A Systematic Literature Review of Adapter-based Approaches to Knowledge-enhanced Language Models

Anonymous ACL submission

Abstract

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 knowledge graphs (KGs). They are frequently combined with adapter modules to reduce the computational load and risk of catastrophic forgetting. In this paper, we conduct a systematic literature review (SLR) on adapter-based approaches to KELMs. We provide an overview of approaches in the field and explore the strengths and potential shortcomings of the multitude of discovered methods. We show that general knowledge and domain-specific approaches have been frequently explored along with various downstream tasks. Furthermore, we discovered that the biomedical domain is the most popular domain-specific field and that the Pfeiffer adapter is the most commonly used adapter type. We outline the main trends and propose promising future directions.

1 Introduction

The field of natural language processing (NLP) has, in recent years, been dominated by the rise of large language models (LLMs). These models are pretrained on large amounts of unstructured textual data, which enables them to solve complex reasoning tasks and generate new text. Still, LLMs can lack awareness of structured knowledge hierarchies, such as relations between concepts. This drawback can lead to inaccurate predictions for downstream tasks relying on structured predictions and so-called "hallucinations" within text generation. This can make LLMs less reliable in practice, which is an especially precarious issue in high-risk domains like healthcare or law.

A potential solution to counteract mispredictions and hallucinations and improve the reliability of LLMs is knowledge enhancement: By leveraging expert knowledge from manually curated knowledge graphs (KGs), structured knowledge can be injected into LLMs. Such knowledge-enhanced language models (KELMs) are a promising approach for higher structured knowledge awareness, better factual accuracy, and less hallucinations (Colon-Hernandez et al., 2021; Wei et al., 2021).

Unfortunately, knowledge enhancement in the form of supervised fine-tuning (SFT) of the whole LLM can be highly computationally expensive, especially for models with billions of parameters. A promising research avenue to overcome this limitation is using lightweight and efficient adapter modules to inject structured knowledge into LLMs. Using adapters for knowledge enhancement helps enhance the task performance of LLMs and is, at the same time, a very computationally efficient solution. Despite the rising popularity of this approach, to the best of our knowledge, a comprehensive overview of adapter-based KELMs is still missing in the NLP research landscape.

To bridge this research gap, we conduct a systematic literature review (SLR) on adapter-based knowledge enhancement of LLMs. Our contributions are: (1) a novel review of adapter-based knowledge enhancement, (2) a quantitative and qualitative analysis of different methods in the field, and (3) a detailed categorization of literature and identification of the most promising trends.

2 Background and Related Work

In this section, we give an overview of related work and existing surveys on knowledge enhancement. Knowledge graphs are the most common external knowledge source, so we start with their overview.

2.1 Knowledge Graphs

Knowledge graphs (KGs) are a structured representation of the world knowledge and have seen a rising prominence in NLP research over the past decade (Schneider et al., 2022). Hogan 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*". Similarly, Ji 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* $(h, r, t) \in \mathcal{F}$". Depending on the source and purpose of a KG, entities and relations can take on various shapes. For example, in the biomedical knowledge graph UMLS (Bodenreider, 2004), 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. Other than UMLS, other examples of popular KGs are DBpedia (Auer et al., 2007) and ConceptNet (Speer et al., 2017).

2.2 Approaches to Knowledge Enhancement

At the time of writing, some reviews had already been published that gave an overview of KELMs and classified different approaches. Colon-Hernandez et al. (2021) review the existing literature and split the approaches to integrate structure knowledge with LMs into three categories: (1) input-centered strategies, centering around altering the structure of the input or selected data, which is fed into the base LLM; (2) architecture-focused approaches, which involve either adding additional layers that integrate knowledge with the contextual representations or modifying existing layers to alter parts like attention mechanisms; (3) output-focused approaches, which work by changing either the output structure or the losses used in the base model. Our study focuses on the second category (2), by examining the adapter-based mechanisms for injecting information into the model, which were shown to be the most promising by the authors.

The second survey by Wei et al. (2021) reviews a large number of studies on KELMs and classifies them using three taxonomies: (1) knowledge sources, (2) knowledge granularity, and (3) application areas. Within (1), the knowledge sources include linguistic knowledge, encyclopedic knowledge, and commonsense and domain-specific knowledge. The second taxonomy (2) acknowledges the common approach of using KGs as a source of knowledge. Levels of granularity mentioned are text-based knowledge, entity knowledge, relation triples, and KG sub-graphs. Lastly, with the third taxonomy (3), the authors discuss how knowledge enhancement can improve natural language generation and understanding. They also review popular benchmarks that can be used for task evaluation of KELMs (Wei et al., 2021).

These two field studies by Colon-Hernandez et al. (2021) and Wei et al. (2021) on the classification of KELM approaches were our starting point for exploring KELMs and initially proved to be very valuable. However, although they address some adapter-based studies like K-Adapter (Wang et al., 2020), most other adapter-based KELMs are missing. This lack of coverage led to our decision to conduct a novel systematic literature search focusing specifically on the adapter-based KELMs, considering their rising popularity and importance.

