Research Article

External Knowledge Integration in Large Language Models: A Survey on Methods, Challenges, and Future Directions

Semantic Web
Vol. 17(4) 1–27

DOI: 10.1177/22104968261453132
journals.sagepub.com/home/swj

Itisha Yadav[1,2], Sirko Schindler[1], Diana Peters[1] and Roman Klinger[2]

Abstract

Large language models (LLMs) have shown effectiveness in various natural language understanding (NLU) tasks. However, they face notable limitations like hallucinations, a lack of contextual knowledge, and outdated or incomplete knowledge when applied across knowledge-intensive domains such as scientific research, biomedical sciences, finance, law, and others. These challenges commonly arise from the scarcity and under-representation of domain-specific data during the training and model alignment phases. Furthermore, Large Language Models (LLMs) struggle to provide nuanced expertise, as their internal knowledge remains static and generalized, hindering their ability to reason accurately or deliver context-aware results in specialized tasks. This survey investigates the integration of external knowledge into LLMs to address these limitations. The focus is on decoder-based LLMs, that is, autoregressive models that generate text sequentially. By investigating parametric and non-parametric approaches, this work discusses methods to enhance model reasoning capabilities, factual accuracy, and adaptability for domain-specific and knowledge-intensive tasks. Additionally, it highlights the potential of integrating external knowledge to improve explainability and ensure more trustworthy outputs. This survey supports software developers and natural language processing (NLP) researchers in designing NLU systems for specialized domains by leveraging pre-trained LLMs. Additionally, the work provides a foundation for advancing LLM-based NLU systems with insights into future research areas.

Keywords

Large language models, natural language understanding, external knowledge integration with LLMs, retrieval augmented generation (RAG), constrained-decoding with LLMs, ontology-guided constrained-decoding with LLMs, knowledge graph construction, knowledge mechanisms in LLM

Received: April 7, 2025; accepted: March 30, 2026

Editor: Guest Editors 2025 LLM GenAI KGs
Solicited reviews: Célian Ringwald, PostDoc at University of Bologna, Italy; George Hannah, PhD Student at University of Liverpool, United Kingdom; one anonymous reviewer

1 Introduction

Large language models (LLMs) have demonstrated strong performance across a range of natural language understanding (NLU) tasks. This is due to their ability to encode vast amounts of knowledge extracted from enormous data crawled from the Internet (Hu et al., 2023). They acquire knowledge through the training phase, during which they process massive corpora of text data to learn statistical patterns and relationships between words, phrases, and concepts. Unlike explicit

[1]Institute of Data Science, German Aerospace Center (DLR), Jena, Thuringia, Germany
[2]Fundamentals of Natural Language Processing, University of Bamberg, Bavaria, Germany

Corresponding Author:
Itisha Yadav, Institute of Data Science, German Aerospace Center (DLR), Jena, Thuringia, Germany.
Email: itisha.yadav@dlr.de

knowledge repositories such as relational databases, the knowledge in LLMs is encoded implicitly in their parameters. This implicit nature means that retrieving specific pieces of information is not straightforward and depends on probabilistic generation rather than deterministic querying (Wang et al., 2024b). Research shows that LLMs contain factual knowledge (Hu et al., 2023), but their knowledge is static, that is, confined to the state of information at the time of training when used in isolation, or "as is." Relying solely on the knowledge embedded in these models' parameters presents several fundamental challenges (Sanu et al., 2024), including hallucinations, outdated data, and a lack of domain-specific context. These challenges can be mitigated by integrating external knowledge with LLMs.[1]

Tasks like sentiment analysis or spelling and grammar correction perform well with generic models because they rely primarily on linguistic pattern detection rather than deep subject understanding, and thus do not always require external knowledge integration. However, knowledge-intensive tasks such as information extraction (IE), which involve identifying relevant entities and relationships from domain-specific documents to create triples for knowledge graph construction, require structured domain modeling and access to specific, context-rich details to produce accurate results. For instance, creating a knowlegde graph (KG) from scientific ontologies and domain-specific documents requires compliance with domain-imposed constraints on entities, relations, and validation rules, which must be effectively incorporated into the LLM's generation process (Gao et al., 2023; Wang et al., 2024b). Moreover, in scientific fields, critical details are often stored in proprietary, confidential documents that are not included in publicly available datasets, including numerical measurements (e.g., temperature, pressure), material properties, procedural descriptions, and safety information, which are derived from technical datasheets, experimental records, and internal reports. Handling these specialized documents requires precise contextual knowledge, encompassing domain-specific concepts, relationships, and constraints, which can be realized via integrating external knowledge with LLMs. In this work, we focus on integrating such externally available knowledge with LLMs to enhance performance on natural language understanding (NLU) tasks, particularly knowledge-intensive ones such as IE, knowledge graph construction (KGC), and knowledge graph population (KGP) for domains poorly represented in the generic dataset. The goal of this integration is to digitize and semantically structure complex textual data and semi-structured documents, thereby enabling downstream reasoning tasks. Beyond these, other knowledge-intensive NLU tasks, such as entity linking, question answering, fact verification, commonsense reasoning, scientific reasoning, and technical document summarization, also benefit substantially from external knowledge integration to ensure factual consistency and domain-specific accuracy.

