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

Anonymous ACL submission

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, dedicated to documenting research in the field of specialized LLM.

1 Introduction

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 have explored methods to enhance LLMs through external or embedded domain expertise, a process often referred to as ***knowledge injection***, as shown in Figure 1. 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 (Cho and Lee, 2025; 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 (Tang et al., 2025a; Xie et al., 2024; Antunes et al., 2024; Zhang et al., 2024b) have achieved advances in material property prediction and discovery. These dedicated models underscore the transformative potential of integrating domain knowledge into LLMs.

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,

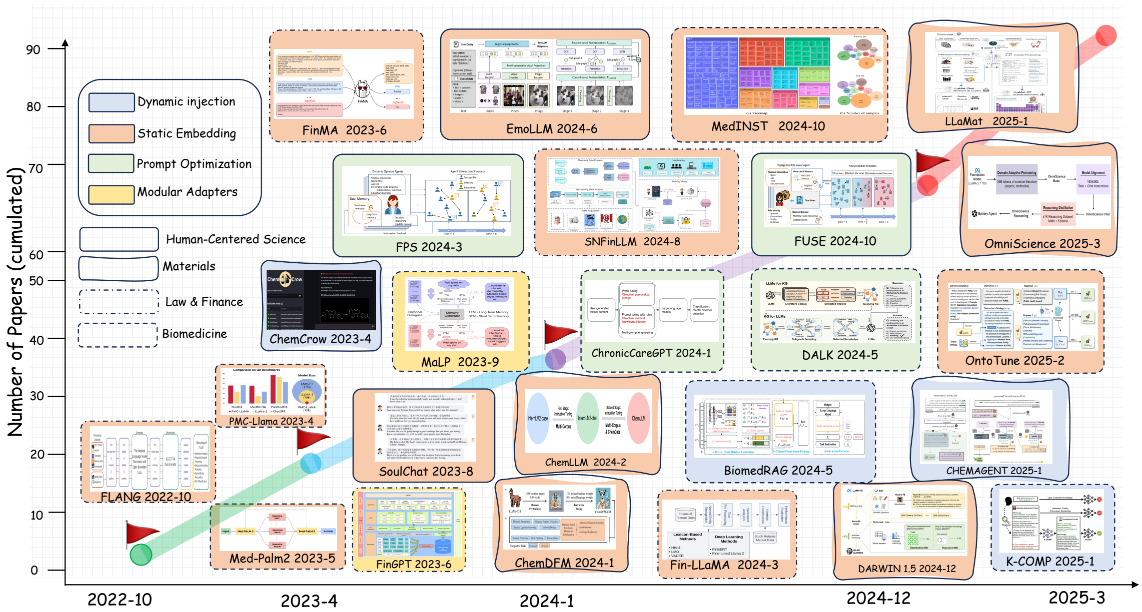

Figure 1: Illustration of Growth Trends in Domain-Specific Knowledge Injection into LLMs. The chart displays the cumulative number of papers published between October 2022 and December 2024. Different colors and border styles represent various injection methods and domains.

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.

We begin in Section 2 with background on domain-specific knowledge and its role in LLMs. Section 3 presents a unified framework of four knowledge injection paradigms: (1) Dynamic Knowledge Injection at inference time; (2) Static Knowledge Embedding during training or fine-tuning; (3) Modular Adapters for parameter-efficient integration; and (4) Prompt Optimization via carefully designed inputs. Section 4 examines these methods across domains such as materials science, chemistry, biology, and law. Section 5 summarizes key datasets, tools, and comparative results. Section B outlines open challenges, including scalability, robustness, and domain transfer. Finally, Section 7 concludes the paper and reflects on future directions.

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. 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 (Singhal et al., 2025; Pei et al., 2024) and medical report summarization rely on integrating knowledge from medical literature like PubMed (Dernoncourt 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 structure and application needs. Knowledge graphs (Liao et al., 2025; Zhang et al., 2024d) encode entities and their relationships in a structured graph, supporting reasoning and inference in tasks like question answering. In contrast, text-based sources like Wikipedia (Jeong et al., 2024) provide rich but unstructured information, useful for tasks requiring broad contextual understanding. Knowledge can also be stored in vector space rather than in text or graph form. For example, soft prompt tuning (Peng et al., 2025; Singhal et al., 2023a) embeds useful information as vectors, which are appended to inputs to guide LLMs on specific tasks. Beyond external forms, knowledge may also emerge internally: chain-of-thought prompting (Sanwal, 2025; Yao et al., 2024) introduces intermediate reasoning steps that help LLMs decompose complex problems and access internal knowledge more effectively—improving performance in tasks involving reasoning, multi-step computation, or decision-making.

2.3 Knowledge Injection Survey

Prior surveys on knowledge-enhanced language models vary in focus and scope. The most relevant works include the following: Cadeddu et al. (2024), who focus on scientific article classification and offer practical insights but lack broader methodological generalization; Wang et al. (2024), who focus on knowledge editing and aim to update internal model knowledge with minimal side effects; and Hu et al. (2023), who adopt a model-centric perspective by classifying knowledge-enhanced models based on task type and knowledge source, though they primarily cover pre-LLM architectures such as BERT and ERNIE. In contrast, our work presents a unified view of knowledge injection in LLMs, emphasizing capability enhancement through external knowledge integration across diverse tasks.

