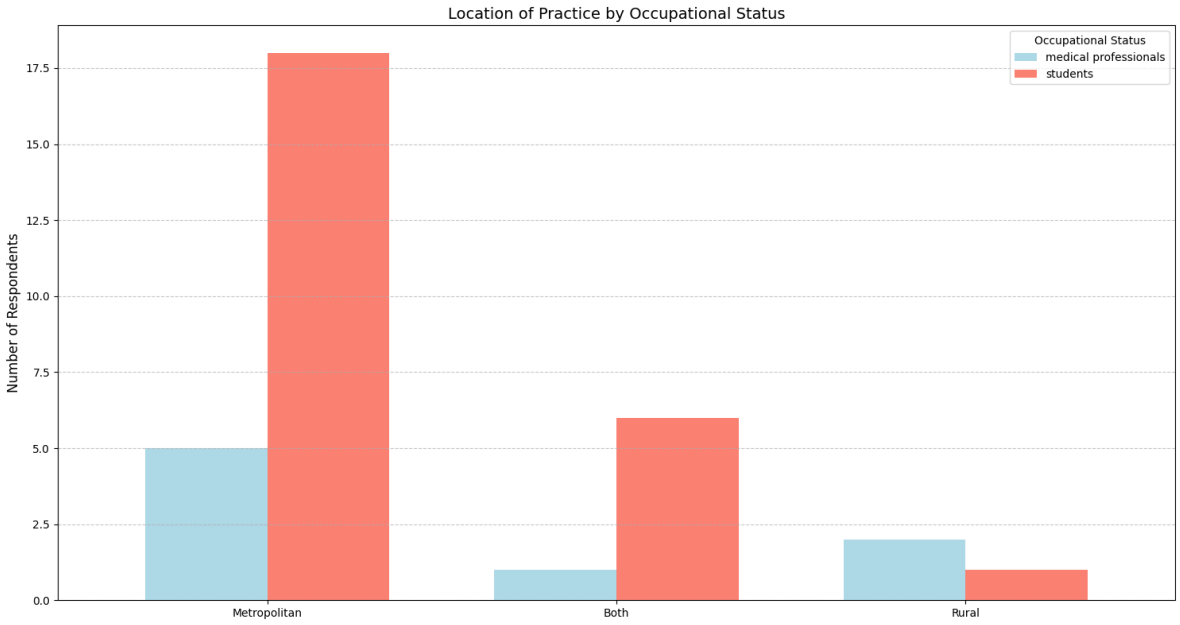

Figure 4.8.: "What is your location of practice?"

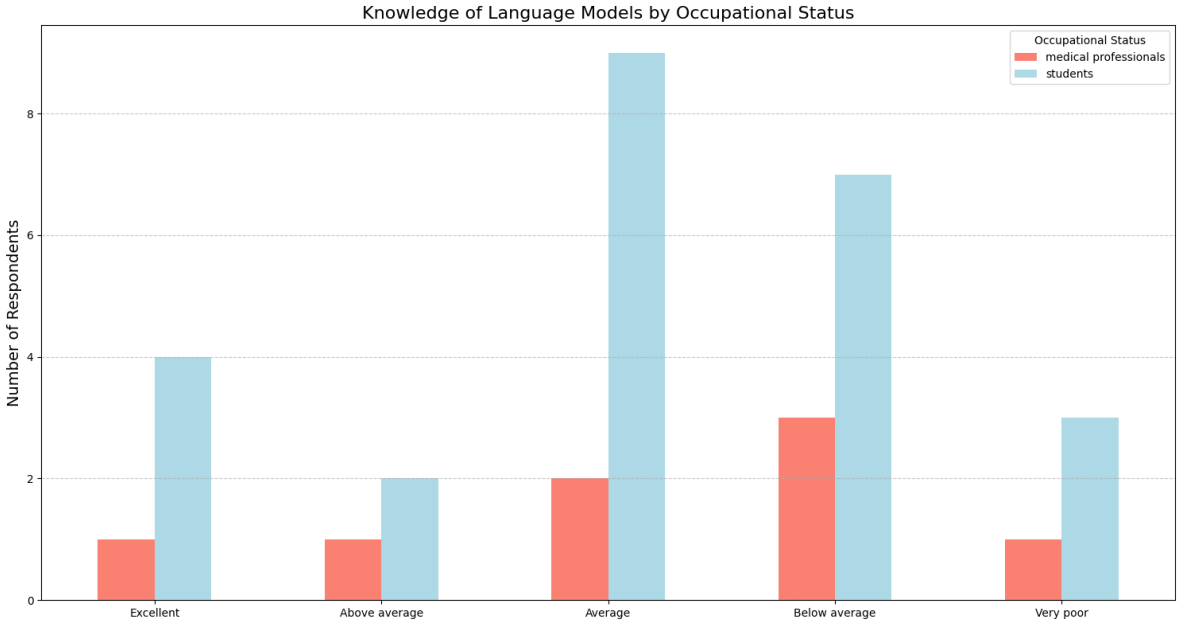

Figure 4.9.: "Relative to your colleagues, how would you rate your knowledge of language models and their application in your field of expertise?"

While almost half of the participants (n=16; 48.5%) mentioned they hadn't utilized LLM tools in their work or for their studies, over a third (n=12; 36.4%) reported a weekly or daily use (see Figure 4.10. Here, we see a significant shift from the 80.9% of participants reporting

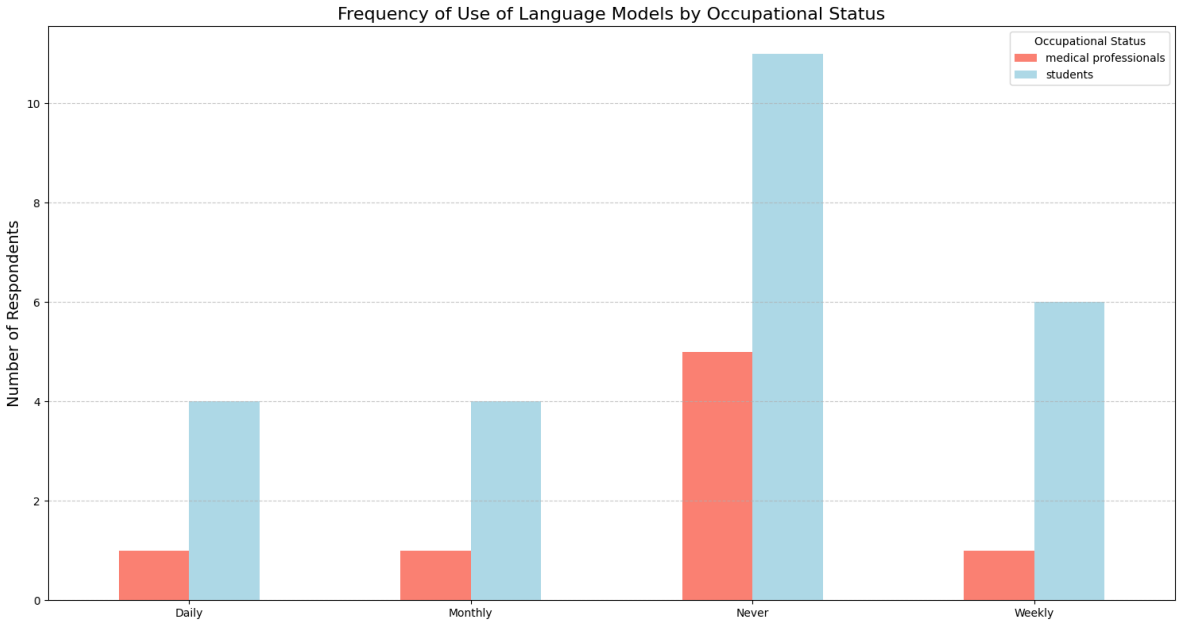

Figure 4.10.: "How often do you use software based on language models for your work/research?"

"Never" in the study by Scheetz, Rothschild, McGuinness, et al. (2021). Even though their study only dates back about two years, this could indicate that today, the medical community is already much more familiar with AI technology. Another interpretation could be that LLMs and NLP have a broader field of applications in the medical field compared to other AI technologies. The following personal use cases of LLMs were given by the participants (exemplary excerpts):

- Anamnesis, transcriptions
- Documentation courses of disease
- E-Mail drafting
- Explanation of intuition of different topics, summarization
- Quick question answering
- Help with coding and translations

4.3.3. Perceived Influence

Figure 4.11 shows that a relative majority of participants anticipated that the integration of LLMs would improve their (preferred) professional domain (n=14, 42.04% agreed or strongly agreed). This number is much smaller than the 71% of respondents in the study by Scheetz, Rothschild, McGuinness, et al., indicating more significant concerns regarding NLP than regarding other AI technologies (or increased general concerns over the past 2.5 years). Notably, students in our survey tend to only "Agree" with the statement significantly more

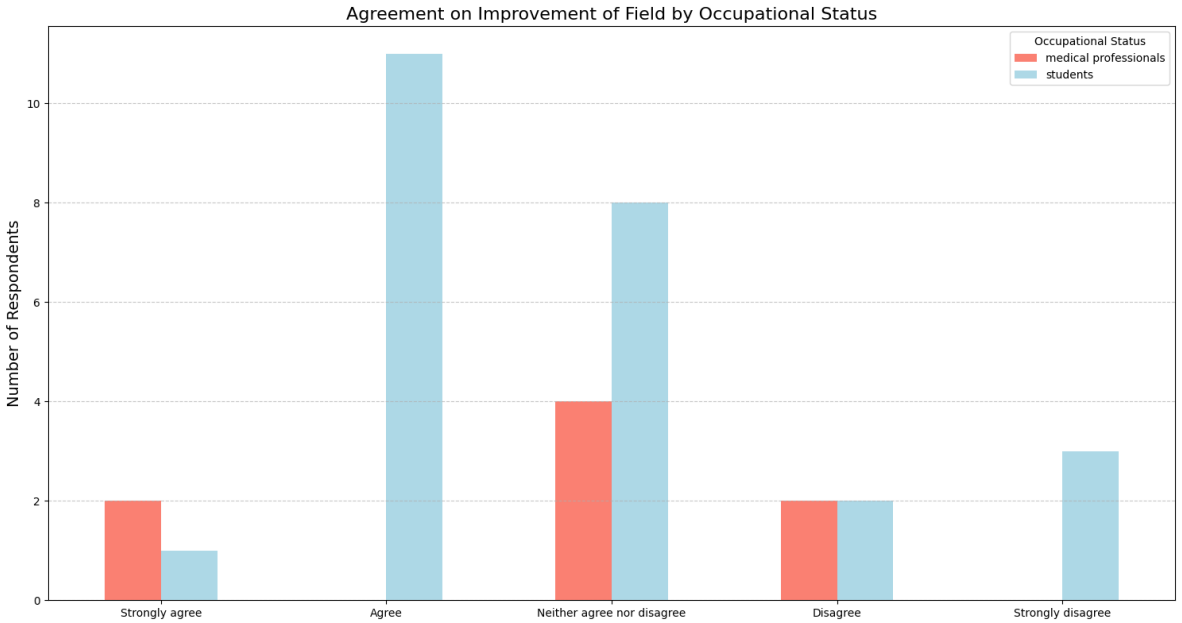

Figure 4.11.: "To what extent do you agree with the following statement: 'the field of [insert specialty] will improve with the introduction of NLP'?"

often than medical professionals, who tend to "Strongly Agree". This might be due to a more confident, experience-based stance of professionals.