1.0.0.1 Survey Purpose. There are several survey papers, as discussed below, that focus on model architecture, model benchmarking, hardware requirements, and explore various applications of LLMs. However, there is a noticeable shortage of literature offering an in-depth and focused discussion on approaches to knowledge integration. For instance, LLM families, such as GPT (Radford et al., 2018), have been compared, examining their internal architectures and datasets for training and fine-tuning, and their applications in diverse fields (Kumar, 2024; Minaee et al., 2024). Additionally, Wang et al. (2024e) shed light on the progression of language models, tracing their development from statistical and neural language models to transformers and, ultimately, LLMs. Another survey discusses open-source LLMs, emphasizing aspects such as data collection, model architectures, and training methodologies (Kukreja et al., 2024). The study by Dong et al. (2025) focuses on "safeguards" and "guardrails" touching the ethical aspects of LLMs usage. Others investigate reinforcement learning using human feedback, fine-tuning on domain-specific datasets, and task-specific fine-tuning (McIntosh et al., 2024; Susnjak et al., 2025) and transfer learning techniques (Sulaiman & Hamzah, 2024). Regarding the synergy between knowledge-based systems and LLMs, Some et al. (2025) provide a comprehensive examination of methods for integrating LLMs with knowledge-based systems. It provides an extensive overview of LLM evolution and architectural paradigms and discusses approaches for combining LLMs with knowledge resources such as knowledge bases, knowledge graphs, and retrieval-augmented generation (RAG) systems. However, it does not offer a detailed methodological analysis of the techniques used to integrate external knowledge into LLMs or address knowledge-intensive NLU tasks for domain-specific datasets. The study presents a broader perspective on what integration entails but lacks a systematic discussion of the underlying mechanisms that operationalize such integration. In essence, that work addresses the question, "how can LLMs be combined with knowledge-based systems in general?"

Building upon this foundation, the present survey extends the discussion from a conceptual overview to a methodological exploration. While prior studies, including Some et al. (2025), provide valuable insights into the general landscape of LLM-knowledge integration, there remains a lack of a comprehensive perspective that individually examines techniques for integrating external knowledge with LLMs for NLU-based systems, their challenges, and future applications. The current survey addresses this gap by examining methods and mechanisms for integrating external knowledge sources into LLMs to improve their performance on knowledge-intensive NLU tasks. It focuses on the question, "how can external, structured, or semi-structured knowledge be injected into LLMs to improve NLU in knowledge-intensive domains?" This

Table 1. Overview of the Selected Literature Corpus (2019–2025) Grouped by Different Categorizations.

Category	Type	Count (%)	Examples/Comments
Paper Type	Survey / Review Papers	21 (11.4%)	Consolidation and meta-analysis studies
	Concept / Method Papers	151 (82.1%)	Theoretical or methodological contributions
	Dataset / Tool Papers	12 (6.5%)	Datasets or software tools required for implementation purposes
Publication Format	Conference Proceedings	70 (38.0%)	ACL, EMNLP, NeurIPS, ICML, AAAI
	Journal Articles	40 (22%)	IEEE, Applied Sciences, ACM Surveys
	arXiv Preprints Only	67 (37%)	Papers only available as preprints
	Tools / Others	7 (4.0%)	GitHub, Zenodo
Peer-Review Status	Peer-Reviewed	109 (59%)	Conference + Journal papers
	Preprints / ArXiv	68 (37%)	Not yet peer-reviewed
	Tools / Other	7 (4.0%)	Software and resources
Total Number:		**184 papers**	

not only enhances the domain-specificity of LLM responses but also addresses critical issues such as bias and misinformation, resulting in more reliable and trustworthy systems for users across various applications. By exploring these techniques, the paper aims to provide insights into improving model robustness and adaptability, thereby facilitating their adoption in real-world scenarios.

1.0.0.2 Intended Audience:. The intended audience for this survey includes researchers, practitioners, and developers working in natural language processing (NLP), machine learning, the Semantic Web, and artificial intelligence who seek to enhance LLMs for knowledge-intensive tasks. These tasks include IE, question answering, KGC, and KGP, particularly in specialized domains where external knowledge integration is necessary for accuracy and contextual understanding. By addressing the challenges and opportunities discussed in this paper, researchers can gain a deeper understanding of how to effectively leverage and integrate external knowledge to improve model accuracy, scalability, and applicability across diverse domains.

1.0.0.3 Scope and Literature Survey Methodology. The literature survey employed a taxonomy of search phrases and keywords, developed through iterative refinement during preliminary analysis. This analysis focused on foundational research in external knowledge representation, such as ontologies and knowledge bases. It also considered the limitations of LLMs in handling specialized scientific content. Additionally, approaches for building domain-specific, knowledge-intensive NLU systems were examined, particularly those that combine formal ontologies with proprietary semi-structured documents to construct knowledge graphs for domains underrepresented in general LLM training data. The scope of this survey encompasses the following:

- An overview of the concept of LLM-based natural language understanding systems.
- Limitations of LLMs in knowledge-intensive tasks.
- Approaches for addressing these limitations through external knowledge integration, categorized into parametric and non-parametric methods.
- Future research directions for advancing knowledge-augmented LLM-based NLU systems.

To facilitate literature retrieval, a Python script was developed to query multiple academic databases using our taxonomy as input. The script systematically crawled scientific papers from open-source databases including Google Scholar,[2] Semantic Scholar,[3] DBLP,[4] and ACL Anthology.[5] Papers from IEEE Xplore[6] and other sources with access restrictions were retrieved manually. The complete taxonomy, the Python script, and the crawling result are publicly available (Yadav, 2025). Research papers published between 2019 and 2024 were considered for inclusion, with a particular emphasis on those from 2022 to 2024 to ensure coverage of recent advances. Studies published in 2025 were manually selected based on their direct relevance to the survey's objectives. From the set of papers retrieved through the Python-based crawling script, relevant studies were manually selected for in-depth research by filtering them based on title or abstract. The Python script facilitated the retrieval of seeding papers, while pertinent additional studies were identified through manual searches. Consequently, a hybrid approach combining automated retrieval and manual searching was employed to collect the literature for the survey. Table 1 provides a quantitative overview of the selected literature corpus, summarizing paper types, publication formats, and peer-review status.

