

Injecting Domain-Specific Knowledge into Large Language Models: A Comprehensive Survey

Zirui Song^{1,2*}, Bin Yan^{1*}, Yuhan Liu³, Miao Fang¹, Mingzhe Li⁴, Rui Yan^{3†}, Xiuying Chen^{2†}

¹Northeastern University

²Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

³Gaoling School of Artificial Intelligence, Renmin University of China

⁴ByteDance

{zirui.song, xiuying.chen}@mbzuai.ac.ae, {yuhan.liu, ruiyan}@ruc.edu.cn

Abstract

Large Language Models (LLMs) have demonstrated remarkable success in various tasks such as natural language understanding, text summarization, and machine translation. However, their general-purpose nature often limits their effectiveness in domain-specific applications that require specialized knowledge, such as healthcare, chemistry, or legal analysis. To address this, researchers have explored diverse methods to enhance LLMs by integrating domain-specific knowledge. In this survey, we provide a comprehensive overview of these methods, which we categorize into four key approaches: dynamic knowledge injection, static knowledge embedding, modular adapters, and prompt optimization. Each approach offers unique mechanisms to equip LLMs with domain expertise, balancing trade-offs between flexibility, scalability, and efficiency. We discuss how these methods enable LLMs to tackle specialized tasks, compare their advantages and disadvantages, evaluate domain-specific LLMs against general LLMs, and highlight the challenges and opportunities in this emerging field. For those interested in delving deeper into this area, we also summarize the commonly used datasets and benchmarks. To keep researchers updated on the latest studies, we maintain an open-source at: [official-repo.com](https://github.com/zirui-song/official-repo), dedicated to documenting research in the field of specialized LLM.

have explored methods to enhance LLMs through external or embedded domain expertise, a process often referred to as *knowledge injection*. This approach aims to bridge the gap between general-purpose language understanding and the stringent requirements of domain-specific tasks, enabling LLMs to perform effectively in highly specialized contexts.

Building on the foundational capabilities of general-purpose LLMs, knowledge injection techniques provide an effective means to address their limitations in handling specialized applications. Compared to the generalized approach of standard LLMs, knowledge injection offers two key advantages: 1) incorporating precise, domain-specific knowledge to improve accuracy and reliability in specialized tasks, and 2) allowing LLMs to dynamically adapt to new information or evolving knowledge bases, ensuring up-to-date expertise. These techniques bridge the gap between general-purpose understanding and domain-specific demands by leveraging both structured and unstructured knowledge sources. As a result, knowledge injection methods have been successfully applied in fields such as healthcare, chemistry, and legal analysis, significantly enhancing LLM performance. For example, biomedical LLMs [Bolton *et al.*, 2024; Yan *et al.*, 2023] have demonstrated superior accuracy in tasks like medical diagnostics and regulatory compliance, while domain-specific models for material science [Xie *et al.*, 2024; Antunes *et al.*, 2024; Zhang *et al.*, 2024a] have achieved advances in material property prediction and discovery. These dedicated models underscore the transformative potential of integrating domain knowledge into LLMs, unlocking solutions to complex, field-specific challenges.

Despite these advancements, early efforts in knowledge injection often treated domains independently, leading to a lack of standardization in methodologies and evaluation. As the volume of research continues to grow rapidly, with applications and studies proliferating across disciplines, the need for a comprehensive review becomes evident. This review aims to summarize the state of knowledge injection techniques, provide a systematic blueprint for future research, and identify key challenges, such as balancing scalability with domain-specific accuracy and enabling efficient, real-time knowledge updates.

1 Introduction

Large Language Models (LLMs) have achieved extraordinary success across various tasks, showcasing remarkable capabilities in reasoning, knowledge representation, and decision-making. However, despite their impressive performance in general-purpose applications, many specialized domains, such as healthcare, chemistry, and legal analysis, demand the integration of domain-specific knowledge to achieve high accuracy and reliability. To address this challenge, researchers

*Equal Contribution

†Corresponding authors.

To assist readers from various backgrounds in understanding knowledge injection techniques for LLMs and to com-