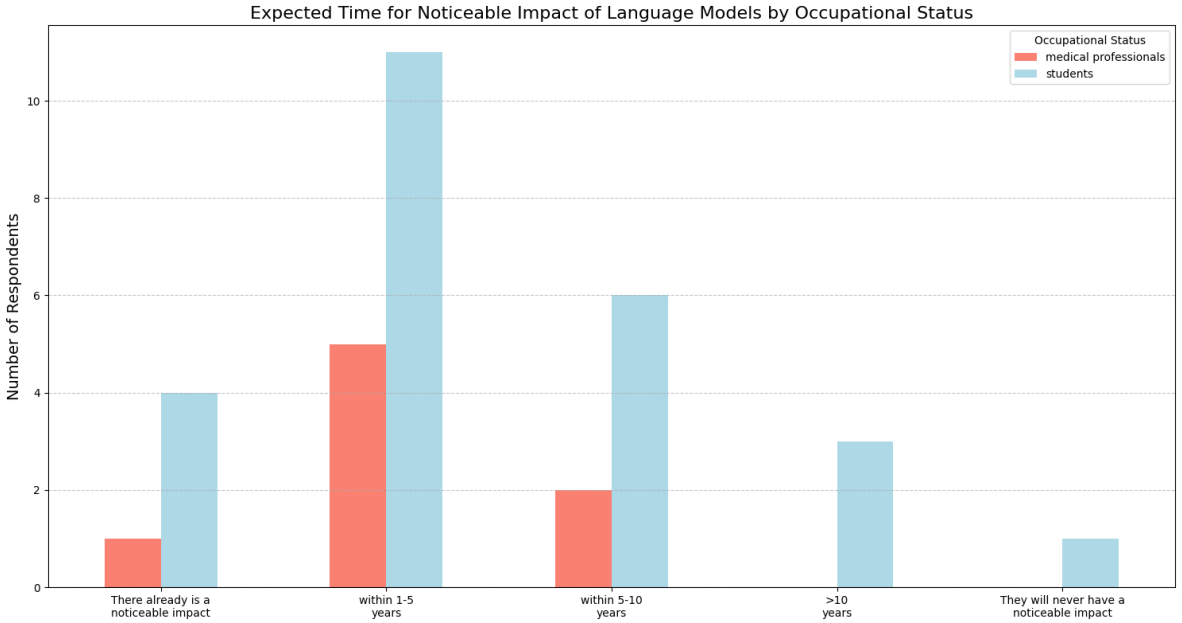

Figure 4.12.: "How long do you think it will be before language models have a noticeable impact on your field expertise? "

Close to half (n=16; 48.5%) estimated that within a span of 5 years or less, LLMs would make a discernible difference in their field (refer to 4.12). Only one person believed that the technology would perpetually remain without significant influence on their (future) profession (similar to the AI survey). Interestingly, several participants who already utilized LLMs (monthly, weekly, or daily) reported that there already is a noticeable impact today (n=5 versus n=0 for participants who have never used LLMs).

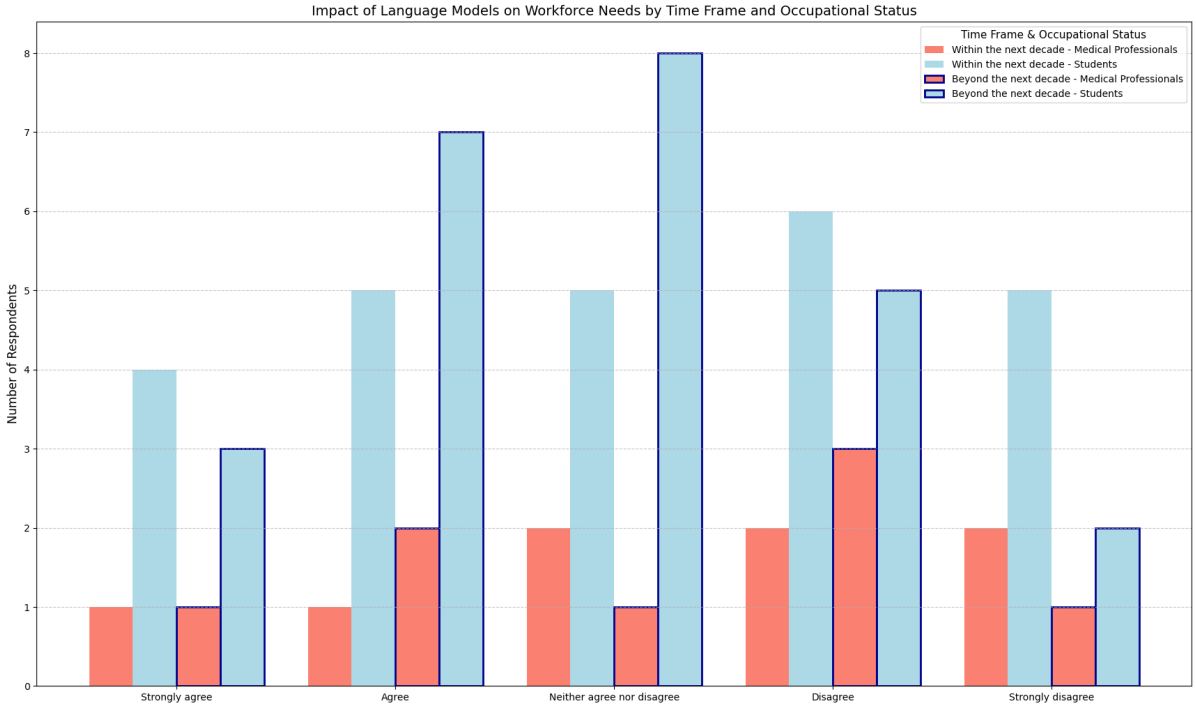

Figure 4.13.: "To what extent will language models have an impact on workforce needs in your area of expertise [within/beyond] the next decade?"

Figure 4.13 focuses on the participant's perception regarding changes in workforce needs. Here, we report strictly different results from the work of Scheetz, Rothschild, McGuinness, et al. (2021), who saw a majority of clinicians (71.1%) expecting a strong impact on workforce needs in the upcoming decade. In contrast, we saw students and professionals (n=15; 48.5%) disagreeing or strongly disagreeing with this statement when considering the impact of LLMs. However, the numbers shift when looking beyond the next decade (only n=9 or 33% still disagree or strongly disagree). Overall, students expect the workforce needs to change to a more considerable extent than professionals. Since the respondents anticipate a decrease in workforce needs (n=16; 48.5%) much more strongly than an increase (n=5; 15%), we have an additional indication that students, in particular, might be slightly concerned about their future job security.

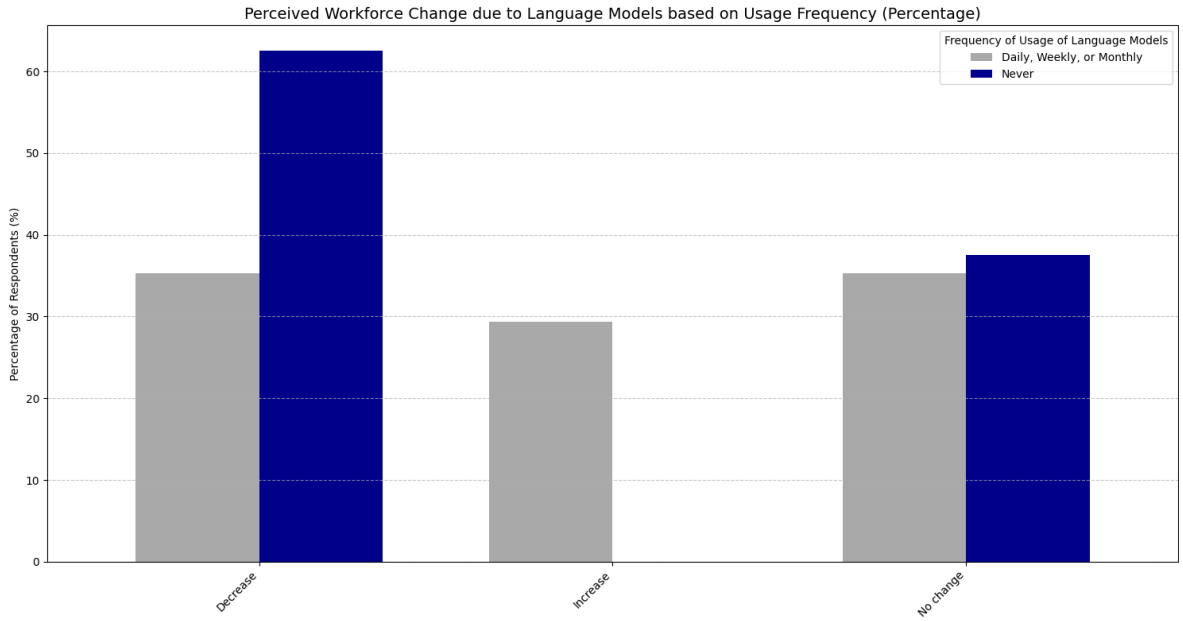

Figure 4.14.: "To what extent will language models have an impact on workforce needs in your area of expertise [within/beyond] the next decade?"

A notable fact is that participants who have already used LLMs themselves more often expect an increase in workforce needs (see Figure 4.14). This finding agrees with the results of Scheetz, Rothschild, McGuinness, et al. and, therefore, seems independent of the underlying AI technology.