While every effort was made to ensure comprehensive coverage, certain methodological constraints should be acknowledged: First, the keyword-based retrieval process may have unintentionally excluded studies that address knowledge

integration under different conceptualizations or terminologies. Second, the rapid pace of progress in LLM research implies that very recent publications, that is, those appearing shortly before or after the completion of this survey, may not yet be indexed in the consulted databases. Third, the focus on English-language literature may have led to the omission of relevant work published in other languages. Notwithstanding these limitations, the resulting corpus is considered sufficiently broad and representative to capture the major trends and directions in contemporary research on LLM–knowledge integration.

This survey extends existing work by systematically examining knowledge integration in LLMs across both parametric and non-parametric paradigms, offering a more holistic view than prior studies that focus mainly on prompting and reasoning (Some et al., 2025) or RAG (Gao et al., 2023). Kindly refer to Table 2 for an overview of the surveyed methods. Our distinct contribution lies in: (a) coverage of both parametric and non-parametric integration techniques, (b) emphasis on methodological aspects of knowledge-intensive NLU tasks, such as IE and KGC, and (c) inclusion of recent developments in retrieval augmented generation (RAG) knowledge editing, and constrained-decoding. Although this survey offers a methodological analysis of knowledge integration strategies, unified performance benchmarking across these diverse approaches remains difficult due to differences in evaluation tasks, datasets, and domain-specific requirements. Our primary focus is on the systematic categorization and in-depth analysis of methods, with performance considerations highlighted where methodologically needed.

2 LLM-Based NLU Systems: An Overview

2.0.0.1 What Are LLM-Based NLU Systems?. LLMs are large neural networks/deep learning models trained on vast corpora of text to capture patterns, semantics, and syntactic structures in natural language. Generally, models with hundreds of millions to several billion parameters are considered large. Examples, include BERT (340M parameters), GPT-3 (175B parameters), and T5 (11B parameters) (Devlin et al., 2019; Radford et al., 2018). These models have demonstrated impressive capabilities in various NLP tasks, including sentiment analysis (Krugmann & Hartmann, 2024), text classification (Fields et al., 2024), machine translation (Gao et al., 2024; Vaswani et al., 2017), named entity recognition (NER) (Luo et al., 2023), and others. NLU is a subfield of natural language processing (NLP) that builds systems to comprehend and interpret human language. It bridges the gap between human language and machine comprehension (Kulkarni, 2023). It involves various tasks, including semantic role labeling, NER, relationship extraction, coreference resolution, intent detection, question answering, reasoning with textual information, and more. The NLU algorithms extract structured patterns from unstructured or semi-structured data[7] (Singh et al., 2018). The emergence of language models such as BERT and the GPT series has led to significant improvements in NLU performance. Integrating LLMs with NLU systems has been shown to substantially improve the semantic understanding of natural language (Huang et al., 2024b), thereby enhancing accuracy. Numerous implementations have been developed to facilitate this integration. For example, McTear et al. (2023) have optimized their RASA-based (Sharma & Joshi, 2020) dialogue system by integrating the GPT-3.5-turbo model (Ye et al., 2023) to engage users in a motivational health coaching offering reflective dialogues. Rajasekharan et al. (2023) combine LLM with Answer Set Programming (ASP) to do qualitative and mathematical reasoning. Mukanova et al. (2024) propose a methodology that uses LLMs for performing the task of ontology enrichment and semantic processing. Other works also include: using LLMs for ontology engineering and knowledge graph creation (Shimizu & Hitzler, 2025). While these examples illustrate key applications of LLMs in NLU, they do not represent an exhaustive list, as research in this area continues to evolve. However, despite their success, they still face limitations in dealing with domain-specific knowledge and ensuring factual correctness in generated responses (Albtosh, 2024; Bouhoun et al., 2024; Kerner, 2024; Nagar et al., 2025).

2.0.0.2 How Do LLMs Perform NLU?. A breakthrough came with the introduction of transformers by Vaswani et al. (2017). Transformers, as originally defined, are encoder-decoder architectures that primarily rely on attention mechanisms. This mechanism models dependencies among elements of a sequence, such as words in a sentence in natural language processing or patches/pixels in an image in computer vision, and it assigns learnable weights that capture their relative relevance. Transformers serve as the fundamental architecture for LLMs, with different models leveraging specific components of the transformer structure. For instance, BERT utilizes only the encoder (Devlin et al., 2019), whereas GPT is built solely on the decoder mechanism (Radford et al., 2018). In contrast, models like BART incorporate both the encoder and decoder components (Lewis et al., 2020a). LLMs perform NLU by learning contextual relationships between words and sentences. Our focus is on decoder-based models, that is, the causal language models that produce the outputs autoregressively. These models are generative language models. In the training phase, they calculate the probability of a word's occurrence based on the previous context words, that is, learn the statistical representation of language (Brown et al., 2020; Yan et al., 2024). In a fine-tuning phase, the pre-trained models are steered towards user-specific responses by

Table 2. Overview of Surveyed Approaches.