3 Adapters

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

3.1 Overview

Broadly speaking, adapters are small bottleneck feed-forward layers inserted within each layer of an LLM (Houlsby et al., 2019). The small amount of additional parameters allows injecting new data or knowledge without fine-tuning 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 tuning). Due to the lightweight nature of adapters, this approach leads to short training times with relatively low computing resource requirements. Adapters used to be utilized mostly for quick and cheap downstream-task fine-tuning but are now increasingly used for knowledge enhancement. Because it is possible to train adapters individually, they can also be used for multi-task training by specializing one adapter for each task or multi-domain knowledge injection by specializing adapters to different

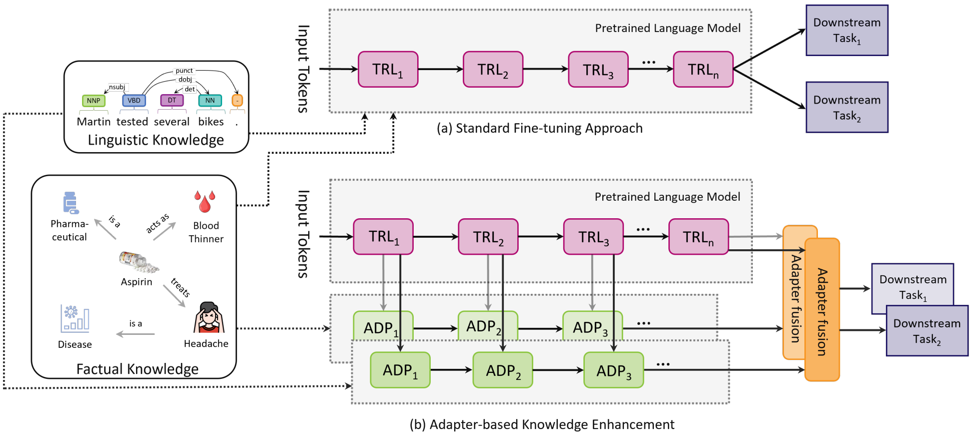

Figure 1: Illustration of a standard fine-tuning versus a knowledge enhancement process. In the example, knowledge from a KG is injected into the model via adapters.

domains (Pfeiffer et al., 2020a).

Leveraging adapters in LLMs also has positive "side effects": Adapters can avoid catastrophic forgetting (the issue when an LLM suddenly deteriorates in performance after fine-tuning) by introducing new task-specific parameters (Houlsby et al., 2019; Pfeiffer et al., 2020a) and, in transfer learning, adapters have even been shown to improve stability and adversarial robustness for various downstream tasks (Han et al., 2021). The specifics of how and where adapters are added to an LLM depend on the adapter type.

3.2 Adapter Types

Houlsby Adapter The Houlsby Adapter (Houlsby 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 et al. (2017) in the computer vision domain. The two main principles stayed the same: Adapters require a relatively small number of parameters 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, while a near-identity initialization is required for stable training of the adapted model (Houlsby 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. In a classical transformer structure (Vaswani et al., 2017), the adapter module is added 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 et al., 2016). Including biases, $2md + d + m$ parameters are added per layer, accounting for only 0.5 to 8 percent of the parameters of the original BERT model used by the authors when setting $m << d$.

Bapna and Firat Adapter In contrast to the Houlsby Adapter, 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 for better understanding of the component positions). Moreover, while Houlsby et al. (2019) re-train layer normalization parameters for every domain, Bapna and Firat (2019) "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".

Pfeiffer Adapter and AdapterFusion. The approaches of Bapna and Firat (2019); Houlsby et al. (2019) did not allow information sharing between tasks. Pfeiffer et al. (2020a) introduce Adapter Fusion, a two-stage algorithm that addresses the sharing of information encapsulated in adapters trained on different tasks. In the first stage, they train the adapters in single-task or multi-task setups for a total of N tasks similar to the Houlsby Adapter, but only keeping the top adapters, sim-

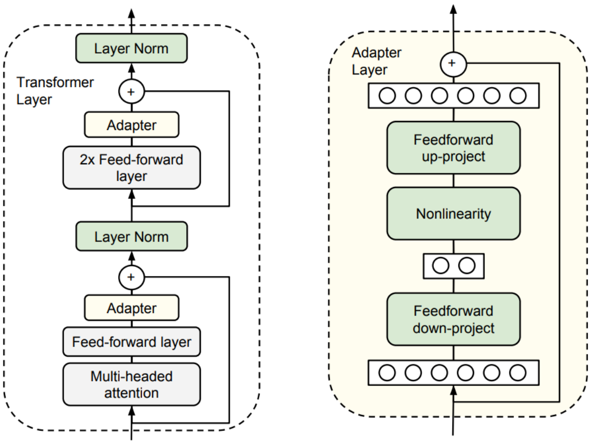

Figure 2: Location of the adapter module in a transformer layer (left) and architecture of the Houlsby Adapter (right). All green layers are trained on fine-tuning data, including the adapter itself, the layer normalization parameters, and the final classification layer (not shown). Image with permission from Houlsby et al. (2019).

ilar 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 et al., 2020a): $\Psi_m \leftarrow \underset{\Psi}{\operatorname{argmin}}\, L_m\left(D_m; \Theta, \Phi_1, \ldots, \Phi_N, \Psi\right)$

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 et al., 2020a). 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. The authors also find that their approach of using only a single adapter after the feed-forward layer performs on par with the Houlsby adapter while requiring only half of the newly introduced adapters (Pfeiffer et al., 2020a). This makes the Pfeiffer adapter an attractive choice for many applications, further proven by its popularity among the papers in our review.