4.3.4. Practical Implications

When asked about the necessary performance levels of LLMs when used by unskilled workers for disease detection and medical advice, a surprisingly large amount of participants (n=17; 42.4%; Figure 4.15) deemed an average specialist's performance (or even worse) as enough. On the other hand, two-thirds of survey participants (n=22; 66.7%; Figure 4.16) believed that LLM systems used by skilled medical specialists for diagnostic decision-support should outperform the average specialist. We expected the results to be the other way around since unskilled workers might be unable to detect mistakes made by bad LLMs. However, the participants could have thought that in areas without access to good health care, any advice would be better than none (a precarious stance). However, medical professionals held LLMs to higher standards compared to students, expecting a much higher performance.

Approvement of Example Workflow The following workflow procedure was suggested: "During a pandemic, a specialist responds to patient questions online. To save time, they generate responses with a language model and then only review them before sending". A relative majority of the participants (n=14; 42.4%) were open to adopting this method, with

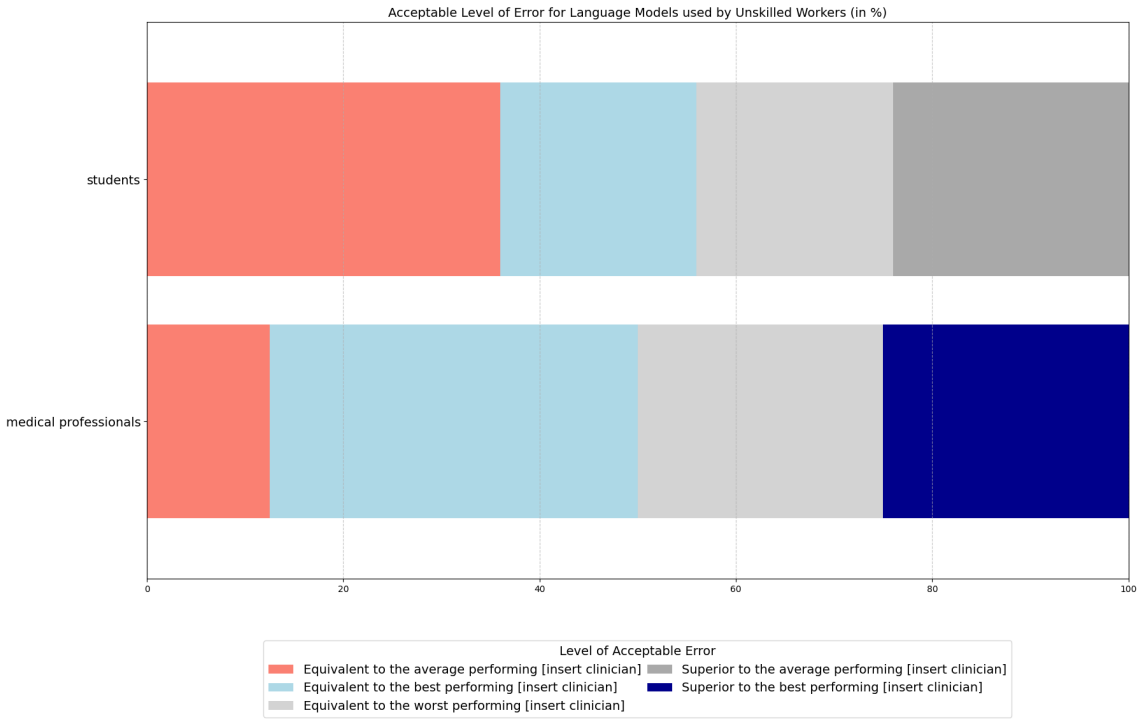

Figure 4.15.: "Language models could be used in the future to detect diseases and give medical advice. If such a language model were to be used by unskilled health workers in your field of expertise, what level of error in the language model's estimates do you think would be acceptable?"

students being almost double as likely to be in favor compared to professionals. We tailored this question specifically to our own model experiments, where we have a focus on models that perform question answering (PubMedQA, BioASQ7b). Therefore, these results validate the relevance of our approach and choice of tasks.

4.3.5. Chances and Risks

Questions 16 and 17 asked the participants to rank potential advantages and concerns of using LLMs in their field. Figures 4.17 and 4.18) show radar plots (following Scheetz, Rothschild, McGuinness, et al. (2021) that visualize the responses. Due to the complexity of the results, we will examine the results in several steps.

Greatest Perceived Advantages and Concerns Students and professionals agree that the most recognized advantage of language model systems in the medical field is the potential to "Reduce time spent by specialists on monotonous tasks" (n=16 "top 1" responses). This implies that medical professionals are eager to streamline repetitive tasks and recognize the potential of language models in assisting with or automating such tasks. There is also strong agreement on the predominant concern, which is data security and privacy (n=18, with n=12

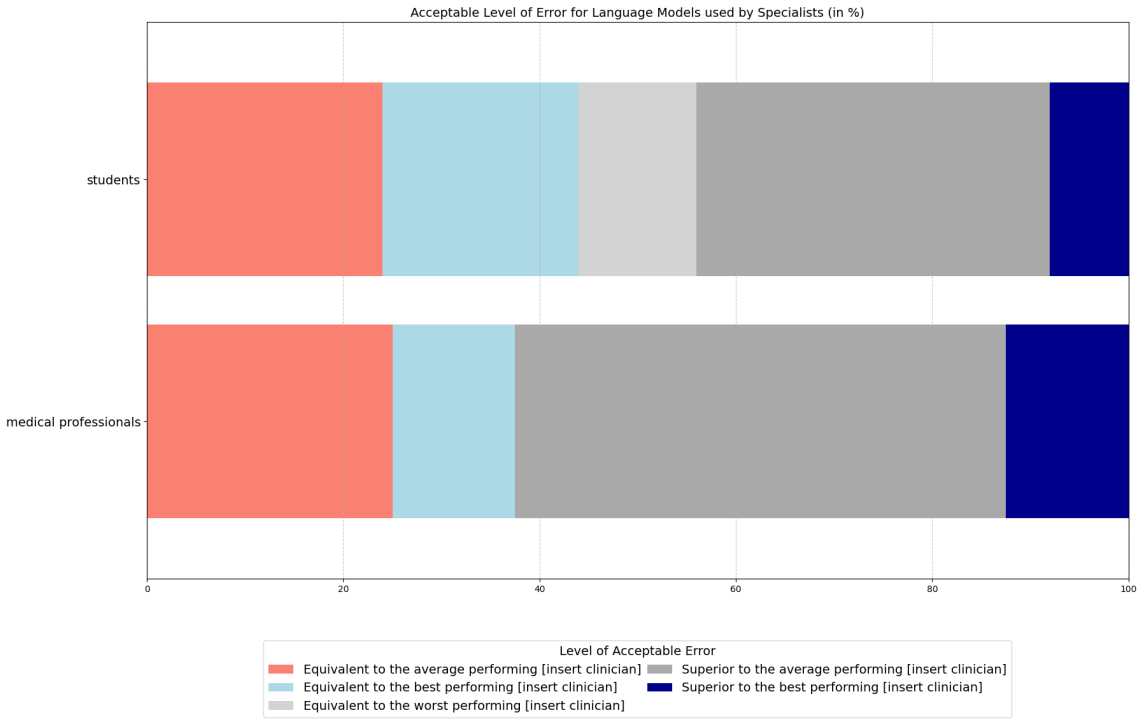

Figure 4.16.: "What level of error would be acceptable for language models that are used by specialists for diagnostic decision-support in your field?"

reporting the issue as their major concern). This highlights the sensibility of data in the medical field, where harsh regulations apply, and the privacy of patients is at stake.

Other Notable Advantages and Concerns "Improved patient access to disease screening" was recognized as a major advantage (n=15 total responses), suggesting the potential role of language models in enhancing healthcare access and screening mechanisms. "More cost-effective health care" was also seen as an important advantage, indicating the perceived financial benefits of integrating language models in healthcare. The trust-based relationship between doctors and patients stands out as an additional significant concern, reflecting the medical community's irreplaceability of such relationships by machines. A notable difference between students and professionals is that the latter see more potential in improving diagnostic confidence with LLMs. On the other hand, students have more concerns about being benchmarked against machines, likely indicating a slight concern for a future secure and stable profession due to higher expectations and possible displacement.

Least Emphasized Advantages and Concerns The relatively lower emphasis on "Improved prediction of disease outcomes" might indicate some skepticism or uncertainty about the predictive capabilities of language models in intricate medical scenarios. Concerns about the "impact on workforce needs" were hardly present, especially for professionals, suggesting

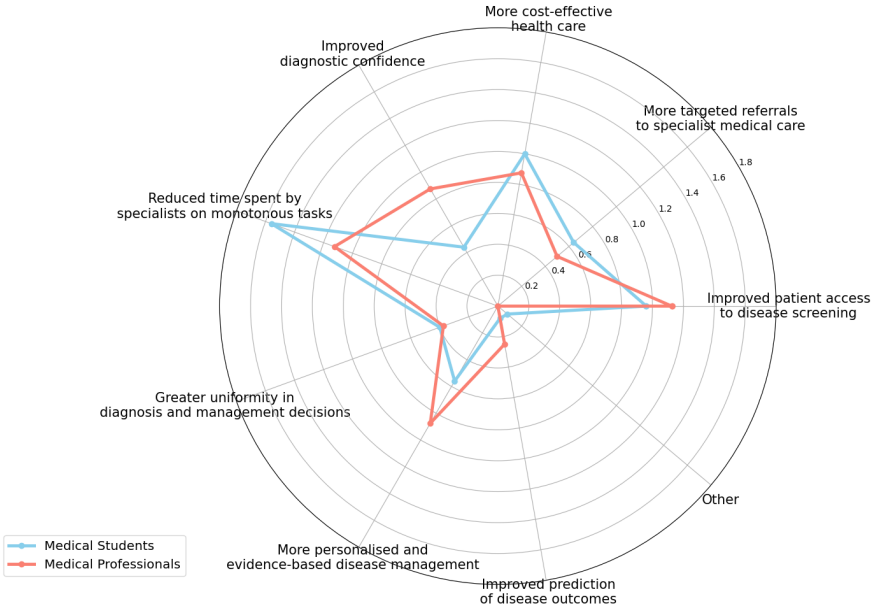

Figure 4.17.: "Which of the following do you perceive as the greatest potential advantage of the use of language model systems in your field?" Participants indicated their top three preferences from a list of set choices. Plot axes represent the average rank for students (blue) and medical professionals (salmon). Higher scores indicate a higher ranking.

confidence in their irreplaceable role as human professionals in the medical domain and perhaps a belief that technology would supplement, rather than replace, human expertise.