Approach	Related Literature
	Parametric Knowledge
Pre-training LLMs	BERT (Devlin et al., 2019), GPT (Radford et al., 2018), MedBERT (Vasantharajan et al., 2022), LEGAL-BERT (Chalkidis et al., 2020), FinBERT (Peng et al., 2021), TAPT (Gururangan et al., 2020), Knowledge Fusion Layer (K-XLNet) (Yan et al., 2021), Structured Knowledge-aware Pre-training (Dong et al., 2023), Learning Knowledge-Enhanced Representations (Zhang et al., 2023), DKPLM (Zhang et al., 2022), KE_PLM (Hu et al., 2024), Knowledge Representation Enhancement (Allen-Zhu & Li, 2025b)
Fine-tuning LLMs	BioBERT (Lee et al., 2019), Knowledge Extraction and KG Construction (Kazemi et al., 2023), Knowledge-AI (Muralidharan et al., 2024), Full Fine-tuning (Devlin et al., 2019), LoRA (Hu et al., 2022), Prefix Tuning (Li & Liang, 2021), Prompt Tuning (Petrov et al., 2023), BitFit (Ben Zaken et al., 2022), Adapter-based Fine-tuning (Poth et al., 2023), KG-Adapter (Tian et al., 2024), Adapter-based KG Integration (Omeliyanenko et al., 2023), InfuserKI (Wang et al., 2024a)
Knowledge Editing	MEND (Pinter & Elhadad, 2023), EasyEdit (Wang et al., 2024c), Knowledge Editing in LLMs (Ishigaki et al., 2024), Knowledge Editing Survey (Wang et al., 2024d), Editing for Bias Mitigation (Ilharco et al., n.d), ROME & MEMIT (Andonian et al., 2022; Meng et al., 2022)
Steering and Styling	sNeuron-TST (Lai et al., 2024), Tell (Zhang et al., 2024c), Focus (Lamb et al., 2025)
Embedding / Graph Methods	Embeddings in LLMs (Zhang et al., 2024a), Graph Embeddings (Jain & Lapata, 2024), KnowFormer (Liu et al., 2024)
Knowledge Distillation	DistilBERT (Sanh et al., 2019), Knowledge Distillation in NeuralNets (Hinton et al., 2015), MiniLM (Wang et al., 2020)
	Non-Parametric Knowledge
Prompting Methods	PEARL (Sun et al., 2024), Chain-of-Thought Prompting (Bosma et al., 2022), ZEP (Rasmussen et al., 2025), Knowledge Mechanism in LLMs 1-3 (Allen-Zhu & Li, 2025a, 2025b)
Knowledge Graphs / Ontologies	Introduction to KGs (Fensel et al., 2020), KG + LLM Integration (Ye et al., 2024), Medical KG + LLM (Wang et al., 2025), GMeLLo (Chen et al., 2024), Knowledge Solver (Feng et al., 2023), Domain-specific KG Retrieval (Jiang et al., 2024), KGQA with Planning-Retrieval-Reasoning (Luo et al., 2024a), KG + LLM Reasoning (Ji et al., 2024; Ma et al., 2025; Wang et al., 2023a), Scalability in KG + LLM Integration (Zhang et al., 2024b)
Constrained-Decoding	Lexically Constrained Decoding (Hokamp & Liu, 2017), Fast Constrained Decoding (Post & Vilar, 2018), Relation-Constrained Generation (Chen et al., 2022), JSON Mode (Tam et al., 2024), Ontologies for Constrained Decoding (Luo et al., 2024b)
Memory-based Systems: Retrieval-Augmented-Generation (RAG)	RAG (Lewis et al., 2020b), RAG for LLMs (Li et al., 2025c), Elasticsearch (Elastic, 2025), ChromaDB (Chroma, 2025), Haystack (deepset, 2025), LangChain (Mavroudis, 2024), Instruct Embeddings (Su et al., 2023), Pre-training RAG Systems (Gao et al., 2023), Task-specific RAG Categorization (Zhao et al., 2024a), HALO: Medical QA RAG (Anjum et al., 2025), Reliability RAG (Hwang et al., 2025), PDF-based LLM-powered RAG (Khan et al., 2024), Extending Context Windows (Chen et al., 2023a), Efficient Long-Context Generation (Hosseini et al., 2025), GraphRAG (Peng et al., 2025)
Other Memory-based Systems	Temporal Knowledge Graphs (Rasmussen et al., 2025), MemGPT (Packer et al., 2023), MemOS (Li et al., 2025d)
Tool Usage / Function Calling	ToolFormer (Cancedda et al., 2023), LangGraph (LangChain-AI, 2025), LangChain (LangChain-AI, 2025), MeCo (Li et al., 2025a), Granite20B (Abdelaziz et al., 2024) Function Calling for KGs (Hertling & Sack, 2024), Self-guided Function Calling (Cui et al., 2025), LLMs as Zero-shot Dialogue State Tracker through Function Calling (Li et al., 2024b), Function Calling with Generic LLMs (Qin et al., 2025)
Reinforcement Learning	Agents and Environment in RL (Sutton & Barto, 1998), InstructGPT (RLHF) (Agarwal et al., 2022), RLKGF (Yan et al., 2025), RLAIF (Bai et al., 2022; Lee et al., 2024), RL for Prompt Optimization (RLPrompt) (Deng et al., 2022), StablePrompt (Kwon et al., 2024)

applying task-specific fine-tuning or instruction fine-tuning. The work by OpenAI (Radford et al., 2018) offers valuable insights into the concepts of generative pre-training and discriminative fine-tuning. They demonstrate how large parameter neural networks, such as the GPT-family models, can offer a task-agnostic architecture for performing NLU tasks with better results than task-specific models. Furthermore, it has been observed that decoder-based GPT-type models excel in knowledge extraction compared to encoder-based models, except when the knowledge is a standalone word or composed of independent words (Allen-Zhu & Li, 2025a). LLMs reliance on large-scale static and generic text corpora can result in issues such as the generation of hallucinated facts or incomplete reasoning, which poses challenges for tasks requiring

factual accuracy or specialized knowledge (Qiang et al., 2024). Developing techniques that incorporate external knowledge sources for access to proprietary or domain-specific data could significantly mitigate these issues, enabling LLMs to deliver more accurate and contextually relevant outputs across various applications and domains (Ye et al., 2024).