K-Adapter Wang 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 and two projection layers. Similar to the approaches above, a skip connection is added but instead applied across the two projection layers. 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.

Adapter architectures for knowledge enhancement exist that differ from the four adapter types mentioned here. For example, the "Parallel Adapter" (He et al., 2021a) 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.

Another popular type of efficient adaptation is the low-rank adaptation LoRA (Hu et al., 2022), and its quantized version QLoRA (Dettmers et al., 2023). These approaches do not add new adapter layers as the previously described ones, but rather enforce a low-rank constraint on the weight updates of the base model's layers. This enables efficient fine-tuning of LLMs and also allows for domain adaption or knowledge enhancement. Despite their popularity, our search string did not match any papers using these approaches, which is likely due to our focus on adapters in form of adapter layers.

4 Methodology

This chapter details the methodology we employed for the systematic literature review. We largely followed the procedure of Kitchenham et al. (2009) for systematic literature reviews in software engineering. The search strategy for the systematic literature review of this thesis included literature that fulfilled the following inclusion criteria:

- Peer-reviewed articles from ACM[1], ACL[2], and IEEE Xplore[3]
- Article abstracts that match 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 that address the topic of adapter-based knowledge-enhanced language models

We also included a limited number of articles not found on the mentioned databases because they were fundamental works on the topic of the SLR and frequently referenced. The SLR was concluded in January 2024 and represents the state of research literature up to this point.

5 Results

This section will present the results of the systematic literature review on adapter-based knowledge enhancement.

5.1 Overview

Source	Initial	Abstract	Full Text
IEEE	28	6	6
ACM	10	6	5
ACL	36	16	13
Others	2	2	2
Total	76	30	26

Table 1: Quantitative overview of the literature sources and the selection process

Table 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 field, 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, 31 articles met all inclusion criteria (and no exclusion criteria). After the full paper screening, 26 papers remained to form the final paper pool of the survey.

Table 2 gives an overview of all papers included in the survey. It includes the information on the adapter type used in the paper, the domain and scope of the paper, and for which downstream NLP tasks it was developed.

5.2 Data Analysis

We will now give a quantitative analysis showcasing and interpreting quantitative distributions, followed by significant qualitative insights from the papers.

5.2.1 Quantitative Analysis

Yearly Distribution There has been a significant increase in publications on adapter-based approaches to knowledge-enhanced language models in recent years (Fig. 3). While only 2 papers were published in 2020, eleven new papers were published in 2023. This trend suggests growing interest and research activity in the domain.

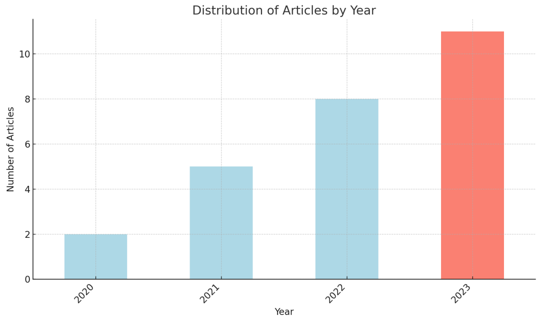

Figure 3: Yearly distribution of publications

Adapter Type Distribution Next, we evaluate the popularity and variety of adapter types used across the papers (Fig. 4). The “Pfeiffer” and "Houlsby" adapter types stand out as the most common, which suggests that the closely related underlying architecture is the most popular methodology in the field. This popularity is likely not only an achievement of the adapter’s performance but also due to the well-established Adapter-Hub platform (Pfeiffer et al., 2020b), 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