3 Paradigms of Knowledge Injection

To systematically understand how domain knowledge is integrated into LLMs, we categorize existing approaches into four paradigms based on **when** the knowledge is incorporated and **how** it interacts with the model, as shown in Figure 2. Specifically, *Static Knowledge Injection* and *Modular Knowledge Adapters* integrate knowledge prior to inference and involve parameter updates—through either full fine-tuning or adapter-based tuning. In contrast, *Dynamic Knowledge Injection* and *Prompt Optimization* inject knowledge at inference time without altering model parameters: the former retrieves external information, while the latter leverages internal knowledge through designed prompts.

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.

We utilize unified notations, as described in Table 1, to systematically represent the processes.

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:

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

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. To improve retrieval quality, commonly used techniques include semantic matching based on sentence embeddings and efficient vector databases for fast similarity search.

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 $\mathcal{K}$:

$$\Delta\theta = \arg\min_\theta \textstyle\sum_{(\mathbf{x_s}, \mathbf{y_s}) \in \mathcal{K}} \mathcal{L}(M(\mathbf{x_s}; \theta), \mathbf{y_s}),$$

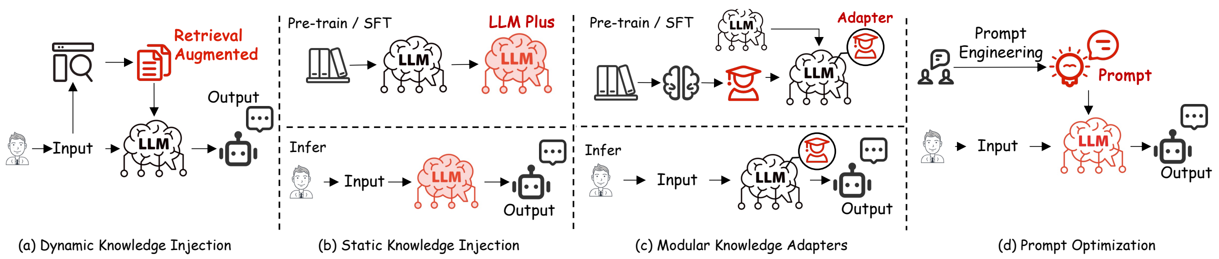

Figure 2: Four knowledge injection paradigms for LLMs. (a) Dynamic Knowledge Injection retrieves external knowledge during inference. (b) Static Knowledge Injection embeds external knowledge into model parameters during fine-tuning. (c) Modular Knowledge Adapters use plug-and-play modules to dynamically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

where $\mathcal{K}$ is the domain-specific knowledge base containing training samples $\mathbf{x_s}$ and $\mathbf{y_s}$, and $\mathcal{L}$ is a 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 enables fast inference by removing the need for additional retrieval steps and often delivers stronger performance. However, it also presents challenges, such as high update costs since fine-tuning is required when domain knowledge changes, and scalability concerns because embedding large or frequently changing knowledge bases demands significant computational resources.

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:

$$\phi = \arg\min_{\phi} \textstyle\sum_{(\mathbf{x_s}, \mathbf{y_s}) \in \mathcal{K}} \mathcal{L}(M(\mathbf{x_s}; \theta, \phi), \mathbf{y_s}),$$

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.

3.4 Prompt Optimization

Prompt optimization refers to the practice of guiding LLMs to perform domain-specific tasks by crafting effective textual prompts. Unlike retrieval-based methods, it relies entirely on the model's internal knowledge and does not require access to external knowledge bases or fine-tuning. The process can be formalized as:

$$\mathbf{p}^* = \arg\min_{\mathbf{p}} \mathcal{L}(M([\mathbf{p}, \mathbf{x}]; \theta), \mathbf{y}^*),$$

where $\mathbf{p}$ is a prompt containing domain-relevant cues, $\mathbf{x}$ is the task input, and θ are the fixed parameters of the LLM.

This paradigm offers practical advantages such as lightweight deployment, no training overhead, and adaptability across domains. However, it also faces key challenges: designing prompts that elicit accurate responses can be non-trivial, and long prompts may reduce the available input space due to context length limitations. Prompt-based approaches can be broadly categorized into manual prompting, prompt tuning, and prefix tuning. These differ in how prompts are constructed or optimized—ranging from discrete, static prompts to learnable embeddings—and have been widely adopted for low-resource domain adaptation.

3.5 Comparison of the Four Paradigms

Dynamic knowledge injection introduces external knowledge at runtime, offering flexibility and

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 choosing injection methods by training cost, inference speed, and constraints.

adaptability without added training cost. However, it depends on efficient retrieval, and inference speed can suffer if retrieval performance is poor. Static knowledge embedding integrates domain expertise during pretraining or fine-tuning, requiring large-scale data and significant computational resources. It adds no inference cost but struggles to adapt to new information and is prone to catastrophic forgetting. Modular adapters offer a middle ground by enhancing domain capabilities through plug-and-play modules that require minimal training data. Only a small number of parameters are trained, reducing cost and preserving inference speed, though performance heavily depends on data quality. Prompt optimization avoids retraining by using well-crafted inputs. It maintains fast inference but relies on significant manual effort and is limited to activating existing knowledge rather than incorporating new information. We summarize these comparisons in Table 2 as a practical guide to help determine the suitable method based on specific requirements and scenarios.

4 Applications

4.1 Finance

In the financial domain, LLM generally follow two main development paths: fine-tuning general-purpose models on financial tasks or training models from scratch using domain-specific corpora.