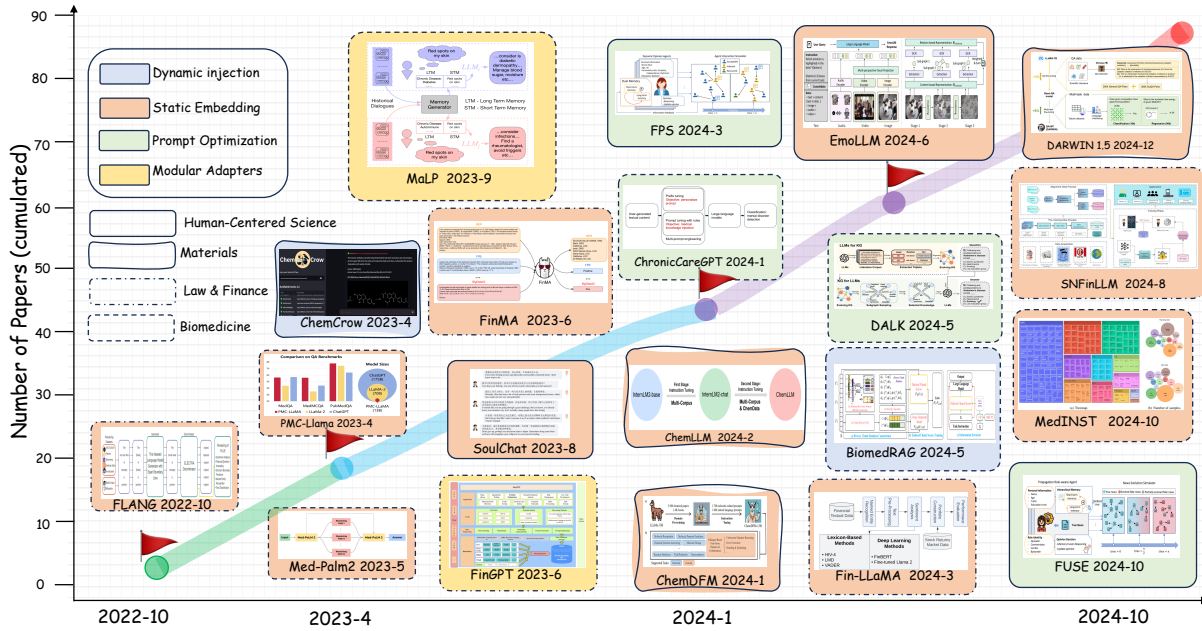


Figure 1: Illustration of Growth Trends in Domain-Specific Knowledge Injection into LLMs. The chart displays the cumulative number of papers published between October 2022 and December 2024. Different colors and border styles represent various injection methods and domains, such as blue with a solid border denoting dynamic injection in the biomedical field.

plement existing surveys by addressing unresolved questions, we organize our survey paper as follows. After providing the *foundational background knowledge* in Section 2, we address a pivotal question: How can domain-specific knowledge be effectively integrated into LLMs? To answer this, we present a *comprehensive framework for categorizing knowledge injection methods* in Section 3. We delve into this topic by discussing: 1) Dynamic Knowledge Injection, which explains how external knowledge is retrieved and incorporated in real-time during inference to enhance reasoning capabilities; 2) Static Knowledge Embedding, which describes how domain knowledge is embedded into the model during training or fine-tuning to make it an inherent part of the model; 3) Adapters, which highlight modular techniques for storing and utilizing external knowledge without altering the primary model’s parameters; and 4) Prompt Optimization, which focuses on how carefully designed prompts enable models to utilize existing knowledge without requiring changes to their architecture. Another perspective for reviewing knowledge injection studies is their application across various domains. In Section 4, we categorize current application domains, such as materials science, chemistry, biology, and law. To guide the identification of appropriate tools and resources, we summarize *commonly used benchmarks, open-source frameworks, and analyses on performance* in Section 5. Based on these summaries, we discuss the *challenges and opportunities in this evolving field* in Section 6. The conclusions are summarized in Section 7.

2 Background

2.1 Domain-Specific Knowledge

Domain-specific knowledge refers to specialized information or expertise pertinent to a specific field or application, distinguishing it from general knowledge that spans across multiple domains. While general knowledge enables models to understand broad contexts, domain-specific knowledge is essential for addressing specialized tasks where precise, field-specific understanding is required. For instance, in scientific text processing [Bran *et al.*, 2023], models must comprehend complex scientific terminologies, concepts, and methodologies to provide accurate and relevant answers. Similarly, in e-commerce search [Zhao *et al.*, 2024a], understanding domain-specific terms such as product categories, technical specifications, or colloquial shopping language is crucial for delivering relevant search results and recommendations. In healthcare applications, LLMs must understand medical terminologies, diagnoses, treatment plans, and drug interactions. For example, biomedical question answering [Pei and others, 2024] and medical report summarization rely on integrating knowledge from medical literature like PubMed [Dermoncourt and Lee, 2017].

To address these needs, researchers have explored various methods for incorporating domain-specific knowledge into LLMs. In this paper, we aim to provide a survey of these various injection methods.

2.2 Knowledge Representation and Encoding

Knowledge can take different forms depending on the structure and application needs. For example, knowledge graphs [Zhang *et al.*, 2024c] represent information as entities and relationships in a graph format, allowing structured

Symbol	Description
\mathbf{x}	Input to LLM
\mathbf{y}	Output of LLM
M	Backbone LLM Function
\mathcal{K}	External domain knowledge base
θ	Parameters of LLM
ϕ	Additional parameters introduced
$\mathcal{R}(\mathbf{x}, \mathcal{K})$	Retrieval function fetches relevant elements of \mathcal{K} given the input \mathbf{x}
$M(\mathbf{x}; \theta)$	Represent LLM takes input \mathbf{x} and produces an output, parameterized by θ
$\Delta\theta$	Offsets to the original LLM’s parameters

Table 1: Summary of Symbols.

reasoning and inference. These are widely used in tasks like question-answering and recommendation systems, where relationships between entities are crucial. Similarly, knowledge in text form, such as Wikipedia [Jeong *et al.*, 2024], provides a vast corpus of unstructured information.

Knowledge can also be stored in vector space instead of readable text or graph formats. For instance, soft prompt tuning [Singhal *et al.*, 2023a] learns useful knowledge in vector form, which is concatenated with the original input to guide LLMs in performing specific downstream tasks. In addition to external representations, knowledge can also emerge from within the model itself. For example, chain-of-thought prompting [Yao *et al.*, 2024] introduces intermediate reasoning steps that help the model break down complex tasks into manageable parts. By explicitly reasoning through these steps, the LLM can utilize its internally stored information more effectively, resulting in better performance on tasks requiring logical reasoning, multi-step computation, or decision-making.

3 Paradigms of Knowledge Injection

This section identifies four major paradigms of knowledge injection: *Dynamic Knowledge Injection*, *Static Knowledge Embedding*, *Adapters*, and *Prompt Optimization*. These paradigms illustrate various mechanisms by which external domain-specific knowledge can be incorporated into LLMs. We utilize unified notations, as described in Table 1, to systematically represent the processes. External knowledge \mathcal{K} is integrated into LLMs either by modifying the original parameters θ , introducing additional parameters ϕ , or leveraging auxiliary mechanisms.