[1] https://dl.acm.org/

[2] https://aclanthology.org/

[3] https://ieeexplore.ieee.org/Xplore/home.jsp

paper & nickname	adapter type	scope	main source	task
K-MBAN (Zou et al., 2022)	K-Adapter	open	T-REx (Wiki)	RC
/ (Moon et al., 2021)	Houlsby	open	WMT20	MT
CSBERT (Yu and Yang, 2023)	Unique	open	diverse	SL
/ (Qian et al., 2022)	Unique	open	AESRC2020	SR
/ (Li et al., 2023)	Houlsby	closed (multiple)	diverse	SF
CPK (Liu et al., 2023)	K-Adapter	closed (biomed)	Wikipedia	RC, ET, QA
CKGA (Lu et al., 2023)	Unique	open	DBpedia	SC
/ (Nguyen-The et al., 2023)	Pfeiffer	open	diverse	SA
KEBLM (Lai et al., 2023)	Pfeiffer	closed (biomed)	UMLS	QA, NLI, EL
/ (Guo and Guo, 2022)	Unique	open	Ch. Lexicon	NER
/ (Tiwari et al., 2023)	Unique	closed (biomed)	Vis-MDD	TS
AdapterSoup (Chronopoulou et al., 2023)	Bapna and Firat	closed (multiple)	diverse	LM
/ (Wold, 2022)	Houlsby	open	ConceptNet	LAMA
/ (Chronopoulou et al., 2022)	Unique	closed (multiple)	diverse	LM
DS-TOD (Hung et al., 2022)	Pfeiffer	closed (multiple)	CCNet	TOD
/ (Emelin et al., 2022)	Houlsby	closed (multiple)	MultiWOZ	TOD
KnowExpert (Xu et al., 2022)	Bapna and Firat	open	WoW	KGD
mDAPT (Kær Jørgensen et al., 2021)	Pfeiffer	closed (multiple)	WMT20	NER, STC
DAKI (Lu et al., 2021)	K-Adapter	closed (biomed)	UMLS	NLI
/ (Majewska et al., 2021)	Pfeiffer	open	VerbNet	EE
/ (Lauscher et al., 2020)	Houlsby	open	ConceptNet	GLUE
TADA (Hung et al., 2023)	Unique	open	CCNet	TOD, NER, NLI
LeakDistill (Vasylenko et al., 2023)	StructAdapt	open	AMR graph	SMATCH
MixDA (Diao et al., 2023)	Houlsby, Pfeiffer	closed (multiple)	diverse	GLUE, TXM
MoP (Meng et al., 2021)	Pfeiffer	closed (biomed)	UMLS	BLURB
K-Adapter (Wang et al., 2020)	K-Adapter	open	T-REx (Wiki)	RCL, ET, QA

Table 2: Overview of the results for the literature survey, including all papers and their references. The task and source 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. For the definition of all task acronyms, see Appendix A.4

less frequently mentioned architectures, suggesting that these approaches are less well-established. Overall, various adapter types are present, indicating a diverse range of methodologies being explored.

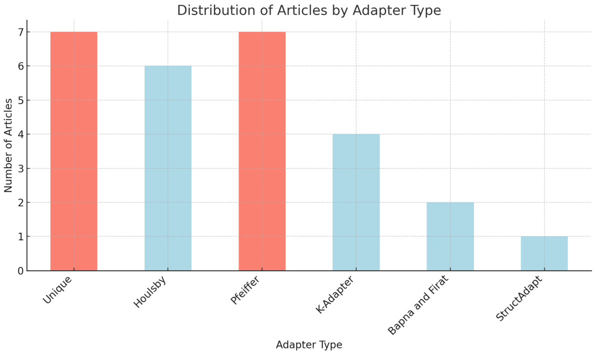

Figure 4: Distribution of adapter types being used in the articles

Domain Analysis Third, we analyze the distribution of papers across the domain scope and coverage to understand domain-specific preferences in the literature (figures given in the appendix). The first plot in Figure 5 shows that the open-domain scope is the most popular, with many papers exploring adapter-based approaches within the open domain. The popularity is likely caused by the interest in creating LLMs with a common-sense understanding or world knowledge.

As illustrated by the second plot in Figure 5, the single- and multi-domain approaches are split evenly within the closed-domain papers.

Finally, the third plot addresses the coverage of the biomedical domain. In absolute numbers, only six 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. The popularity likely comes down to the availability of large biomedical KGs, and medicine historically being one of the most active research fields in general science (Cimini et al., 2014).

Task and Source Distribution A highly diverse range of tasks and sources is being explored throughout the papers, which signifies the versatility and potential of adapter-based approaches across different NLP tasks and domains. 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. This could be because these tasks are the most demanding regarding structural knowledge requirements. The knowledge source distributions show very little overlap.

5.2.2 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, we still provide an overview of the most common patterns.

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 et al., 2017) for common-sense knowledge, and DBpedia (Auer et al., 2007) for encyclopedic world knowledge. Two example works that use these KGs are Wold (2022) and the CKGA ("knowledge graph-based adapter") by Lu et al. (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 et al., 2019) and see increasing performance while only adding 2.1% of new parameters to the original models. CKGA (Lu et al., 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, where the corresponding knowledge is extracted with graph convolutional networks (Lu et al., 2023).

Linguistic Knowledge Instead of only including factual knowledge, some works also inject linguistic knowledge into adapters (Majewska et al., 2021; Zou et al., 2022; Yu and Yang, 2023; Wang et al., 2020). While LLMs already encode a range of syntactic and semantic properties of language, Majewska 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 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 et al. (2022) on machine reading comprehension (MRC). They proposed the K-MBAN model to integrate linguistic and factual external knowledge into LLMs through K-Adapters.

Domain-specific Knowledge Chronopoulou et al. (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 et al. (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 various 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. Earlier, we saw that the biomedical domain is the most prevalent among the closed-domain approaches to adapter-based KELMs. We will briefly examine the relevant works in the following.

