SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY

TECHNICAL UNIVERSITY OF MUNICH

Master's Thesis in Informatics

Towards Adapter-based Multi-domain Adaptation and Evaluating Factuality in Language Models

Hasan Selim Yagci

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY

TECHNICAL UNIVERSITY OF MUNICH

Master's Thesis in Informatics

Towards Adapter-based Multi-domain Adaptation and Evaluating Factuality in Language Models

Adapter-basierte Mehrdomänenanpassung und Evaluierung der Faktualität in Sprachmodellen

Author:	Hasan Selim Yagci
Supervisor:	Prof. Dr. Georg Groh
Advisor:	Alexander Fichtl, M.Sc.
Submission Date:	31.10.2024

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

Munich, 31.10.2024 Hasan Selim Yagci

Abstract

Large language models (LLMs) have transformed natural language processing (NLP) by enabling a broad range of applications that require generating human-like, contextually relevant text. With excessive pretraining on diverse data, these models demonstrate remarkable adaptability across general tasks. However, in domain-specific applications, such as finance, law, or medicine, LLMs frequently generate responses that, while plausible, are factually incorrect. This issue is particularly concerning in contexts where accuracy is crucial. Further, maintaining domain relevance and factual precision within these large, complex models is computationally challenging, particularly when addressing the need to update domain knowledge continually. This thesis addresses these challenges through adapter-based fine-tuning for multi-domain generative question answering (QA), designed to improve modularity, efficiency, and adaptability, guided by three main research questions. First, it examines how domain-specific adapters can improve the performance of pretrained language models when applied to single-domain generative QA settings. Second, it investigates the potential for a multi-adapter configuration to enable LLMs to integrate knowledge across distinct domains, allowing models to effectively handle diverse inputs. Finally, it evaluates whether existing automated metrics for factuality are sufficiently robust to assess model performance in a multi-domain, open-ended generative QA setting. The results indicate that the adapter-based approach achieves performance comparable to conventional fine-tuning methods for domain adaptation, with the added benefits of quicker model updates and memory-efficient integration in environments with limited resources. Yet, current automatic metrics fall short in evaluating the factuality of generated, open-ended answers, particularly when response length and complexity vary across domains. Experiments with LLM-based and multi-aspect prompting metrics reveal limitations, underscoring the need for more reliable metrics to assess factual accuracy in multi-domain generative QA. In summary, this thesis lays the groundwork for an efficient, adaptable strategy for multi-domain language models, advancing the development of modular, domain-knowledge-rich generative QA systems that fulfill the needs of domain-specific NLP applications while retaining general-purpose capabilities.

Kurzfassung

Große Sprachmodelle (LLMs) haben die Verarbeitung natürlicher Sprache (NLP) revolutioniert, indem sie eine Vielzahl von Anwendungen ermöglichen, die die Erzeugung menschlich wirkender, kontextuell relevanter Texte erfordern. Durch umfangreiches Pretraining auf vielfältigen Daten zeigen diese Modelle eine bemerkenswerte Anpassungsfähigkeit in allgemeinen Aufgaben. In domänenspezifischen Anwendungen, wie etwa in den Bereichen Finanzen, Recht oder Medizin, generieren LLMs jedoch häufig Antworten, die zwar plausibel, aber faktisch inkorrekt sind. Dieses Problem ist insbesondere in Kontexten besorgniserregend, in denen Genauigkeit von entscheidender Bedeutung ist. Darüber hinaus ist es rechnerisch herausfordernd, die domänenspezifische Relevanz und die faktische Präzision innerhalb dieser großen, komplexen Modelle aufrechtzuerhalten, insbesondere wenn es darum geht, das notwendige Domänenwissen kontinuierlich zu aktualisieren. Diese Dissertation adressiert diese Herausforderungen durch adapterbasiertes Fine-Tuning für generative Fragenbeantwortung (QA) in mehreren Domänen, das darauf abzielt, Modularität, Effizienz und Anpassungsfähigkeit zu verbessern, geleitet von drei zentralen Forschungsfragen. Erstens wird untersucht, wie domänenspezifische Adapter die Leistung von vortrainierten Sprachmodellen in generativen QA-Einstellungen mit einer einzigen Domäne verbessern können. Zweitens wird das Potenzial einer Multi-Adapter-Konfiguration untersucht, um LLMs zu ermöglichen, Wissen über verschiedene Domänen hinweg zu integrieren, wodurch Modelle in der Lage sind, vielfältige Eingaben effektiv zu verarbeiten. Schließlich wird evaluiert, ob bestehende automatisierte Metriken zur Faktualität ausreichend robust sind, um die Modellleistung in einer generativen QA-Umgebung mit mehreren Domänen und offenen Fragen zu bewerten. Die Ergebnisse zeigen, dass der adapterbasierte Ansatz eine Leistung erreicht, die mit herkömmlichen Fine-Tuning-Methoden zur Domänenanpassung vergleichbar ist, wobei zusätzliche Vorteile wie schnellere Modellaktualisierungen und speichereffiziente Integration in Umgebungen mit begrenzten Ressourcen bestehen. Dennoch sind die aktuellen automatischen Metriken unzureichend, um die Faktualität generierter, offener Antworten zu bewerten, insbesondere wenn die Länge und Komplexität der Antworten über die Domänen hinweg variieren. Experimente mit LLM-basierten und multi-aspektualen Prompting-Metriken zeigen Einschränkungen auf und unterstreichen die Notwendigkeit zuverlässigerer Metriken zur Bewertung der faktischen Genauigkeit in der generativen QA über mehrere Domänen. Zusammenfassend legt diese Dissertation die Grundlage für eine effiziente, anpassungsfähige Strategie für Sprachmodelle in mehreren Domänen und fördert die Entwicklung modularer, wissensreicher generativer QA-Systeme, die den Anforderungen domänenspezifischer NLP-Anwendungen gerecht werden und gleichzeitig allgemeine Fähigkeiten beibehalten.

Contents

1. Introduction

The emergence of large language models (LLMs) has transformed the field of artificial intelligence and natural language processing (NLP). These advanced models excel at generating human-like text, enabling a wide range of applications from chatbots to content creation. As organizations increasingly seek to utilize these capabilities, the demand for domain-specific models has grown (S. Wu et al. 2023; Chalkidis et al. 2020; Fan et al. 2023). While general-purpose LLMs perform well across many tasks, they face significant limitations when adaptability is required.

In specialized fields such as finance or medicine, without sufficient guidance, LLMs can generate responses that appear plausible yet contain factually unsupported content, also known as hallucinations (L. Huang et al. 2023), while the need for accurate and context-specific outputs is critical. Moreover, a system that generates clinical summaries (G. Wang et al. 2023), for example, must not only integrate vast amounts of specialized knowledge but also be flexible enough to update this information frequently, as knowledge in these specialized domains is ever-changing. However, general-purpose LLMs often struggle to keep pace with the latest knowledge, resulting in outputs that can quickly become outdated or irrelevant (K. Huang et al. 2024).

The computational demands of training and maintaining LLMs further hinder their use in specialized tasks (Wan et al. 2024). As the complexity of these applications increases, so does the model's resource consumption, making it less practical for environments with limited computing power. These practical constraints including hardware limitations, data availability, bandwidth restrictions, and generation time significantly affect the performance of these models (Z. Lu et al. 2024). Moreover, relying on retrieval-based methods or external access to knowledge sources for generating content also raises concerns about consistency, reliability, and privacy (C. V. Nguyen et al. 2024).

In light of these considerations, this thesis aims to explore methods that enhance the adaptability, modularity, and efficiency of language models. By focusing on the shortcomings of general-purpose models, we seek to bridge the gap between their broad capabilities and the specialized needs of various applications, ultimately advancing the field of NLP in multi-domain text generation settings.

1.1. Research Questions

Building upon the insights discussed in the preceding section, this thesis seeks to address several research questions (RQs):

RQ1 How well domain-specific adapters improve the performance of pretrained language models in generative question-answering tasks for a single domain?

RQ2 Can language models acquire multi-domain knowledge in a multiple-adapter setting where each adapter is trained on distinct domain-specific knowledge?

RQ3 Are current automatic evaluation metrics for factuality comprehensive enough to assess the performance over a multi-domain generative QA setting?

1.2. Contribution and Outline

In this thesis, we investigate the adapter fine-tuning approach within generative settings, a domain that has received limited attention. Specifically, we focus on the generative question answering (QA) task, where answers are presented in long-form text. Our exploration includes a multi-domain approach, training both domain and task adapters within a multi-adapter framework utilizing a hard-gating mechanism rather than fusing adapters. We also assess the effectiveness of common evaluation metrics for text generation, emphasizing the unique requirements of QA tasks, particularly in knowledge-sensitive domains where factuality is crucial. Our research aims to establish a foundation for future studies on more efficient multi-domain adaptation and adapter-based methods in generative QA tasks. Chapter 2 provides a comprehensive background and related work on language models, parameter-efficient fine-tuning, domain adaptation, question answering, and factuality evaluation. Chapter 3 outlines the methodology of our experiments, detailing the datasets, models, training stages, inference pipeline, and evaluation setups, including the metrics used. Chapter 4 presents the results of our experiments, along with a discussion of the main findings, while Chapter 5 concludes the thesis by addressing limitations and suggesting avenues for further research.

2. Background and Related Work

In this chapter, we provide an overview of the key concepts relevant to this thesis, alongside a discussion of recent works closely related to the research questions. Given the vast scope of Natural Language Processing (NLP) research, this chapter focuses specifically on transformer-based language models, parameter-efficient methods, domain adaptation, question answering and factuality evaluation. We followed a semi-systematic approach in the literature review for this thesis. The methodology is described in Appendix A.1.

2.1. Language Models

2.1.1. Overview

Early neural language models (NLMs) based on variations of recurrent neural networks (RNNs) architecture, such as long short-term memory (LSTM) and gated recurrent unit (GRU), were applied to many NLP tasks including text generation, text-to-text generation, and text classification (Minaee et al. 2024). Then, with its innovative attention mechanism, Transformer architecture has significantly advanced NLP research, enabling impressive performance of Pretrained Language Models (PLMs) on both standard benchmarks and open-ended generation tasks (Chang and Bergen 2023).

Attention is one of the crucial elements of the Transformer's performance, as it assigns weights to all input representations and identifies the most important parts of it. Transformer's ability to scale through parallelization is also another key factor as it facilitates the development of much larger models (C. Zhou et al. 2023). These advancements laid the groundwork for the development of highly capable PLMs such as GPT-4 (OpenAI 2024), Gemini (Google 2024), and OPT (S. Zhang et al. 2022), which have widespread use across various areas of Machine Learning.

While the distinction between PLMs and Large Language Models (LLMs) has been introduced in recent literature (Z. Lu et al. 2024), primarily based on model scale, the criteria for what qualifies as *large* is evolving rapidly with continued advancements. Given this ongoing shift, this thesis will employ PLMs as a comprehensive term that refers to deep neural networks trained on extensive, unlabeled corpora to acquire general language understanding. These models can then be fine-tuned for specific downstream tasks, enabling them to adapt effectively to diverse applications across natural language processing domains.

2.1.2. Standard Architectures

PLMs can be classified based on their architecture, particularly in how they handle input encoding and output decoding. The three primary types are masked language models, causal language models, and prefix language models, employing distinct attention masking strategies to optimize for various language processing tasks.

Masked LM

Masked LMs use a Transformer encoder to capture bidirectional context by predicting masked tokens from surrounding information, with BERT (Devlin et al. 2019) as a notable example for natural language understanding tasks. However, due to a mismatch in pretraining objectives, Masked LMs are rarely used directly for text generation, which requires sequential prediction (Junyi Li et al. 2021). Instead, they often serve as encoders in generation models, leveraging their strong contextual representations.

Causal LM

Causal LMs use a diagonal mask matrix to predict each word based only on previous tokens, making them suitable for tasks like language modeling and text generation. GPT (Radford and Narasimhan 2018) was the first causal LM designed for generation, with GPT-2(Radford, J. Wu, et al. 2019) and GPT-3(Brown et al. 2020) showing that larger models improve zero-shot and prompt-based generation capabilities. CTRL (Keskar et al. 2019) extended this approach, enabling controlled text generation through style or task codes. However, causal LMs' unidirectional encoding limits their ability to handle tasks needing full context. In tasks like summarization and translation, they are generally outperformed by models optimized for sequence-to-sequence generation (Junyi Li et al. 2021).

GPT Family Generative Pretrained Transformers (GPT), developed by OpenAI[1], are a decoder-only model family that includes models such as GPT-1, GPT-2, GPT-3, InstructGPT, ChatGPT, GPT-4, CODEX, and WebGPT. GPT-2 (Radford, J. Wu, et al. 2019) demonstrated that language models could learn specific natural language tasks without explicit supervision when trained on large datasets, improving upon GPT-1 with modifications like changes to layer normalization, residual path adjustments, and an expanded vocabulary of 50,257 tokens and a larger context size of 1,024 tokens, which enhanced its ability to handle longer text sequences. Released in March 2023, GPT-4 (OpenAI 2024) is the most advanced model in the GPT family, featuring multimodal capabilities that allow it to process both text and image inputs. While it remains less capable than humans in some complex tasks, GPT-4 achieves human-level performance on various benchmarks, including scoring in the top 10% on a simulated bar exam (Minaee et al. 2024). Like its predecessors, GPT-4 was pretrained on extensive text corpora and fine-tuned using reinforcement learning from human feedback

[1] `https://openai.com/`

(RLHF). Unlike earlier models, like GPT-2, GPT-4 is closed-source and available only through API, which limits their accessibility for independent research.

LLaMA Family LLaMA is a collection of foundation language models, released by Meta[2]. The earliest set of LLaMA (Touvron et al. 2023) models was introduced in February 2023 as a family of transformer-based autoregressive causal LMs, featuring parameter sizes ranging from 7 billion to 65 billion. Pretrained on trillions of tokens from publicly available datasets, LLaMA adopts a transformer architecture similar to GPT-3 but incorporates several key differences, including the SwiGLU activation function instead of ReLU, rotary positional embeddings instead of absolute positional embeddings, and root-mean-squared layer normalization rather than standard layer normalization (Minaee et al. 2024). Notably, the open-source LLaMA-13B model outperforms the proprietary GPT-3 model (175 billion parameters) on most benchmarks, establishing it as a strong baseline for research in large language models (LLMs).

Encoder-Decoder

Encoder-decoder LMs consist of both encoder and decoder Transformer layers. During pretraining, models like ProphetNet (Qi et al. 2020) use a sequence with a masked segment as input for the encoder, while the decoder autoregressively generates the masked tokens. T5 (Raffel et al. 2023) improves this by randomly replacing spans in the source text with special tokens for sequential prediction. BART (M. Lewis et al. 2019) employs a denoising autoencoder approach, learning to reconstruct the original text from various corrupted inputs, such as sentence permutation and token deletion. This architecture effectively leverages both encoding and decoding processes for a variety of text generation tasks.