For fine-tuning, PIXIU(Xie et al., 2023) adapts LLaMA using 136K financial instruction samples, equipping the model to handle diverse finance-related scenarios. Instruct-FinGPT(Zhang et al., 2023) focuses on sentiment classification by fine-tuning on 10K samples from two financial sentiment datasets. FinGPT (Yang et al., 2023) proposes an end-to-end framework for developing FinLLMs, efficiently fine-tuning LLaMA and ChatGLM with 50K samples via LoRA, significantly reducing computational costs. In contrast, scratch-trained FinLLMs aim for deep domain alignment from the ground up. BloombergGPT (Wu et al., 2023b) uses 5B Bloomberg-specific tokens (0.7% of its total corpus) to specialize in financial applications. XuanYuan 2.0 (Zhang and Yang, 2023) is the largest Chinese financial chatbot, trained on 366B tokens and fine-tuned on 13B. Fin-T5 (Lu et al., 2023) leverages a 300GB Chinese financial corpus using the T5 architecture, while SNFinLLM (Zhao et al., 2024a) enhances inference through real-time financial data injection.

In summary, this field showcases a rich diversity in training strategies, from lightweight tuning to comprehensive, end-to-end development.

4.2 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., 2023a) 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. GeneGPT(Jin et al., 2024) leverages a code-pretrained LLM to address GeneTuring tests by calling NCBI Web APIs, combining in-context learning with an augmented decoding algorithm capable of identifying and executing API requests. Med-PaLM(Singhal et al., 2023a) extends the capabilities of Flan-PaLM (Chung et al., 2024) through the use of vector prompts—dense representations designed to store and retrieve medical domain knowledge during inference.

Overall, biomedical LLMs lead in combining static pretraining, instruction tuning, and tool integration, reflecting a shift toward hybrid reasoning

Domain	Model	Paradigms	Knowledge Source	Link
Finance	FLANG (Shah et al., 2022)	Static Knowledge Embedding	Financial PhraseBank,FiQA 2018 Task-1, News Headline Classification, Named Entity Recognition, Structure Boundary Detection,Question Answering	Link
	BloomBergGPT (Wu et al., 2023b)	Static Knowledge Embedding	Finance dataset (web, news, filings, press, Bloomberg), Public dataset (the Pile, C4, Wikipedia)	\
	FinMA (Xie et al., 2023)	Static Knowledge Embedding	FPB,FiQA-SA,Headline,NER,FinQA, ConvFinQA,BigData22,ACL18,CIKM18	Link
	FinGPT (Zhang et al., 2023)	Modular Knowledge Adapters	Financial news, Company filings and announcements, Social media discussions, Trends	Link
	Fin-LLaMA (Konstantinidis et al., 2024)	Static Knowledge Embedding	fin-llama-dataset	Link
	SNFinLLM (Zhao et al., 2024a)	Static Knowledge Embedding	FinEval, FinanceIQ,qEQA,FinC,KQA,MRC,cMRC	\
	Fino1 (Qian et al., 2025)	Static Knowledge Embedding	FinQA, TATQA , DocMath (Simpshort and Compshort), DocFinQA , and BizBench-QA	Link
Biomedicine	PMC-LLaMA (Wu et al., 2023a)	Static Knowledge Embedding	PMC-OA, MedC-I, PubMedQA, MedMCQA, USMLE	Link
	Med-PaLM 2 (Singhal et al., 2023b)	Static Knowledge Embedding	MultiMed	Link
	DALK (Li et al., 2024a)	Dynamic Knowledge Injection Prompt Optimization	MedQA, MedMCQA, MMLU, QA4MRE	Link
	ChronicCareGPT (Liu et al., 2024b)	Prompt Optimization	eRisk	Link
	SA-MDKIF (Xu et al., 2024c)	Modular Knowledge Adapters	MedQuA,emrQA, PubMedQA, MedQA	\
	MaLP (Zhang et al., 2024c)	Modular Knowledge Adapters	HealthCareMagic-100k, iCliniq	Link
	BioMedLM (Bolton et al., 2024)	Static Knowledge Embedding	PubMed,MedMCQA,MedQA,MMLU,BioASQ	Link
	BiomedRAG (Li et al., 2024b)	Dynamic Knowledge Injection	CHEMPROT,DDI,ade-corpus-v2,MTsample,ADInt,UMLS	Link
	MedINST (Han et al., 2024)	Static Knowledge Embedding	MedINST	Link
	K-COMP (Cho and Lee, 2025)	Dynamic Knowledge Injection	MedCorp corpus	\
	OntoTune (Liu et al., 2025)	Static Knowledge Embedding	SemEval2018 Task 9 dataset	Link
Materials	ChemCrow (Bran et al., 2023)	Dynamic Knowledge Injection	18 expert-designed tools	Link
	ChemDFM (Zhao et al., 2024b)	Static Knowledge Embedding	SciQ,PIQA,PubChem,ARC,USPTO	Link
	ChemLLM (Zhang et al., 2024a)	Static Knowledge Embedding	ChemData,ChemBench	Link
	CrystaLLM (Antunes et al., 2024)	Static Knowledge Embedding	Materials Project, OQMD, NOMAD	Link
	ScholarChemQA (Chen et al., 2024)	Static Knowledge Embedding	AG News,Yahoo Answers ,Yelp-5,Amazon-5	Link
	DARWIN 1.5 (Xie et al., 2024)	Static Knowledge Embedding	FAIR datasets	Link
	ChemAgent (Tang et al., 2025b)	Dynamic Knowledge Injection Prompt Optimization	college chemistry textbooks:Quantum chemistry (quan), hemistry kinetics (matter) Quantum mechanics (chemmc), Physical chemistry (atkins)	Link
	LLaMat (Mishra et al., 2025)	Static Knowledge Embedding	MatBookQA,MaScQA,MatSciInstruct	\
	OmniScience (Prabhakar et al., 2025)	Static Knowledge Embedding	daring-anteater dataset, s1K dataset	\
Mental Health	MeChat (Qiu et al., 2023)	Dynamic Knowledge Injection	SMILECHAT, PsyQA	Link
	MindChat (Xin Yan, 2023)	Static Knowledge Embedding	Multi-turn psychological dialogue data	Link
	SoulChat (Chen et al., 2023)	Static Knowledge Embedding	Long-text counseling sessions	Link
	EmoLLM (Yang et al., 2024)	Static Knowledge Embedding Modular Knowledge Adapters	CPsyCounD	Link
Education	EduChat (Dan et al., 2023)	Static Knowledge Embedding	Textbooks Data, Open QA Data, Emotional Support Data, Socratic Teaching Data	Link
	QiaoBan (Weixiang et al., 2023)	Prompt Optimization	Children's emotional education dialogue data	Link
	HiTA (Liu et al., 2024a)	Dynamic Knowledge Injection	Educator curated database	\
	SocraticLM (Liu et al., 2024c)	Modular Knowledge Adapters	SocraTeach dataset	\
	CyberQ (Agrawal et al., 2024)	Static Knowledge Embedding Dynamic Knowledge Injection	AISecKG, Q&A	\
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	\
	FPS (Liu et al., 2024e)	Prompt Optimization	Fake News Dataset, Big Five Personality Traits	Link
	FUSE (Liu et al., 2024f)	Prompt Optimization	True News Dataset, Big Five Personality Traits	\