Comparison with AI Survey Overall, we see a lot of parallels between our findings and those of Scheetz, Rothschild, McGuinness, et al. (2021). For both NLP and AI in general, the medical community seems to agree that reducing the time spent on monotonous tasks (especially emphasized by radiologists in the related study) and improving patient access to disease screening (especially endorsed by ophthalmologists and dermatologists) are significant advantages that the technologies could bring to the medical domain. An interesting finding is that in the AI survey, medical professionals are much more concerned over the divestment of health care to large technology and data companies. Scheetz, Rothschild, McGuinness, et al. (2021) refer to a general mistrust in big tech companies and an increase in the trading of health data (P. Hunter, 2016). One could speculate that there is more trust in the German law system to prevent such things from happening in healthcare in Germany.

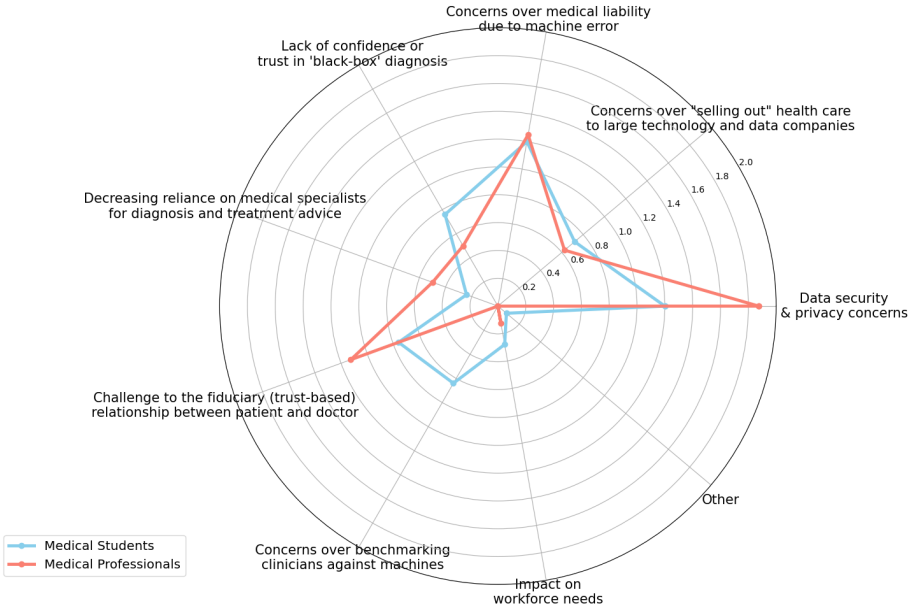

Figure 4.18.: "Which of the following do you perceive as concerns to the utilization of language models in your field?" Participants indicated their top three concerns from a list of set choices. Again, plot axes represent the average rank for students (blue) and medical professionals (salmon). Higher scores indicate a higher ranking.

4.3.6. Preparedness of Medical Centres and Universities

A relative majority of participants (n=15; 45.5% believe that their university/clinic/medical center is not adequately equipped to deal with the introduction of LLMs (versus n=10; 30.3% who believe they are adequately equipped). However, these results are much more positive than that of the AI survey, where only 13.8% felt like their institution was prepared. The results are quite similar amongst both students and professionals (see Figure 4.19. We asked the participants to make suggestions on what has to change for better preparedness and received a large amount of (sometimes very detailed) responses. We factor this as an indication that the topic is important to the participants and they would like to be involved in the process. We report reoccurring themes here:

- Addressing the general lack of digitalization in Germany
- Give university courses and seminars for students and training for professionals to increase technology awareness and proficiency
- Employ NLP experts and more IT personnel at institutions
- Support and encourage technology openness

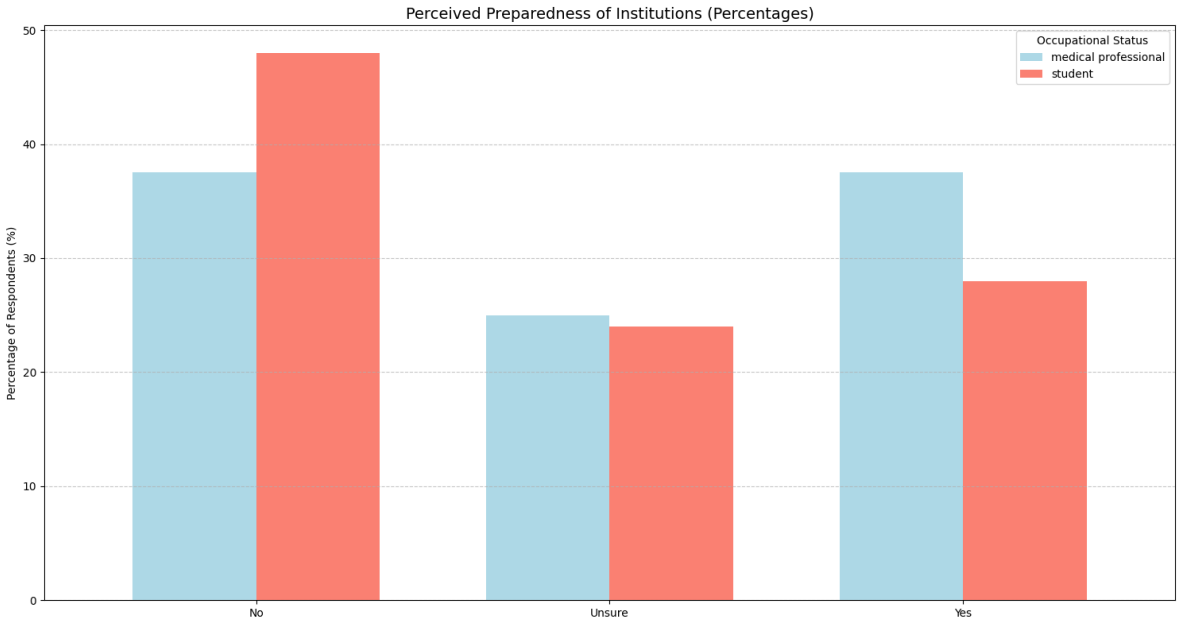

Figure 4.19.: "Do you think that your university/clinic/medical centre is adequately equipped to deal with the introduction of language models in your area of expertise?"

Some answers were particularly elaborate. We will share them in the following (with some edits for readability; originals can be found in Appendix C):

> It would help if part of our didactic training covered new technology, such as language models, and how they can be beneficial in practice. This could be through any form of education that can be regularly updated and revised to match the pace of quickly updating language models. (psychiatrist in training/residency)

> For me (a biomedical engineer), training needs to be more rigorous so that those using the technology can detect potential issues. I do think that a free chatbot replacing a $400 doctor's visit to tell you that you need bed rest and plenty of water is viable, but I think that it will take a decade to become mainstream. (student, specializing in brain-computer interfaces)

> More information about possible use cases should be provided. Also, experts should be hired who, in collaboration with doctors, can devise customized strategies or concepts tailored to individual needs." (pathologist, 10-20 years of experience, translated from German)

4.3.7. Discussion

Our survey on the use of LLMs in biomedicine revealed several key insights and now allows us to give a preliminary answer to RQ 3, "Is the research on biomedical KELMs relevant to

medical professionals, and what factors hinder or support the deployment of the technology in practice?":

The results show that the medical community is gaining familiarity with NLP and LLMs, with many already incorporating the technology into their weekly or daily tasks, such as documentation, transcribing, and information gathering. This growing adoption signals openness to the potential of LLMs to streamline monotonous tasks and enhance patient access to healthcare. However, alongside this positivity, there are prevalent concerns. Data security and the impact on the doctor-patient trust relationship stand out as significant apprehensions. This dual perspective underscores the cautious optimism of the medical community towards the evolving role of language models in healthcare. While students and professionals both see the benefits of LLMs, they differ in their expectations; professionals anticipate higher performance standards, while students express concerns about future job security and being benchmarked against machines. Regarding institutional preparedness for LLM integration, many feel their institutions lag behind. Suggestions to bridge this gap include higher investments in digitalization, offering specialized training, and fostering a culture of technological openness.

In essence, the results of the survey show that the research on biomedical LLMs matters and is indeed relevant. The medical community recognizes the potential advantages of LLMs, ranging from improving efficiency to better patient care. However, addressing concerns about data security, trust, and institutional readiness will be essential for their broader acceptance and successful integration. We are confident that the results of this survey, together with the results of the upcoming survey at MRI, can guide where future research funding and efforts should be directed to ensure that biomedical NLP evolves in a manner most beneficial to the healthcare system.

5. Conclusion

It has been a challenging endeavor to understand the nuances of adapter-based knowledge-enhancement in BioNLP and extract the numerous insights from the literature, model experiments, and research survey. Nevertheless, we addressed these challenges thoroughly and in-depth in our work and succeeded in answering the three RQs. This chapter first summarizes our results while highlighting our contributions to the field and then proceeds to address shortcomings and possible future research opportunities.

5.1. Thesis summary

Systematic Literature Review Delving into the systematic literature review, it became clear that adapter-based knowledge-enhancement is an emerging and fast-growing domain with its flexibility and potential recognized across various domains and tasks. The majority of the 12 unique open-domain approaches represented the interest in creating LLMs with a common-sense understanding or "world knowledge". At the same time, the distinction of the biomedical domain as the most frequently explored closed-domain area signified the paramount importance of precise knowledge representation in this intricate field.

In essence, we found that adapter-based knowledge-enhancement represents a unique and flexible way to integrate vast, diverse knowledge into LLMs without expensive fine-tuning. This fact was backed up by the qualitative analysis of the examined papers, where we discovered ingenious and unique solutions for the integration of linguistic as well as factual knowledge sources. With our review, we contribute a novel and extensive resource for other researchers to refer to in the future.

Model Experiments From our extensive model experiments, it became evident that our methods can compete with and even improve similar approaches in the field. With an accuracy of 92.14, the BioLinkBERT-base model injected with the OntoType20Rel KG was able to slightly outperform all other related works on the BioASQ7b task (an increase of 0.7). On other tasks, our best models stayed in the middle field but often came in second and bested the base models. While we found that the SOTA of models like BioGPT cannot be reached with BERT-based models anymore, the true value of our adapter-based method lies in its resource efficiency. Our models offer a valid opportunity to research groups and institutions that are constrained by limited computing resources. Adapter-based KELMs such as ours enable them to leverage advanced LLMs without the burden of exceedingly expensive training.