Prefix LM

Prefix LMs use hybrid attention, with bidirectional encoding for input tokens and left-to-right, unidirectional attention for output tokens, on a single Transformer. This mixed attention mask allows input tokens to attend to each other, while output tokens attend only to the input and previous tokens. UniLM (Dong et al. 2019), the first prefix LM, applied this setup for conditional generation tasks. Later, UniLMv2 (Bao et al. 2020) and GLM (Du et al. 2022) incorporated permuted language modeling to capture richer dependencies. However, separate encoder-decoder structures with cross-attention better capture conditional relationships, outperforming single-transformer prefix models in conditional generation tasks (Junyi Li et al. 2021).

2.1.3. Model Pretraining

Pretraining is a crucial initial phase in the training process of language models, enabling them to develop foundational language understanding. This stage typically involves training

[2] `https://ai.meta.com/`

on vast amounts of unlabeled text data using self-supervised learning techniques. Common approaches include next token prediction (autoregressive language modeling) and masked language modeling (MLM), both of which help the model acquire general language representations applicable to various language tasks (Minaee et al. 2024). In recent developments, Mixture of Experts (MoE) models (Shazeer et al. 2017) have gained significance reducing computational resources required for pretraining, allowing for larger model scales or datasets to be used without increasing the computation requirements.

Commonly used open-source pretraining datasets for training language models include:

- The Pile (Gao et al. 2020) (825 billion tokens): A comprehensive dataset that merges smaller corpora from various domains to provide diverse training material.
- C4 (Raffel et al. 2023) (1.4 trillion tokens): A cleaned version of Common Crawl's[3] web crawl corpus, which is composed over billions of web pages, metadata, and text extracts spanning years, in over 40 languages.
- FineWeb-Edu (Penedo et al. 2024) (1.3 trillion tokens): A collection of educational texts filtered from the FineWeb dataset, outperforming all publicly available web datasets on several educational benchmarks.
- Cosmopedia (Ben Allal et al. 2024) (25 billion tokens): This dataset consists of synthetic texts, including textbooks, blog posts, stories, and WikiHow[4] articles, generated by the Mistral 7B model (Jiang et al. 2023), contributing to a wide range of knowledge.

2.1.4. Fine-Tuning PLMs

Fine-tuning is a key technique in transfer learning, where the knowledge gained from pretrained neural networks is leveraged to solve new, domain-specific tasks. In the context of NLP, fine-tuning involves adapting large pretrained language models, such as BERT (Devlin et al. 2019) or GPT, to particular tasks like text classification, sentiment analysis, and named entity recognition. By starting from a pretrained model, fine-tuning allows models to learn task-specific representations more efficiently, significantly reducing both the training time and the amount of task-specific data required (H. Zheng et al. 2023). This approach often results in enhanced performance, as the model builds upon the general language understanding gained during pretraining and applies it to solve specialized problems.

It is important to note that most generative PLMs are initially pretrained using language modeling objectives before being fine-tuned on task-specific objectives for text generation tasks. This divergence between the pretraining and fine-tuning stages can influence the performance of PLMs in downstream tasks. Prompt tuning (P. Liu et al. 2021) has emerged as a solution to this performance limitation by transforming the downstream task into a language modeling task and making it consistent with the pretraining objectives.

[3] `https://commoncrawl.org/`
[4] `https://wikihow.com/`

2.1.5. Limitations of PLMs

As PLMs become more prevalent in various applications, it is essential to acknowledge their limitations. Key challenges include efficiency concerns, hallucination, and catastrophic forgetting. Understanding these limitations is crucial for advancing the effectiveness of language models.

Efficiency

Large models, although powerful, can be costly and inefficient in several areas, such as exhibiting high latency. In response to these challenges, there is an increasing trend in research towards the development of Small Language Models (Z. Lu et al. 2024). These models serve as cost-effective alternatives to larger models, particularly for specific tasks that do not require the extensive capabilities of large-scale PLMs. In parallel, there is a growing interest in exploring cost-effective approaches, including parameter-efficient training, knowledge distillation, and quantization, to enhance the training and usage of the larger models. In section 2.2, we provide a deeper insight into the parameter-efficient approach and Adapters.

Catastrophic Forgetting

A significant challenge in fine tuning models is catastrophic forgetting (Y. Luo et al. 2024), which occurs when a model forgets previously learned information as it adapts its parameters for new tasks, particularly in continual learning settings or when the pretraining data are inaccessible. This can lead to reduced performance on earlier tasks, impacting the model's overall effectiveness. It is observed that as the model scale increases, the intensity of catastrophic forgetting also increases. One possible explanation for this trend is that larger language models start with stronger performance, which leads to more significant declines in performance during continual instruction tuning. Y. Luo et al. 2024 further proposed that incorporating diverse instruction tuning can help alleviate the catastrophic forgetting phenomenon.

Hallucination

Hallucination in language models refers to instances where the generated content is either nonsensical or unfaithful to the input source (Maynez et al. 2020). This phenomenon can be categorized into two main types: intrinsic and extrinsic hallucinations (W. Li et al. 2022). Intrinsic hallucinations occur when the generated content conflicts with the provided source, while extrinsic hallucinations arise when the model generates content that cannot be verified against the source. In addition to intrinsic and extrinsic hallucinations, hallucinations in language models can also be distinguished based on factuality and faithfulness (L. Huang et al. 2023). Factual hallucination involves discrepancies with verifiable real-world facts, resulting in fabrications or factual inconsistencies. Faithfulness hallucination refers to divergence from user instructions or the context provided, impacting the accuracy and coherence of the

Type	Question	Answer	Explanation
Insufficient Domain Knowledge	What causes phenylketonuria?	Phenylketonuria is caused by **a mutation in the hemoglobin gene**, which results in an unusual interaction with hemoglobin, affecting its ability to transport oxygen in the bloodstream.	Phenylketonuria (PKU) is actually caused by **mutations in the gene that helps make an enzyme called phenylalanine hydroxylase**. Due to the lack of specialized medical knowledge in LLM's training data, it may sometimes provide incorrect information on medical topics.
Outdated Knowledge	Which city hosted the most recent Winter Olympics?	The most recent city to host the Winter Olympics was **Pyeongchang, South Korea, in 2018**.	The most recent city to host the Winter Olympics was **Beijing, in 2022**. This is due to the fact that LLM stores outdated knowledge, and the answer to this question exhibits a time-shift phenomenon.

Table 2.1.: Example of knowledge boundaries (L. Huang et al. 2023). In bold text, the problematic parts are highlighted.

generated content.

Hallucinations in language models arise from various factors, one of which is their knowledge boundaries (L. Huang et al. 2023). While extensive pretraining corpora provide a wealth of factual knowledge, these models have limitations in two areas: a lack of up-to-date information and insufficient specialized domain knowledge. Illustrating these two common types of hallucination behavior, Table 2.1 presents two question-answer pairs. The given explanations in the table highlight the need for models that are both specialized in domain knowledge and capable of easily updating their knowledge.

2.2. Parameter Efficient Fine Tuning

2.2.1. Overview

As the parameter size of the pretrained language models (PLM) becomes greater, scaling up to over 100 billion, the computational resource requirements are making it harder to utilize such LLMs. The need for task-specific full fine-tuning, especially with models like Falcon-180B (Almazrouei et al. 2023), significantly increases computational demands, potentially requiring at least 5120GB of computational unit (Xu et al. 2023). Besides the training requirements, scaling up model size also increases the inference latency which challenges the operational

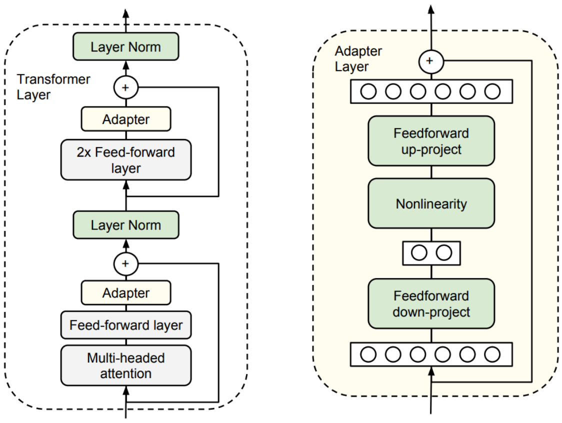

Figure 2.1.: Architecture of the adapter module proposed by Houlsby et al. 2019 and its integration with the Transformer, known as the *Houlsby* adapter. This is one of two commonly used bottleneck adapter architectures, alongside Pfeiffer, Vulić, et al. 2020's *Pfeiffer* adapter.

cost of such enormous LLMs. These high resource requirements indicate the importance of developing methods that improve the efficiency of language models (Wan et al. 2024).

One significant framework to address the efficiency need is called parameter efficient fine tuning (PEFT). In general, PEFT updates only a limited number of additional parameters or adjusts a subset of the pretrained parameters, preserving the knowledge embedded in the PLM while adapting it to the target task, language, or domain. This helps minimize the risk of catastrophic forgetting. Additionally, since the fine-tuned dataset is typically smaller than the pretrained dataset, full fine-tuning may lead to overfitting, an issue that PEFT mitigates by selectively updating or not updating certain pretrained parameters (Xu et al. 2023). The prominent approaches such as Sequential Adapters (Houlsby et al. 2019), Prefix Tuning (X. L. Li and Liang 2021) and Low-rank Adaptation (Hu et al. 2021) have shown that PEFT can successfully reduce the number of trainable parameters while maintaining comparable performance to the full fine-tuning.

2.2.2. Adapters

Training the adapter modules injects additional parameters into Transformer blocks and freezes the rest of the parameters during the fine tuning process. This structure makes the adapter-based methods preserve old knowledge from the base language model's pretraining and allowing parameters of the original network to be shared across multiple tasks (Houlsby et al. 2019).

More specifically for the sequential adapters, adapter modules are inserted after multi-head attention and feed-forward layers of the Transformer (see Figure 2.1). They include a down-projection, a non-linear activation function, an up-projection, and a residual connection, which makes each adapter a low-rank module. Given an input X, the output of a sequential adapter with a non-linear activation function can be expressed by (2.1).

$$X = (ReLU(X * W_{down})) * W_{up} + X, W_{down} \in R^{d \times k}, W_{up} \in R^{k \times d}. \tag{2.1}$$

During adapter fine-tuning, only the parameters W_{up} and W_{down} are adjusted to adapt to the chosen downstream task. This bottleneck architecture helps limit the parameter size between 0.5 - 8 of the size of the core model. With typical hyper-parameter settings, adapters can be up to 60 faster than full model fine-tuning (Rücklé et al. 2021).

Other than sequential adapters, many adapter-based approaches have been proposed as the modules present various architectural choices. AdapterDrop (Rücklé et al. 2021) improves inference efficiency by removing adapters in each Transformer layer that are not crucial for the given task. Parallel Adapter (He et al. 2022) integrates the adapter network in parallel with both the attention and feed-forward layers, enabling more efficient incorporation of the adapter module into the Transformer.

For a wide range of NLU tasks, most language models have relied on distributional information from large corpora. Adapters, however, also offer an important method for incorporating knowledge from more structured sources, such as knowledge bases. Unlike traditional fine-tuning, adapters help prevent catastrophic forgetting of the knowledge acquired during pretraining (Lauscher et al. 2020). In one recent work C. Liu et al. 2023 combines medical knowledge adapters and continual learning to balance abilities of acquiring the new knowledge and remembering the old knowledge.

Multi-adapter Frameworks

The following paragraphs review recent advancements in multi-adapter frameworks, presenting a range of approaches from diverse adapter training and composition techniques to applications in downstream tasks. These works illustrate the evolving strategies for using the modularity of adapters across varied tasks and domains in NLP.

AdapterSoup The AdapterSoup (Chronopoulou et al. 2023) is one of the prominent work in ensembling adapters. It has demonstrated enhanced performance in novel domains by effectively combining multiple adapters trained across different domains. Employing text clustering, AdapterSoup identifies optimal adapter combinations during testing. Specifically, their domain clustering technique aligns with the approach outlined by Aharoni and Goldberg 2020, wherein they encode 100 samples from all training domains using a PLM and fit a Gaussian Mixture Model, with the number of components corresponding to the number of training domains. During testing, an adapter is added to the *soup* if at least 10% of the test samples map to the corresponding domain cluster. The variety of training domains is emphasized as a critical factor in the success of this multiple adapter framework. Their experimental setup utilizes the small version of GPT-2, incorporating an adapter into each transformer block after the feed-forward network layer, and training the adapters solely for language modeling. The overall results show that the model performs well on a specific domain while preserving the initial performance of the PLM across unknown domains as a lower bound and further improves out-of-domain generalizability.

Mixture-of-Domain-Adapters The recent work on Mixture-of-Domain-Adapters (Diao et al. 2023) explores the emerging paradigm of combining multiple domain-specific adapters to enhance model performance. This study focuses on two key aspects: generalizability within a known domain, particularly for unseen examples, and adversarial robustness when mixing adapters from very different tasks. An important consideration raised in this research is that not all mixtures of adapters yield superior performance, highlighting the necessity of understanding when and what to mix. The findings indicate that weight-space merging can result in performance degradation concerning both generalizability and adversarial robustness, emphasizing the need for careful selection in the mixture of adapters to optimize model efficacy.

Gated Adapters Connected to the emerging paradigm of combining multiple adapters to enhance model performance, Klimaszewski, Belligoli, Kumar, et al. 2023 propose gated adapters for multi-domain machine translation. This study introduces a novel gating mechanism designed to improve translation quality through knowledge distillation, enabling soft gating via multiple adapter triggering. The effectiveness of this method is evaluated through experiments involving two language pairs: English to Polish and English to Greek, with each pair encompassing six distinct domains. Performance metrics utilized in this study include COMET (Rei et al. 2020), BLEU (Papineni et al. 2002), and chrF (Popović 2015), providing a comprehensive assessment of the model's translation capabilities across varied linguistic contexts.

Domain Knowledge Integration from Multiple Sources The Diverse Adapters for Knowledge Integration (DAKI) framework (Q. Lu, Dou, and T. H. Nguyen 2021) addresses the limitations of knowledge-enhanced PLMs by differentiating between unstructured free texts and structured external knowledge sources. Recognizing the challenges posed by training models from scratch and optimizing entire architectures, the authors highlight that many existing approaches often incorporate single-source knowledge, neglecting the potential benefits of multiple sources and formats. To enhance knowledge integration, they employ an attention-based knowledge controller module that adaptively adjusts the activation levels of various adapters, thus improving the model's performance on tasks involving biomedical information, such as those derived from UMLS and Wikipedia.

2.3. Domain Adaptation

2.3.1. Overview

A domain is characterized by a corpus originating from a specific source and may vary in aspects such as topic, genre, style, and level of formality, among others (Klimaszewski, Belligoli, Kumar, et al. 2023). The general-purpose language models trained on massive data sourced from many domains have drawbacks with regard to flexibility. Their direct application to domain-specific downstream tasks yields sub-optimal performance due to changes in vocabulary and style of language (J. Lee et al. 2019), especially for the domains that are not represented well in the training data. A common strategy to achieve better performance for specific domains involves pretraining these language models over domain-specific data using objectives such as Masked Language Modelling (MLM) and Autoregressive Language Modelling. In domains of biomedical literature, finance, and law, PubMedBERT (Gu et al. 2021), FLANG-BERT (Shah et al. 2022), and LEGALBERT (Chalkidis et al. 2020) achieved state-of-the-art results. However, obtaining labels for each task is costly and time-intensive. To this end, domain adaptation offers more efficient and faster adaptation to domains.