Biomedical Knowledge We have found the works of DAKI (Lu et al., 2021), MoP (Meng et al., 2021), and KEBLM (Lai et al., 2023) to be the most impactful. 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. Lu et al. (2021) leverage data from the UMLS meta-thesaurus and UMLS Semantic Network groups concepts, but also from Wikipedia articles for diseases as proposed by He et al. (2020). Meng et al. (2021) recognize that KGs like UMLS, which can be several gigabytes large, are very expensive to train on in their entirety. They propose to use a "Mixture of Partitions" (MoP), which splits the KG into sub-graphs and combines later with AdapterFusion (Pfeiffer et al., 2020a). Finally, the KEBLM framework's trademark is that it allows the inclusion of a variety of knowledge types from multiple sources into biomedical LLMs. 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 et al., 2023). For completeness, we refer to Kær Jørgensen et al. (2021) for information on the m-DAPT framework, which addresses multilingual domain adaptation for biomedical LLMs and KeBioSum (Xie et al., 2022), who state their work is the first study exploring knowledge injection for biomedical extractive summarization.

Performance Insights Performance of adapter-based KELMs on downstream tasks is consistently shown in papers from our survey to be better than with base LMs. For example, Diao et al. (2023) show an increase of +9% on Common-sense QA Talmor et al. (2019) with their mixture-of-adapters approach, while (Kær Jørgensen et al., 2021) improve financial text classification on OMP-9 (Schabus et al., 2017) by +4%. While the task variation across domains is too diverse to be shown systematically in our survey, we report in detail on performance comparison in the biomedical domain in Appendix A.2. Table 3 shows the performance over five common biomedical tasks, covering text classification, QA, NLI, and NER. It shows that adapter-based KELMs consistently improved the performance in almost all instances. For example, MoP (Meng et al., 2021) and KEBLM (Lai et al., 2023) improve the performance on PubMedQA (Jin et al., 2019) for around +7% and +8%, respectively. Another interesting insight is found by He et al. (2021b), who show that adapter-based tuning mitigates forgetting issues better than regular fine-tuning since it yields representations with less deviation from those generated by the initial pre-trained language model.

6 Current and Future Trends

In this section, we outline the most important findings and trends of the review and point out the promising future directions:

- Adapter-based KELMs are a recent development in NLP, but interest in them is rising fast, with a linear yearly increase of published papers. We predict the growth trend to continue.

- Various adapter architectures exist and are advanced by researchers to be more efficient while preserving task performance. This peaked with the Pfeiffer adapter, which is the most popular type. We expect future work to focus their updates on adapter architecture by overcoming the latency of sequential data processing in adapters and enabling hardware parallelism.

- Research focuses on the open domain – injecting general world knowledge into models. Within the closed domain, the biomedical domain is the most popular, owing to the existence of large biomedical KGs. We foresee the potential to apply adapter-based KELMs to other highly structured domains, such as the legal or financial domain (documents with rigid structure).

- A wide array of downstream tasks is being explored. The biggest improvement in task performance is seen in knowledge-intensive tasks like question answering and text classification, with a smaller improvement for reasoning tasks like entailment recognition. Generative tasks, other than dialogue modeling, are rather unexplored. We envision a future popular use case that could use knowledge enhancement to improve the factuality and informativeness of generated text.

7 Conclusion

In this paper, we conducted a systematic literature review on approaches to enhancing language models with external knowledge using adapter modules. We portrayed which adapter-based approaches exist and how they compare to each other. We showed there is a steady growth of interest in this domain with each new year and highlighted the most popular adapter architectures (with "Pfeiffer" as the predominant one). We discovered there is a balance in popularity between open-domain approaches, focusing on integrating general world knowledge into models, and closed-domain focusing on specialized fields, with biomedical as the most popular domain. With our review, we contribute a novel and extensive resource for this nascent yet fast-growing field and we hope it will be a useful entry point for other researchers in the future.

Limitations

The methodology of a systematic literature review follows a strict search string and exclusion criteria. Therefore, it is possible that we excluded some relevant work on adapter-based KELMs. Moreover, while we tried to report on our survey as comprehensively as possible, there are several aspects we could not include in this work. Also, some of the reviewed articles were not given an adequate qualitative analysis in this work due to space constraints, leading to potentially missing insights and a non-complete representation of the state of research on adapter-based knowledge enhancement. Additionally, due to the variety of applications and domains, we were not able to give precise guidelines on what methods to use under which circumstances. Still, we aimed to report on the most common patterns and trends discovered in the literature, which can serve as a basis for future research.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *ArXiv*, abs/1607.06450.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2015. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Conference on Empirical Methods in Natural Language Processing*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Alexandra Chronopoulou, Matthew Peters, and Jesse Dodge. 2022. Efficient hierarchical domain adaptation for pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1336–1351, Seattle, United States. Association for Computational Linguistics.

Alexandra Chronopoulou, Matthew Peters, Alexander Fraser, and Jesse Dodge. 2023. AdapterSoup: Weight averaging to improve generalization of pretrained language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2054–2063, Dubrovnik, Croatia. Association for Computational Linguistics.

Giulio Cimini, Andrea Gabrielli, and Francesco Labini. 2014. The scientific competitiveness of nations. *PloS one*, 9.