This fact is particularly important for medical professionals who may not have the means to customize expensive LLMs to cater to their individual needs swiftly. Another important finding to the medical domain stems from our qualitative probing, where we witnessed by what means the injection of knowledge (in our case, from OntoChem's KGs) can improve factual accuracy. Moreover, we believe that our findings and those of related works can still have value for the current SOTA of BioNLP. New approaches that work for smaller LLMs could also work for larger ones. Therefore, we encourage research teams with more resources to incorporate adapters where they find it applicable.

Research Survey Our survey targeting medical professionals and students shed light on the community's perception of LLMs and BioNLP. The survey showed a growing adoption of NLP technology for various tasks, from documentation and anamnesis to information sourcing and translations. This finding paints a picture of a community that is receptive to change. The most anticipated and recognized advantage of LLMs in the medical field was the potential to reduce the time clinicians have to spend on monotonous tasks. However, this enthusiasm is tinged with concerns, primarily surrounding data security and the sanctity of the fiduciary doctor-patient relationship. These sentiments underscore the balance that needs to be struck between innovation and trust as we venture further into the realm of BioNLP in healthcare. Overall, the survey gives a promising outlook for the future since most participants anticipate that integrating LLMs will improve their professional domain.

5.2. Shortcomings

Unfortunately, our research did not come without certain challenges and limitations. In the following, we list the shortcomings of our work:

Incomplete KGs A portion of the data provided to us by OntoChem was not usable due to incomplete ID mappings. As a result, only a fraction of the available knowledge was integrated into the experimental segment of this thesis, which has likely led to less thoroughly connected KGs. Moreover, the experiment results lack an investigation into the changes in recall and precision that come with the injection of knowledge. While it would be hard to compare such insights to other works since these metrics are rarely reported, these results would still give a more thorough and accurate report of our experiments.

Ethical Concerns Medical students and professionals indicated many concerns regarding ethical questions and the development and use of LLMs in biomedicine. While our methodology and models will likely not be used in practice without further research and improvements, we did not specifically address the medical community's concerns in our work. There are no guarantees that our methods and models might be reproduced and used without our knowledge, which might lead to unsafe medical advice, misinformation, or data security and privacy breaches. We tried to improve the overall model performance and factual accuracy to reduce hallucinations, but there is no way to entirely eliminate the risk of wrong predictions

and other critical issues. Therefore, if readers should consider using our models not just for further research but in practice, we urge them to refrain from doing so.

Survey Response Rates Lastly, the low response rate from medical professionals for our survey did constrain the richness and reliability of our findings in the third part of the thesis: We mainly reached medical students and could not gain more insights from experienced professionals. The main hurdles were the stressful schedules of doctors and nurses and the general difficulty connected with reaching medical professionals on a larger scale. An additional significant shortcoming in this regard was that we could not conduct the survey at the TUM MRI before the end of this thesis, which could have increased our responses from medical professionals tenfold or more.

5.3. Further Research

During our work on this thesis, we have encountered multiple possible avenues of future research that we did not have time to explore ourselves. These opportunities are given in the following for others to explore.

OntoChem Potential The data provided by OntoChem included much more than entities and relations and would allow for additional experiments. For instance, every data triplet comes with the source sentence from which it was extracted. Drawing inspiration from works like K-Adapter (R. Wang, Tang, Duan, et al., 2020) and other papers discussed in the literature survey, linguistic knowledge could be extracted from these sentences (e.g., by dependency parsing). This linguistic knowledge could then be used in additional adapters to enhance our models further.

KG Merging Moreover, the idea of merging OntoChem Data with sub-graphs from MSI (Ruiz, Zitnik, & Leskovec, 2021), UMLS (Bodenreider, 2004), or PubChem(Kim, Chen, Cheng, et al., 2020) presents a promising direction. Similarly, a potential game-changer could be the training of medical professionals to curate the KGs and sub-graphs with OntoChems FactFinder and MoP's graph partitioning (Meng, Liu, Clark, et al., 2021). This way, the resulting KELMs would be tailored directly by those who use them.

Approach Variation An additional future research direction is exploring and incorporating methodologies of some of the other adapter-based works we discussed in the systematic literature review. It would, for example, be possible to examine the strategy of AdapterSoup (Chronopoulou, Peters, Fraser, & Dodge, 2023) and combine it with our methods. Moreover, the intriguing prospect of allowing adapters to pool their knowledge with hypernetworks, as suggested by Karimi Mahabadi, Ruder, Dehghani, and Henderson (2021), is worthy of exploration. Additionally, further investigations into possible improvements in model explainability through adapters would be interesting, for example, by routing and tracing what knowledge in which adapter is most active for specific tasks.

A. General Addenda

A.1. Appendix A: OntoChem Knowledge Extraction Process

For data source transparency, we will give an overview of OntoChems knowledge extraction process and the ideas behind it in this section of the appendix. All information stems directly from Irmer, Bobach, and Böhme (2019) and personal discussions with the authors Dr. Claudia Bobach, Dr. Matthias Irmer, and Dr. Felix Berthelmann, COO of OntoChem.

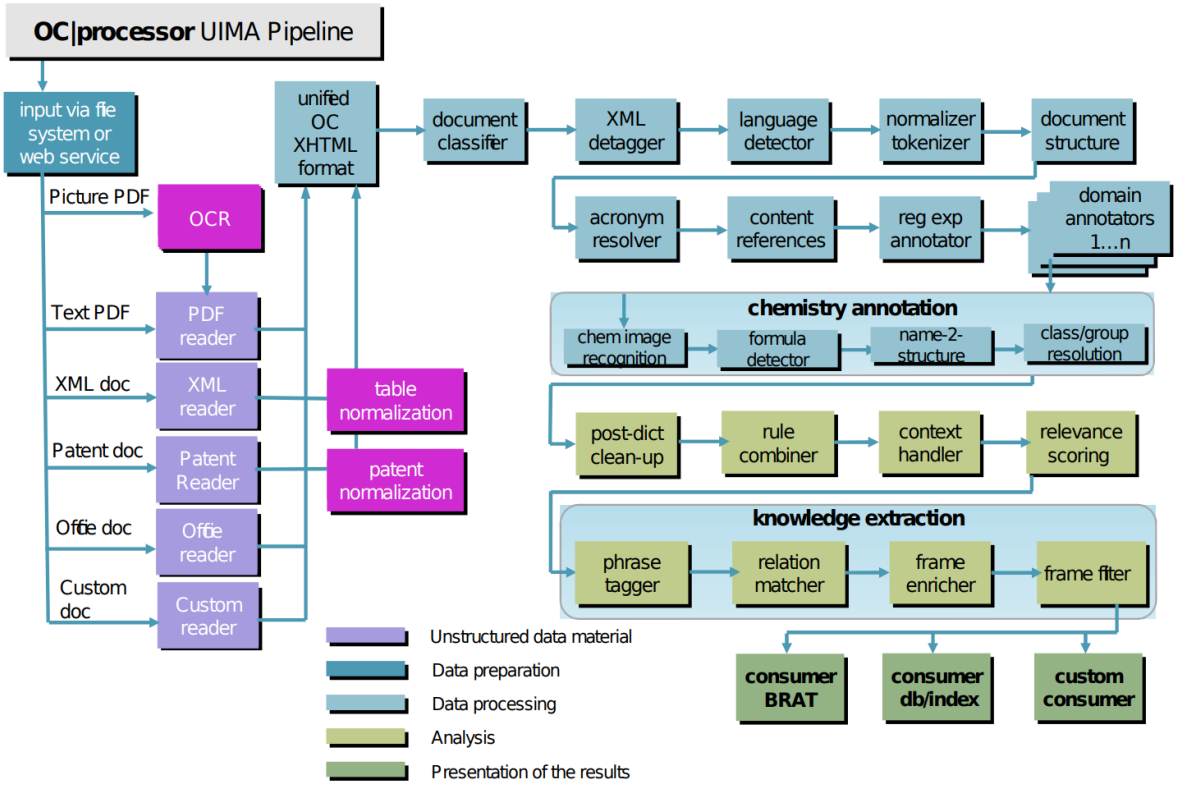

Figure A.1.: UIMA processing pipeline

OntoChem employs frame-based methods to extract knowledge from text documents effectively. A frame is a data structure encapsulating information about a specific object or concept. It is constructed through slots, which describe particular attributes of the frame. These slots can store attribute values or pointers to other frames. The utility of frames

extends beyond the representation of named entities: They provide a structured way to depict categories and individual objects using recursive attribute-value structures and act as a bridge between the textual entities and the complex facts they signify, facilitating efficient knowledge extraction. They can represent intricate details in the chemical realm, such as relations between chemical compounds and diseases or specific compound properties.

Regarding OntoChems extraction process, figure A.1 gives an excellent overview of the different stages and components. According to Dr. Claudia Bobach and Dr. Matthias Irmer, their pattern-based method leads to a very high precision compared to pure deep learning approaches.

A.2. Appendix B: Model Experiment Details

The run specifics that varied from experiment to experiment can be found in Tables A.1 and A.2. For all other parameters, we followed Meng, Liu, Clark, et al. (2021). Training times varied depending on the task (from approximately 3 hours for BioASQ7b to up to 20 hours for MedNLI). We used the following library versions and hardware:

- python_version: 3.8.18
- framework: huggingface, version: 1.1.1
- gpu: 1 Tesla V100-SXM2-16GB, 17179869184 memory
- os: "Linux-5.15.109+-x86_64-with-glibc2.35"

settings v task >	**HoC**	**PubMedQA**	**BioASQ7b**	**MedNLI**
repeat_runs	5	10	10	3
epochs	20	30	25	20
patience	3	4	5	3
batch_size	16	4	4	8
learning_rate	1e-5	0.5e-5	0.5e-5	0.5e-5
max_seq_len	128	512	512	256

Table A.1.: Changing hyperparameters of experiment runs

model v task >	HoC	PubMedQA	BioASQ7b	MedNLI
PubMedBERT-base +Onto20Rel	1695545510, 1695551194, 1695631414, 1695635952, 1695545284	1695302987, 1695303496, 1695305525, 1693487106, 1693486969, 1693486875, 1693486786, 1693486671, 1693486519, 1693486438	1695294922, 1695296457, 1695297480, 1695298979, 1695300121, 1695294957, 1695296728, 1695298346, 1695299850, 1695301350	1693438247, 1693348293, 1693158923
PubMedBERT-base+OntoType20Rel	1696003589, 1696010016, 1695978691, 1695983965, 1696003582	1695301624, 1695302090, 1695303096, 1695302987, 1695303496, 1695305525, 1695305786, 1695306529, 1695306846, 1695307756	1696281308, 1696282974, 1696284209, 1696285747, 1696287304, 1696288728, 1696290290, 1696292218, 1696293848, 1696295337	1696268190, 1696276332, 1696288411
BioLinkBERT-base +Onto20Rel	1695323423. 1695389154, 1695395947, 1695309647, 1695315963	1693512957, 1693514490, 1693516826, 1693518324, 1693519586, 1693501875, 1693503037, 1693504978, 1693506333, 1693507926	1695208870, 1695210546, 1695212437, 1695213810, 1695215559, 1695208205, 1695209378, 1695210768, 1695212272, 1695214143	1692534250, 1692534254, 1692534298
BioLinkBERT-base+OntoType20Rel	1695720138, 1695724845, 1695646163, 1695652457, 1692547237	1695068916, 1695070871, 1692544809, 1692545800, 1692546444, 1692546956, 1692547340, 1693609462, 1693649763, 1693655551	1694284328, 1694285491, 1694286633, 1694288370, 1694289991, 1694346616, 1694347940, 1694348954, 1694350207, 1694351700	1692535324, 1692535524, 1692535833

Table A.2.: Seeds of the runs used in the experiment section

A.3. Appendix C: Research Survey Documentation

In the following pages, the questionnaire and introductory document are provided. We make all responses to the survey available under https://github.com/alexander-fichtl/thesis_survey_results.