3.1 Dynamic Knowledge Injection

We define dynamic knowledge injection as the process of first retrieving information from external knowledge bases or knowledge graphs and then combining it with the input for use in LLMs:

$$\mathbf{y} = M(\mathbf{x}, \mathcal{R}(\mathbf{x}, \mathcal{K}); \theta), \quad (1)$$

where \mathbf{x} represents the original input, \mathcal{R} denotes the retrieval function, \mathcal{K} is the external knowledge base, and θ are the model parameters, which remain unchanged. This paradigm offers several advantages, including ease of updating (hence the term "dynamic injection") and the ability to incorporate

new knowledge without retraining the model. However, it also presents challenges, such as dependency on the quality of the knowledge base \mathcal{K} , the retrieval function \mathcal{R} , and limitations imposed by the maximum input length of the LLM.

3.2 Static Knowledge Embedding

Compared with dynamic knowledge retrieval, static knowledge embedding involves embedding knowledge into the model’s parameters through full or partial fine-tuning, making it less flexible to changes. Concretely, the model learns new parameters $\Delta\theta$ that encode domain knowledge from \mathcal{K} :

$$\Delta\theta = \arg \min_{\theta} \sum_{(\mathbf{x}_s, \mathbf{y}_s) \in \mathcal{K}} \mathcal{L}(M(\mathbf{x}_s; \theta), \mathbf{y}_s),$$

where \mathcal{K} is the domain-specific knowledge base containing training samples \mathbf{x}_s and \mathbf{y}_s , and \mathcal{L} is a typical supervised training loss function. After optimization, the updated parameters $\Delta\theta$ are obtained.

At inference time, no further retrieval or external knowledge calls are required:

$$\mathbf{y} = M(\mathbf{x}; \Delta\theta).$$

This paradigm provides fast inference, as it eliminates additional retrieval steps and often results in stronger performance. However, it also comes with challenges, such as costly updates—requiring fine-tuning whenever domain knowledge changes—and scalability issues, as embedding large or frequently changing knowledge bases can be computationally expensive.

3.3 Modular Knowledge Adapters

To address the costly updates associated with static knowledge embedding, another paradigm, known as modular knowledge adapters, introduces *small*, trainable modules that can be inserted into or operate alongside the base model to store domain-specific knowledge while saving computational resources. In this approach, the original parameters θ of the LLM typically remain frozen, preserving the model’s general-purpose capabilities. Given a knowledge dataset \mathcal{K} , the adapter parameters ϕ are trained by minimizing the following objective:

$$\phi = \arg \min_{\phi} \sum_{(\mathbf{x}_s, \mathbf{y}_s) \in \mathcal{K}} \mathcal{L}(M(\mathbf{x}_s; \theta, \phi), \mathbf{y}_s),$$

where $M(\mathbf{x}_s; \theta, \phi)$ represents the base model’s generation function enhanced with the new adapter parameters. At inference time, the enhanced model generates outputs as:

$$\mathbf{y} = M(\mathbf{x}; \theta, \phi).$$

This paradigm offers a parameter-efficient method to adapt LLMs to specific domains without modifying the original model weights. By freezing the base model’s parameters, the approach seeks to preserve previously acquired knowledge while enabling the seamless incorporation of new domain-specific information. However, this approach also introduces challenges, such as the need to design new architectural components and determine appropriate hyperparameters, including the size and number of adapters. These additional elements can increase the overall complexity of the model and its training process.

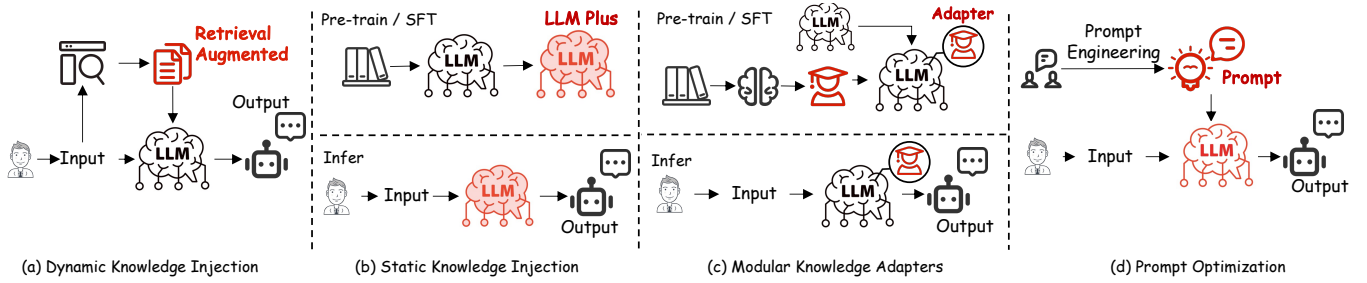


Figure 2: Four knowledge injection paradigms for LLMs. (a) Dynamic Knowledge Injection retrieves external knowledge during inference for enhanced reasoning. (b) Static Knowledge Injection embeds external knowledge into model parameters during fine-tuning. (c) Modular Knowledge Adapters use plug-and-play modules to dynamically adapt to tasks or updates. (d) Prompt Optimization utilizes precise prompts to guide the LLM without altering its parameters.

3.4 Prompt Optimization

Unlike previous approaches, prompt optimization does not retrieve knowledge from external sources. Instead, it focuses on fully leveraging or guiding the LLM to utilize its internal, pre-existing knowledge. The process can be formalized as:

$$\mathbf{y} = M([\mathbf{p}, \mathbf{x}]; \theta),$$

where \mathbf{p} represents a textual prompt containing implicit domain knowledge or specific instructions.