3 Challenges of LLM-Based NLU Systems

Since LLMs form the core of modern NLU systems, any shortcomings in LLMs naturally propagate to the NLU systems built on top of them. In the following, we discuss typical limitations of LLMs that emerge when they are applied to domain-specific or knowledge-intensive tasks. These challenges directly affect LLM-based NLU systems, reducing their ability to reliably process, interpret, and reason over specialized information.

3.1 Hallucinations Leading to Limitations in Factual Accuracy

One of the primary limitations of LLM-based NLU systems is their tendency to generate grammatically correct but factually incorrect information, a phenomenon known as "hallucination" or "confabulation" (Lewis et al., 2020b). Hallucination results in the generation of misinformation, making LLMs unreliable and ultimately reducing the accuracy of the output responses. In healthcare, for example, LLM hallucinations can have serious consequences, such as providing medically incorrect guidance, which could lead to harmful patient outcomes (Maes, 2025). In scientific fields, these inaccuracies may limit the effectiveness of LLMs in extracting knowledge and structuring data. One reason for hallucination is the lack of factual information from proprietary data (Lewis et al., 2020b). When performing tasks such as IE, LLMs can produce extraneous or "hallucinated" content—for example, appending interpretations or qualitative judgments to a value extracted from a technical datasheet (e.g., stating that a measured temperature is "high" or "low"). Although the extracted key values may be accurate, LLMs often generate supplementary information that is not explicitly present in the source document, potentially introducing semantic noise or misinformation. Another prevalent form of hallucination involves numerical distortion—either by altering decimal values or generating entirely spurious numbers absent from the input source. Additionally, when processing technical datasheets (semi-structured documents), LLMs may generate misinformation as the generated information is neither explicitly nor implicitly present in the datasheets used for training. For instance, consider the following use-case. The prompt that produced the output below included a task instruction for *information extraction*, along with a machine-parsed version of the datasheet, originally in PDF format. The input structure is provided below for clarity[8] :

Prompt Structure:

```
Task: Ontology-based information extraction from technical datasheets
Instruction:
Extract relevant entities and relationships from the given technical datasheet.
Input Context:
[Parsed PDF Datasheet]
- Tables and floating-text representing experimental measurements.
[Domain Ontology]
- Associated domain ontology defining target concepts.
Expected Output:
Structured information aligned with the specified ontology.
```

The model-generated output: *"There are apparent asymmetries in the bias applied to some of the samples (upper samples have a slight bias-rand), suggesting an inconsistency in the manufacturing process that could affect the quality of the final product."* In this output, the clause following "suggesting" constitutes a *hallucination*, as it introduces content not present in the source datasheet and thereby adds noise to the extracted information. Such hallucinations are particularly problematic for tasks such as *information extraction* and *knowledge graph construction*, where the accuracy and reliability of results depend on maintaining strict alignment with the underlying data sources. Relying on hallucinated data in these tasks can lead to inaccuracies and misrepresentations, ultimately affecting the reliability and applicability of LLM-generated knowledge. Therefore, understanding and investigating hallucinations in LLMs is important for their seamless application.

At a broader level, hallucinations are classified as intrinsic (resulting from internal parameters of the model) and extrinsic (due to integrated external knowledge) (Cleti & Jano, 2024). These categories can be further divided into subtypes,

including fact-based hallucinations, where incorrect information is generated, and faithfulness hallucinations, wherein the output of the LLMs fails to align with the input prompt. Additionally, coherence hallucinations are characterized by the generation of incoherent text. Other types include relevance hallucinations, where the LLM produces an out-of-domain or irrelevant response, and sensibility hallucinations, which involve the generation of nonsensical text (Huang et al., 2025). Researchers or practitioners of NLP have different perspectives on hallucinations. The survey by Cleti and Jano (2024) highlights the dual nature of hallucinations across diverse domains like healthcare/science and art/design. For fact-based fields like healthcare and scientific research, hallucinations lead to the generation of incorrect facts that are not acceptable. On the other hand, for philosophical and abstract fields like arts and design, hallucinations can foster creativity by generating unconventional or unseen outputs that inspire innovations. Therefore, the issue of hallucinations and ways of mitigating it depends on the respective domain, or at least the broader categorization of the domain.

One perspective on why hallucinations persist in LLMs is that existing guardrail mechanisms are inadequate and fail to effectively prevent them. According to Pantha et al. (2024), generic guardrails face significant challenges when applied to specialized domains. Ideally, LLMs should rely on guardrails to issue alerts when they lack domain-specific understanding, signaling their inability to generate a reliable response. However, in reality, these mechanisms are often inadequate. The problem is that while some LLMs recognize their limitations and trigger alerts, many fail to detect their own uncertainty and instead generate misleading responses with unwarranted confidence. This failure stems from the lack of domain-specific understanding, preventing LLMs from correctly identifying when their outputs should be restricted. As a result, hallucinations persist. The inadequacy of generic guardrails highlights the need for domain-specific interventions. Research has explored customized guardrails designed to mitigate hallucinations in scientific applications. Proposed frameworks and methodologies address key challenges, including time sensitivity, contextualization of knowledge, and intellectual property concerns. By integrating domain-aware constraints, these solutions aim to enhance the trustworthiness and reliability of LLM-generated content in specialized fields.