Pedro Colon-Hernandez, Catherine Havasi, Jason B. Alonso, Matthew Huggins, and Cynthia Breazeal. 2021. Combining pre-trained language models and structured knowledge. *ArXiv*, abs/2101.12294.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Shizhe Diao, Tianyang Xu, Ruijia Xu, Jiawei Wang, and Tong Zhang. 2023. Mixture-of-domain-adapters: Decoupling and injecting domain knowledge to pre-trained language models' memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5113–5129, Toronto, Canada. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *ArXiv*, abs/1811.01241.

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Hady Elsahar. 2017. T-Rex : A Large Scale Alignment of Natural Language with Knowledge Base Triples [NIF SAMPLE].

Denis Emelin, Daniele Bonadiman, Sawsan Alqahtani, Yi Zhang, and Saab Mansour. 2022. Injecting domain knowledge in language models for task-oriented dialogue systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11962–11974. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael R. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1 – 23.

Qian Guo and Yi Guo. 2022. Lexicon enhanced chinese named entity recognition with pointer network. *Neural Computing and Applications*.

Wenjuan Han, Bo Pang, and Ying Nian Wu. 2021. Robust transfer learning with pretrained language models through adapters. *ArXiv*, abs/2108.02340.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021a. Towards a unified view of parameter-efficient transfer learning. *ArXiv*, abs/2110.04366.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. 2021b. On the effectiveness of adapter-based tuning for pretrained language model adaptation.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614, Online. Association for Computational Linguistics.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutiérrez, S. Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2020. Knowledge graphs. *ACM Computing Surveys (CSUR)*, 54:1 – 37.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Chia-Chien Hung, Lukas Lange, and Jannik Strötgen. 2023. TADA: Efficient task-agnostic domain adaptation for transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 487–503, Toronto, Canada. Association for Computational Linguistics.

Chia-Chien Hung, Anne Lauscher, Simone Ponzetto, and Goran Glavaš. 2022. DS-TOD: Efficient domain specialization for task-oriented dialog. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 891–904. Association for Computational Linguistics.

Shaoxiong Ji, Shirui Pan, E. Cambria, Pekka Marttinen, and Philip S. Yu. 2020. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33:494–514.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Rasmus Kær Jørgensen, Mareike Hartmann, Xiang Dai, and Desmond Elliott. 2021. mDAPT: Multilingual domain adaptive pretraining in a single model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3404–3418. Association for Computational Linguistics.

Barbara Kitchenham, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. 2009. Systematic literature reviews in software engineering – a systematic literature review. *Information and Software Technology*, 51(1):7–15. Special Section - Most Cited Articles in 2002 and Regular Research Papers.

Tuan Manh Lai, ChengXiang Zhai, and Heng Ji. 2023. Keblm: Knowledge-enhanced biomedical language models. *Journal of Biomedical Informatics*, 143:104392.

Anne Lauscher, Olga Majewska, Leonardo F. R. Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020. Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Bo Li, Dongseong Hwang, Zhouyuan Huo, Junwen Bai, Guru Prakash, Tara N. Sainath, Khe Chai Sim, Yu Zhang, Wei Han, Trevor Strohman, and Francoise Beaufays. 2023. Efficient domain adaptation for speech foundation models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

C. Liu, S. Zhang, C. Li, and H. Zhao. 2023. Cpk-adapter: Infusing medical knowledge into k-adapter with continuous prompt. In *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pages 1017–1023, Los Alamitos, CA, USA. IEEE Computer Society.

Guojun Lu, Haibo Yu, Zehao Yan, and Yun Xue. 2023. Commonsense knowledge graph-based adapter for aspect-level sentiment classification. *Neurocomputing*, 534:67–76.

Qiuhao Lu, Dejing Dou, and Thien Huu Nguyen. 2021. Parameter-efficient domain knowledge integration from multiple sources for biomedical pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3855–3865. Association for Computational Linguistics.

Olga Majewska, Ivan Vulić, Goran Glavaš, Edoardo Maria Ponti, and Anna Korhonen. 2021. Verb knowledge injection for multilingual event processing. In *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)*, pages 6952–6969. Association for Computational Linguistics.

Zaiqiao Meng, Fangyu Liu, Thomas Hikaru Clark, Ehsan Shareghi, and Nigel Collier. 2021. Mixture-of-partitions: Infusing large biomedical knowledge graphs into bert. *ArXiv*, abs/2109.04810.

Hyeonseok Moon, Chanjun Park, Sugyeong Eo, Jaehyung Seo, and Heuiseok Lim. 2021. An empirical study on automatic post editing for neural machine translation. *IEEE Access*, 9:123754–123763.

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2020. Results of the seventh edition of the bioasq challenge. In *Machine Learning and Knowledge Discovery in Databases*, pages 553–568, Cham. Springer International Publishing.

Maude Nguyen-The, Soufiane Lamghari, Guillaume-Alexandre Bilodeau, and Jan Rockemann. 2023. Leveraging sentiment analysis knowledge to solve emotion detection tasks. In *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, pages 405–416. Springer Nature Switzerland.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *ArXiv*, abs/1909.01066.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterfusion: Non-destructive task composition for transfer learning. *ArXiv*, abs/2005.00247.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020b. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54.