Building on the concept of domain adaptation, in this section, we highlight key approaches that enable effective knowledge transfer from pretrained models to specific domains. Continual learning, prompting, and prefix tuning are important strategies that facilitate this adaptation. The following subsections will explore these approaches and their contributions to enhancing domain adaptation.

Continual Learning

Continual learning aims to adapt PLMs to specific domains by extending the original pretraining task using large, unlabeled corpora (Guo and H. Yu 2022). During this phase, the PLM continues training on domain-specific datasets with original language modeling objectives. When domain-specific data is limited, this objective often acts as a regularization term within downstream task loss functions.

Various approaches exist based on the purpose of continual pretraining. One of them is the pretrain-finetune approach (Guo and H. Yu 2022). Research has shown (Gururangan et al. 2020)that it is beneficial to customize a pretrained model to align with the domain of a target task by conducting an additional phase of pretraining, applicable to both high- and low-resource contexts. For example, the cross-domain adapted BERT (Rietzler et al. 2019) demonstrates that performing domain-specific language modeling followed by supervised task-specific fine tuning can significantly enhance performance in aspect-based sentiment classification. Vocabulary adaptation is also part of the continual learning paradigm. Texts in specialized domains often contain many unique terms that are not present in general-domain corpora and thus may not be adequately represented in the vocabulary of PLMs (Gu et al. 2021). Adapting PLMs with domain-specific vocabularies can facilitate their adaptation to new fields. For instance, training the embeddings of special tokens in GPT-2 enables its application in task-oriented dialogue systems without training new dialogue-specific modules (Budzianowski and Vulić 2019).

Continual pretraining is closely linked to lifelong learning or continual adaptation (Parisi et al. 2019), wherein a general model is persistently fine-tuned to accommodate new domains while retaining previously acquired knowledge. For example, recent works have adopted this approach by training models like BART across different domains for text generation (Thompson et al. 2019). However, despite its simplicity and ease of deployment, this learning paradigm often incurs the challenge of catastrophic forgetting. Blindly continuing to pretrain a given PLM on a target domain can lead to this issue (Guo and H. Yu 2022). Studies have shown (T. Yu, Z. Liu, and Fung 2021) that the dissimilarity between pretraining data and the target domain task can diminish the effectiveness of BART in tasks like abstractive summarization. Further efforts have been made to enhance domain adaptation, alternating continual pretraining approach. Task-adaptive pretraining is proposed which shows the benefits of using far smaller and much more task-relevant pretraining corpus (Gururangan et al. 2020). Another method has highlighted the limits of domain-agnostic pretraining and proposed using tailored pretraining objectives to exploit the specific linguistic features of the target domain (Nair and Modani 2023), which performed well in the legal domain.

Prompting

Recently, prompt methods have emerged as a novel paradigm for adapting PLMs to downstream tasks. Prompting involves supplying PLMs with additional contextual information about the data, such as task descriptions, which enhances the input data and guides the model's responses. When a prompt is well-designed for a specific task, a pre-trained language model (PLM) can accurately generate the appropriate label words based on its inherent language modeling knowledge. This approach has shown remarkable few-shot performance across various datasets (Brown et al. 2020).

Prefix Tuning

Due to the limitations imposed by the context window length, soft prompt methods have been introduced to address this issue (Guo and H. Yu 2022). The key distinction between prompt tuning and prefix tuning (X. L. Li and Liang 2021) lies in the integration of parameters: prefix tuning incorporates prefix parameters across all layers of the model, while prompt tuning exclusively adds these parameters to the model's input embeddings. This approach allows subsequent tokens to attend to a small, continuous, task-specific vector, called "prefix". These prefix prompts, along with a task-specific linear head are then trained, offering a more resource-efficient alternative for model adaptation.

Prefix Domain Adaptation In recent work, Jonathan Li, Bhambhoria, and Zhu 2022 proposed the prefix domain adaptation method, a promising approach inspired by traditional domain adaptation techniques. This method leverages a deep prompt trained for the masked language modeling task on a large, domain-specific corpus to create a compact, domain-adapted initialization (around 0.1% of the base model size). Such a prompt is both efficient to store and share, enabling versatile use across tasks. Following unsupervised pretraining, the deep prompt supports downstream tasks with a randomly initialized, task-specific head, such as for classification, further fine-tuned through prefix tuning for task-specific objectives. This approach represents one of the earliest applications of domain-specific prefix prompts initialized via unsupervised pretraining, enhancing adaptability and generalization in downstream performance.

2.3.2. Further Works

More recent works employed retrieval-based methods for enhancing the performance of LLMs in specific domains (Peng et al. 2023). Despite their remarkable performance, they suffer from several limitations. Handling complex queries can lead to incomplete information retrieval. Furthermore, retrieved information can be lack of capturing intended meaning fully. They can retrieve chunks that contain the correct keywords of the query without understanding the question, leading to incorrect responses. Prompt length, again, poses another limitation for the in-context learning settings (F. Yang et al. 2023).

One promising recent work is proposed by F. Yang et al. 2023. The main domain adaptation strategy in their work is transferring knowledge from a smaller LLM fine-tuned on domain-specific documentation to a larger LLM at run-time. The generation of relevant domain knowledge reduces the risk of data leakage during fine-tuning unlike methods involving retrieval. This approach is advantageous particularly when handling restricted and sensitive domain-specific data.

2.4. Question Answering

2.4.1. Overview

Question Answering (QA) is a core task in NLP, aimed at developing systems capable of responding to questions in natural language. It is difficult to determine a single criterion for classifying QA tasks. These systems are often classified based on multiple dimensions such as the format of the answer, the skill that the task requires, the modality of the evidence/knowledge, and the domain (Rogers, Gardner, and Augenstein 2023). To simplify this partially ambiguous classification, we will mainly refer to the types of QA as extractive and generative (C. Zhou et al. 2023).

With this classification in mind, the modality of the evidence/knowledge source serves as a crucial factor influencing performance across various QA tasks. The evidence may be in the form of unstructured text, tables, structured knowledge bases (KBs), images, videos, and cases where no direct evidence is provided (Rogers, Gardner, and Augenstein 2023).

QA models can also be part of a conversational system, where the model not only answers individual questions but also maintains a coherent dialogue across multiple turns, making the interaction more natural and context-aware. However, in this section, our focus is on settings where there is an explicit question that the model is supposed to answer in a single turn.

Extractive QA

Since answers are directly derived from a provided context, enabling models to develop reading comprehension skills, extractive QA can be regarded as a specific instance of a text classification task. A common answer format for this type is an extracted span from the provided context, such as in Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al. 2016), RACE (Lai et al. 2017) and QASPER (Dasigi et al. 2021). Whereas in multiple-choice format, a set of predefined options is presented, from which the correct answer is selected. CommonsenseQA is a well-known example of multiple-choice type (Talmor et al. 2019). Another answer format is categorical, involving choosing from a set of fixed categories such as yes and no. BoolQ dataset (Clark et al. 2019) and a subcategory of PubMedQA (Jin et al. 2019) belong to this subcategory.

Retrieval-based Approaches

Retrieval-based QA involves sourcing relevant documents or text segments from large datasets, knowledge bases, or the internet, which augment the input query. In a foundational work, D. Chen et al. 2017 presented the DrQA model, which combines a document retriever to select relevant paragraphs from Wikipedia and a document reader to generate answers. The document reader utilizes a standard machine reading comprehension framework, allowing for the flexible integration of other models. Similarly, Dense Passage Retrieval (DPR) (Karpukhin et al. 2020) employs dense vector representations for retrieving paragraphs relevant to a query from large datasets, enhancing semantic information by leveraging dense vectors to calculate similarity scores. More recently, retrieval-augmented generation (RAG) has gained significant traction, particularly in knowledge-intensive (P. Lewis et al. 2021) and open-domain QA (Guu et al. 2020), which retrieve relevant text documents based on the input sequence and use them as additional context when generating the output, unlike retriever-reader models.

Domain-specificity

QA systems also differ based on their specific domains. They may operate in open-domain settings, where questions cover a broad range of topics, or in domain-specific settings, focusing on specialized fields such as medicine (Singhal et al. 2023) or law (Louis, Dijck, and Spanakis 2023). Some systems are designed for multi-domain adaptation (Engländer et al. 2024), enabling them to handle multiple specialized areas simultaneously.

2.4.2. Generative QA

Extractive or retrieval-based models often face challenges in effectively transferring knowledge from pretrained language models, as these models do not closely mimic the pretraining task. In contrast, generative QA models facilitate a more effective transfer of pretrained knowledge, given that their tasks align more closely with the pretext tasks (Zafar et al. 2024). Additionally, generative QA models tend to generate more engaging and human-like responses, making them particularly advantageous for end systems that prioritize user interaction.

It is important to understand that generative QA systems act as open-ended text generation tasks, which leads to specific challenges. One major issue is their tendency to hallucinate, generating responses that may lack essential information. This problem is especially concerning in fields like medicine, where accurate information is more important than fluency (Junyi Li et al. 2021). On the other hand, even when hallucinations are avoided, responses may still fail to encompass all potential references relevant to a question. This implies that a comprehensive evaluation approach for generative QA tasks cannot rely solely on human-annotated references. As a potential solution, Farea et al. 2022 suggests enhancing the evaluation process with a range of alternative expressions to better capture the diverse outputs these systems can generate.

2.5. Factuality Evaluation

Input data for text generation tasks often contains factual information, making it essential for the generated content to align with these original facts (Junyi Li et al. 2021). Evaluating factual accuracy is important for assessing the performance of language models, especially in applications where reliable information is critical. In this section, we outline approaches for evaluating general text generation quality, followed by a review of methodologies specific to factuality evaluation, including current challenges and metrics in the field. This background provides a comprehensive basis for understanding how factuality is assessed in generated outputs.

2.5.1. Evaluating Generated Text

Text generation evaluation is generally divided into two main approaches: human evaluation and automatic evaluation. Automatic evaluation can be further broken down into three categories: lexical, semantic, and LLM-based. Lexical evaluation focuses on word-level properties, such as accuracy in word choice and grammar. Semantic evaluation assesses the quality of meaning representation in the generated text, ensuring that it conveys the intended message. LLM-based evaluation utilizes large language models to assess the quality of the generated text, offering a more sophisticated, model-driven evaluation metric.

Human Evaluation

Human rating is the most widely used method for human evaluation in text generation tasks. It involves a well-defined process where a group of domain-specific human raters assesses the generated answers based on a set of clear, objective criteria (Srivastava and Memon 2024). These criteria typically include factors such as factual correctness, relevance to the question and its intent, completeness and informativeness of the response, clarity, and fluency of the language used, as well as the originality of the answer.

Human evaluation remains the gold standard for assessing the quality of generative question-answering systems due to the nuanced language comprehension capabilities of human evaluators. These evaluators can assess not only the factual accuracy of a system's responses but also their contextual appropriateness, capturing subtleties that automated methods may overlook (Srivastava and Memon 2024). However, this method has limitations. Inter-rater reliability (IRR) (McHugh 2012), a statistical measure of agreement among evaluators, is crucial for ensuring consistency and validity. Low IRR can introduce variability and reduce the generalizability of the results, with factors such as rater training, subjectivity, and unclear evaluation criteria influencing outcomes. Additionally, human evaluation is time-consuming and resource-intensive, especially when applied to large-scale datasets.

Lexical Matching

Metrics such as BLEU (Papineni et al. 2002), ROUGE (Lin 2004), and METEOR (Banerjee and Lavie 2005) operate by calculating the degree of overlap between n-grams in the generated output and reference texts, providing a quantifiable measure of similarity. This approach has proven effective in machine translation, where the objective is often to generate text that closely resembles human-written translations (Srivastava and Memon 2024). While these metrics can effectively represent fluency and surface similarity, they may fail to capture other crucial dimensions of evaluating generated text.

Semantic Similarity

In response to limitations of lexical evaluation metrics, approaches based on semantic similarity have gained traction, shifting the focus from surface-level comparisons to deeper evaluations of meaning. Semantic evaluation is often framed as a classification task where the goal is to determine whether a reference answer and a predicted answer are semantically equivalent (Srivastava and Memon 2024). This type of evaluation has proven valuable in measuring the effectiveness of various natural language generation (NLG) tasks, such as machine translation, image captioning, dialogue systems, and text summarization. One notable example is BERTScore (T. Zhang et al. 2020), which employs pretrained contextual embeddings from BERT to assess sentences by comparing the cosine similarities of their tokens' embeddings.

A. Chen et al. 2019 argued that while semantic similarity metrics are effective for evaluating tasks such as summarization and translation, their applicability to QA evaluation is limited, as they often show poor correlation with human judgments. Although works such as C. Wang et al. 2023's, attempted to enhance semantic similarity assessment by incorporating both questions and answers in the similarity computation, they ultimately found that these metrics remained insufficient for accurately evaluating QA performance. They emphasized the metrics' sensitivity to the choice of similarity threshold, which can introduce inconsistencies in evaluation outcomes.

LLM-based Evaluation

Given the remarkable advancements in text comprehension and instruction-following abilities of recent LLMs, a growing body of research (Yang Liu et al. 2023; Chiang and H.-Y. Lee 2023; Fu et al. 2023) has explored using LLMs as evaluators for assessing the quality of responses in various NLG tasks, such as open-ended generation and summarization (Shen et al. 2023). This approach has demonstrated that LLMs can effectively simulate human evaluation patterns, thus offering a scalable and efficient alternative to the labor-intensive and costly nature of traditional human evaluations.

Although LLM-based evaluations show promise, they also have significant limitations. LLMs are heavily influenced by their prompts, which affect the generated outputs. It is examined that "LLM-as-a-judge" approaches tend to bring several biases and challenges (L. Zheng et al. 2023), including positional bias, where preference for certain positions can distort evaluation results; verbosity bias, which is the tendency to favor longer, verbose responses over more concise answers that may be clearer or more accurate; and self-enhancement bias, where models exhibit a bias toward their own generated answers. Additionally, LLMs face limitations in grading mathematical problems and logical reasoning (Cobbe et al. 2021). These issues underscore the difficulties of relying solely on LLMs for evaluative judgments.

GPTScore Utilizing the emergent abilities of generative pretrained models such as zero-shot instruction and in-context learning, the GPTScore framework achieves multi-aspect and training-free evaluation (Fu et al. 2023). Its evaluation protocol consists of two components: task specifications outlining the context of text generation and aspect definitions detailing desirable criteria (i.e. faithfulness). The results highlight the effectiveness of this new evaluation paradigm, revealing several key observations. First, the reliability of generative models improves when tasks and aspects are clearly defined, allowing for greater flexibility across various criteria. Additionally, incorporating examples alongside in-context learning enhances the evaluation process, resulting in more accurate assessments. Furthermore, the findings indicate that certain correlations exist among different evaluation aspects, suggesting that combining definitions of highly correlated aspects can improve overall evaluation performance and lead to a more nuanced understanding of generated texts.