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

in specialized AI.

4.3 Materials

In contrast to the biomedical domain, the field of materials science and chemistry has largely focused on static knowledge embedding. Many models rely on domain-specific corpora to fine-tune general models for improved task performance. Darwin 1.5 (Xie et al., 2024) adopts a two-stage training strategy using natural language inputs to enhance performance in materials discovery. ScholarChemQA (Chen et al., 2024) constructs a chemistry QA dataset to fine-tune BERT and LLaMA, improving chemical reasoning. Recently, some efforts have begun to explore dynamic knowledge integration. ChemCrow (Bran et al., 2023) augments LLMs with chemistry tools for applications like synthesis and drug discovery. ChemAgent (Tang et al., 2025b) shows that well-designed planning prompts can guide models through complex execution tasks by leveraging internal reasoning.

While still in early stages, the field is transitioning from static embedding toward interactive and tool-augmented reasoning, indicating strong potential for future developments

4.4 Human-Centered Science

Human-centered science focuses on understanding and assisting human behaviors, needs, and decisions. This interdisciplinary domain includes mental health, education, social behavior prediction, and legal reasoning—each benefiting from personalized and context-aware LLMs.

In ***mental health***, datasets like PsyQA (Sun et al., 2021) provide a foundation for training models in psychological counseling scenarios. 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. In contrast, MeChat (Qiu et al., 2023) employs dynamic knowledge injection to adapt to real-time inputs, 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) applies psychological and pedagogical theories via static knowledge embedding to support tasks like Q&A, writing feedback, and emotional guidance. Similarly, QiaoBan (Weixiang et al., 2023) uses prompt optimization to tailor model behavior to children's psychological and emotional needs. Domain-specific education and interactive teaching have also seen advancements. 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 Social-LLM (Jiang and Ferrara, 2023) combine static knowledge embedding and dynamic knowledge injection to analyze human behavior in social networks. 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: official-repo.

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., 2024d) for Dynamic Knowledge Injection, StructTuning (Liu et al., 2024d) for Static Knowledge Embedding, K-Adapter (Wang et al., 2021) for Modular Knowledge Adapters, and Self-Lift (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 a good 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 Knowledge Source

We summarize commonly used knowledge sources for domain-specific LLMs in Table 3, referring to datasets that provide the external knowledge used in various injection methods—including training corpora for static embedding or adapter tuning, and retrieval or prompt design resources for dynamic knowledge injection. Biomedicine includes numerous high-quality datasets, such as PubMed, PubMedQA (Jin et al., 2019), and BioASQ (Tsatsaronis et al., 2012), which 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 4 Paradigms

Model	Category	MedQA	PubMedQA	MedMCQA
GPT-4 (Medprompt)	Prompt Optimization	90.2	82.0	79.1
GPT-4	General	90.2	80.4	73.7
Med-PaLM 2	Static Knowledge	85.4	81.8	72.3
Flan-PaLM (3-shot)	Dynamic Knowledge	67.6	79.0	57.6
PMC-LLaMA	Static Knowledge	56.3	77.9	56.0
BioMedLM	Static Knowledge	50.3	74.4	–
LLaMA (MedAdapter)	Knowledge Adapters	37.4	63.6	32.0

Table 4: Model performance across four knowledge paradigms on medical benchmarks.