Survey on the Use of Language Models in Medicine and Biomedical Research

In recent years, there has been significant progress in the development of large language models (LLMs). In addition to well-known LLMs such as ChatGPT, there also exists an increasing number of specialized language models for use and research in medicine and biomedicine. Our research at the Technical University of Munich deals with the applications and relevance of these specialized models. Your insights help us assess the impact and importance of ongoing research.

This survey consists of 20 short questions and takes a maximum of 10 minutes to complete.

If you are unfamiliar with LLMs and their applications in the (bio)medical field, please read the brief introduction provided for you and then return to the survey.

If you have any questions or comments regarding the survey, please email us at: alexander.fichtl@tum.de

We appreciate your help and thank you for your time!

*Indicates required question

Informed Consent

All information given on this survey will be completely anonymous. All data is used exclusively for research purposes at the Technical University of Munich. If you wish to have your data deleted after you have already submitted the survey, please contact: alexander. chtl@tum.de.

1. Do you wish to participate? *

Mark only one oval.

- ◯ Yes *Skip to question 2*
- ◯ No *Skip to section 2 (Declined Participation)*

Declined Participation

You have declined to participate in this survey or do not meet the requirements for participation. Thank you for your time. You may close the browser or click submit below.

Occupation

2. What is your current occupational status? *

Mark only one oval.

- ◯ Working as a doctor, (bio)medical specialist, teaching or research staff
  Skip to question 3
- ◯ Ongoing studies in the eld of (bio)medicine
  Skip to section 4 (Information for students and trainees)
- ◯ Ongoing training in (bio)medicine
  Skip to section 4 (Information for students and trainees)
- ◯ None of the options apply *Skip to section 2 (Declined Participation)*

Information for students and trainees

You indicated that you are still studying or in training. In this survey, you will be asked questions about your area of expertise and your experiences there. When answering these questions, please simply refer to your desired area of expertise. Answer all other questions, if appropriate, based on your experiences and assessments of your studies or your training.

Questions 1-5

1. What is your (aspired) specialty? *

2. Years of practice in your field of expertise? *

Mark only one oval.

- ◯ Currently in training/Student
- ◯ <5 years
- ◯ 5-10 years
- ◯ 10-20 years
- ◯ 20-30 years
- ◯ >30 years

3. How often do you use software based on language models for your work/research?

Mark only one oval.

- ◯ Never
- ◯ Monthly
- ◯ Weekly
- ◯ Daily
- ◯ Unsure

4. What applications do you use language models for?

5. What is your location of practice? *

Mark only one oval.

- ◯ Rural
- ◯ Metropolitan
- ◯ Both

Questions 6- 10

6. Relative to your colleagues, how would you rate your knowledge of language models and its application in your field of expertise? *

Mark only one oval.

	1	2	3	4	5	
Excellent	◯	◯	◯	◯	◯	Very poor

7. How long do you think it will be before language models have a noticeable impact on your field expertise?

Mark only one oval.

- ◯ There already is a noticeable impact
- ◯ within 1-5 years
- ◯ within 5-10 years
- ◯ >10 years
- ◯ They will never have a noticeable impact

8. To what extent will language models have an impact on workforce needs in your area of expertise within the next decade? *

Mark only one oval.

	1	2	3	4	5	
To a great extent	◯	◯	◯	◯	◯	Not at all

9. To what extent will language models have an impact on workforce needs in your field of expertise? *

Mark only one oval.

	1	2	3	4	5	
To a great extent	◯	◯	◯	◯	◯	Not at all

10. In what direction will the workforce needs change due to language models? *

Mark only one oval.

- ◯ Increase
- ◯ Decrease
- ◯ No change

Questions 11-15

11. Do you think that your university/clinic/medical centre is adequately equipped to deal with the introduction of language models in your area of expertise? *

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Unsure

12. What do you think that your university/clinic/medical centre should do in preparation for the deployment of language models in your area of expertise?

13. Language models could be used in the future to detect diseases and give medical advice. If such a language model were to be used by **unskilled** health workers in your field of expertise, what level of error in the language model's estimates do you think would be acceptable? *

Mark only one oval.

- Equivalent to the worst performing [insert clinician]
- Equivalent to the average performing [insert clinician]
- Superior to the average performing [insert clinician]
- Equivalent to the best performing [insert clinician]
- Superior to the best performing [insert clinician]

14. What level of error would be acceptable for language models that are used by **specialists** for diagnostic decision-support in your field? *

Mark only one oval.

- Equivalent to the worst performing [insert clinician]
- Equivalent to the average performing [insert clinician]
- Superior to the average performing [insert clinician]
- Equivalent to the best performing [insert clinician]
- Superior to the best performing [insert clinician]

15. Would you consider using the following clinical workflow? During a pandemic, a specialist responds to patient questions online. To save time, they generate responses with a language model and then only review them before sending. *

Mark only one oval.

- Yes
- No
- Unsure

Questions 16-20

16. Which of the following do you perceive as the greatest potential advantage of the use of language models systems in your field? (only rank your top 3 preferences where 1=greatest advantage)

Answer options:

- Improved patient access to disease screening
- More targeted referrals to specialist medical care
- More cost-effective health care
- Improved diagnostic confidence
- Reduced time spent by specialists on monotonous tasks

- Greater uniformity in diagnosis and management decisions
- More personalised and evidence-based disease management
- Improved prediction of disease outcomes
- Other

17. Which of the following do you perceive as concerns to the utilisation of language models in your field? (only rank your top 3 preferences where 1=greatest drawback)

Answer options:

- Data security & privacy concerns
- Concerns over "selling out" health care to large technology and data companies
- Concerns over medical liability due to machine error
- Lack of confidence or trust in 'black-box' diagnosis
- Decreasing reliance on medical specialists for diagnosis and treatment advice
- Challenge to the fiduciary (trust-based) relationship between patient and doctor
- Concerns over benchmarking clinicians against machines
- Impact on workforce needs
- Other

18. Which professional group do you think will be most impacted by the introduction of NLP in your field of expertise? (e.g. nurses, doctors, other)

19. To what extent do you agree with the following statement: "the field of [insert specialty] will improve with the introduction of NLP"?*

Mark only one oval.

	1	2	3	4	5	
Strongly agree	◯	◯	◯	◯	◯	Strongly disagree

20. How do you think the majority of professionals in your field of expertise would answer the previous question compared to yourself? *

Mark only one oval.

- ◯ They would agree more strongly on average
- ◯ They would disagree more strongly on average
- ◯ They would give a similar response on average

Comments and Interview Participation

Would you be willing to take part in an interview on the subject of the survey? If yes, please provide your preferred contact details below. If you have any other thoughts, ideas or concerns you would like to share with us but do not wish to participate in an interview, you can also do so in the box below.

Survey Submission

Thank you for participating in this survey. Please click the submit button below to complete the survey.

Technical University of Munich (TUM)
School of Computation, Information and Technology (CIT)
Chair of Software Engineering for Business
Information Systems (sebis)
Website: www.matthes.in.tum.de
Contact: alexander.fichtl@tum.de

Biomedical Large Language Models

What are Language Models?

Language models are, in essence, word prediction models. They are trained on large amounts of unstructured text, usually from online sources, to solve specific tasks by statistically predicting the words in the solutions. These tasks can range from extracting desired information from a list of documents ("Information Extraction") to answering questions as a chatbot. Recently, these models have grown ever larger and cost more and more resources to train, hence the name "Large Language Models" (LLMs).

LLMs that are trained on biomedical texts, usually research papers and patient data, can be used to solve tasks in the biomedical domain. Research on these biomedical LLMs has increased over the last couple of years, and they are being used by more and more clinicians and researchers in practice.

What biomedical LLMs can do already

- Decipher and explain medical jargon in documents [Go22]
- Extract, structure, and learn from various sources, including clinical information from health records
- Answer publicly posed patient questions, even with more empathy and quality than physicians (under certain circumstances) [Ay23]
- Pass benchmarks like professional medical board exams[1]
- Be a reference and support tool for medical professionals

What biomedical LLMs cannot do

- Replace clinicians or medical professionals in general
- Give medical advice that does not have to be counter-checked
- Take responsibility

[1] Examples from the "MedQA" [Ji20] and "MedNLI" [Ro18] dataset are provided in the Appendix

Both lists given here list are, of course, not exhaustive. The fast progress in the field of Natural Language Processing will likely enable biomedical LLMs to solve new and more difficult tasks in the future. At the same time, new regulations, like the AI Act in the European Union, will limit the research on and applications of LLMs for safety and privacy protection.

Main takeaways

Language models are word prediction models that usually learn from collections of text documents. Some of them are specifically trained in closed domains, like the biomedical domain. Our research focuses on such biomedical language models. These models can assist medical professionals in their everyday tasks but are not intended to replace professionals.

If you want to learn more about our work, you can find our research questions here. We will also share our findings under the same link when our research is completed.

You can now return to the survey if you still have it open. Otherwise, you can click here to restart it.

APPENDIX

MedQA *[Ji20]* example task:

Question given to the Language model:

> An ECG is most likely to show which of the following findings in this patient?