2.5.2. Factuality

Recent advanced LLMs achieve impressive text fluency, primarily due to extensive pretraining and ranking-based methods like reinforcement learning from human feedback (Ziegler et al. 2020). However, they remain prone to hallucinating (Tian et al. 2023). Efforts to improve factuality in text generation have included fine-tuning with auto-generated factual preference rankings (Tian et al. 2023), using factual-nucleus sampling (N. Lee et al. 2023), and training models to self-assess for faithfulness (Kadavath et al. 2022). Despite these advancements, considerable gaps remain, particularly in developing more effective evaluation metrics. Common metrics, whether n-gram-based or using neural models like BERTScore (T. Zhang et al. 2020), often struggle to accurately measure factuality (Aharoni, Narayan, et al. 2023). Additionally, entailment-based Natural Language Inference scores commonly used for faithfulness evaluation (Maynez et al. 2020) lack interpretability and may inherit biases from the underlying PLMs.

Recent Approaches on Factuality Evaluation

FACTSCORE FACTSCORE (Factual precision in Atomicity Score) decomposes generated content into atomic facts and calculates the percentage of these facts that are verifiable against a reliable knowledge source (Min et al. 2023). To validate this metric, the authors conducted a comprehensive human evaluation of biographies generated by several leading language models, including InstructGPT (Ouyang et al. 2022), ChatGPT (OpenAI n.d.), and the retrieval-augmented PerplexityAI[5], highlighting the need for such granular assessment (e.g., ChatGPT achieved only 58% factual accuracy). Recognizing the high cost of human evaluation, the paper also introduces an automated model to estimate FACTSCORE by using retrieval and a robust language model, achieving accuracy within 2% of human ratings. It is important to note that FACTSCORE has limitations in scenarios where the generated content involves nuanced, open-ended, or debatable facts.

SAFE Building on the FACTSCORE approach, SAFE (Wei et al. 2024) proposed a comprehensive evaluation approach that utilizes an LLM to assess the factuality of atomic facts by referencing information retrieved from the Google search engine.

SelfCheckGPT This work approaches the issue of factuality through hallucination in model outputs (Manakul, Liusie, and Gales 2023). The authors introduce a sampling-based method to detect hallucinated versus factual responses in LLMs without relying on external resources, making it suitable for black-box systems. The underlying concept of SelfCheckGPT is that if a model has adequately learned a concept, multiple sampled responses should remain similar and factually consistent; in contrast, hallucinated content often results in divergent or contradictory responses. By sampling and comparing multiple outputs, SelfCheckGPT measures informational consistency to identify potential hallucinations. This

[5] `https://www.perplexity.ai/`

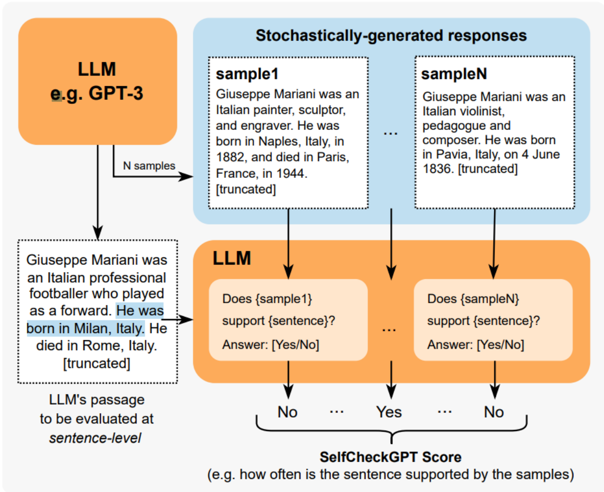

Figure 2.2.: SelfCheckGPT with Prompt, as presented in the Manakul, Liusie, and Gales 2023 paper. Each sentence generated by the LLM is compared against sampled responses without relying on an external database.

approach incorporates five methods to assess consistency: BERTScore, question-answering, n-gram overlap, natural language inference (NLI), and LLM prompting. Evaluation of GPT-3-generated text demonstrated SelfCheckGPT's effectiveness, often surpassing even grey-box techniques, and establishing a robust baseline for hallucination detection in LLMs.

3. Methodology

This chapter details the approaches used to investigate the research questions introduced in section 1.1, beginning with an overview of the datasets and their preparation for our experiments. Next, the experimental setup outlines the configurations and procedures applied to evaluate different models. Finally, the evaluation section describes the metrics and criteria used to assess model performance, setting the groundwork for the results in the following chapter.

3.1. Datasets

3.1.1. Overview

To demonstrate the robustness of our approach in various domain adaptation cases, we selected a diverse set of domains: biomedical, legal, and finance. These domains differ significantly in terms of style, level of formality, and domain-specific terminology (Klimaszewski, Belligoli, and Stergiadis 2023). The selected datasets vary in terms of the type of knowledge they encompass, their sources, and the inclusion of synthetic textual data. This diversity is intended to represent a broader range of data, thereby improving the model's ability to generalize to in-domain shifts.

3.1.2. Preparation Process

The datasets are preprocessed to ensure a consistent training setup across experiments. This preprocessing includes several steps: truncating long text fields, standardizing naming conventions, removing duplicates, eliminating empty values, and shuffling entries. Additionally, for some datasets, we implement further modifications to align with our research focus. Notably, we establish a lower bound for reference answer text length to mitigate the prevalence of brief, token-based responses, as we aim to work with more open-ended, longer-form answers for our experiments.

During the dataset selection process, we utilize open-source platforms, including Kaggle[1], Hugging Face[2], and Papers with Code[3]. However, we encountered challenges in accessing datasets specific to the legal and finance domains, which were less readily available. This

[1]`https://www.kaggle.com/`
[2]`https://huggingface.co/datasets`
[3]`https://paperswithcode.com/`

Question	Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?
Context	(Objective) Recent studies have demonstrated that statins have pleiotropic effects, including anti-inflammatory effects and atrial fibrillation (AF) preventive effects [...] (Methods) 221 patients underwent CABG in our hospital from 2004 to 2007. 14 patients with preoperative AF and 4 patients with concomitant valve surgery [...] (Results) The overall incidence of postoperative AF was 26%. Postoperative AF was significantly lower in the Statin group compared with the Non-statin group (16% versus 33%, p=0.005). Multivariate analysis demonstrated that independent predictors of AF [...]
Long Answer	Our study indicated that preoperative statin therapy seems to reduce AF development after CABG.
Answer	yes

Table 3.1.: An instance from PubMedQA dataset.

limitation prompted us to combine several datasets, enhancing the potential generalizability of the models to be trained.

3.1.3. Domain and Characteristics

We utilized datasets with diverse sizes, characteristics, formulations, evidence, and source types. This selection serves multiple purposes, addressing the knowledge-intensive nature, specialized terminology, and subjective elements of open-ended questions and answers. By incorporating a broad range of data, we aim to assess the robustness of our approach across varied scenarios. The summary of the datasets can be seen in Table 3.4.

Biomedical Domain

PubMedQA PubMedQA (Jin et al. 2019) is a biomedical QA dataset designed for answering research questions derived from article titles, with long answers extracted from the conclusion sections of the articles (see Table 3.1). We adapted this dataset for a generative QA task by using the long answers as reference answers and applying an open generative format where no context is provided. The context column of the dataset, along with other datasets, was used during domain adapter training to fine-tune the model for domain-specific adaptation. In a second approach, we separated the results from the context to use it in a closed generative format, where context was provided to guide the answer generation.

Other Biomedical Datasets We incorporate several additional datasets as unlabeled data or labeled question-answer pairs in the domain adapter training process:

- The BioASQ corpus [4], annotated by biomedical experts, comprises multiple question-

[4] `http://participants-area.bioasq.org/datasets/`

Question	Could my parents change my birth name without me knowing?
Answer	This generally requires a court order (everything depends on jurisdiction: this is a state matter, not a federal matter). As a minor, the courts could allow your parents to change from Dweezil to William without involving you, until you are old enough that the judge thinks you might be able to have reasonable input into the matter. Once you're over 18, your parents can't change your name – you would have to do that, at least if you are mentally competent.

Table 3.2.: A open-ended QA instance from Jonathan Li, Bhambhoria, and Zhu 2022's Law Stack Exchange dataset, highlighting the ambiguity and variable length of possible responses.

answering tasks, where questions are typically scientific in nature (Krithara et al. 2023). BioASQ provides various types of QA instances along with relevant context, some of which we utilize as unlabeled data during domain adapter training.

- We also augment our training sets with samples from datasets curated by Fries et al. 2022, which provides an open library[5] of biomedical data loaders built using Hugging Face's datasets library. This includes MedNLI (Romanov and Shivade 2018), a dataset annotated by medical professionals that focuses on natural language inference tasks based on patients' medical histories.

Legal Domain

Legal Community QA Datasets Legal forums, such as Legal Advice Reddit[6] and Law Stack Exchange[7], offer valuable data sources for legal NLP tasks, containing vast collections of community-based question-answer pairs that support domain adaptation. For our work, we utilized two datasets processed and curated by Jonathan Li, Bhambhoria, and Zhu 2022, sourcing these forums and containing legal questions paired with high-quality answers. While questions in the Legal Advice Reddit dataset are often informal, these datasets provide substantial legal insights from expert responses, effectively supporting domain adaptation. In contrast, the Law Stack Exchange dataset is generally more formal, with questions often theoretical or hypothetical (see Table 3.2), creating a balanced set when combined with the other.

Further Unlabeled Domain Data In addition to the community-sourced datasets mentioned in the previous paragraph, we utilized several other legal domain datasets to enhance our domain adapter training with more extensive, unlabeled data:

[5] `https://huggingface.co/bigbio`
[6] `https://www.reddit.com/r/legaladvice/`
[7] `https://law.stackexchange.com/`

Question	Why are big companies like Apple or Google not included in the Dow Jones Industrial Average (DJIA) index?
Answer	That is a pretty exclusive club and for the most part they are not interested in highly volatile companies like Apple and Google. Sure, IBM is part of the DJIA, but that is about as stalwart as you can get these days. The typical profile for a DJIA stock would be one that pays fairly predictable dividends, has been around since money was invented, and are not going anywhere unless the apocalypse really happens this year.

Table 3.3.: An opinion-based QA instance from FiQA dataset.

- The European Court of Human Rights (ECHR) dataset (Chalkidis et al. 2020) offers a collection of legal case facts.
- The CUAD dataset (Hendrycks et al. 2021), developed by The Atticus Project with input from numerous legal professionals, provides resources specifically for legal contract review.
- The COLIEE dataset (Kano et al. 2017) includes a series of legal questions, each accompanied by multiple alternative responses, further supporting our domain adaptation efforts.

Financial Domain

For the financial domain, we used unlabeled data from a finance corpus (H. Yang, X.-Y. Liu, and C. D. Wang 2023) covering economic news, phrases and topics, and a finance term encyclopedia from Investopedia[8] curated by Z. Zhou, Ma, and H. Liu 2021, for domain adapter training. For task adapter training and evaluation, we included labeled data from the FiQA dataset (Maia et al. 2018), featuring finance-focused question-answer pairs and augmented with stock trade query-context pairs (Tang and Y. Yang 2024) to enhance the variety of sources.

3.2. Experimental Setup

3.2.1. Overview

The experimental setup of our methodology is structured around three key components: the training of domain-specific knowledge adapters, the training of task-specific adapters, and the selection of the appropriate domain-task adapter tuple during inference. This two-stage approach leverages the modularity and flexibility of adapter architecture, allowing for more effective adaptation to varying domains and tasks. By decoupling domain and task

[8] `https://www.investopedia.com/`

Dataset	Domain	Stage	Size
PubMedQA	Biomedical	1, 2	270,000
BioASQ	Biomedical	1	40,000
TradeEncylopedia	Finance	1	5700
PhraseBank	Finance	1	2300
Investopedia	Finance	1	220,000
FiQA	Finance	2	6600
TradeQA	Finance	2	7000
Law Stack Exchange	Legal	2	25,000
Legal Advice	Legal	1, 2	165,000
ECHR	Legal	1	23,000
CUAD	Legal	1	84,000
OpusLaw	Legal	1	475,000

Table 3.4.: Dataset summary, domain, size, and their utilization stage in training domain and task adapters.

adaptations, we enhance the performance and efficiency of the model, allowing each adapter to focus on its specific function while enabling effective combination during inference.

Model and Adapter Configuration Details

- We use two different PLMs that vary in parameter size and architecture as our base adapter models, GPT-2 and Llama 3.2. For the GPT-2 model, we choose the 355M parameter version (medium) to investigate the approach on a smaller PLM. For the Llama 3.2 model, the pretrained text-only version with 1B parameters is used. To make training further efficient, we use quantization with the help of *bitsandbytes* wrapper that Dettmers et al. 2022 released. The pretrained checkpoints of the models can be seen in Appendix A.2.
- We configured Pfeiffer architecture sequential bottleneck adapters, as proposed in Pfeiffer, Vulić, et al. 2020, placing an adapter layer only after the feed-forward layer in each Transformer block.
- We set the reduction factor of layers to 16, which reduced the trainable parameter size up to 1% of original models (see A.2)

3.2.2. Target Domain Adaptation

For each experiment domain, we trained Pfeiffer adapters on unlabeled domain-specific data using a causal language modeling objective while keeping the model weights frozen. The purpose of this stage was to enable the adapter modules to acquire domain-specific knowledge on top of the base model's general knowledge. These trained adapters were then retained for the task adaptation stage, where they were added to the model without their prediction heads.

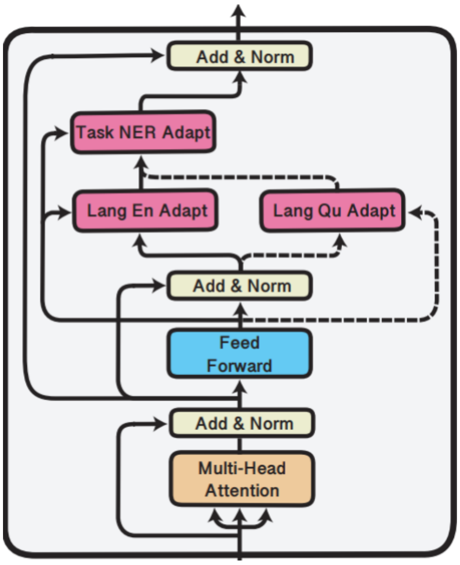

Figure 3.1.: The MAD-X framework. They stack task-specific adapters on top of source language adapters when training on a downstream task, aiming multi-task cross-lingual transfer (Pfeiffer, Vulić, et al. 2020).