3.2 Incompleteness and Outdated Data

LLMs struggle with processing or generating up-to-date information for the NLU task, as their training data is static and may not capture up-to-date domain-specific knowledge. This limitation arises because most pre-trained LLMs rely on datasets collected at a specific time, making them unable to reflect real-time changes, temporal trends, or evolving domain-specific knowledge (Fan et al., 2024; Gao et al., 2023; Lewis et al., 2020b). For example, an LLM-based system used in the manufacturing sector for tasks such as reasoning, design assistance, or technical question-answering may miss recent innovations in materials, automation methods, or production standards (Li et al., 2024a). This lack of timely updates can lead to outdated or incomplete insights, limiting the system's usefulness in fast-moving industrial contexts. Similarly, in healthcare, where medical treatments, clinical guidelines, and biomedical discoveries evolve rapidly, an outdated LLM may generate recommendations based on outdated information, with potentially serious consequences. To overcome these challenges, it is crucial to integrate mechanisms that enable LLMs to access external, up-to-date knowledge bases or adapt dynamically to evolving datasets. The strategies for handling outdated knowledge enable models to bridge the gap between static training data and the real-time knowledge needed for accurate, reliable decision-making across diverse domains.

A common challenge with LLMs, closely related to the problem of incomplete or outdated knowledge, is their inconsistent response generation, that is, producing varied outputs for the same input. Even when the required external knowledge is available, the reliability of an NLU system also depends on the model's ability to generate stable and predictable outputs. Inconsistency arises from several factors, including limited or imbalanced training data, stochastic sampling strategies, sensitivity to prompt phrasing, training data biases, and architectural design choices (Zhao et al., 2024b). Mathematically, LLMs are deterministic: given identical inputs and fixed model parameters (temperature = 0, fixed random seed), they produce the same output. However, most real-world applications intentionally introduce stochastic sampling through hyperparameters like temperature and top-p to promote diverse text generation. This causes LLMs to probabilistically sample tokens at each generation step, resulting in output variability across repeated prompts. Beyond stochasticity, another inherent source of inconsistency lies in the stateless nature of LLMs. Unlike systems such as recommender engines or chatbots that depend on historical user interactions or time-dependent profiles, LLMs operate without memory of previous exchanges. This statelessness improves privacy and security by avoiding storage of user-specific data, but it also presents difficulties for tasks requiring continuity or persistent reasoning. Without retained context, responses may appear arbitrary or disconnected when the input lacks explicit contextual cues. As a result, the stateless nature of LLMs contributes to the unpredictability of their outputs, leading to inconsistent responses (Liu et al., 2023, p. 7). Such variability poses a significant concern, as it affects the reliability and trustworthiness of LLMs across a wide range of applications (Saxena et al., 2024).

3.3 Lack of Contextual Understanding and Related Ambiguities

LLMs are sensitive to the noisy language of prompts. The lack of contextual understanding can contribute to linguistic ambiguities, as LLMs rely on statistical patterns inferred from the text rather than explicit reasoning (Keluskar et al., 2024). For instance, words with multiple meanings, such as "bank" (which could refer to a financial institution or the side of a river), can be misinterpreted if the model lacks sufficient domain-specific knowledge or contextual clues. This limitation becomes particularly evident in specialized fields or nuanced conversations where precise understanding is critical. Consequently, semantic disambiguation in LLMs remains an active research area (Yang et al., 2023). Keluskar et al. (2024) analyze ambiguities in LLM-generated responses within the context of open-domain question answering. They argue that ambiguity primarily arises due to a lack of context, which includes not only textual cues but also social and psychological ones. Their study highlights the need to integrate external knowledge sources, such as knowledge graphs, to enhance clarity and disambiguation. Their results demonstrate that contextual enrichment, that is, contextual information to LLMs, significantly reduces disambiguation and increases accuracy. Unlike scientific experts, LLMs struggle to understand the nuances of experimental setups or domain-specific methodologies. Humans can interpret the meaning of a document, identify and link related data across tables, and infer implicit relationships and contextual insights. LLMs, however, are generally unable to perform such reasoning reliably, limiting their ability to extract and contextualize knowledge from complex scientific texts. This often results in superficial or incomplete responses, such as misinterpreting parameters like temperature or pressure in experimental setups, because the context is not clearly defined.

Although LLMs possess relevant knowledge, they often struggle to apply it when prompts contain ambiguous entity types. Therefore, understanding different types of ambiguities is crucial for improving LLM-powered NLU tasks. Ambiguities can be categorized into three primary types: semantic, syntactic, and lexical (Zait & Zarour, 2018). Semantic ambiguity, also known as referential ambiguity, pertains to multiple interpretations of a word, phrase, or sentence. For example, the question, "What is the home stadium of the Cardinals?" yields different answers depending on whether it refers to the Arizona Cardinals (football) or the St. Louis Cardinals (baseball) (Keluskar et al., 2024). As this example illustrates, even humans struggle to resolve meaning without sufficient context, making it even more challenging for LLMs (Zait & Zarour, 2018). Lexical ambiguity arises at the word level and is related to parts-of-speech tagging. For instance, the word silver can function as a noun, adjective, or verb: "She bagged two silver (noun) medals.", "She made a silver (adjective) speech.", and "His worries had silvered (verb) his hair." (Anjali & Babu, 2014). Lastly, syntactic ambiguity concerns the grammar and structure of a sentence. A widely cited example is: "I saw the man with the telescope." This can be interpreted as either "I saw the man [who was holding the telescope]." or "I used the telescope to see the man." Different types of ambiguities contribute to vagueness, fuzziness, and uncertainty in language, leading to confusion in LLM responses. Augmenting contextual information or extending LLMs with external knowledge bases can mitigate these issues (Singh & Patil, 2024). Additionally, LLMs struggle with self-verification, demonstrating a lack of self-consistency. This highlights a key challenge in polysemy resolution, emphasizing the need for further research into entity-type ambiguities and the broader complexities of language understanding (Sedova et al., 2024).