Yanmin Qian, Xun Gong, and Houjun Huang. 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.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 506–516.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1241–1244, New York, NY, USA. Association for Computing Machinery.

Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. A decade of knowledge graphs in natural language processing: A survey. In *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)*, pages 601–614, Online only. Association for Computational Linguistics.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Xian Shi, Fan Yu, Yizhou Lu, Yuhao Liang, Qiangze Feng, Daliang Wang, Yanmin Qian, and Lei Xie. 2021. The accented english speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6918–6922.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995. PMLR.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhisek Tiwari, Manisimha Manthena, Sriparna Saha, Pushpak Bhattacharyya, Minakshi Dhar, and Sarbajeet Tiwari. 2022. Dr. can see: Towards a multi-modal disease diagnosis virtual assistant. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 1935–1944, New York, NY, USA. Association for Computing Machinery.

Abhisek Tiwari, Anisha Saha, Sriparna Saha, Pushpak Bhattacharyya, and Minakshi Dhar. 2023. Experience and evidence are the eyes of an excellent summarizer! towards knowledge infused multi-modal clinical conversation summarization. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 2452–2461, New York, NY, USA. Association for Computing Machinery.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Pavlo Vasylenko, Pere Lluís Huguet Cabot, Abelardo Carlos Martínez Lorenzo, and Roberto Navigli. 2023. Incorporating graph information in transformer-based AMR parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1995–2011, Toronto, Canada. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings*.

Xiaokai Wei, Shen Wang, Dejiao Zhang, Parminder Bhatia, and Andrew O. Arnold. 2021. Knowledge enhanced pretrained language models: A compreshensive survey. *ArXiv*, abs/2110.08455.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Sondre Wold. 2022. The effectiveness of masked language modeling and adapters for factual knowledge injection. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, pages 54–59, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Qianqian Xie, Jennifer Amy Bishop, Prayag Tiwari, and Sophia Ananiadou. 2022. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowledge-Based Systems*, 252:109460.

Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2022. Retrieval-free knowledge-grounded dialogue response generation with adapters. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 93–107. Association for Computational Linguistics.

Shichuan Yu and Yan Yang. 2023. A new feature fusion method based on pre-training model for sequence labeling. In *2023 6th International Conference on Data Storage and Data Engineering (DSDE)*, pages 26–31.

Dongsheng Zou, Xiaotong Zhang, Xinyi Song, Yi Yu, Yuming Yang, and Kang Xi. 2022. Multiway bidirectional attention and external knowledge for multiple-choice reading comprehension. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 694–699.

A Supplementary Survey Data

A.1 Domain Distribution

See Figure 5.

A.2 Performance Comparison (Biomedical)

Table 3 gives an overview of the downstream task performance of several papers that are included in this survey. The focus lies on the biomedical domain so that the task overlap is high enough for an insightful comparison. The scores are reported for five downstream tasks, namely HoC (Baker et al., 2015), PubMedQA (Gu et al., 2020), BioASQ7b (Nentidis et al., 2020), MedNLI (Romanov and Shivade, 2018), and NCBI (Dogan et al., 2014), as well as three common biomedical language models (SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2019), and PubMedBERT (Gu et al., 2020)). Performances across the different tasks and models vary strongly

A.3 Methodology

Articles on the following topics were excluded:

- Articles published before February 2, 2019
- 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

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

The collected data was tabulated to show:

- Source and publication dates of the studies
- Adapter architectures and knowledge sources 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)

A.4 Acronyms

- AESRC2020: Accented English Speech Recognition Challenge 2020 (Shi et al., 2021)
- BioNLP: Biomedical Natural Language Processing
- BLURB: Biomedical Language Understanding and Reasoning Benchmark (Gu et al., 2020)
- CCNet: Common Crawl Net (Wenzek et al., 2020)
- EE: Event Extraction
- EL: Entity Linking
- ES: Extractive Summarization
- ET: Entity Typing
- GLUE: General Language Understanding Evaluation (Wang et al., 2019)
- IE: Information Extraction
- KELM: Knowledge-Enhanced Language Model
- KGD: Knowledge-grounded Dialogue
- LAMA: Concept-Net Split of LAMA Probe (Petroni et al., 2019)
- LM: Language Modeling
- LLM: Large Language Model
- MT: Machine Translation
- MultiWOZ: Multi-Domain Wizard-of-Oz dataset (Budzianowski et al., 2018)
- NER: Named Entity Recognition