Prompt optimization offers significant advantages, including eliminating dependency on external domain knowledge bases and avoiding training. However, it also presents challenges, as designing effective prompts can be both complex and time-consuming. Additionally, long prompts may reduce the available context window, potentially affecting model efficiency and performance.

3.5 Comparison of the Four Paradigms

Paradigm	Training Cost	Inference Speed	Limitations
Dynamic Injection	None, but requires retrieval module	Slower due to retrieval latency	Relies heavily on retrieval quality
Static Embedding	High (requires pretraining or fine-tuning)	No extra cost	Fixed knowledge; risks catastrophic forgetting
Modular Adapters	Low (train small subset of parameters)	Almost unaffected	Sensitive to training data quality
Prompt Optimization	None	Almost unaffected	Labor-intensive; limited to pre-existing knowledge

Table 2: Guidance on selecting knowledge injection paradigms based on training cost, inference speed, and limitations.

Dynamic knowledge injection integrates external knowledge at runtime, offering flexibility and adaptability to new information without increasing training cost. However, it requires an effective retrieval module, and the inference speed depends heavily on retrieval performance, which can slow down the overall process. Static knowledge embedding embeds domain expertise during pretraining or fine-tuning, requiring large-scale domain-specific data and significant training resources, including GPUs and time. While it incurs

no additional inference cost, its limitations lie in the potential risks of catastrophic forgetting and its inability to adapt to evolving information. Modular adapters serve as a middle ground, allowing plug-and-play components to enhance domain-specific capabilities with minimal training data. Only a small subset of parameters needs to be trained, which reduces training costs, and inference speed is largely unaffected. However, the quality of training data significantly impacts the performance of this method. Prompt Optimization, on the other hand, avoids retraining entirely by leveraging carefully crafted inputs. It has no impact on inference speed but relies on extensive human effort to find optimal prompts. This method is limited in its ability to utilize new knowledge and primarily activates pre-existing knowledge.

We summarize these comparisons in Table 2 as a practical guide to help determine the most suitable method based on specific task requirements and scenarios.

4 Applications

4.1 Biomedicine

The biomedicine domain benefits from a wealth of specialized corpora, such as PubMed [Dernoncourt and Lee, 2017] and MedQA [Jin *et al.*, 2021], enabling the development of LLMs specifically trained on biomedical texts. These models often follow the static knowledge embedding approach, leveraging the domain-specific richness of biomedical data. For instance, PMC-LLaMA [Wu *et al.*, 2023] extends the LLaMA 7B model through further pretraining on 4.9 million PubMed Central articles curated from the S2ORC dataset [Lo *et al.*, 2020], completing five epochs to embed biomedical knowledge effectively. Similarly, Med-PaLM 2 [Singhal *et al.*, 2023b] builds on PaLM 2 via instruction fine-tuning. This fine-tuning incorporates a diverse mix of medical question-answering datasets, including MedQA, MedMCQA [Pal *et al.*, 2022], and HealthSearchQA [Singhal *et al.*, 2023a].

Beyond foundational models, integrating external tools and knowledge can further enhance performance. For example, GeneGPT [Jin *et al.*, 2024] utilizes an LLM pretrained on code tasks to tackle GeneTuring tests by employing NCBI Web APIs. This approach combines in-context learning with an augmented decoding algorithm that identifies and executes API calls. Similarly, Med-PaLM [Singhal *et al.*, 2023a] in-

roduces vector prompts—representations that store and retrieve medical domain knowledge—to extend the capabilities of Flan-PaLM [Chung *et al.*, 2024].

4.2 Finance

Fine-tuned financial LLMs have demonstrated significant advancements in adapting general-purpose models to domain-specific tasks through task-specific training. PIXIU [Xie *et al.*, 2023] fine-tunes LLaMA on 136K instruction samples tailored to financial tasks, enabling the model to handle a wide range of domain-relevant scenarios. Instruct-FinGPT [Zhang *et al.*, 2023] fine-tunes LLaMA on 10K instruction samples derived from two financial sentiment analysis datasets, focusing primarily on finance classification tasks. FinGPT [Yang *et al.*, 2023] introduces a comprehensive end-to-end framework for training and deploying FinLLMs in the financial industry. Utilizing the LoRA technique, FinGPT fine-tunes open-source LLMs like LLaMA and ChatGLM with approximately 50K task-specific samples, achieving efficient fine-tuning without full model retraining.

In contrast, scratch-trained financial LLMs aim to create models specifically designed for financial tasks from the ground up. BloombergGPT [Wu and others, 2023] leverages a subset of 5 billion tokens from Bloomberg-specific data, representing only 0.7% of its total training corpus, to tailor its model to financial applications. XuanYuan 2.0 [Zhang and Yang, 2023] combines 366 billion tokens for pre-training with an additional 13 billion tokens for fine-tuning, creating the largest Chinese financial chat model. Similarly, Fin-T5 [Lu *et al.*, 2023] introduces a Chinese financial pre-training language model built on the T5 architecture, utilizing a 300GB financial corpus. Furthermore, SNFinLLM [Zhao *et al.*, 2024a] dynamically incorporates real-time financial data during inference to enhance decision-making capabilities, demonstrating the value of domain-specific pretraining and adaptability in financial LLMs.

4.3 Materials

In contrast to biomedicine, where significant efforts have been devoted to static knowledge embedding due to the availability of extensive corpora, research in materials and chemistry has primarily focused on utilizing task-related tools that align with the dynamic knowledge injection paradigm.