To compare knowledge injection paradigms in a practical setting, we focus on the biomedical domain due to its popularity and the availability of benchmarks such as MedQA, PubMedQA, and MedMCQA, as shown in Table 4. Although the models differ in architecture, we align backbones when possible. For example, both PMC-LLaMA and MedAdapter use LLaMA-13B. SOTA models like GPT-4 are closed-source, making prompt optimization the only feasible adaptation strategy. Despite no domain-specific training, GPT-4 with Medprompt achieves strong performance, showing the effectiveness of prompt methods for closed models. Among open models, MedAdapter underperforms compared to PMC-LLaMA, suggesting that full fine-tuning may outperform adapter-based methods for some tasks. Performance differences across paradigms also highlight the importance of pretraining corpus and task alignment, particularly in static injection approaches. Furthermore, in Appendix A, we also compare knowledge injection paradigms in the finance domain and obtain similar conclusions to those in the medical domain.

6 Challenges and Opportunities

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., 2024b). For example, in healthcare or legal analysis, conflicting treatment protocols or contradictory legal precedents could arise (Dayton, 2012), 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 and validation modules help ensure retrieved knowledge fits the model's reasoning.

Cross-Domain Knowledge Transfer. Cross-domain knowledge transfer involves equipping LLMs with the ability to generalize knowledge across diverse and distinct fields (Li et al., 2025). While this significantly expands their applicability, it also introduces challenges due to the complexity and diversity of domain-specific terminologies, ontologies, and reasoning patterns (Montero et al., 2004). For example, transferring knowledge from chemistry to healthcare might require reconciling differing data structures and reasoning frameworks (Schroeder et al., 2018). Overcoming these challenges requires 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, standardized cross-domain benchmarks can enable consistent evaluation and drive innovation in knowledge transfer methods. We provide more discussions in Appendix B.

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 exploration of knowledge-enhanced LLMs for domain-specific challenges.

Limitation

Despite providing a comprehensive review of current methods and applications for domain-specific knowledge injection in LLMs, this survey has certain limitations. First, while we strive to cover several key domains such as finance, biomedicine, and materials science, some less-studied or emerging areas (for example, low-resource languages, cross-cultural education, and niche disciplines) receive relatively limited attention. Second, our focus is primarily on summarizing methodological principles and representative models from existing literature. Due to substantial variation in model architectures, application domains, training data, and evaluation protocols, we were only able to conduct targeted comparisons under controlled conditions within the biomedical domain, using commonly adopted datasets. A more systematic and broad-based empirical evaluation across methods remains an important direction for future work. Nevertheless, we hope this survey serves as a useful reference and provides a clear roadmap for ongoing research in knowledge-enhanced LLMs.

References

Garima Agrawal. 2023. Aiseckg: Knowledge graph dataset for cybersecurity education. *Proc. of AAAI*.

Garima Agrawal, Kuntal Pal, Yuli Deng, Huan Liu, and Ying-Chih Chen. 2024. Cyberq: Generating questions and answers for cybersecurity education using knowledge graph-augmented llms. In *Proc. of AAAI*.

Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. 2024. Crystal structure generation with autoregressive large language modeling. *Preprint*, Nature Communications:2307.04340.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. Biomedlm: A 2.7b parameter language model trained on biomedical text.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *Nature Machine Intelligence*.

Andrea Cadeddu, Alessandro Chessa, Vincenzo De Leo, Gianni Fenu, Enrico Motta, Francesco Osborne, Diego Reforgiato Recupero, Angelo Salatino, and Luca Secchi. 2024. A comparative analysis of knowledge injection strategies for large language models in the scholarly domain. *Engineering Applications of Artificial Intelligence*, 133:108166.

Xiuying Chen and 1 others. 2024. Scholarchemqa: Unveiling the power of language models in chemical research question answering. *Communications Chemistry*.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Proc. of EMNLP Findings*.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2024. Lift yourself up: Retrieval-augmented text generation with self-memory. *Proc. of NeurIPS*.

Jeonghun Cho and Gary Geunbae Lee. 2025. K-comp: Retrieval-augmented medical domain question answering with knowledge-injected compressor. *Preprint*, arXiv:2501.13567.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *JMLR*.

Giovanni Ciatto, Federico Sabbatini, Andrea Agiollo, Matteo Magnini, and Andrea Omicini. 2024. Symbolic knowledge extraction and injection with sub-symbolic predictors: A systematic literature review. *ACM Computing Surveys*, 56(6):1–35.

Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, and 1 others. 2023. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*.

Kim Dayton. 2012. Standards for health care decision-making: legal and practical considerations. *Utah L. Rev.*, page 1329.

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *IJCNLP 2017*.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. *arXiv preprint arXiv:2210.03329*.

Mahmoud El-Haj, Ahmed AbuRa'ed, Marina Litvak, Nikiforos Pittaras, and George Giannakopoulos. 2020. The financial narrative summarisation shared task (FNS 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 1–12, Barcelona, Spain (Online). COLING.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.

Wenhan Han, Meng Fang, Zihan Zhang, Yu Yin, Zirui Song, Ling Chen, Mykola Pechenizkiy, and Qingyu Chen. 2024. Medinst: Meta dataset of biomedical instructions.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2208–2222.

Zejiang Hou, Julian Salazar, and George Polovets. 2022. Meta-learning the difference: preparing large language models for efficient adaptation. *Transactions of the Association for Computational Linguistics*, 10:1249–1265.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1413–1430.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proc. of AACL*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Julie Jiang and Emilio Ferrara. 2023. Social-llm: Modeling user behavior at scale using language models and social network data. *arXiv preprint arXiv:2401.00893*.