3.2.3. Task Adaptation

To model the QA task as causal language modeling, inputs are structured in the format: "answer the following question: [question] answer: [answer]". Following the work of MAD-X (Pfeiffer, Vulić, et al. 2020), where task adapters are stacked on top of language adapters (see Figure 3.1), and of M2QA (Engländer et al. 2024), we use a stacked composition of adapters for task adaptation, where task adapters are added on top of previously trained domain adapters. This setup involves freezing the base model and domain adapter parameters, allowing only task-specific layers to be fine-tuned. Training is performed for three to five epochs with early stopping, depending on the dataset size for each task.

3.2.4. Adapter Selection

In this section, we outline the adapter selection mechanisms implemented for our multi-domain QA framework, detailing approaches designed to ensure the most relevant domain-specific adapters are activated during inference. First, we implemented a method involving a trained domain classifier, which predicts the relevant adapter for each input. Separately, we leveraged sentence transformers to generate domain-specific embeddings, enabling selection through similarity comparisons at inference. Additionally, we reviewed strategies from recent research that combine and weight multiple adapters dynamically, often using soft-gating mechanisms to adjust adapter contributions.

Training Domain Classifier

This approach involved training a DistilRoBERTa (Sanh et al. 2019) model on a balanced dataset constructed from question samples across three domains (biomedical, legal, and finance), with a fourth label representing open-domain questions. The balanced set ensured that the classifier was exposed equally to each domain and open-domain samples. During inference, this small classifier was used to predict the domain of the input question, aiding in the activation of the appropriate domain adapter.

Leveraging Sentence Transformers

In our adapter selection mechanism, we drew inspiration from the gating mechanism commonly used in MoE models. In this general framework, there is a step for sampling a few instructions from each expert, followed by calculating the distance between the input prompt and the *expert* centroids Adapting these principles, we designed a method that selects domain adapters based on the distance between input query and average text embeddings of each *domain*.

Our approach involves the following steps:

- Dataset Subset Selection: We randomly choose a subset of questions from the training QA datasets (described in the previous section) for each domain: biomedical, legal, and finance.
- Embedding Creation: Using the Sentence Transformers framework (Reimers and Gurevych 2019), we encode the text embeddings of these questions with a pretrained model such as *paraphrase-MiniLM-L6-v2*[9] from HuggingFace.
- Averaging Embedding: For each domain, we compute the average of the question embeddings, creating a single representative embedding for each domain. These averages are stored for use during inference.
- Testing Similarity: To evaluate how well these average embeddings represent their respective domains, we calculate the cosine similarity between the saved averages and the embeddings of additional sample questions, which were not included in the averaging step. The results and details of this testing process are provided in the Results chapter.

At inference time, we use the stored average embeddings to classify the domain of an input question:

- Domain Classification: We compare the embedding of the input question with the three average embeddings, selecting the domain with the highest similarity score, provided it exceeds a set lower-bound threshold. This threshold helps identify questions that may not belong to any of the three predefined domains.

[9] `https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2`

- Handling Out-of-Domain Questions: If no domain meets the threshold, indicating potential out-of-domain input, we explore two strategies: either using no active adapters and relying on the model's base knowledge, or averaging the three domain adapters using the Adapters library's built-in Average function to handle the input.

This method allows for flexible domain selection and ensures robust handling of domain ambiguity during inference.

Soft Gating Strategies

Recent works have proposed alternative mechanisms for selecting and combining adapters in various setups:

- Diao et al. 2023 propose a mixture-of-adapters (MoA) gate to dynamically integrate outputs from the attention layer, feed-forward network (FFN), and domain adapters. The MoA gate processes the attention output through a multi-layer perceptron (MLP) to generate intermediate weights. These weights are then used to assign varying importance to the FFNs and domain adapters. The final output is computed as a weighted sum of the outputs from these components.
- Q. Lu, Dou, and T. H. Nguyen 2021 propose a knowledge controller module to adaptively integrate the outputs of the knowledge adapters. The controller functions as an independent adapter augmented with additional linear layers. It assigns varying importance weights to each knowledge adapter, instead of a simple concatenation of their outputs (R. Wang et al. 2020). At each layer where adapters are positioned, attention weights are calculated for each adapter, enabling a weighted sum of the hidden states to be processed and yield the final output of the knowledge controller for that layer.
- In another recent work, Sparse Mixture of Adapters to Mitigate Multiple Dataset Biases, a sparse gating network, similar to MoE, is proposed, which computes weights for each sub-adapter and only keeps the top k weights to multiply with the output of the sub-adapters (Yanchen Liu et al. 2023).

3.2.5. Inference Pipeline

The inference pipeline in our experimental setup loads both the trained domain and task adapters into the base model. For each input query, we calculate its embedding and compare it against stored average embeddings for each domain, selecting the domain with the highest similarity score. The chosen domain adapter is then activated, with the task adapter stacked on top. We set the generation parameters to include a low temperature (0.1) and apply a repetition penalty, alongside limiting output to a maximum token length. To handle cases where the model might continue generating content beyond the answer—such as beginning a new question-answer sequence—we implemented a process to isolate only the answer segment. In the case of the LLaMA model, a custom stopping function was also integrated to ensure the generation stops appropriately after completing the answer.

3.2.6. Supporting Platforms and Resources

It is important to acknowledge some of the platforms and resources we primarily utilized during the experiments:

- Hugging Face Hub[10] serves as a comprehensive hub for transformer models and a robust datasets library, providing a vast collection of pretrained models and datasets that facilitated the training and evaluation processes. Its well-documented resources enabled efficient experimentation of state-of-the-art language models for our methodology.
- AdapterHub (Pfeiffer, Rücklé, et al. 2020) is a framework designed to simplify the integration, training, and utilization of adapters and other efficient fine-tuning techniques for Transformer-based language models.
- Google Colab[11] is a hosted Jupyter Notebook service that offers access to a variety of computing resources, which is essential for training language models that demand significant computational power. We run most of our experiments on the Colab environment using a T4 GPU with 16GB RAM.

3.3. Evaluation

3.3.1. Metrics

This section outlines the evaluation metrics used to assess our models' performance. We employed perplexity, ROUGE, BERTScore, FactCC, and a GPT-4 prompting-based approach, focusing on various aspects such as relevance, support, and factuality. Different settings involved combinations of generated answers over context passages or gold answers, allowing for a thorough evaluation of model performance.

Perplexity

Perplexity is a metric used to evaluate language models by measuring how well a model predicts a sample of text; a lower score means the model predicts word sequences more accurately. It is commonly used in assessing language modeling tasks. In our study, we used perplexity to evaluate how effective our domain adapter training was in comparison to the base model.

[10] `https://huggingface.co/`
[11] `https://research.google/resources/`

ROUGE

ROUGE compares the n-grams, word sequences, or longest common subsequences between a generated text and one or more reference texts, focusing on how many n-grams from the reference appear in the generated output (Lin 2004). ROUGE includes variations: ROUGE-N considers n-gram overlap and ROUGE-L, the longest common subsequence. We report on ROUGE-1 for informativeness and ROUGE-L, for fluency on lexical level.

BLEU

BLEU assesses how closely a generated text matches one or more reference texts by comparing overlapping n-grams (sequences of words) between them (Papineni et al. 2002). It calculates the proportion of n-gram matches, with matches counted regardless of their position in the text, making the evaluation position independent. BLEU also employs brevity penalties to discourage excessively short generations, balancing n-gram precision with fluency (Farea et al. 2022).

BERTScore

BERTScore measures the similarity between generated and reference texts by comparing tokens through contextual embeddings rather than exact word matches (T. Zhang et al. 2020). It calculates token-level similarity using models like BERT, capturing deeper semantic relationships. Originally successful in machine translation, we adapted BERTScore for generative QA, to evaluate contextual relevance between generated and reference answers and explore its relationship to the factuality aspect.

FactCC

Kryscinski et al. 2020 propose a weakly-supervised, model-based approach, called FactCC, for verifying factual consistency and identifying conflicts between source documents and generated summaries. The training data is generated through rule-based transformations applied to sentences from the source documents. The factual consistency model is trained jointly on three tasks: consistency prediction, determining if each summary sentence is factually consistent; span extraction for support, which extracts a supporting span from the source document for each consistency prediction; and inconsistency span extraction, where inconsistent summary sentences yield the corresponding inconsistent span from the summary itself. Although developed as a metric for summaries, we applied and evaluated its effectiveness in a generative QA setting, using the provided context as the source for each question.

Prompt-based approach

LLMs are also another alternative source of auxiliary information as they possess an implicit memory encompassing a vast array of knowledge (Roberts, Raffel, and Shazeer 2020). In this

Zero-shot Prompt: *You are evaluating generated text, in three scoring objectives: first score indicating* **relevancy** *of the generated answer to the question, the second score indicating how much of generated text can be* **supported** *by the reference answer, and the third score indicating* **factuality** *evaluation based on your knowledge regardless of the reference. Give the scores from 0 to 10, and do not explain.*

Table 3.5.: Example prompt used in multi-aspect factuality evaluation.

context, we implemented a prompt-based approach using the GPT-4 model via OpenAI's text completion API, focusing on key aspects related to hallucination—specifically, support, relevance, and factuality (see Table 3.5). We experimented with both zero-shot and few-shot prompting methods, where we either provided no example or included a few scored examples to guide the model's assessment. To ensure stable results, we maintained a low-temperature setting across evaluations. When dataset context was available, it was incorporated into the prompt for more context-aware assessments; otherwise, we referenced only the provided answers while specifying the domain context to help the model focus on relevance and factuality without relying on additional context. We also experimented with multi-aspect prompts to help the model better distinguish between evaluation criteria. This approach was particularly beneficial for aspects not directly related to the main focus, such as completeness, coherence, and fluency, supporting a more nuanced assessment across multiple dimensions.

3.3.2. Evaluating Selection Strategy

To evaluate the effectiveness of our adapter selection mechanism, we conducted a test using a random sample of 200 questions from each domain, labeled with their respective domain names. The mechanism operates by leveraging the average embeddings previously generated and stored for each domain type. For each test sample, we calculated the similarity score between the question embedding and the stored average embedding of each domain. The domain with the highest similarity score was returned as the predicted label, enabling an accuracy assessment based on how often the mechanism successfully matched the question to its correct domain. This straightforward approach allowed us to quantify the accuracy of our domain selection strategy in classifying questions according to domain-specific embeddings.

3.3.3. Baseline Models

We also incorporated some open-source, domain-specific language models into our evaluation setup:

- BioGPT (R. Luo et al. 2022), a domain-specific generative Transformer model tailored for the biomedical field, was pretrained on an extensive corpus of biomedical literature. Its performance was evaluated across six biomedical NLP tasks, consistently outperforming existing models in most cases. Notably, BioGPT achieved F1 scores of 44.98%, 38.42%, and 40.76% on the BC5CDR, KD-DTI, and DDI end-to-end relation extraction tasks, respectively, and set a new benchmark with 78.2% accuracy on PubMedQA. Additionally, a case study on text generation demonstrated BioGPT's ability to produce fluent and accurate descriptions of biomedical terms, further highlighting its potential for biomedical language processing. In our experiments, we tested the BioGPT model on 200 sample question-long answer pairs, using the same generation hyperparameters as those applied to other methods. This setup allowed for a direct performance comparison with our domain and task adapter-activated GPT-2 model, offering insights into the relative effectiveness of adapter-based fine-tuning for domain-specific generation tasks.

4. Results

This chapter provides the results of our experimental stages, starting with performance metrics across various tasks and examples of generated outputs displayed in tables. Following this, we discuss and analyze both the strengths and shortcomings observed in the generated outputs, as well as in the overall experimental results, connecting these findings to our research questions.

4.1. Research Questions

To provide clarity and continuity in our analysis, we will revisit the research questions (RQ) guiding this study before presenting detailed results. This reminder of our central questions serves to orient the reader, linking the upcoming findings to the objectives that motivated each experiment:

RQ1 How well domain-specific adapters improve the performance of pretrained language models in generative question-answering tasks for a single domain?

RQ2 Can language models acquire multi-domain knowledge in a multiple-adapter setting where each adapter is trained on distinct domain-specific knowledge?

RQ3 Are current automatic evaluation metrics for factuality in generative QA comprehensive enough to assess the performance over a multiple-domain generative QA setting?

4.2. Findings

This section presents the results of our experiments, systematically addressing the research questions posed in this thesis. Specifically,

- To address RQ1 and RQ2 (see section 4.1), we report performance gains across each evaluation domain in Table 4.1 and Table 4.2, using all metrics outlined in Section 3.3, These comparisons span two configurations: the PLM with only task adapters activated, and the PLM with both domain and task adapters activated in a stacked arrangement. Since our selection mechanism activates adapters specific to each domain, the resulting performance improvements can also be examined individually, providing a clear view of the isolated impact of domain-specific adapter activation.
- Building on RQ2, we reported the results of our adapter selection mechanism, summarized in Table 4.3, indicating high accuracy in identifying and activating the correct domain adapters during inference.

Domain	Model	ROUGE-1	ROUGE-L	BLEU	BERTScore	FactCC	Relevance	Support	Factuality
Biomed	TA	26	16	36	75	48	50	48	45
	TA+DA	25	17	33	80	52	46	49	53
Legal	TA	13	8	25	73	49	38	23	37
	TA+DA	20	12	23	73	53	40	27	42
Finance	TA	28	23	34	82	52	68	62	52
	DA+TA	37	31	44	86	60	75	70	60

Table 4.1.: Evaluation results for GPT2 model with domain and task adapters in comparison to base model evaluations

Domain	Model	ROUGE-1	ROUGE-L	BLEU	BERTScore	FactCC	Relevance	Support	Factuality
Biomed	TA	19	10	26	70	52	60	42	59
	TA+DA	28	14	25	78	65	58	50	58
Finance	TA	42	32	47	86	88	85	70	80
	DA+TA	39	29	35	86	91	85	74	83
Legal	TA	16	9	39	73	57	32	19	36
	TA+DA	19	10	43	76	55	55	34	59

Table 4.2.: Evaluation results for Llama model with domain and task adapters in comparison to base model evaluations

- Perplexity scores before and after each stage of training, reported in Table 4.4 for each domain, provide insight into the model's familiarity with domain-specific terminology.
- In the biomedical domain, we report evaluation results using a fully fine-tuned model, BioGPT (see section 3.3, to benchmark the effectiveness of our domain adaptation approach (see Table 4.5). To ensure a focused comparison of domain adaptation performance, we isolated the task adapter during this evaluation.
- Notably, in Tables 4.5, 4.1, and 4.2, the scores that specified as relevance, support, and factuality purpose a more fine-grained and multi-aspect evaluation as discussed in 3.3. They provide further detailed insights related to all research questions in this context.

The results in Tables 4.5, 4.1, and 4.2 provide a solid foundation for analyzing RQ3. We specifically examine the inconsistencies in metrics across different models, domains, and training stages, along with the challenges encountered in evaluating the specific aspects of quality. In the following section, we analyze this further, along with the odd behaviors observed and the quality of both good and poor generated outputs.

4.3. Analysis

This section provides a detailed analysis of the evaluation results, examining the connections outlined to the research questions in the previous section and offering insights into the performance of our adapter-based multi-domain approach.