For instance, Xie *et al.* [2024] demonstrated how Darwin 1.5 leverages natural language inputs and a two-stage training strategy to achieve significant improvements in materials discovery and design tasks. Bran *et al.* [2023] introduced ChemCrow, a framework that augments LLMs with chemistry-expert-designed tools for downstream tasks such as organic synthesis and drug discovery. There are also studies on prompt optimization Tang and others [2025], demonstrating that better-designed planning prompts can effectively utilize the model’s internal knowledge to orchestrate complex tasks. This approach capitalizes on the planning and execution capabilities of multiple LLMs to achieve autonomy in chemical experimentation.

More recently, there has been growing interest in exploring static knowledge embedding and modular knowledge adapters within the chemistry domain. For example, Chen

and others [2024] curated a QA dataset to fine-tune pretrained models like BERT and LLMs such as Llama, aiming to enhance their performance in chemistry-related tasks. Similarly, Xie *et al.* [2024] introduced Darwin 1.5, an open-source large language model tailored for materials science.

4.4 Human-Centered Science

The last domain we introduce is human-centered science, which encompasses a wide range of applications such as psychological counseling, financial forecasting, social behavior prediction, and legal reasoning. All these domains center on understanding and interacting with human needs, behaviors, and decision-making processes.

In *mental health*, datasets like PsyQA [Sun *et al.*, 2021] provide a foundation for training models in psychological counseling scenarios. For instance, SoulChat [Chen *et al.*, 2023], a model fine-tuned on over 100,000 long-text counseling sessions using static knowledge embedding, is designed for empathic conversations. Similarly, MeChat [Qiu *et al.*, 2023] employs dynamic knowledge injection to adapt to real-time inputs, significantly enhancing its emotional support capabilities. These advancements demonstrate the potential of human-centered science in addressing complex, real-world challenges through personalized and context-aware solutions.

In the *education domain*, LLMs have shown immense potential in addressing challenges such as personalized learning, curriculum alignment, and interactive teaching. Personalized learning, for example, requires models to adapt to individual needs, providing tailored feedback and emotional support. EduChat [Dan *et al.*, 2023] tackles this by leveraging educational theories from psychology and pedagogy through static knowledge embedding, enabling tasks like open Q&A, composition correction, and emotional support. Similarly, QiaoBan [Weixiang *et al.*, 2023] focuses on child-centered education by using prompt optimization to adapt the model’s behavior based on child psychology and emotional well-being, catering specifically to young learners. Domain-specific education and interactive teaching have also seen advancements through LLMs. CyberQ [Agrawal *et al.*, 2024] blends static knowledge embedding and dynamic knowledge injection via AISecKG [Agrawal, 2023], generating Q&A based on cybersecurity best practices. Interactive teaching, on the other hand, benefits from models like SocraticLM [Liu *et al.*, 2024c], which employs adapters fine-tuned on the SocraTeach dataset to engage students in critical thinking and problem-solving.

For *social sciences*, models like SocialLLM [Jiang and Ferrara, 2023] combine static knowledge embedding and dynamic knowledge injection to analyze human behavior in social networks. Adapters facilitate large-scale data integration while prompt optimization guides the model to focus on specific social behavior patterns. Models like FPS [Liu *et al.*, 2024e] and FUSE [Liu *et al.*, 2024f] use prompt optimization to simulate the spread and evolution of fake news in social networks, helping understand misinformation’s impact.

A summary of the mainstream models and their information is provided in Table 3. More models across various domains can be found at: [Survey-official-repo](#).