4 Approaches of Integrating External Knowledge Into LLM-Based NLU Systems

In this survey, external knowledge refers to information supplied through domain-specific documents, structured resources like knowledge bases or relational databases, and ontologies that formally represent domain expertise. Such specialized knowledge covers factual details—concepts, relations, and constraints that are unique to a domain, but also extends to insights from unstructured sources, including narrative text or document passages. Both structured and unstructured knowledge play essential roles in enabling reasoning within a specific domain. For example, interpreting technical datasheets can greatly benefit from integrating domain-specific ontologies with a defined set of entities and relations to achieve a comprehensive understanding of the mentioned concepts. In specialized domains such as science and engineering, integrating external knowledge can reduce hallucinations, enhance reasoning, improve factual accuracy, and boost overall task performance. For instance, the work by Wang and Li (2023) identifies challenges in using LLMs for manufacturing applications and recommends integrating an external knowledge base to generate industry-relevant insights. Another example of external knowledge integration can be seen in text-to-query systems, where LLMs translate natural language inputs into formal query languages such as SQL or SPARQL. In both cases, the model must understand and reason over the underlying structured schema—tables and relations in databases or entities and predicates in knowledge graphs, to generate syntactically correct and semantically meaningful queries. These systems demonstrate how LLMs integrate with structured external knowledge sources to produce actionable outputs, generating knowledge that extends beyond their pre-trained parameters.

4.0.0.1 Classification of knowledge:. In the context of LLMs, data refers to text/corpora from books, websites, articles, among others, and knowledge is derived by processing the data during the training phase, that is, the data is processed to create knowledge (da Costa & Oliveira e Souza Filho, 2024). LLMs store internal knowledge within their parameters, which is derived during the training phase from large-scale, unstructured, and unlabeled datasets. For larger LLMs with over 100 billion parameters, this training data encompasses vast amounts of internet-scale information. The knowledge stored in model parameters is represented by numerical values known as weights, which are learned probabilistically through the backpropagation algorithm. This form of knowledge retention is referred to as *implicit memory* or *internal knowledge*. However, this knowledge is inherently limited as it only reflects data available up to the model's last training timestamp. Additionally, the training corpus is generic, often lacking specialized domain-specific information due to confidentiality restrictions that prevent such documents from being included in publicly available datasets. Furthermore, even within general datasets, highly detailed or technical documents may be disproportionately represented. Therefore, LLMs often requires external data to perform domain-specific or knowledge-intensive tasks, which serves as the additional contextual knowledge. Unlike implicit memory, this external data is maintained outside the model, typically in knowledge bases, relational databases, or structured file systems. When LLMs consume and process such external information, it is referred to as *external knowledge*. Integrating external knowledge with LLMs is crucial for reducing hallucinations, enhancing contextual understanding, and ensuring access to up-to-date information. This integration can be done in two ways, at a broader level: (i) parametric knowledge integration, that is, knowledge is encoded within the parameters of the LLMs, (ii) non-parametric knowledge integration, where the knowledge is not encoded within the model weights, but supplied from a storage system outside the LLM, like relational or knowledge databases. Table 2 summarizes the parametric and non-parametric methods, and the following sections explain these methods in detail.

4.1 Parametric Knowledge

As mentioned earlier, parametric knowledge is implicitly embedded within the model's architecture (Allen-Zhu & Li, 2025a). Therefore, parametric methods for integrating knowledge with LLMs involve modifying their internal parameters and encoding/embedding information during training or fine-tuning.

4.1.1 Pre-Training LLMs. The development of LLMs typically follows a hierarchical training continuum that transitions from general-purpose learning to task- and fact-specific adaptation. At the foundation lies pre-training, where models are trained from scratch on large, unlabeled, and diverse corpora using self-supervised objectives (e.g., masked or causal language modeling) to acquire general linguistic and world knowledge. Foundational examples include encoder-based models such as BERT (Devlin et al., 2019) and decoder-based models such as GPT (Radford et al., 2018). Following the general pre-training phase, domain-adaptive pre-training (DAPT) continues training a pre-trained model on unlabeled, domain-specific corpora to better align its internal representations with specialized knowledge (e.g., biomedical, legal, or financial domains) (Chalkidis et al., 2020; Peng et al., 2021; Vasantharajan et al., 2022). For example, LEGAL-BERT (Chalkidis et al., 2020) and FinBERT (Peng et al., 2021) are further trained on unstructured textual data from their respective domains—LEGAL-BERT on diverse English legal texts such as legislation, court cases, and contracts, and FinBERT on the Reuters TRC2 financial news corpus ((2004), NIST). Both models demonstrate superior performance compared to the original BERT model on downstream tasks within their domains. In continuation of DAPT, task-adaptive pre-training (TAPT) further refines the model using unlabeled text drawn directly from the downstream task domain, enabling better alignment with the stylistic and distributional characteristics of the task data. This approach, formalized by Gururangan et al. (2020), has been shown to improve model robustness and task-specific generalization before supervised fine-tuning.

Building upon these stages, several studies have explored enhancing the model by incorporating external structured knowledge into the language modeling process. For instance, Yan et al. (2021) introduce a knowledge fusion layer on top of the transformer architecture to integrate knowledge graph information during pre-training without altering the underlying model structure. Similarly, Dong et al. (2023) propose a structured knowledge–aware pre-training framework that embeds structured knowledge into the model using the masked language modeling (MLM) objective from BERT, enabling it to learn representations of complex subgraphs for improved performance on Knowledge Base Question Answering (KBQA) tasks. Related approaches that inject knowledge into the pre-training, DAPT, and TAPT phases include (Hu et al., 2024; Zhang et al., 2022, 2023), all of which demonstrate that structured knowledge integration enhances model understanding and reasoning capabilities.