Bowen Jin, Jinsung Yoon, Zhen Qin, Ziqi Wang, Wei Xiong, Yu Meng, Jiawei Han, and Sercan O Arik. 2025. Llm alignment as retriever optimization: An information retrieval perspective. *arXiv preprint arXiv:2502.03699*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *EMNLP*.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2024. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*.

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. 2022. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4764–4780.

Thanos Konstantinidis and 1 others. 2024. Finllama: Financial sentiment classification for algorithmic trading applications.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Dawei Li and 1 others. 2024a. Dalk: Dynamic co-augmentation of llms and kg to answer alzheimer's disease questions with scientific literature. *Proc. of EMNLP findings*.

Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. 2024b. Biomedrag: A retrieval augmented large language model for biomedicine. *Preprint*, arXiv:2405.00465.

Yunzhe Li, Junting Wang, Hari Sundaram, and Zhining Liu. 2025. A zero-shot generalization framework for llm-driven cross-domain sequential recommendation. *arXiv preprint arXiv:2501.19232*.

Xiaoxuan Liao, Binrong Zhu, Jacky He, Guiran Liu, Hongye Zheng, and Jia Gao. 2025. A fine-tuning approach for t5 using knowledge graphs to address complex tasks. *Preprint*, arXiv:2502.16484.

Chang Liu, Loc Hoang, Andrew Stolman, and Bo Wu. 2024a. Hita: A rag-based educational platform that centers educators in the instructional loop. In *International Conference on Artificial Intelligence in Education*.

Haoxin Liu, Wenli Zhang, Jiaheng Xie, Buomsoo Kim, Zhu Zhang, and Yidong Chai. 2024b. Few-shot learning for chronic disease management: Leveraging large language models and multi-prompt engineering with medical knowledge injection. *arXiv preprint arXiv:2401.12988*.

Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024c. Socraticlm: Exploring socratic personalized teaching with large language models. In *Proc. of NeurIPS*.

Kai Liu, Ze Chen, Zhihang Fu, Rongxin Jiang, Fan Zhou, Yaowu Chen, Yue Wu, and Jieping Ye. 2024d. Structure-aware domain knowledge injection for large language models. *arXiv preprint arXiv:2407.16724*.

Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024e. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. *Proc. of IJCAI*.

Yuhan Liu, Zirui Song, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2024f. From a tiny slip to a giant leap: An llm-based simulation for fake news evolution. *arXiv preprint arXiv:2410.19064*.

Zhiqiang Liu, Chengtao Gan, Junjie Wang, Yichi Zhang, Zhongpu Bo, Mengshu Sun, Huajun Chen, and Wen Zhang. 2025. Ontotune: Ontology-driven self-training for aligning large language models. *Preprint*, arXiv:2502.05478.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13452–13460.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proc. of ACL*.

Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. 2023. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Vaibhav Mishra, Somaditya Singh, Dhruv Ahlawat, Mohd Zaki, Vaibhav Bihani, Hargun Singh Grover, Biswajit Mishra, Santiago Miret, Mausam, and N. M. Anoop Krishnan. 2025. Foundational large language models for materials research. *Preprint*, arXiv:2412.09560.

Susana Montero, Paloma Díaz, and Ignacio Aedo. 2004. A semantic representation for domain-specific patterns. In *International Symposium on Metainformatics*, pages 129–140. Springer.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*.

Parth Patwa, Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, and 1 others. 2022. Benchmarking multi-modal entailment for fact verification. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*.

Qizhi Pei and 1 others. 2024. Biot5+: Towards generalized biological understanding with iupac integration and multi-task tuning. *ACL findings*.

Zhiyuan Peng, Xuyang Wu, Qifan Wang, and Yi Fang. 2025. Soft prompt tuning for augmenting dense retrieval with large language models. *Knowledge-Based Systems*, 309:112758.

Vignesh Prabhakar, Md Amirul Islam, Adam Atanas, Yao-Ting Wang, Joah Han, Aastha Jhunjhunwala, Rucha Apte, Robert Clark, Kang Xu, Zihan Wang, and Kai Liu. 2025. Omniscience: A domain-specialized llm for scientific reasoning and discovery. *Preprint*, arXiv:2503.17604.

Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Han Yi, Jimin Huang, Qianqian Xie, and Jianyun Nie. 2025. Fino1: On the transferability of reasoning enhanced llms to finance. *Preprint*, arXiv:2502.08127.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *Proc. of EMNLP findings*.

Ruiyang Ren, Yuhao Wang, Kun Zhou, Wayne Xin Zhao, Wenjie Wang, Jing Liu, Ji-Rong Wen, and Tat-Seng Chua. 2025. Self-calibrated listwise reranking with large language models. In *Proceedings of the ACM on Web Conference 2025*, pages 3692–3701.

Manish Sanwal. 2025. Layered chain-of-thought prompting for multi-agent llm systems: A comprehensive approach to explainable large language models. *Preprint*, arXiv:2501.18645.

Lianne Schroeder, Joshua Bierdz, Donald J Wink, Maripat King, Patrick L Daubenmire, and Ginevra A Clark. 2018. Relating chemistry to healthcare and more: implementation of more in a survey organic and biochemistry course for prehealth students. *Journal of chemical education*, 95(1):37–46.