Domain	Model	Paradigms	Knowledge Source	Link
Biomedicine	PMC-LLaMA [Wu <i>et al.</i> , 2023]	Static Knowledge Embedding	PMC-OA, MedC-I, PubMedQA, MedMCQA, USMLE	Link
	Med-PaLM 2 [Singhal <i>et al.</i> , 2023b]	Static Knowledge Embedding	MultiMed	Link
	DALK [Li and others, 2024]	Dynamic Knowledge Injection Prompt Optimization	MedQA, MedMCQA, MMLU, QA4MRE	Link
	ChronicCareGPT [Liu <i>et al.</i> , 2024b]	Prompt Optimization	eRisk	Link
	SA-MDKIF [Xu <i>et al.</i> , 2024b]	Modular Knowledge Adapters	MedQuA, emrQA, PubMedQA, MedQA	Link
	MaLP [Zhang <i>et al.</i> , 2024b]	Modular Knowledge Adapters	HealthCareMagic-100k, iCliniq	Link
	BioMedLM [Bolton <i>et al.</i> , 2024]	Static Knowledge Embedding	PubMed, MedMCQA, MedQA, MMLU, BioASQ	Link
	BiomedRAG [Li <i>et al.</i> , 2024]	Dynamic Knowledge Injection	CHEMPROT, DDI, ade-corpus-v2, MTsample, ADInt, UMLS	Link
Finance	MedINST [Han <i>et al.</i> , 2024]	Static Knowledge Embedding	MedINST	Link
	FLANG [Shah and others, 2022]	Static Knowledge Embedding	Financial PhraseBank, FiQA 2018 Task-1, News Headline Classification, Named Entity Recognition, Structure Boundary Detection, Question Answering	Link
	BloomBergGPT [Wu and others, 2023]	Static Knowledge Embedding	Finance dataset (web, news, filings, press, Bloomberg), Public dataset (the Pile, C4, Wikipedia)	Link
	FinMA [Xie <i>et al.</i> , 2023]	Static Knowledge Embedding	FPB, FiQA-SA, Headline, NER, FinQA, ConvFinQA, BigData22, ACL18, CIKM18	Link
	FinGPT [Zhang <i>et al.</i> , 2023]	Modular Knowledge Adapters	Financial news, Company filings and announcements, Social media discussions, Trends	Link
	Fin-LLaMA [Konstantinidis and others, 2024]	Static Knowledge Embedding	fin-llama-dataset	Link
Materials	SNFinLLM [Zhao <i>et al.</i> , 2024a]	Static Knowledge Embedding	FinEval, FinanceQ, qEQA, FinC, KQA, MRC, cMRC	Link
	ChemCrow [Bran <i>et al.</i> , 2023]	Dynamic Knowledge Injection	18 expert-designed tools	Link
	ChemDFM [Zhao <i>et al.</i> , 2024b]	Static Knowledge Embedding	SciQ, PIQA, PubChem, ARC, USPTO	Link
	ChemLLM [Zhang and others, 2024]	Static Knowledge Embedding	ChemData, ChemBench	Link
	CrystaLLM [Antunes <i>et al.</i> , 2024]	Static Knowledge Embedding	Materials Project, OQMD, NOMAD	Link
	ScholarChemQA [Chen and others, 2024]	Static Knowledge Embedding	AG News, Yahoo Answers, Yelp-5, Amazon-5	Link
	DARWIN 1.5 [Xie <i>et al.</i> , 2024]	Static Knowledge Embedding	FAIR datasets	Link
	ChemAgent [Tang and others, 2025]	Dynamic Knowledge Injection Prompt Optimization	Google API, Docker container, Internet, Hardware API document, Physical world hardware	Link
	MeChat [Qiu <i>et al.</i> , 2023]	Dynamic Knowledge Injection	SMILECHAT, PsyQA	Link
	MindChat [Xin Yan, 2023]	Static Knowledge Embedding	Multi-turn psychological dialogue data	Link
Mental Health	SoulChat [Chen <i>et al.</i> , 2023]	Static Knowledge Embedding	Long-text counseling sessions	Link
	EmoLLM [Yang <i>et al.</i> , 2024]	Static Knowledge Embedding Modular Knowledge Adapters	CPsyCounD	Link
Education	EduChat [Dan <i>et al.</i> , 2023]	Static Knowledge Embedding	Textbooks Data, Open QA Data, Emotional Support Data, Socratic Teaching Data	Link
	QiaoBan [Weixiang <i>et al.</i> , 2023]	Prompt Optimization	Children’s emotional education dialogue data	Link
	HiTA [Liu <i>et al.</i> , 2024a]	Dynamic Knowledge Injection	Educator curated database	Link
	SocraticLM [Liu <i>et al.</i> , 2024c]	Modular Knowledge Adapters	SocraTeach dataset	Link
	CyberQ [Agrawal <i>et al.</i> , 2024]	Static Knowledge Embedding Dynamic Knowledge Injection	AISeckG, Q&A	Link
Social Science	SocialLLM [Jiang and Ferrara, 2023]	Static Knowledge Embedding Prompt Optimization	Covid-Political, Election2020, COVID-Morality, Ukr-Rus-Suspended, Ukr-Rus-Hate, Immigration-Hate-08, Immigration-Hate-05	Link
	FPS [Liu <i>et al.</i> , 2024e]	Prompt Optimization	Fake News Dataset, Big Five Personality Traits	Link
	FUSE [Liu <i>et al.</i> , 2024f]	Prompt Optimization	True News Dataset, Big Five Personality Traits	Link

Table 3: Summary of the domain-specific knowledge injection studies. We categorize current work according to their research domain and knowledge injection method.

5 Tools, Resources, and Analysis

5.1 Knowledge Injection Framework

In this section, we provide a detailed introduction to four open-source frameworks categorized under different knowledge injection methods to facilitate understanding and application: KnowGPT [Zhang *et al.*, 2024c] for Dynamic Knowledge Injection, StructTuning [Liu *et al.*, 2024d] for Static Knowledge Embedding, K-Adapter [Wang *et al.*, 2021] for Modular Knowledge Adapters, and SelfLift [Cheng *et al.*, 2024] for Prompt Optimization.

KnowGPT *dynamically* combines knowledge graphs with prompt optimization by leveraging reinforcement learning to extract highly relevant subgraphs from the knowledge graph. These subgraphs are represented as triples and transformed into natural language prompts that language models can interpret and utilize via diverse prompt templates. The KnowGPT framework significantly reduces the API call costs of LLMs while enhancing their performance in domain-specific tasks.

StructTuning uses a structure-aware approach to *embed* domain knowledge into pre-trained models with a two-stage strategy: Structure-Aware Continual Pre-Training encodes knowledge into the model’s parameters, and Structure-Aware Supervised Fine-Tuning refines understanding through structured QA tasks. This framework demonstrates significant performance improvements in knowledge-driven tasks such as relation classification and question answering, achieving a balance between generality and efficiency.

K-Adapter stores knowledge within *adapter* modules. Its core method involves freezing the original model parameters and assigning an independent, task-specific adapter for each type of knowledge. These adapters are inserted as independent modules into the intermediate layers of the model to generate enhanced representations of specific knowledge. This design effectively mitigates the issue of catastrophic forgetting, preventing newly injected knowledge from overwriting the model’s pre-existing knowledge.

Finally, SelfLift iteratively employs a retrieval-augmented

generator to create an *unbounded memory pool* and uses a memory selector to choose one output as memory for the subsequent generation round. This is an excellent demonstration of prompt optimization, where the model’s outputs are dynamically refined and reused to enhance its overall performance and coherence in subsequent tasks.