Accuracy of Selection Mechanism High accuracy scores in Table 4.3, highlight the effectiveness of using average embeddings from sentence transformers to guide adapter selection based on domain similarity. However, these findings are limited by the uniformity of our test samples, as they share a similar in-domain distribution to the initial average question set used for selection. Future experiments with more varied distributions across question samples could further validate the robustness of this approach and better assess its adaptability to diverse domain-specific inquiries.

	Biomedical	Finance	Legal
Accuracy	95	90	90

Table 4.3.: Accuracy scores of similarity with average text embeddings domain selection mechanism evaluation.

Perplexity Higher perplexity scores indicate that the model finds token predictions more 'surprising,' particularly when specialized terminology is dense, as seen in the Table 4.4 legal domain. We observed a decrease in perplexity after each stage of training, reflecting an improvement in the model's ability to predict domain-relevant tokens accurately. Given that the Llama model is much more capable of text generation than the GPT2 model, the differences between models were as expected.

	Biomedical			Finance			Legal		
	Base	DA	TA	Base	DA	TA	Base	DA	TA
GPT2	31	20	16	32	19	9	38	24	19
Llama3.2	21	12	9	27	14	12	19	9	5

Table 4.4.: Perplexity scores before and after each stage of adapter training with GPT2 model.

Effect of Domain Adapters The results in Table 4.5 support the fact that the domain adapter model achieved nearly comparable performance in answer generation to the fully fine-tuned BioGPT, despite requiring less than 1% of the original model's parameters—in this case, GPT-2.

Lexical Matching As anticipated, the results from lexical matching metrics, specifically ROUGE and BLEU, did not fully capture the quality of generated responses in the open-ended answer generation setting for either base model configuration. However, after incorporating domain and task-specific adapter training, these scores showed an upward trend, indicating that the model learned relevant n-grams during domain adapter fine-tuning. Interestingly, BLEU scores were consistently higher than ROUGE, reflecting the differing natures of these metrics and the specific evaluation context. Given that the generated outputs generally

Model	ROUGE-1	ROUGE-L	BLEU	BERTScore	FactCC	GPT4-v1	GPT4-v2	GPT4-v3
BioGPT	21	16	39	78	64	42	27	44
GPT2+DA	25	17	33	80	52	47	31	52

Table 4.5.: Comparison of the GPT model with domain adapter and the BioGPT model.

contain significantly fewer tokens than the reference answers or contexts, this disparity in length impacts recall and precision, leading to a more pronounced gap between ROUGE and BLEU scores.

BERTScore and FactCC Utilizing BERTScore helped us examine the limitations of the semantic similarity metric, which we anticipated would be more comprehensive than lexical matching for evaluating generative QA tasks. While this metric effectively assesses semantic alignment between the generated and reference answers, it falls short in capturing deeper aspects such as relevance, faithfulness, and factual accuracy. Despite these limitations, the semantic similarity metric provides some insight into the domain relevance of the generated answers, proving to be more effective than perplexity in this regard. On the other hand, FactCC, a model-based approach specifically designed for identifying conflicts, demonstrated inconsistent results when evaluated using only the reference answers. This inconsistency is understandable, as the metric primarily assesses the quality of summaries. However, in the biomedical domain, when provided with contextual information, the results from FactCC were more aligned with our LLM-based approach. One downside of the FactCC metric is that it tends to produce scores that are either overly confident (high) or excessively low. As a result, we attempted to interpret these scores as binary labels, but this approach ultimately led to a distribution centered around the score of 50.

4.3.1. LLM-based Evaluation Results

The observed variations in results across domains reflect several factors: the specific formulation of questions in each domain's training and evaluation data, the types of expected answers, and the evidence used for evaluating generation quality. For example, in the legal domain, evaluating outputs generated for ill-defined questions (see Table 4.7) proved particularly challenging when compared to reference answers. In such cases, the generated responses may have been of poor quality, yet they did not directly conflict with the reference answers, complicating the assessment of their overall effectiveness, as discussed in Section 2.4.

We also observed some high-quality outputs. Table 4.6 exemplifies a good example of our model's behavior; the particular scoring for this instance further demonstrates that, despite a lower support score from the reference text, the model assigns a higher factuality score. Especially in the Biomedical and Finance domains, for knowledge-centered questions, our test results show relatively better generative performance. One example of this can be seen in Table 4.9; yet, the relatively trivial and structured nature of the questions in these instances may overshadow the good performance, making it difficult to generalize the results to other

Question	Can prisoners of war be exchanged against their will?
Answer	In the context of international law, prisoners of war (POWs) are individuals who have been captured by an enemy and held in captivity. The Geneva Convention on the Law of Non-Applicable Warfare provides a framework for how POWs should be treated during warfare. This includes provisions related to the exchange of POWs between belligerents.
Reference	One option would be for the PoW to claim asylum as a refugee in the "capturing" country or, for example, via an international human rights organisation: Refugees are defined and protected in international law. The 1951 Refugee Convention is a key legal document and defines a refugee as: "someone who is unable or unwilling to return to their country of origin owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group, or political opinion." That said, things like political objectives, diplomatic negotiations, military strategies, and the prisoner's "value" (on both sides) may influence decisions around which prisoners are exchanged and when.

Table 4.6.: A generated answer example in Legal domain, from Llama model. The evaluation result for this instance shows that support from the reference text is lower, however model gives a higher factuality score.

contexts.

Other difficult types of questions to evaluate, particularly in limited context settings, include those that require listing (see Table 4.8). This difficulty aligns with typical patterns in generative QA, especially for open-ended responses. This challenge likely explains why our LLM-based evaluation metric (see section 3.3), which assesses quality based solely on the given question, more accurately captured the generation quality than other metrics. Having analyzed the challenges of evaluation and the complexities involved, we still encountered numerous instances of hallucinated answers, One example can be seen in Table 2.1.

4.3.2. Other Observations

Effect of Model Size Overall using the LLaMA model as the base yielded better results than the GPT-2 model, likely due to its larger parameter size, though still smaller than typical 7B parameter models. However, the absolute gains were less pronounced compared to experiments with GPT-2. Given LLaMA's inherent capacity to generate coherent and relevant text, due to its extensive intrinsic knowledge, adding an additional layer of domain-specific knowledge had a comparatively smaller impact. Nevertheless, we observed consistent improvements, particularly in the legal and finance domains, where domain-relevant adaptation remained beneficial.

Question	Can my business partner use the company's attorney in a partnership dispute, if the fees are paid by the company? Am I entitled to know discussion since it was paid for by company funds?
Answer	Partners have certain rights and obligations regarding their businesses that may be affected by your situation. If you're concerned about how your business is being managed or if there's any potential conflict of interest between you and your business partner, it might be beneficial to review the terms of your partnership agreement or consult with an attorney who specializes in corporate law.
Reference	Typically no because that is a conflict of interest for the attorney. The company is the client, not the shareholder and representation of the shareholder individually is usually a conflict.

Table 4.7.: A poorly stated question instance from the Legal domain, posing a challenging evaluation.

Question	What are some popular statistical arbitrage models used by traders?
Answer	Some popular statistical arbitrage models include the Fama-French three-factor model, the Jegadeesh and Titman two-factor model, and the Sharpe ratio model. Traders use these models to identify potential entry points for trading pairs.
Reference	Popular statistical arbitrage models include mean-reversion models, momentum-based models, Kalman filter models, machine learning algorithms, and pairs trading strategies based on statistical factors.

Table 4.8.: A listing type of question in Finance domain and generated output from Llama model with domain adapter activated. It shows the reference answer would not be enough to assess the quality of the answer.

Behaviour of the model An observable behavior in language models is their tendency to generate new questions that closely resemble the original query but fail to provide a meaningful answer. This phenomenon can be interpreted as the model's implicit admission of uncertainty, essentially indicating, "I have no idea." (Peinl and Wirth 2023). Unlike human respondents, these models lack the ability to acknowledge their limitations or confess their lack of knowledge. Our results indicate that some generation instances still exhibit this behavior, generating redundant questions in instances where they struggle to provide accurate or relevant responses, highlighting their challenges in delivering coherent and informative answers.

Question	What is a stock index?
Answer	A stock index is a measurement of the value of a section of the stock market, calculated from the prices of individual stocks. It provides an overview of how the market as a whole is performing and helps traders make informed decisions about investments. Examples include the S&P 500 Index.
Reference	A stock index is a measurement of the value of a section of the stock market, representing a particular market or sector.
Question	Implementation of epidural analgesia for labor: is the standard of effective analgesia reachable in all women?
Answer	Epidural analgesia during labor can be implemented effectively and safely with a high degree of patient satisfaction. The use of epidural analgesia should therefore become routine practice, especially among low-risk patients. However, further research is needed to identify factors that may influence the effectiveness of epidural analgesia.
Reference	Present audit shows that the process of implementation of labor analgesia was quick, successful and safe, notwithstanding the identification of one cluster of women with suboptimal response to epidural analgesia that need to be further studies, overall pregnant womens'adhesion to labor analgesia was satisfactory.

Table 4.9.: A comparison of two knowledge-intensive type instances from the Finance and Biomedical domain. It indicates the level of expertise required might have a significant role in handling such instances.

Question	Do tumor-infiltrating immune cell profiles and their change after neoadjuvant chemotherapy predict response and prognosis of breast cancer?
Hallucinated Answer	Tumor-infiltrating immune cell profiles predict breast cancer response and prognosis after neoadjuvant chemotherapy. *Increased CD8+ T cells and shifts in immune phenotypes correlate with better outcomes and lower recurrence rates, serving as valuable biomarkers for personalized treatment strategies.*
Reference Answer	Breast cancer immune cell subpopulation profiles, determined by immunohistochemistry-based computerized analysis, identify groups of patients characterized by high response (in the pre-treatment setting) and poor prognosis (in the post-treatment setting). Further understanding of the mechanisms underlying the distribution of immune cells and their changes after chemotherapy may contribute to the development of new immune-targeted therapies for breast cancer.

Table 4.10.: An hallucinated answer, generated for PubMedQA instance.

5. Conclusion

This thesis has explored some of the key concepts, under-explored task variants, and challenges in multi-adapter domain adaptation, providing insights into efficient performance and factuality evaluation.

Specifically, we investigated adapter fine-tuning methods on a generative question-answering task, focusing on adapting a general-purpose pretrained model to specific domains, i.e. biomedical, legal, and finance, using sequential bottleneck adapters. Our approach employed causal language modeling, leveraging two base models, GPT-2 and LLaMA 3.2, to compare the performance of adapters across varied parameter sizes. For each domain, we trained specialized domain adapters while keeping the base model weights frozen, allowing efficient parameter updates. In the second stage of training, task adapters were trained, being stacked on top of the domain adapters, with only the task adapter weights updated.

To handle a possible multi-domain test time distribution of inputs, we implemented a selection mechanism based on sentence transformers, average text embeddings, and sentence similarity. This allowed us to dynamically activate the appropriate domain and task adapters, and consequently, active parameters that will influence the output, based on the input question. Through evaluations, we assessed the effectiveness of this layered adapter setup in generating open-ended answers relevant to each domain.

Performance of Adapter-based Training Training language models using domain-specific knowledge with adapters has demonstrated comparable performance to fully fine-tuned models on downstream generative QA tasks across multiple evaluation metrics. This holds true even when trained with limited unstructured data and on constrained computational resources. Separating domain adaptation from task adaptation and adopting a modular approach allows our model to perform similarly in multi-domain settings as it does in single-domain scenarios when the number of domains is restricted.

Limitations of Evaluation Metrics While we observe overall gains with this approach, conventional metrics struggle to effectively assess the quality of generative QA, particularly for open-ended answers. LLM-based evaluations exhibit inconsistent behavior across domains and are sensitive to answer length. Prompting the model with multi-aspect requirements improves assessment capabilities; however, this approach also relies heavily on the quality of the prompts.

The promise of Adapter-based Multi-domain Adaptation Overall, the adapter-based multi-domain adaptation approach for generative QA tasks has yielded promising results and leveraged its additional advantages while maintaining comparable performance to fully fine-tuned models. This method is parameter-efficient, easily deployable in memory-constrained environments, and facilitates straightforward updates.

Despite these contributions, several limitations emerged, and there remain multiple avenues for further exploration, which will be addressed in the following sections.

5.1. Limitations

In this section, we discuss limitations ranging from methodological choices to data-related issues, are discussed to ensure a transparent understanding of the scope and applicability of the results.

5.1.1. Dataset Limitations

One limitation encountered during this research was the availability and quality of datasets for domain-specific generative question-answering tasks. In terms of quantity, the number of datasets available was limited, and those that were available often varied significantly in size, impacting the robustness of task adaptation. Moreover, the quality of these datasets presented further challenges. In several cases, the provided gold answers were not factually accurate or faithful to the source, which could negatively influence model training. Additionally, we observed that some datasets contained a high proportion of irrelevant data, including out-of-domain QA samples or numeric QA instances, which fell outside the scope of our training objectives. These issues collectively limited the performance of the model.

5.1.2. Adapters Library

The use of adapters in our experiments presented several challenges, largely due to the limited available resources for adapting and modifying their configurations. Documentation on how to alter the architecture of adapters or explore different hyperparameter setups was sparse. Beyond the material provided by AdapterHub, which itself did not extensively cover generative models, there was a notable lack of comprehensive guidance or external work addressing adapter-based approaches in generative contexts. This scarcity of resources made it more difficult to explore innovative configurations

5.1.3. Computational Resources

Despite the parameter efficiency of adapters, computational limitations posed significant challenges during our experiments. Access to sufficient computational resources restricted the size of the datasets, the number of training steps, and the extent of hyperparameter tuning we could perform. It also influenced our ability to experiment with larger core models, particularly those with 7B parameters and above. Many of the experiments were conducted on the Google Colab environment, which often led to disruptions due to inconsistent GPU access and run-time expiration. These constraints limited the scope of variations we could explore in terms of model architecture, dataset size, hyperparameter configurations, and large-scale evaluations.

5.2. Future Work

As NLP technology advances, there are numerous opportunities to refine existing methods, address current limitations, and expand the applications of adapter-based techniques and factuality evaluation. These directions will guide ongoing efforts to improve the adaptability and reliability of language models across diverse tasks.

Heterogeneous Knowledge Adapters Future work could focus on creating heterogeneous knowledge adapters that combine structured sources (i.e. knowledge graphs) with unstructured textual data across domains. Throughout our experiments, we observed a lack of factual data in our domain adaptation datasets. Training multiple adapters using various knowledge bases can enhance the effectiveness of this modular approach. This is particularly important in tasks such as open-ended and long-form question answering since questions might require multiple document levels of information. One recent study (Friedman, Dodge, and D. Chen 2021) proposes a similar approach by training multi-adapters, each specialized in a specific dataset. This could be adapted so that each adapter becomes an expert in a knowledge base, individual datasets, or domain corpora, thereby utilizing heterogeneous sources within a single model.