Raj Sanjay Shah and 1 others. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *Preprint*, arXiv:2211.00083.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *Advances in Neural Information Processing Systems*, 35:13158–13173.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023a. Large language models encode clinical knowledge. *Nature*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, and 1 others. 2023b. Towards expert-level medical question answering with large language models. *Nature Medicine*.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. *Proc. of ACL findings*.

Xiangru Tang, Tianyu Hu, Muyang Ye, Yanjun Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, Arman Cohan, and Mark Gerstein. 2025a. Chemagent: Self-updating library in large language models improves chemical reasoning. *Preprint*, arXiv:2501.06590.

Xiangru Tang and 1 others. 2025b. Chemagent: Self-updating library in large language models improves chemical reasoning. *arXiv preprint arXiv:2501.06590*.

George Tsatsaronis and 1 others. 2012. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *Proc. of AAAI*.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Proc. of ACL Findings*.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37.

Zhao Weixiang, Wang Shilong, Tong Yanpeng, Lu Xin, Li Zhuojun, Zhao Yanyan, Wang Chenxue, Zheng Tian, and Qin Bing. 2023. Qiaoban: A parental emotion coaching dialogue assistant for better parent-child interaction.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-llama: Further fine-tuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Shijie Wu and 1 others. 2023b. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: a large language model, instruction data and evaluation benchmark for finance. In *Proc. of ICONIP*.

Tong Xie, Yuwei Wan, Yixuan Liu, Yuchen Zeng, Wenjie Zhang, Chunyu Kit, Dongzhan Zhou, and Bram Hoex. 2024. Darwin 1.5: Large language models as materials science adapted learners. *Preprint*, arXiv:2412.11970.

Dong Xue Xin Yan. 2023. Mindchat: Psychological large language model.

Haike Xu, Zongyu Lin, Yizhou Sun, Kai-Wei Chang, and Piotr Indyk. 2024a. Sparsecl: Sparse contrastive learning for contradiction retrieval. *arXiv preprint arXiv:2406.10746*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Tianhan Xu, Zhe Hu, Ling Chen, and Bin Li. 2024c. Sa-mdkif: A scalable and adaptable medical domain knowledge injection framework for large language models. *arXiv preprint arXiv:2402.00474*.

Shweta Yadav, Mourad Sarrouti, and Deepak Gupta. 2021. Nlm at mediqa 2021: Transfer learning-based approaches for consumer question and multi-answer summarization. In *proceedings of the 20th workshop on biomedical language processing*, pages 291–301.

Xi Yan, Cedric Möller, and Ricardo Usbeck. 2023. Biomedical entity linking with triple-aware pre-training. *arXiv preprint arXiv:2308.14429*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Qu Yang, Mang Ye, and Bo Du. 2024. Emollm: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2406.16442*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Proc. of NeurIPS*.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *FinLLM at IJCAI*.

Di Zhang and 1 others. 2024a. Chemllm: A chemical large language model. *Preprint*, arXiv:2402.06852.

Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. 2024b. Honeycomb: A flexible llm-based agent system for materials science.

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024c. Llm-based medical assistant personalization with short-and long-term memory coordination. In *Proc. of AACL*.

Qinggang Zhang, Junnan Dong, Hao Chen, Daochen Zha, Zailiang Yu, and Xiao Huang. 2024d. Knowgpt: Knowledge graph based prompting for large language models. In *Proc. of NeurIPS*.

Xuanyu Zhang and Qing Yang. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proc. of CIKM*.

Shujuan Zhao, Lingfeng Qiao, Kangyang Luo, Qian-Wen Zhang, Junru Lu, and Di Yin. 2024a. Snfinllm: Systematic and nuanced financial domain adaptation of chinese large language models. *arXiv preprint arXiv:2408.02302*.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pages 4442–4457.

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. 2024b. Chemdfm: A large language foundation model for chemistry. *Preprint*, arXiv:2401.14818.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

A Performance Comparison of 4 Paradigms

Model	Category	FPB	FiQA-SA	TFNS
GPT-4	General	83.3	63.0	80.8
GPT-4(finetune)	Static Knowledge	87.8	88.7	88.3
FinGPT	Knowledge Adapters	88.2	87.4	90.3
FinBERT	Static Knowledge	88.0	59.6	73.3
BloombergGPT	Static Knowledge	51.1	75.1	-
Llama2-7B	General	39.0	80.0	29.6
FLANG	Static Knowledge	91.9	3.4	-

Table 5: Performance comparison of representative models under the four knowledge injection paradigms on financial benchmarks.

To systematically compare the effectiveness of different knowledge injection paradigms in practical settings, we focus on two representative domains: biomedicine and finance.

In the biomedical domain, which is widely studied and rich in benchmark datasets (e.g., MedQA, PubMedQA, and MedMCQA), we evaluate the performance of various models (see Table 4). Although the models differ in architecture, we align their backbones where possible—for example, both PMC-LLaMA and MedAdapter use LLaMA-13B. For closed-source models like GPT-4, prompt engineering is the only feasible adaptation strategy. Nevertheless, GPT-4 combined with Medprompt achieves strong performance, demonstrating the effectiveness of prompt-based knowledge injection. Among open models, MedAdapter underperforms compared to fully fine-tuned models such as PMC-LLaMA, suggesting that full fine-tuning may be more effective than adapter-based methods for certain tasks. Models with static knowledge (e.g., MedBERT) show substantial variance across tasks, underscoring the importance of alignment between pretraining corpora and downstream objectives.