5.2 Datasets and Benchmarks

We summarize commonly used datasets or benchmarks for domain-specific LLM study in Table 3, observing significant variation in dataset richness across domains. Biomedicine boasts numerous high-quality datasets, such as PubMed, PubMedQA [Jin *et al.*, 2019], and BioASQ Tsatsaronis and others [2012], that support tasks such as question answering and clinical summarization. In contrast, materials and chemistry have more limited resources, and datasets like USPTO and Enzymes focus on chemical reactions. Miscellaneous datasets are scattered across other domains, such as PsyQA and SmileChat in mental health, SocraTeach, and Children’s emotional education dialogue data dataset in education. This diversity underscores the effort to tailor LLMs to specialized fields while emphasizing the need for broader curation of benchmarks in underrepresented domains.

5.3 Performance Comparison of Domain-specific LLM and General-domain LLM

Model	Domain Type	Size	MedQA	PubMedQA	MedMCQA
Med-Gemini	Specific	–	91.1	–	–
GPT-4	General	–	90.2	80.4	73.7
Med-PaLM 2	Specific	–	85.4	81.8	72.3
PMC-LLaMA	Specific	13B	56.3	77.9	56.0
BioMedLM	Specific	2.7B	50.3	74.4	–
Llama 2	General	70B	43.7	74.3	35.0
Galactica	General	120B	44.4	77.6	52.9

Table 4: Performance comparison of domain-specific and general-domain model performance on medical benchmarks.

Since there are also general-domain strong LLMs, it is necessary to discuss the comparison between domain-specific LLMs and general ones to determine if a specific knowledge injection process is essential. Here, we take the biomedical domain as an example due to significant research efforts in this field, as shown in Table 4. The results are collected from corresponding papers or *paperswithcode.com*.

First, we can observe that closed-source LLMs are currently the most effective models, while the performance gap between general-domain and domain-specific LLMs is relatively narrow. For example, both GPT-4 and Med-Gemini [Saab *et al.*, 2024] achieve outstanding performance, with scores higher than 90 on the MedQA dataset. However, due to the lack of transparency in closed-source LLMs, open-source LLM efforts should not be overlooked. In this field, domain-specific LLMs often outperform general-domain models. For instance, PMC LLaMA-13B outperforms LLaMA2-70B by more than 10 points on the MedQA dataset. This demonstrates the value of domain-specific LLMs in achieving superior performance on specialized tasks. While general-domain models can deliver strong results, incorporating domain-specific knowledge allows for

significant improvements, particularly in open-source initiatives. This underscores the importance of investing in domain-specific LLMs to address the unique challenges of specialized fields.

6 Challenges and Opportunities

6.1 Integrated Knowledge Consistency

Knowledge injection allows LLMs to incorporate and integrate different domain-specific knowledge. However, retrieved knowledge may conflict with the model’s pre-trained representations or other retrieved facts, leading to inconsistencies in outputs [Xu *et al.*, 2024a]. For example, in healthcare or legal analysis, conflicting treatment protocols or contradictory legal precedents could arise, resulting in unreliable decisions and undermining the system’s trustworthiness.

To address this, future research must focus on detecting inconsistencies, resolving conflicts, and maintaining consistency in integrated knowledge. Conflicts can be addressed by prioritizing reliable sources, applying domain-specific rules, or using ensemble techniques to balance multiple perspectives. Alignment algorithms and validation modules can further ensure that retrieved knowledge aligns with the model’s reasoning processes and is reliable before influencing outputs. These efforts are essential to enhance the reliability and applicability of knowledge-enhanced LLMs in complex, high-stakes domains.

6.2 Cross-Domain Knowledge Transfer

Cross-domain knowledge transfer involves equipping LLMs with the ability to generalize knowledge across diverse and distinct fields. While this significantly expands their applicability, it also introduces challenges due to the complexity and diversity of domain-specific terminologies, ontologies, and reasoning patterns. For example, transferring knowledge from chemistry to healthcare might require reconciling differing data structures and reasoning frameworks.

Overcoming these challenges necessitates advancements in modular knowledge representation and transfer learning techniques. Future efforts could explore hybrid approaches that blend static embeddings with dynamic retrieval, enabling LLMs to adapt knowledge flexibly across domains without compromising depth. Additionally, creating standardized cross-domain benchmarks and datasets could facilitate systematic evaluation and foster innovation in multi-domain knowledge transfer methodologies.

7 Conclusion

LLMs enhanced by domain-specific knowledge have shown remarkable potential and garnered increasing research interest. This survey systematically reviews LLM knowledge injection systems, exploring knowledge representation methods, integration strategies, and mechanisms for preserving model generality. We also summarize applications across biomedicine, chemistry, and computational social science domains. By highlighting standard datasets, benchmarks, challenges, and future opportunities, we aim to provide a valuable resource that inspires further exploration of knowledge-enhanced LLMs for domain-specific challenges.