Context (patient information) given to the language model:

- Age: 64 years
- Gender: F, self-identified
- Ethnicity: unspecified
- Site of Care: emergency department History
- Reason for Visit/Chief Concern: “My chest hurts, especially when I take a deep breath.”
- History of Present Illness: 2-hour history of chest pain pain described as “sharp” pain rated 6/10 at rest and 10/10 when taking a deep breath
- Past Medical History: rheumatoid arthritis major depressive disorder
- Medications: methotrexate, folic acid, fluoxetine
- Ect.

Options given to the Language Model:

- 'A': 'S waves in lead I, Q waves in lead III, and inverted T waves in lead III'

- 'B': 'Diffuse, concave ST-segment elevations'
- 'C': 'Sawtooth-appearance of P waves'
- 'D': 'Peaked T waves and ST-segment elevations in leads V1-V6'

State-of-the-art accuracy on MedQA and similar questions: **85.4%**

MedNLI *[Ro18]* example task: Predict whether two sentences (a premise and a hypothesis) entail or contradict each other or if they are of neutral status:

Examples from MedNLI development dataset:

1. **Premise:** ALT , AST , and lactate were elevated as noted above
   Hypothesis: patient has abnormal lfts
   Correct Label: entailment

2. **Premise:** Chest x-ray showed mild congestive heart failure
   Hypothesis: The patient complains of cough
   Correct label: neutral

3. **Premise:** During hospitalization, patient became progressively more dyspnic requiring BiPAP and then a NRB
   Hypothesis: The patient is on room air
   Correct label: contradiction

State-of-the-art accuracy on MedNLI and similar questions: **86.6%**

REFERENCES

[Mi23] Microsoft (2023)t: *BioGPT: generative pre-trained transformer for biomedical text generation and mining*
[Go22] Gordon, R (2022).: *Large language models help decipher clinical notes*
[Ay23] Ayers, J., Poliak, A., Dredze, M., et al. (2023). *Comparing Physician and Artificial Intelligence Chatbot Responses to Patient* Questions Posted to a Public Social Media Forum
[Ji20] Jin, D., Pan, E., Oufattole, N., Weng, W., Fang, H., & Szolovits, P. (2020). What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. ArXiv, abs/2009.13081.
[Ro18] Romanov, A., & Shivade, C.P. (2018). Lessons from Natural Language Inference in the Clinical Domain. Conference on Empirical Methods in Natural Language Processing.

List of Figures

List of Tables

Acronyms

BLURB BLURB. 6, 7, 16, 18–21, 26, 32, 34–36, 69

EE Event Extraction. 26

EL Entity Linking. 26

ES Extractive Summarization. 26

ET Entity Typing. 26

GLUE GLUE. 26

IE Information Extraction. 5

KELM Knowledge-Enhanced Language Model. iv, v, 1–3, 7–10, 13–16, 20, 21, 24, 28, 30, 33, 37, 48, 50, 52

KGD Knowledge-grounded Dialogue. 26

LAMA Concept-Net Split of LAMA Probe (Petroni, Rocktäschel, Lewis, et al., 2019). 26, 29

LLM Large Language Model. iv, 1, 2, 4–8, 10, 11, 14, 20, 22, 27, 29–33, 38–40, 42–45, 47–49

LM Language Modeling. 26

MT Machine Translation. 26

NER Named Entity Recognition. 5, 6, 26

NLI Natural Language Inference. 26, 37

QA Question Answering. 5, 6, 26

RC Reading Comprehension. 26

RCL Relation Classification. 26

RE Relation Extraction. 5, 6

SA Sentiment Analysis. 26

SC Sentiment Classification. 26

SF Speech Foundation. 26

SL Sequence Labelling. 26

SOTA State-of-the-art. 6, 7

SR Speech Recognition. 26

STC Sentence Classification. 26

TC Text Classification. 5

TOD Task-Oriented dialogue. 26

UMLS Unified Medical Language System. 2, 8, 31, 32, 37, 52

Bibliography

Almutiri, T., & Nadeem, F. (2022). Markov models applications in natural language processing: A survey. *International Journal of Information Technology and Computer Science (presumed, based on the DOI structure)*. https://doi.org/https://doi.org/10.5815/ijitcs.2022.02.01

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007a). Dbpedia: A nucleus for a web of open data. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, & P. Cudré-Mauroux (Eds.), *The semantic web* (pp. 722–735). Springer Berlin Heidelberg.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007b). Dbpedia: A nucleus for a web of open data. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, & P. Cudré-Mauroux (Eds.), *The semantic web* (pp. 722–735). Springer Berlin Heidelberg.

Ba, J., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *ArXiv*, *abs/1607.06450*. https://api.semanticscholar.org/CorpusID:8236317

Baker, S., Silins, I., Guo, Y., Ali, I., Högberg, J., Stenius, U., & Korhonen, A. (2015). Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, *32*(3), 432–440. https://doi.org/10.1093/bioinformatics/btv585

Bapna, A., & Firat, O. (2019). Simple, scalable adaptation for neural machine translation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1538–1548. https://doi.org/10.18653/v1/D19-1165

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620. https://doi.org/10.18653/v1/D19-1371

Bodenreider, O. (2004). The unified medical language system (umls): Integrating biomedical terminology. *Nucleic acids research*, *32*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308795/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T. J., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *ArXiv*, *abs/2005.14165*. https://api.semanticscholar.org/CorpusID:218971783

Chronopoulou, A., Peters, M., & Dodge, J. (2022). Efficient hierarchical domain adaptation for pretrained language models. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1336–1351. https://doi.org/10.18653/v1/2022.naacl-main.96

Chronopoulou, A., Peters, M., Fraser, A., & Dodge, J. (2023). AdapterSoup: Weight averaging to improve generalization of pretrained language models. *Findings of the Association for Computational Linguistics: EACL 2023*, 2054–2063. https://aclanthology.org/2023.findings-eacl.153

Cimini, G., Gabrielli, A., & Labini, F. (2014). The scientific competitiveness of nations. *PloS one*, *9*. https://doi.org/10.1371/journal.pone.0113470

Colon-Hernandez, P., Havasi, C., Alonso, J. B., Huggins, M., & Breazeal, C. (2021). Combining pre-trained language models and structured knowledge. *ArXiv*, *abs/2101.12294*. https://api.semanticscholar.org/CorpusID:231728366

Davis, R. (2021, November).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, *abs/1810.04805*. https://api.semanticscholar.org/CorpusID:52967399

Emelin, D., Bonadiman, D., Alqahtani, S., Zhang, Y., & Mansour, S. (2022). Injecting domain knowledge in language models for task-oriented dialogue systems. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11962–11974. https://doi.org/10.18653/v1/2022.emnlp-main.820

Gonz'alez-Carvajal, S., & Garrido-Merch'an, E. C. (2020). Comparing bert against traditional machine learning text classification. *ArXiv*, *abs/2005.13012*. https://api.semanticscholar.org/CorpusID:218900594

Goodwin, T., & Demner-Fushman, D. (2020). Enhancing question answering by injecting ontological knowledge through regularization. *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 56–63. https://doi.org/10.18653/v1/2020.deelio-1.7

Green, B. F., Wolf, A. K., Chomsky, C., & Laughery, K. (1961). Baseball: An automatic question-answerer. *Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference*, 219–224. https://doi.org/10.1145/1460690.1460714

Gu, Y., Tinn, R., Cheng, H., Lucas, M. R., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, *3*, 1–23. https://api.semanticscholar.org/CorpusID:220919723

Guo, Q., & Guo, Y. (2022). Lexicon enhanced chinese named entity recognition with pointer network. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-022-07287-1

Han, W., Pang, B., & Wu, Y. N. (2021). Robust transfer learning with pretrained language models through adapters. *ArXiv*, *abs/2108.02340*. https://api.semanticscholar.org/CorpusID:236460041

He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., & Neubig, G. (2021). Towards a unified view of parameter-efficient transfer learning. *ArXiv*, *abs/2110.04366*. https://api.semanticscholar.org/CorpusID:238583580

He, R., Liu, L., Ye, H., Tan, Q., Ding, B., Cheng, L., Low, J.-W., Bing, L., & Si, L. (2021). On the effectiveness of adapter-based tuning for pretrained language model adaptation.

He, Y., Zhu, Z., Zhang, Y., Chen, Q., & Caverlee, J. (2020). Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4604–4614. https://doi.org/10.18653/v1/2020.emnlp-main.372

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutiérrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2020). Knowledge graphs. *ACM Computing Surveys (CSUR)*, *54*, 1–37. https://api.semanticscholar.org/CorpusID:235716181

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for nlp. *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:59599816

Huang, X., Lin, J., & Demner-Fushman, D. (2006). Evaluation of pico as a knowledge representation for clinical questions. *AMIA Annu Symp Proc*, 359–363.

Hung, C.-C., Lauscher, A., Ponzetto, S., & Glavaš, G. (2022). DS-TOD: Efficient domain specialization for task-oriented dialog. *Findings of the Association for Computational Linguistics: ACL 2022*, 891–904. https://doi.org/10.18653/v1/2022.findings-acl.72

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Hunter, P. (2016). The big health data sale. *EMBO reports*, *17*(8), 1103–1105. https://doi.org/https://doi.org/10.15252/embr.201642917

Irmer, M., Bobach, C., & Böhme, T. (2019). *Oc processor annotation modules for knowledge extraction*. Version 1.2. Halle (Saale), Germany. https://ontochem.com/

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2020). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, *33*, 494–514. https://api.semanticscholar.org/CorpusID:211010433

Jin, Q., Dhingra, B., Liu, Z., Cohen, W., & Lu, X. (2019). PubMedQA: A dataset for biomedical research question answering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2567–2577. https://doi.org/10.18653/v1/D19-1259

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd draft). https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf

Kær Jørgensen, R., Hartmann, M., Dai, X., & Elliott, D. (2021). MDAPT: Multilingual domain adaptive pretraining in a single model. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3404–3418. https://doi.org/10.18653/v1/2021.findings-emnlp.290

Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2022). Ammu: A survey of transformer-based biomedical pretrained language models. *Journal of Biomedical Informatics*, *126*, 103982. https://doi.org/https://doi.org/10.1016/j.jbi.2021.103982

Karimi Mahabadi, R., Ruder, S., Dehghani, M., & Henderson, J. (2021). Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th*

International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 565–576. https://doi.org/10.18653/v1/2021.acl-long.47

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B., Thiessen, P., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. (2020). Pubchem in 2021: New data content and improved web interfaces. *Nucleic Acids Research*, *49*. https://doi.org/10.1093/nar/gkaa971