In Table 5, we extend this comparison to the financial domain, evaluating models on benchmarks such as FPB(Malo et al., 2014), FiQA-SA(Maia et al., 2018), and TFNS(El-Haj et al., 2020). The findings closely mirror those in the biomedical domain. Finetuned GPT-4 consistently outperforms others, confirming the value of injecting domain-specific knowledge into general-purpose LLMs. Static knowledge models like FinBERT and FLANG perform well on certain tasks but show significant variability, again emphasizing the crucial role of corpus-task alignment. FinGPT, which adopts lightweight adapter-based knowledge injection, achieves competitive performance while maintaining adaptability. In contrast, LLaMA2-7B lags behind across most tasks, reinforcing the necessity of targeted knowledge injection for domain-intensive applications. The consistency of observations across both domains suggests that the effectiveness of knowledge injection depends on a careful balance of architectural design, adaptation strategy, and corpus alignment to support complex, high-stakes tasks.

B Detailed Discussions on Challenges & Opportunities

B.1 Integrated Knowledge Consistency

While knowledge injection empowers LLMs to reason with external facts, it introduces a crucial consistency problem: injected knowledge may contradict either the model's internal representations or other pieces of retrieved information (Xu et al., 2024b). In high-stakes domains such as healthcare and law, even minor inconsistencies can lead to significant consequences—for instance, conflicting drug dosages from different clinical guidelines (Dayton, 2012; Zhao et al., 2025), or divergent legal interpretations across jurisdictions (Guha et al., 2023).

Recent research proposes techniques such as post-retrieval contradiction detection (Xu et al., 2024a), and confidence-aware re-ranking (Ren et al., 2025) to address these issues. Some frameworks, like MedRAG(Zhao et al., 2025), apply weighted retrieval and ensemble voting to prioritize reliable sources. Others explore neural symbolic consistency checking, where injected knowledge is aligned to pre-defined ontologies or verified using structured reasoning paths (Ciatto et al., 2024).

Another growing area involves alignment-aware reranking, where retrieved documents are filtered based on their alignment with the LLM's intermediate beliefs (Jin et al., 2025). Future directions may include interactive consistency resolution (e.g., user-in-the-loop conflict selection), as well as integrating factual calibration modules (Dong et al., 2022) that explicitly monitor factuality during decoding. These methods collectively aim to make knowledge-enhanced LLMs more robust, explainable, and reliable in dynamic or sensitive environments.

B.2 Cross-Domain Knowledge Transfer

Cross-domain transfer is a central challenge in building generalized yet specialized LLMs. As LLMs are exposed to knowledge from diverse domains, they must navigate incompatible ontologies, varied domain languages, and distinct reasoning structures (Li et al., 2025). For instance, transferring concepts from chemistry to healthcare involves not only bridging terminology gaps but also adapting to different causal assumptions and data formats (Schroeder et al., 2018).

Several strategies have been proposed to manage this complexity. Adapter-based modularization (He et al., 2021) allows domain-specific components to be trained separately and selectively activated. Meta-learning approaches(Hou et al., 2022) help models rapidly adapt to new domains with minimal supervision. Additionally, continual pretraining on mixed-domain corpora (Jin et al., 2022) offers a scalable method to improve robustness without catastrophic forgetting.

Standardized datasets such as CrossNER(Liu et al., 2021) for multilingual named entity recognition, MultiLexSum(Shen et al., 2022) for cross-domain summarization, and MEDIQA-QA (Yadav et al., 2021) for biomedical QA serve as valuable testbeds for cross-domain evaluation. Future work may explore retrieval-augmented transfer, where dynamic selection of domain-relevant knowledge supports adaptive reasoning, or domain-invariant embedding learning, enabling LLMs to generalize across tasks without explicit supervision.

B.3 Scalability and Efficiency of Knowledge Integration

As LLMs are increasingly augmented with large-scale external knowledge such as entire knowledge graphs, document corpora, or real-time retrieval APIs, the computational and memory cost of incorporating this knowledge becomes a bottleneck. Efficient integration remains a key challenge, especially when operating under low-resource or latency-constrained settings.

Techniques such as sparse retrieval (Lee et al., 2019), memory compression (Zhong et al., 2024), and caching strategies have been proposed to reduce overhead. Modular architectures (e.g., adapters or plug-in modules) also allow partial activation of knowledge, improving scalability. Future research could explore task-aware pruning of external knowledge, knowledge distillation from retrieval-based pipelines into compact models, and efficient routing mechanisms to select only relevant knowledge for each input.

B.4 Evaluation and Hallucination Detection

Evaluating knowledge-enhanced LLMs remains difficult due to the lack of standardized benchmarks and automatic metrics for factual consistency, coverage, and reasoning depth. Moreover, LLMs often hallucinate facts even when augmented with accurate knowledge (Ji et al., 2023), making it hard to trust outputs in high-stakes tasks.

Recent work explores metrics like FactScore (Min et al., 2023), entailment-based verification (Patwa et al., 2022), and human-in-the-loop evaluation schemes. However, few of these methods scale across domains or languages. There is a growing need for task-specific, fine-grained evaluation metrics that capture whether the model used the retrieved knowledge effectively and truthfully. Additionally, incorporating hallucination detection as an internal module, such as through consistency checks between generation and source knowledge, may help reduce risk and improve interoperability.