References

- Garima Agrawal, Kuntal Pal, Yuli Deng, Huan Liu, and Ying-Chih Chen. Cyberq: Generating questions and answers for cybersecurity education using knowledge graph-augmented llms. In *Proc. of AAAI*, 2024.
- Garima Agrawal. Aiseckg: Knowledge graph dataset for cybersecurity education. *Proc. of AAAI*, 2023.
- Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling, 2024.
- Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. Biomedlm: A 2.7b parameter language model trained on biomedical text, 2024.
- Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *Nature Machine Intelligence*, 2023.
- Xiuying Chen et al. Scholarchemqa: Unveiling the power of language models in chemical research question answering. *Communications Chemistry*, 2024.
- Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. Soulchat: Improving llms’ empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Proc. of EMNLP Findings*, 2023.
- Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. Lift yourself up: Retrieval-augmented text generation with self-memory. *Proc. of NeurIPS*, 2024.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *JMLR*, 2024.
- Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, et al. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*, 2023.
- Franck Dernoncourt and Ji Young Lee. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *IJCNLP 2017*, 2017.
- Wenhan Han, Meng Fang, Zihan Zhang, Yu Yin, Zirui Song, Ling Chen, Mykola Pechenizkiy, and Qingyu Chen. Medinst: Meta dataset of biomedical instructions, 2024.
- Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proc. of AACL*, 2024.
- Julie Jiang and Emilio Ferrara. Social-llm: Modeling user behavior at scale using language models and social network data. *arXiv preprint arXiv:2401.00893*, 2023.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *EMNLP*, 2019.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 2021.
- Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 2024.
- Thanos Konstantinidis et al. Finllama: Financial sentiment classification for algorithmic trading applications, 2024.
- Dawei Li et al. Dalk: Dynamic co-augmentation of llms and kg to answer alzheimer’s disease questions with scientific literature. *Proc. of EMNLP findings*, 2024.
- Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. Biomedrag: A retrieval augmented large language model for biomedicine, 2024.
- Chang Liu, Loc Hoang, Andrew Stolman, and Bo Wu. Hita: A rag-based educational platform that centers educators in the instructional loop. In *International Conference on Artificial Intelligence in Education*, 2024.
- Haoxin Liu, Wenli Zhang, Jiaheng Xie, Buomsoo Kim, Zhu Zhang, and Yidong Chai. Few-shot learning for chronic disease management: Leveraging large language models and multi-prompt engineering with medical knowledge injection. *arXiv preprint arXiv:2401.12988*, 2024.
- Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. Socraticlm: Exploring socratic personalized teaching with large language models. In *Proc. of NeurIPS*, 2024.
- Kai Liu, Ze Chen, Zhihang Fu, Rongxin Jiang, Fan Zhou, Yaowu Chen, Yue Wu, and Jieping Ye. Structure-aware domain knowledge injection for large language models. *arXiv preprint arXiv:2407.16724*, 2024.
- Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. *Proc. of IJCAI*, 2024.
- Yuhan Liu, Zirui Song, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. From a tiny slip to a giant leap: An llm-based simulation for fake news evolution. *arXiv preprint arXiv:2410.19064*, 2024.
- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. S2orc: The semantic scholar open research corpus. In *Proc. of ACL*, 2020.
- Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*, 2023.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject

- multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, 2022.
- Qizhi Pei et al. Biot5+: Towards generalized biological understanding with iupac integration and multi-task tuning. *ACL findings*, 2024.
- Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *Proc. of EMNLP findings*, 2023.
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.
- Raj Sanjay Shah et al. When flue meets flang: Benchmarks and large pre-trained language model for financial domain, 2022.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 2023.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *Nature Medicine*, 2023.
- Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. Psyqa: A chinese dataset for generating long counseling text for mental health support. *Proc. of ACL findings*, 2021.
- Xiangru Tang et al. Chemagent: Self-updating library in large language models improves chemical reasoning. *arXiv preprint arXiv:2501.06590*, 2025.
- George Tsatsaronis et al. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *Proc. of AAAI*, 2012.
- Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Proc. of ACL Findings*, 2021.
- Zhao Weixiang, Wang Shilong, Tong Yanpeng, Lu Xin, Li Zhuojun, Zhao Yanyan, Wang Chenxue, Zheng Tian, and Qin Bing. Qiaoban: A parental emotion coaching dialogue assistant for better parent-child interaction. 2023.
- Shijie Wu et al. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2023.
- Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: a large language model, instruction data and evaluation benchmark for finance. In *Proc. of ICONIP*, 2023.
- Tong Xie, Yuwei Wan, Yixuan Liu, Yuchen Zeng, Wenjie Zhang, Chunyu Kit, Dongzhan Zhou, and Bram Hoex. Darwin 1.5: Large language models as materials science adapted learners, 2024.
- Dong Xue Xin Yan. Mindchat: Psychological large language model, 2023.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*, 2024.
- Tianhan Xu, Zhe Hu, Ling Chen, and Bin Li. Sa-mdkif: A scalable and adaptable medical domain knowledge injection framework for large language models. *arXiv preprint arXiv:2402.00474*, 2024.
- Xi Yan, Cedric Möller, and Ricardo Usbeck. Biomedical entity linking with triple-aware pre-training. *arXiv preprint arXiv:2308.14429*, 2023.
- Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*, 2023.
- Qu Yang, Mang Ye, and Bo Du. Emollm: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2406.16442*, 2024.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Proc. of NeurIPS*, 2024.
- Di Zhang et al. Chemllm: A chemical large language model, 2024.
- Xuanyu Zhang and Qing Yang. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proc. of CIKM*, 2023.
- Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *FinLLM at IJCAI*, 2023.
- Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. Honeycomb: A flexible llm-based agent system for materials science, 2024.
- Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. Llm-based medical assistant personalization with short-and long-term memory coordination. In *Proc. of AACL*, 2024.
- Qinggang Zhang, Junnan Dong, Hao Chen, Daochen Zha, Zailiang Yu, and Xiao Huang. Knowgpt: Knowledge graph based prompting for large language models. In *Proc. of NeurIPS*, 2024.
- Shujuan Zhao, Lingfeng Qiao, Kangyang Luo, Qian-Wen Zhang, Junru Lu, and Di Yin. Sfnllm: Systematic and nuanced financial domain adaptation of chinese large language models. *arXiv preprint arXiv:2408.02302*, 2024.
- Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, Kai Yu, and Xin Chen. Chemdfm: A large language foundation model for chemistry, 2024.