Leveraging Domain Adapters in Factuality Evaluation In this thesis, we also observed the lack of effective metrics for assessing factuality in generated responses. To advance factuality evaluation, future research should explore the integration of a mixture or combination of domain adapters into a factuality assessing framework. By employing specialized adapters, this approach may allow for more efficient and modular incorporation of domain knowledge, enhancing the assessment of factual accuracy across various domains in one model. Moreover, this can be integrated with a retriever approach to obtain context for verification using a multi-adapter fact-checking model.

A. General Addenda

A.1. Literature Review

To support our methodology, we conducted a literature review on multi-adapter frameworks, domain adaptation, generative question answering and the evaluation of factuality. We followed the procedure based on the guidelines established by Kitchenham et al. 2009 for systematic literature reviews in software engineering. In addition, we employed snowballing based on the guidelines by Wohlin 2014, a method in which the reference lists of papers or the citations to those papers are used to identify additional relevant studies.

Following the procedure,

- We searched articles from sources such as IEEE Xplore[1], ACM Digital Library[2], and ACL Anthology[3]
- Our search terms contain keywords like "multi-adapter" or "adapter-based" in relation to "domain adaptation," "generative question answering," and "factuality evaluation."
- Recent articles were prioritized to ensure the relevance of findings.
- Excluded works were non-peer-reviewed, outdated publishing, duplicate articles, and studies not related to domain adaptation, parameter efficient fine tuning, question answering, and factuality in language models.
- We applied an iterative process of forward and backward snowballing, continuously exploring references and citations of each selected work.

A.2. Implementation Details

During the experiments, we have set a specific seed and hyperparameters that can be important for repeating our work. For transformers, torch, and dataset shuffles, we use the same seed, which is 42. We also set this seed when we evaluated with GPT4 model.

Other hyperparameters and specifications:

[1] `https://ieeexplore.ieee.org/`
[2] `https://dl.acm.org/`
[3] `https://aclanthology.org/`

- During the adapter trainings, we used different ranges for epoch size from 2-5 depending on dataset size and overfitting. We set batch size between 16 and 32, and adopted a learning rate of 1e-4 as suggested in Pfeiffer, Rücklé, et al. 2020.
- For the checkpoints of the base models we have experimented with: openai-community/gpt2-medium[4] and meta-llama/Llama-3.2-1B[5]
- For adapters we specify a reduction factor of 16 which resulted in training only 3,171,840 parameters of the GPT2 model and 8,423,424 parameters for the Llama model.

[4] https://huggingface.co/openai-community/gpt2-medium
[5] https://huggingface.co/meta-llama/Llama-3.2-1B

List of Figures

List of Tables

Bibliography

Aharoni, Roee and Yoav Goldberg (July 2020). "Unsupervised Domain Clusters in Pretrained Language Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 7747–7763. DOI: 10.18653/v1/2020.acl-main.692. URL: https://aclanthology.org/2020.acl-main.692.

Aharoni, Roee, Shashi Narayan, et al. (July 2023). "Multilingual Summarization with Factual Consistency Evaluation". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 3562–3591. DOI: 10.18653/v1/2023.findings-acl.220. URL: https://aclanthology.org/2023.findings-acl.220.

Almazrouei, Ebtesam et al. (2023). *The Falcon Series of Open Language Models*. arXiv: 2311.16867 [cs.CL]. URL: https://arxiv.org/abs/2311.16867.

Banerjee, Satanjeev and Alon Lavie (June 2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ed. by Jade Goldstein et al. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 65–72. URL: https://aclanthology.org/W05-0909.

Bao, Hangbo et al. (2020). *UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training*. arXiv: 2002.12804 [cs.CL]. URL: https://arxiv.org/abs/2002.12804.

Ben Allal, Loubna et al. (2024). *Cosmopedia*. URL: https://huggingface.co/datasets/HuggingFaceTB/cosmopedia.

Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. arXiv: 2005.14165 [cs.CL]. URL: https://arxiv.org/abs/2005.14165.

Budzianowski, Paweł and Ivan Vulić (Nov. 2019). "Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems". In: *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Ed. by Alexandra Birch et al. Hong Kong: Association for Computational Linguistics, pp. 15–22. DOI: 10.18653/v1/D19-5602. URL: https://aclanthology.org/D19-5602.

Chalkidis, Ilias et al. (2020). *LEGAL-BERT: The Muppets straight out of Law School*. arXiv: 2010.02559 [cs.CL]. URL: https://arxiv.org/abs/2010.02559.

Chang, Tyler A. and Benjamin K. Bergen (2023). *Language Model Behavior: A Comprehensive Survey*. arXiv: 2303.11504 [cs.CL]. URL: https://arxiv.org/abs/2303.11504.

Chen, Anthony et al. (Nov. 2019). "Evaluating Question Answering Evaluation". In: *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Ed. by Adam Fisch et al. Hong Kong, China: Association for Computational Linguistics, pp. 119–124. DOI: 10.18653/v1/D19-5817. URL: https://aclanthology.org/D19-5817.

Chen, Danqi et al. (2017). *Reading Wikipedia to Answer Open-Domain Questions*. arXiv: 1704.00051 [cs.CL]. URL: https://arxiv.org/abs/1704.00051.
Chiang, Cheng-Han and Hung-Yi Lee (2023). *Can Large Language Models Be an Alternative to Human Evaluations?* arXiv: 2305.01937 [cs.CL]. URL: https://arxiv.org/abs/2305.01937.
Chronopoulou, Alexandra et al. (May 2023). "AdapterSoup: Weight Averaging to Improve Generalization of Pretrained Language Models". In: *Findings of the Association for Computational Linguistics: EACL 2023*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 2054–2063. DOI: 10.18653/v1/2023.findings-eacl.153. URL: https://aclanthology.org/2023.findings-eacl.153.
Clark, Christopher et al. (June 2019). "BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions". 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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2924–2936. DOI: 10.18653/v1/N19-1300. URL: https://aclanthology.org/N19-1300.
Cobbe, Karl et al. (2021). *Training Verifiers to Solve Math Word Problems*. arXiv: 2110.14168 [cs.LG]. URL: https://arxiv.org/abs/2110.14168.
Dasigi, Pradeep et al. (2021). *A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers*. arXiv: 2105.03011 [cs.CL]. URL: https://arxiv.org/abs/2105.03011.
Dettmers, Tim et al. (2022). *8-bit Optimizers via Block-wise Quantization*. arXiv: 2110.02861 [cs.LG]. URL: https://arxiv.org/abs/2110.02861.
Devlin, Jacob et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.
Diao, Shizhe et al. (2023). *Mixture-of-Domain-Adapters: Decoupling and Injecting Domain Knowledge to Pre-trained Language Models Memories*. arXiv: 2306.05406 [cs.CL]. URL: https://arxiv.org/abs/2306.05406.
Dong, Li et al. (2019). *Unified Language Model Pre-training for Natural Language Understanding and Generation*. arXiv: 1905.03197 [cs.CL]. URL: https://arxiv.org/abs/1905.03197.
Du, Zhengxiao et al. (2022). *GLM: General Language Model Pretraining with Autoregressive Blank Infilling*. arXiv: 2103.10360 [cs.CL]. URL: https://arxiv.org/abs/2103.10360.
Engländer, Leon et al. (2024). *M2QA: Multi-domain Multilingual Question Answering*. arXiv: 2407.01091 [cs.CL]. URL: https://arxiv.org/abs/2407.01091.
Fan, Angela et al. (2023). *Large Language Models for Software Engineering: Survey and Open Problems*. arXiv: 2310.03533 [cs.SE]. URL: https://arxiv.org/abs/2310.03533.
Farea, Amer et al. (2022). *Evaluation of Question Answering Systems: Complexity of judging a natural language*. arXiv: 2209.12617 [cs.CL]. URL: https://arxiv.org/abs/2209.12617.
Friedman, Dan, Ben Dodge, and Danqi Chen (2021). *Single-dataset Experts for Multi-dataset Question Answering*. arXiv: 2109.13880 [cs.CL]. URL: https://arxiv.org/abs/2109.13880.
Fries, Jason Alan et al. (2022). *BigBIO: A Framework for Data-Centric Biomedical Natural Language Processing*. arXiv: 2206.15076 [cs.CL]. URL: https://arxiv.org/abs/2206.15076.
Fu, Jinlan et al. (2023). *GPTScore: Evaluate as You Desire*. arXiv: 2302.04166 [cs.CL]. URL: https://arxiv.org/abs/2302.04166.

Gao, Leo et al. (2020). *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*. arXiv: 2101.00027 [cs.CL]. URL: https://arxiv.org/abs/2101.00027.

Google, Gemini Team (2024). *Gemini: A Family of Highly Capable Multimodal Models*. arXiv: 2312.11805 [cs.CL]. URL: https://arxiv.org/abs/2312.11805.

Gu, Yu et al. (Oct. 2021). "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing". In: *ACM Transactions on Computing for Healthcare* 3.1, pp. 1–23. ISSN: 2637-8051. DOI: 10.1145/3458754. URL: http://dx.doi.org/10.1145/3458754.

Guo, Xu and Han Yu (2022). *On the Domain Adaptation and Generalization of Pretrained Language Models: A Survey*. arXiv: 2211.03154 [cs.CL]. URL: https://arxiv.org/abs/2211.03154.

Gururangan, Suchin et al. (July 2020). "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 8342–8360. DOI: 10.18653/v1/2020.acl-main.740. URL: https://aclanthology.org/2020.acl-main.740.

Guu, Kelvin et al. (2020). *REALM: Retrieval-Augmented Language Model Pre-Training*. arXiv: 2002.08909 [cs.CL]. URL: https://arxiv.org/abs/2002.08909.

He, Junxian et al. (2022). *Towards a Unified View of Parameter-Efficient Transfer Learning*. arXiv: 2110.04366 [cs.CL]. URL: https://arxiv.org/abs/2110.04366.

Hendrycks, Dan et al. (2021). *CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review*. arXiv: 2103.06268 [cs.CL]. URL: https://arxiv.org/abs/2103.06268.

Houlsby, Neil et al. (2019). *Parameter-Efficient Transfer Learning for NLP*. arXiv: 1902.00751 [cs.LG]. URL: https://arxiv.org/abs/1902.00751.

Hu, Edward J. et al. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv: 2106.09685 [cs.CL]. URL: https://arxiv.org/abs/2106.09685.

Huang, Kaifeng et al. (2024). *Lifting the Veil on the Large Language Model Supply Chain: Composition, Risks, and Mitigations*. arXiv: 2410.21218 [cs.SE]. URL: https://arxiv.org/abs/2410.21218.

Huang, Lei et al. (2023). *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. arXiv: 2311.05232 [cs.CL]. URL: https://arxiv.org/abs/2311.05232.

Jiang, Albert Q. et al. (2023). *Mistral 7B*. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

Jin, Qiao et al. (Nov. 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)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 2567–2577. DOI: 10.18653/v1/D19-1259. URL: https://aclanthology.org/D19-1259.

Kadavath, Saurav et al. (2022). *Language Models (Mostly) Know What They Know*. arXiv: 2207.05221 [cs.CL]. URL: https://arxiv.org/abs/2207.05221.

Kano, Yoshinobu et al. (2017). "Overview of COLIEE 2017". In: *COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment*. Ed. by Ken Satoh et al. Vol. 47. EPiC Series in Computing. EasyChair, pp. 1–8. DOI: 10.29007/fm8f. URL: /publications/paper/Fglr.

Karpukhin, Vladimir et al. (2020). *Dense Passage Retrieval for Open-Domain Question Answering*. arXiv: 2004.04906 [cs.CL]. URL: https://arxiv.org/abs/2004.04906.

Keskar, Nitish Shirish et al. (2019). *CTRL: A Conditional Transformer Language Model for Controllable Generation*. arXiv: 1909.05858 [cs.CL]. URL: https://arxiv.org/abs/1909.05858.

Kitchenham, Barbara et al. (2009). "Systematic literature reviews in software engineering – A systematic literature review". In: *Information and Software Technology* 51.1. Special Section - Most Cited Articles in 2002 and Regular Research Papers, pp. 7–15. ISSN: 0950-5849. DOI: https://doi.org/10.1016/j.infsof.2008.09.009. URL: https://www.sciencedirect.com/science/article/pii/S0950584908001390.

Klimaszewski, Mateusz, Zeno Belligoli, Satendra Kumar, et al. (2023). "Gated Adapters for Multi-Domain Neural Machine Translation". In: *ECAI 2023 - 26th European Conference on Artificial Intelligence*. Vol. 372. Frontiers in Artificial Intelligence and Applications, pp. 1264–1271.

Klimaszewski, Mateusz, Zeno Belligoli, and Emmanouil Stergiadis (Sept. 2023). "Gated Adapters for Multi-Domain Neural Machine Translation". In: ISBN: 9781643684369. DOI: 10.3233/FAIA230404.

Krithara, Anastasia et al. (2023). "BioASQ-QA: A manually curated corpus for Biomedical Question Answering". In: *Scientific Data* 10, p. 170. URL: https://doi.org/10.1038/s41597-023-02068-4.

Kryscinski, Wojciech et al. (Nov. 2020). "Evaluating the Factual Consistency of Abstractive Text Summarization". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, pp. 9332–9346. DOI: 10.18653/v1/2020.emnlp-main.750. URL: https://aclanthology.org/2020.emnlp-main.750.

Lai, Guokun et al. (2017). *RACE: Large-scale ReAding Comprehension Dataset From Examinations*. arXiv: 1704.04683 [cs.CL]. URL: https://arxiv.org/abs/1704.04683.

Lauscher, Anne et al. (Nov. 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*. Ed. by Eneko Agirre, Marianna Apidianaki, and Ivan Vulić. Online: Association for Computational Linguistics, pp. 43–49. DOI: 10.18653/v1/2020.deelio-1.5. URL: https://aclanthology.org/2020.deelio-1.5.

Lee, Jinhyuk et al. (Sept. 2019). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4. Ed. by Jonathan Wren, pp. 1234–1240. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btz682. URL: http://dx.doi.org/10.1093/bioinformatics/btz682.

Lee, Nayeon et al. (2023). *Factuality Enhanced Language Models for Open-Ended Text Generation*. arXiv: 2206.04624 [cs.CL]. URL: https://arxiv.org/abs/2206.04624.

Lewis, Mike et al. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. arXiv: 1910.13461 [cs.CL]. URL: https://arxiv.org/abs/1910.13461.

Lewis, Patrick et al. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv: 2005.11401 [cs.CL]. URL: https://arxiv.org/abs/2005.11401.

Li, Jonathan, Rohan Bhambhoria, and Xiaodan Zhu (Dec. 2022). "Parameter-Efficient Legal Domain Adaptation". In: *Proceedings of the Natural Legal Language Processing Workshop 2022*. Ed. by Nikolaos Aletras et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 119–129. DOI: 10.18653/v1/2022.nllp-1.10. URL: https://aclanthology.org/2022.nllp-1.10.

Li, Junyi et al. (2021). *Pretrained Language Models for Text Generation: A Survey*. arXiv: 2105.10311 [cs.CL]. URL: https://arxiv.org/abs/2105.10311.