- NLI: Natural Language Inference
- NLP: Natural Language Processing
- OOD: Out-of-domain Detection
- QA: Question Answering
- RC: Reading Comprehension
- RE: Relation Extraction
- RCL: Relation Classification
- SA: Sentiment Analysis
- SC: Sentiment Classification
- SF: Speech Foundation
- SL: Sequence Labelling
- SMATCH: Semantic Match Score (Cai and Knight, 2013)
- SOTA: State-of-the-art
- SR: Speech Recognition
- STC: Sentence Classification
- TC: Text Classification
- TOD: Task-Oriented dialogue
- T-REx (wiki): A Large Scale Alignment of Natural Language with Knowledge Base Triples (Elsahar, 2017)
- UMLS: Unified Medical Language System
- VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon (Schuler, 2006)
- Vis-MDD: Visual Medical Disease Diagnosis (Tiwari et al., 2022)
- WMT20: Workshop on Machine Translation 2020 (Barrault et al., 2020)
- WoW: Wizard-of-Wikipedia (Dinan et al., 2018)

A.5 Performance Comparison (Biomedical)

Table 3 gives an overview of the downstream task performance of several papers that are included in this survey. The focus lies on the biomedical domain so that the task overlap is high enough for an insightful comparison. The scores are reported for five downstream tasks, namely HoC (Baker et al., 2015) (text classification), PubMedQA (Gu et al., 2020) (QA), BioASQ7b (Nentidis et al., 2020) (QA), MedNLI (Romanov and Shivade, 2018) (NLI), and NCBI (Dogan et al., 2014) (disease entity recognition), as well as three common biomedical language models (SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2019), and PubMedBERT (Gu et al., 2020)).

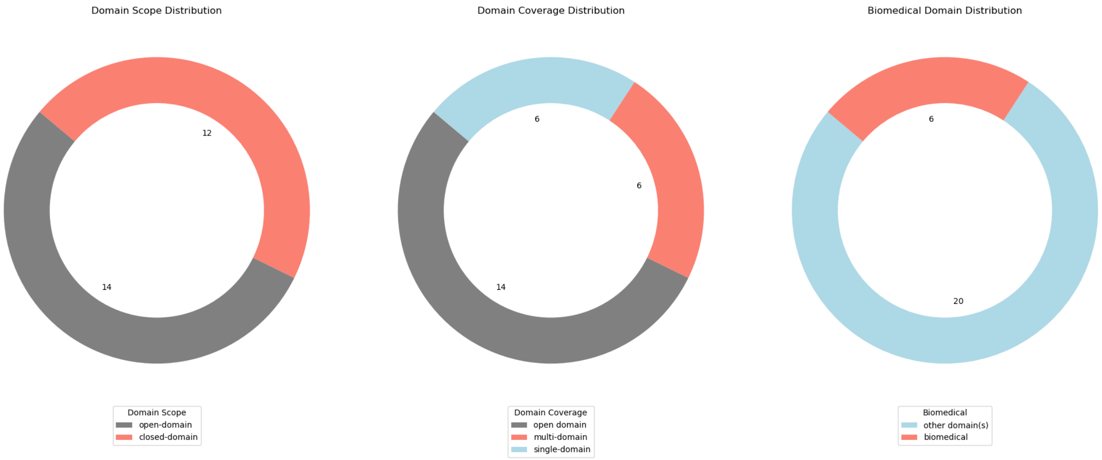

Figure 5: Distribution of domain scope, coverage, and the biomedical domain

↓ **model|dataset** →	**HoC**	**PubMedQA**	**BioASQ7b**	**MedNLI**	**NCBI**
SciBERT-base	$80.52_{\pm 0.60}$	$57.38_{\pm 4.22}$	$75.93_{\pm 4.20}$	$81.19_{\pm 0.54}$	88.57
+ ***MoP***	$81.79^{\dagger}_{\pm 0.66}$ ↑	$54.66_{\pm 3.10}$	$78.50^{\dagger}_{\pm 4.06}$ ↑	$81.20_{\pm 0.37}$ ↑	/
+ ***KEBLM***	/	59.00↑	/	82.14↑	**93.50**↑
+ ***DAKI***	/	/	/	/	/
+ ***CPK***	/	/	/	/	/
BioBERT-base	$81.41_{\pm 0.59}$	$60.24_{\pm 2.32}$	$77.50_{\pm 2.92}$	$82.42_{\pm 0.59}$	88.30
+ ***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 ↑	93.20↑
+ ***DAKI***	/	/	/	83.41 ↑	89.01↑
+ ***CPK***	/	/	/	81.65	88.42↑
PubMedBERT-base	$82.25_{\pm 0.46}$	$55.84_{\pm 1.78}$	$87.71_{\pm 4.25}$	$84.18_{\pm 0.19}$	87.82
+ MoP	$\mathbf{83.26}^{\dagger}_{\pm 0.32}$ ↑	$62.84^{\dagger}_{\pm 2.71}$ ↑	$\mathbf{90.64}^{\dagger}_{\pm 2.43}$ ↑	$\mathbf{84.70}_{\pm 0.19}$ ↑	/

Table 3: Performance reports for tasks with highest overlap in the biomedical domain. The metric for HoC is Micro F1; for NCBI, it is F1, while for the other three, it is accuracy. The best results for every task are in bold. "↑" denotes that improvements are observed compared to the base model. "†" denotes a statistically significant better result over the base model (T-test, $p < 0.05$), but not all papers report their scores. The baseline performance of the models is taken from the original papers if given. Otherwise, the scores are taken from the MoP results.