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering – a systematic literature review [Special Section - Most Cited Articles in 2002 and Regular Research Papers]. *Information and Software Technology*, *51*(1), 7–15. https://doi.org/https://doi.org/10.1016/j.infsof.2008.09.009

Lai, T. M., Zhai, C., & Ji, H. (2023). Keblm: Knowledge-enhanced biomedical language models. *Journal of Biomedical Informatics*, *143*, 104392. https://doi.org/https://doi.org/10.1016/j.jbi.2023.104392

Lauscher, A., Majewska, O., Ribeiro, L. F. R., Gurevych, I., Rozanov, N., & Glavaš, G. (2020). Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 43–49. https://doi.org/10.18653/v1/2020.deelio-1.5

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Li, B., Hwang, D., Huo, Z., Bai, J., Prakash, G., Sainath, T. N., Chai Sim, K., Zhang, Y., Han, W., Strohman, T., & Beaufays, F. (2023). Efficient domain adaptation for speech foundation models. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096330

Lu, G., Yu, H., Yan, Z., & Xue, Y. (2023). Commonsense knowledge graph-based adapter for aspect-level sentiment classification. *Neurocomputing*, *534*, 67–76. https://doi.org/https://doi.org/10.1016/j.neucom.2023.03.002

Lu, Q., Dou, D., & Nguyen, T. H. (2021). Parameter-efficient domain knowledge integration from multiple sources for biomedical pre-trained language models. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3855–3865. https://doi.org/10.18653/v1/2021.findings-emnlp.325

Majewska, O., Vulić, I., Glavaš, G., Ponti, E. M., & Korhonen, A. (2021). Verb knowledge injection for multilingual event processing. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6952–6969. https://doi.org/10.18653/v1/2021.acl-long.541

Marburger-Bund. (2022). Zu wenig personal, zu viel bürokratie, unzulängliche digitalisierung. *marburger-bund*. https://www.marburger-bund.de/sites/default/files/files/2022-08/2%5C%20-%5C%20Pressemitteilung_0.pdf

Marcus, G. (2019, December). Deep understanding: The next challenge for ai. https://slideslive.com/38922748/deep-understanding-the-next-challenge-for-ai

Meng, Z., Liu, F., Clark, T. H., Shareghi, E., & Collier, N. (2021). Mixture-of-partitions: Infusing large biomedical knowledge graphs into bert. *ArXiv*, *abs/2109.04810*. https://api.semanticscholar.org/CorpusID:237485295

Moon, H., Park, C., Eo, S., Seo, J., & Lim, H. (2021). An empirical study on automatic post editing for neural machine translation. *IEEE Access*, *9*, 123754–123763. https://doi.org/10.1109/ACCESS.2021.3109903

Nentidis, A., Bougiatiotis, K., Krithara, A., & Paliouras, G. (2020). Results of the seventh edition of the bioasq challenge. In P. Cellier & K. Driessens (Eds.), *Machine learning and knowledge discovery in databases* (pp. 553–568). Springer International Publishing.

Nguyen-The, M., Lamghari, S., Bilodeau, G.-A., & Rockemann, J. (2023). Leveraging sentiment analysis knowledge to solve emotion detection tasks. In J.-J. Rousseau & B. Kapralos (Eds.), *Pattern recognition, computer vision, and image processing. icpr 2022 international workshops and challenges* (pp. 405–416). Springer Nature Switzerland.

Nwagwu, W. E. (2022). The rise and rise of natural language processing research, 1958-2021. https://doi.org/10.21203/rs.3.rs-2265814/v1

Overhagea, M. J., & McCallie, D. (2020). Physician time spent using the electronic health record during outpatient encounters [PMID: 31931523]. *Annals of Internal Medicine*, *172*(3), 169–174. https://doi.org/10.7326/M18-3684

pandas development team, T. (2020, February). *Pandas-dev/pandas: Pandas* (Version latest). Zenodo. https://doi.org/10.5281/zenodo.3509134

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:202786778

Peng, Y., Yan, S., & Lu, Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. *BioNLP@ACL*. https://api.semanticscholar.org/CorpusID:189762009

Peters, M. E., Neumann, M., Logan, R., Schwartz, R., Joshi, V., Singh, S., & Smith, N. A. (2019). Knowledge enhanced contextual word representations. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 43–54. https://doi.org/10.18653/v1/D19-1005

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases? *ArXiv*, *abs/1909.01066*. https://api.semanticscholar.org/CorpusID:202539551

Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., & Gurevych, I. (2020). Adapterfusion: Non-destructive task composition for transfer learning. *ArXiv*, *abs/2005.00247*. https://api.semanticscholar.org/CorpusID:218470208

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., & Gurevych, I. (2020). Adapterhub: A framework for adapting transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 46–54.

Phan, L., Anibal, J. T., Tran, H. T., Chanana, S., Bahadroglu, E., Peltekian, A., & Altan-Bonnet, G. (2021). Scifive: A text-to-text transformer model for biomedical literature. *ArXiv*, *abs/2106.03598*. https://api.semanticscholar.org/CorpusID:235358786

Qian, Y., Gong, X., & Huang, H. (2022). Layer-wise fast adaptation for end-to-end multi-accent speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2842–2853. https://doi.org/10.1109/TASLP.2022.3198546

Rebuffi, S.-A., Bilen, H., & Vedaldi, A. (2017). Learning multiple visual domains with residual adapters. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 506–516.

Romanov, A., & Shivade, C. (2018). Lessons from natural language inference in the clinical domain. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1586–1596. https://doi.org/10.18653/v1/D18-1187

Ruiz, C., Zitnik, M., & Leskovec, J. (2021). Identification of disease treatment mechanisms through the multiscale interactome. *Nature Communications*, *12*, 1–15. https://doi.org/ https://doi.org/10.1038/s41467-021-21770-8

Scheetz, J., Rothschild, P., McGuinness, M., Hadoux, X., Soyer, H. P., Janda, M., Condon, J. J., Oakden-Rayner, L., Palmer, L. J., Keel, S., & et al. (2021). A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Nature News*. https://www.nature.com/articles/s41598-021-84698-5

Schneider, P., Schopf, T., Vladika, J., Galkin, M., Simperl, E., & Matthes, F. (2022). A decade of knowledge graphs in natural language processing: A survey. *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 601–614. https://aclanthology.org/2022.aacl-main.46

Singhal, K., Azizi, S., Tu, T., Mahdavi, S., Wei, J., Chung, H. W., Scales, N., Tanwani, A. K., Cole-Lewis, H. J., Pfohl, S. J., Payne, P. A., Seneviratne, M. G., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P. A., y Arcas, B. A., Webster, D. R., … Natarajan, V. (2022). Large language models encode clinical knowledge. *Nature*, *620*, 172–180. https://api.semanticscholar.org/CorpusID:255124952

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H. J., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P. A., Prakash, S., Green, B., Dominowska, E., y Arcas, B. A., … Natarajan, V. (2023). Towards expert-level medical question answering with large language models. *ArXiv*, *abs/2305.09617*. https://api.semanticscholar.org/CorpusID:258715226

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge, 4444–4451.

Stickland, A. C., & Murray, I. (2019). BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Pro-*

ceedings of the 36th international conference on machine learning (pp. 5986–5995, Vol. 97). PMLR. https://proceedings.mlr.press/v97/stickland19a.html

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *NIPS*. https://api.semanticscholar.org/CorpusID:13756489

Wang, B., Xie, Q., Pei, J., Tiwari, P., Li, Z., & Fu, J. (2021). Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*. https://api.semanticscholar.org/CorpusID:238634270

Wang, R., Tang, D., Duan, N., Wei, Z., Huang, X., Ji, J., Cao, G., Jiang, D., & Zhou, M. (2020). K-adapter: Infusing knowledge into pre-trained models with adapters. *Findings*. https://api.semanticscholar.org/CorpusID:211031933

Wei, X., Wang, S., Zhang, D., Bhatia, P., & Arnold, A. O. (2021). Knowledge enhanced pretrained language models: A compreshensive survey. *ArXiv*, *abs/2110.08455*. https://api.semanticscholar.org/CorpusID:239015890

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Wold, S. (2022). The effectiveness of masked language modeling and adapters for factual knowledge injection. *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, 54–59. https://aclanthology.org/2022.textgraphs-1.6

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

Wu, H., Wang, M., Wu, J., Francis, F., Chang, Y.-H., Shavick, A., Dong, H., Poon, M. T. C., Fitzpatrick, N., Adam P. Levine, L. T. S., Handy, A., Karwath, A., Gkoutos, G. V., Chelala, C., Shah, A. D., Stewart, R., Collier, N., Alex, B., Whiteley, W., … Dobson, R. J. B. (2022). A survey on clinical natural language processing in the united kingdom from 2007 to 2022. *Digital Medicine*. https://doi.org/10.1038/s41746-022-00730-6

Xie, Q., Bishop, J. A., Tiwari, P., & Ananiadou, S. (2022). Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowledge-Based Systems*, *252*, 109460. https://doi.org/https://doi.org/10.1016/j.knosys.2022.109460

Xu, Y., Ishii, E., Cahyawijaya, S., Liu, Z., Winata, G. I., Madotto, A., Su, D., & Fung, P. (2022). Retrieval-free knowledge-grounded dialogue response generation with adapters. *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, 93–107. https://doi.org/10.18653/v1/2022.dialdoc-1.10

Yang, L. F., Chen, H., Li, Z., Ding, X., & Wu, X. (2023). Chatgpt is not enough: Enhancing large language models with knowledge graphs for fact-aware language modeling. *ArXiv*, *abs/2306.11489*. https://api.semanticscholar.org/CorpusID:259203671

Yasunaga, M., Leskovec, J., & Liang, P. (2022). Linkbert: Pretraining language models with document links. *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:247793456

Yu, S., & Yang, Y. (2023). A new feature fusion method based on pre-training model for sequence labeling. *2023 6th International Conference on Data Storage and Data Engineering (DSDE)*, 26–31. https://doi.org/10.1109/DSDE58527.2023.00012

Zhang, Z., Wu, Y., Hai, Z., Li, Z., Zhang, S., Zhou, X., & Zhou, X. (2019). Semantics-aware bert for language understanding. *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:202539891

Zou, D., Zhang, X., Song, X., Yu, Y., Yang, Y., & Xi, K. (2022). Multiway bidirectional attention and external knowledge for multiple-choice reading comprehension. *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 694–699. https://doi.org/10.1109/SMC53654.2022.9945587