Li, Wei et al. (2022). *Faithfulness in Natural Language Generation: A Systematic Survey of Analysis, Evaluation and Optimization Methods*. arXiv: 2203.05227 [cs.CL]. URL: https://arxiv.org/abs/2203.05227.

Li, Xiang Lisa and Percy Liang (2021). *Prefix-Tuning: Optimizing Continuous Prompts for Generation*. arXiv: 2101.00190 [cs.CL]. URL: https://arxiv.org/abs/2101.00190.

Lin, Chin-Yew (July 2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: https://aclanthology.org/W04-1013.

Liu, Chen et al. (2023). "CPK-Adapter: Infusing Medical Knowledge into K-Adapter with Continuous Prompt". In: *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 1017–1023. URL: https://api.semanticscholar.org/CorpusID:262078574.

Liu, Pengfei et al. (2021). *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. arXiv: 2107.13586 [cs.CL]. URL: https://arxiv.org/abs/2107.13586.

Liu, Yanchen et al. (2023). *SMoA: Sparse Mixture of Adapters to Mitigate Multiple Dataset Biases*. arXiv: 2302.14413 [cs.CL]. URL: https://arxiv.org/abs/2302.14413.

Liu, Yang et al. (2023). *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*. arXiv: 2303.16634 [cs.CL]. URL: https://arxiv.org/abs/2303.16634.

Louis, Antoine, Gijs van Dijck, and Gerasimos Spanakis (2023). *Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models*. arXiv: 2309.17050 [cs.CL]. URL: https://arxiv.org/abs/2309.17050.

Lu, Qiuhao, Dejing Dou, and Thien Huu Nguyen (Nov. 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*. Ed. by Marie-Francine Moens et al. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 3855–3865. DOI: 10.18653/v1/2021.findings-emnlp.325. URL: https://aclanthology.org/2021.findings-emnlp.325.

Lu, Zhenyan et al. (2024). *Small Language Models: Survey, Measurements, and Insights*. arXiv: 2409.15790 [cs.CL]. URL: https://arxiv.org/abs/2409.15790.

Luo, Renqian et al. (Sept. 2022). "BioGPT: generative pre-trained transformer for biomedical text generation and mining". In: *Briefings in Bioinformatics* 23.6. ISSN: 1477-4054. DOI: 10.1093/bib/bbac409. URL: http://dx.doi.org/10.1093/bib/bbac409.

Luo, Yun et al. (2024). *An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning*. arXiv: 2308.08747 [cs.CL]. URL: https://arxiv.org/abs/2308.08747.

Maia, Macedo et al. (2018). "WWW'18 Open Challenge: Financial Opinion Mining and Question Answering". In: *Companion Proceedings of the The Web Conference 2018*. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee, pp. 1941–1942. ISBN: 9781450356404. DOI: 10.1145/3184558.3192301. URL: https://doi.org/10.1145/3184558.3192301.

Manakul, Potsawee, Adian Liusie, and Mark J. F. Gales (2023). *SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models*. arXiv: 2303.08896 [cs.CL]. URL: https://arxiv.org/abs/2303.08896.

Maynez, Joshua et al. (July 2020). "On Faithfulness and Factuality in Abstractive Summarization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 1906–1919. DOI: 10.18653/v1/2020.acl-main.173. URL: https://aclanthology.org/2020.acl-main.173.

McHugh, Mary L (2012). "Interrater reliability: the kappa statistic". In: *Biochem. Med. (Zagreb)* 22.3, pp. 276–282.

Min, Sewon et al. (2023). *FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation*. arXiv: 2305.14251 [cs.CL]. URL: https://arxiv.org/abs/2305.14251.

Minaee, Shervin et al. (2024). *Large Language Models: A Survey*. arXiv: 2402.06196 [cs.CL]. URL: https://arxiv.org/abs/2402.06196.

Nair, Inderjeet and Natwar Modani (May 2023). "Exploiting Language Characteristics for Legal Domain-Specific Language Model Pretraining". In: *Findings of the Association for Computational Linguistics: EACL 2023*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 2516–2526. DOI: 10.18653/v1/2023.findings-eacl.190. URL: https://aclanthology.org/2023.findings-eacl.190.

Nguyen, Chien Van et al. (2024). *A Survey of Small Language Models*. arXiv: 2410.20011 [cs.CL]. URL: https://arxiv.org/abs/2410.20011.

OpenAI (2024). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774.

– (n.d.). *Introducing ChatGPT*. https://openai.com/index/chatgpt/. Accessed: 2024-09-30.

Ouyang, Long et al. (2022). "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. URL: https://openreview.net/forum?id=TG8KACxEON.

Papineni, Kishore et al. (July 2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational*

Linguistics. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040.

Parisi, German I. et al. (2019). "Continual lifelong learning with neural networks: A review". In: *Neural Networks* 113, pp. 54–71. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2019.01.012. URL: https://www.sciencedirect.com/science/article/pii/S0893608019300231.

Peinl, René and Johannes Wirth (2023). *Evaluation of medium-large Language Models at zero-shot closed book generative question answering*. arXiv: 2305.11991 [cs.CL]. URL: https://arxiv.org/abs/2305.11991.

Penedo, Guilherme et al. (2024). *The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale*. arXiv: 2406.17557 [cs.CL]. URL: https://arxiv.org/abs/2406.17557.

Peng, Baolin et al. (2023). *Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback*. arXiv: 2302.12813 [cs.CL]. URL: https://arxiv.org/abs/2302.12813.

Pfeiffer, Jonas, Andreas Rücklé, et al. (2020). "AdapterHub: A Framework for Adapting Transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*. Online: Association for Computational Linguistics, pp. 46–54. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.7.

Pfeiffer, Jonas, Ivan Vulić, et al. (2020). *MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer*. arXiv: 2005.00052 [cs.CL]. URL: https://arxiv.org/abs/2005.00052.

Popović, Maja (Sept. 2015). "chrF: character n-gram F-score for automatic MT evaluation". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar et al. Lisbon, Portugal: Association for Computational Linguistics, pp. 392–395. DOI: 10.18653/v1/W15-3049. URL: https://aclanthology.org/W15-3049.

Qi, Weizhen et al. (2020). *ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training*. arXiv: 2001.04063 [cs.CL]. URL: https://arxiv.org/abs/2001.04063.

Radford, Alec and Karthik Narasimhan (2018). "Improving Language Understanding by Generative Pre-Training". In: URL: https://api.semanticscholar.org/CorpusID:49313245.

Radford, Alec, Jeff Wu, et al. (2019). "Language Models are Unsupervised Multitask Learners". In: URL: https://api.semanticscholar.org/CorpusID:160025533.

Raffel, Colin et al. (2023). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. arXiv: 1910.10683 [cs.LG]. URL: https://arxiv.org/abs/1910.10683.

Rajpurkar, Pranav et al. (Nov. 2016). "SQuAD: 100,000+ Questions for Machine Comprehension of Text". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by Jian Su, Kevin Duh, and Xavier Carreras. Austin, Texas: Association for Computational Linguistics, pp. 2383–2392. DOI: 10.18653/v1/D16-1264. arXiv: 1606.05250 [cs.CL]. URL: https://aclanthology.org/D16-1264.

Rei, Ricardo et al. (2020). *COMET: A Neural Framework for MT Evaluation*. arXiv: 2009.09025 [cs.CL]. URL: https://arxiv.org/abs/2009.09025.

Reimers, Nils and Iryna Gurevych (Nov. 2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods* *in Natural Language Processing*. Association for Computational Linguistics. URL: http://arxiv.org/abs/1908.10084.

Rietzler, Alexander et al. (2019). *Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification*. arXiv: 1908.11860 [cs.CL]. URL: https://arxiv.org/abs/1908.11860.

Roberts, Adam, Colin Raffel, and Noam Shazeer (Nov. 2020). "How Much Knowledge Can You Pack Into the Parameters of a Language Model?" In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, pp. 5418–5426. DOI: 10.18653/v1/2020.emnlp-main.437. URL: https://aclanthology.org/2020.emnlp-main.437.

Rogers, Anna, Matt Gardner, and Isabelle Augenstein (Feb. 2023). "QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension". In: *ACM Computing Surveys* 55.10, pp. 1–45. ISSN: 1557-7341. DOI: 10.1145/3560260. URL: http://dx.doi.org/10.1145/3560260.

Romanov, Alexey and Chaitanya Shivade (Aug. 21, 2018). "Lessons from Natural Language Inference in the Clinical Domain". In: *arXiv:1808.06752 [cs]*. arXiv: 1808.06752. URL: http://arxiv.org/abs/1808.06752 (visited on 08/27/2018).

Rücklé, Andreas et al. (2021). *AdapterDrop: On the Efficiency of Adapters in Transformers*. arXiv: 2010.11918 [cs.LG]. URL: https://arxiv.org/abs/2010.11918.

Sanh, Victor et al. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *ArXiv* abs/1910.01108.

Shah, Raj Sanjay et al. (2022). *WHEN FLUE MEETS FLANG: Benchmarks and Large Pre-trained Language Model for Financial Domain*. arXiv: 2211.00083 [cs.CL]. URL: https://arxiv.org/abs/2211.00083.

Shazeer, Noam et al. (2017). *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer*. arXiv: 1701.06538 [cs.LG]. URL: https://arxiv.org/abs/1701.06538.

Shen, Chenhui et al. (2023). *Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization*. arXiv: 2305.13091 [cs.CL]. URL: https://arxiv.org/abs/2305.13091.

Singhal, Karan et al. (2023). *Towards Expert-Level Medical Question Answering with Large Language Models*. arXiv: 2305.09617 [cs.CL]. URL: https://arxiv.org/abs/2305.09617.

Srivastava, Akchay and Atif Memon (2024). *Towards Robust Evaluation: A Comprehensive Taxonomy of Datasets and Metrics for Open Domain Question Answering in the Era of Large Language Models*. arXiv: 2406.13232 [cs.CL]. URL: https://arxiv.org/abs/2406.13232.

Talmor, Alon et al. (June 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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4149–4158. DOI: 10.18653/v1/N19-1421. URL: https://aclanthology.org/N19-1421.

Tang, Yixuan and Yi Yang (2024). *Do We Need Domain-Specific Embedding Models? An Empirical Investigation*. arXiv: 2409.18511 [cs.CL]. URL: https://arxiv.org/abs/2409.18511.

Thompson, Brian et al. (June 2019). "Overcoming Catastrophic Forgetting During Domain Adaptation of Neural Machine Translation". 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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2062–2068. DOI: 10.18653/v1/N19-1209. URL: https://aclanthology.org/N19-1209.

Tian, Katherine et al. (2023). *Fine-tuning Language Models for Factuality*. arXiv: 2311.08401 [cs.CL]. URL: https://arxiv.org/abs/2311.08401.

Touvron, Hugo et al. (2023). *LLaMA: Open and Efficient Foundation Language Models*. arXiv: 2302.13971 [cs.CL]. URL: https://arxiv.org/abs/2302.13971.

Wan, Zhongwei et al. (2024). *Efficient Large Language Models: A Survey*. arXiv: 2312.03863 [cs.CL]. URL: https://arxiv.org/abs/2312.03863.

Wang, Cunxiang et al. (2023). *Evaluating Open-QA Evaluation*. arXiv: 2305.12421 [cs.CL]. URL: https://arxiv.org/abs/2305.12421.

Wang, Guangyu et al. (2023). *ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation*. arXiv: 2306.09968 [cs.CL]. URL: https://arxiv.org/abs/2306.09968.

Wang, Ruize et al. (2020). *K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters*. arXiv: 2002.01808 [cs.CL]. URL: https://arxiv.org/abs/2002.01808.

Wei, Jerry et al. (2024). *Long-form factuality in large language models*. arXiv: 2403.18802 [cs.CL]. URL: https://arxiv.org/abs/2403.18802.

Wohlin, Claes (2014). "Guidelines for snowballing in systematic literature studies and a replication in software engineering". In: EASE '14. London, England, United Kingdom: Association for Computing Machinery. ISBN: 9781450324762. DOI: 10.1145/2601248.2601268. URL: https://doi.org/10.1145/2601248.2601268.

Wu, Shijie et al. (2023). *BloombergGPT: A Large Language Model for Finance*. arXiv: 2303.17564 [cs.LG]. URL: https://arxiv.org/abs/2303.17564.

Xu, Lingling et al. (2023). *Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment*. arXiv: 2312.12148 [cs.CL]. URL: https://arxiv.org/abs/2312.12148.

Yang, Fangkai et al. (Dec. 2023). "Empower Large Language Model to Perform Better on Industrial Domain-Specific Question Answering". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Ed. by Mingxuan Wang and Imed Zitouni. Singapore: Association for Computational Linguistics, pp. 294–312. DOI: 10.18653/v1/2023.emnlp-industry.29. URL: https://aclanthology.org/2023.emnlp-industry.29.

Yang, Hongyang, Xiao-Yang Liu, and Christina Dan Wang (2023). *FinGPT: Open-Source Financial Large Language Models*. arXiv: 2306.06031 [q-fin.ST]. URL: https://arxiv.org/abs/2306.06031.

Yu, Tiezheng, Zihan Liu, and Pascale Fung (June 2021). "AdaptSum: Towards Low-Resource Domain Adaptation for Abstractive Summarization". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, pp. 5892–5904. DOI: 10.18653/v1/2021.naacl-main.471. URL: https://aclanthology.org/2021.naacl-main.471.

Zafar, Aizan et al. (2024). "KI-MAG: A knowledge-infused abstractive question answering system in medical domain". In: *Neurocomputing* 571, p. 127141. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2023.127141. URL: https://www.sciencedirect.com/science/article/pii/S092523122301264X.

Zhang, Susan et al. (2022). *OPT: Open Pre-trained Transformer Language Models*. arXiv: 2205.01068 [cs.CL]. URL: https://arxiv.org/abs/2205.01068.

Zhang, Tianyi et al. (2020). *BERTScore: Evaluating Text Generation with BERT*. arXiv: 1904.09675 [cs.CL]. URL: https://arxiv.org/abs/1904.09675.

Zheng, Hongling et al. (2023). *Learn From Model Beyond Fine-Tuning: A Survey*. arXiv: 2310.08184 [cs.AI]. URL: https://arxiv.org/abs/2310.08184.

Zheng, Lianmin et al. (2023). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. arXiv: 2306.05685 [cs.CL]. URL: https://arxiv.org/abs/2306.05685.

Zhou, Ce et al. (2023). *A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT*. arXiv: 2302.09419 [cs.AI]. URL: https://arxiv.org/abs/2302.09419.

Zhou, Zhihan, Liqian Ma, and Han Liu (Aug. 2021). "Trade the Event: Corporate Events Detection for News-Based Event-Driven Trading". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, pp. 2114–2124. DOI: 10.18653/v1/2021.findings-acl.186. URL: https://aclanthology.org/2021.findings-acl.186.

Ziegler, Daniel M. et al. (2020). *Fine-Tuning Language Models from Human Preferences*. arXiv: 1909.08593 [cs.CL]. URL: https://arxiv.org/abs/1909.08593.