In the context of (large) language models, NLU can be viewed as a process of manipulating and extracting knowledge, where relevant information is identified and adapted to meet task-specific requirements. One illustrative example is IE, which involves deriving structured signals, such as entities or relations, from data, retrieving pertinent facts encoded within the model, or classifying sentences into predefined categories (Allen-Zhu & Li, 2025b). Such operations highlight how

LLMs leverage and transform knowledge to support a broad range of NLU tasks. However, despite their ability to store vast amounts of information, LLMs often struggle to extract and manipulate knowledge effectively. For instance, a model trained on the fact "Abraham Lincoln was born in Hodgenville, K.Y." may correctly answer the direct question "Where was Abraham Lincoln born?" but fail to respond to the inverse query "Who was born in Hodgenville, K.Y.?" unless explicitly trained with bidirectional mappings. This type of retrieval is called reverse retrieval. This limitation highlights that memorizing knowledge alone does not guarantee effective knowledge extraction. To address this gap, Allen-Zhu and Li (2025b) propose strategies to enhance LLMs performance by refining knowledge representation processes. One approach involves rewriting pre-training data using small auxiliary models to generate diverse permutations of knowledge, thereby improving retrieval flexibility. Another method focuses on incorporating fine-tuning data during pre-training, enhancing the model's ability to reason over stored information.

4.1.2 Fine-Tuning LLMs. Unlike pre-training, DAPT, and TAPT, fine-tuning uses labeled, task-specific datasets and supervised learning objectives to adapt the model to specific applications, such as question answering, text classification, and summarization. The BERT paper (Devlin et al., 2019) distinguishes pre-training and fine-tuning for larger language models. During fine-tuning, the parameters of a pre-trained language model are updated using labeled examples to optimize its performance on downstream NLP tasks (Howard & Ruder, 2018; Soudani et al., 2024). For instance, Lee et al. (2019) first adapt BERT to the biomedical domain through additional pre-training on biomedical text, resulting in BioBERT, and subsequently fine-tune it on labeled, task-specific datasets for downstream NLP tasks such as NER, relation extraction, and question answering. LLMs, when fine-tuned, can be used to extract factual knowledge. Kazemi et al. (2023) fine-tune LLMs to do knowledge extraction and KGC. In general, fine-tuning has been effective in overcoming performance limitations on domain-specific NLU. The work of Muralidharan et al. (2024) examines the effectiveness of LLMs in understanding and extracting information in the scientific domain. They propose a methodology called Knowledge-AI, which fine-tunes LLMs for downstream NLU tasks, such as question answering, NER, summarization, and text generation.

Exploring different approaches to fine-tuning LLMs is essential for understanding their practical implications, scalability, and resource requirements. Fine-tuning strategies differ primarily in the number of parameters updated during training, influencing computational efficiency, memory consumption, and model adaptability. Full fine-tuning modifies all parameters of an LLM, typically achieving strong task-specific adaptation. However, this approach is computationally expensive, requiring high-end GPU clusters, particularly for models with more than 100 billion parameters (Rajabzadeh et al., 2024). To address these constraints, the NLP community has increasingly turned to the Parameter-Efficient Fine-Tuning (PEFT) paradigm, which updates only a small fraction of parameters, or introduces a small number of additional trainable parameters—while keeping the core model weights frozen. This reduces hardware requirements and accelerates training, making fine-tuning feasible on modest computational setups. Within this paradigm, several techniques have emerged, including *Low-Rank Adaptation (LoRA) (Hu et al., 2022), prefix-tuning (Li & Liang, 2021), prompt-tuning (Petrov et al., 2023), BitFit (Ben Zaken et al., 2022), and adapter-based fine-tuning (Poth et al., 2023)*. Among these, adapter-based methods are one of the earliest and most prominent forms of PEFT. These methods introduce small, trainable neural modules, known as adapters, within the transformer layers of a pre-trained model. During fine-tuning, only the adapter parameters are updated.

For instance, Tian et al. (2024) combine knowledge graphs (KGs) with LLMs at the parameter level to improve reasoning. They note that simply adding KG information via prompting can introduce inconsistencies and lead the model to overly rely on its prior knowledge. To address this, they introduce a KG-adapter that leverages PEFT-based fine-tuning to embed both node and relation representations from the KG into the model. This approach enhances reasoning performance on KGQA. Similarly, Omeliyanenko et al. (2023) introduce an adapter-based architecture for integrating KG knowledge into LLMs for link prediction tasks. Their approach inserts lightweight, trainable layers between the transformer blocks of a pre-trained model, enabling task-specific adaptation without full model retraining (Wang et al., 2021). Extending these foundations, Wang et al. (2024a) propose an infuser-guided adapter integration framework—an optimized variant of adapter-based fine-tuning. This approach effectively mitigates catastrophic forgetting during knowledge integration. To ensure the added modules do not interfere with the model's internal representations, they introduce infuser-based adapters, which selectively inject knowledge only when the model lacks it. This selective infusion mechanism delivers superior performance compared to conventional adapter-based methods, indicating a promising direction for adaptive, knowledge-aware LLM fine-tuning.

Fine-tuning LLMs requires significant resources and computational power and comes with its own set of challenges. Kazemi et al. (2023) highlight a downside known as "Frequency Shock", where fine-tuned models tend to overpredict rare entities while underpredicting common ones in the training data, ultimately degrading performance. Ghosal et al. (2024) address another issue: "attention imbalance", where the attention mechanism unevenly prioritizes specific tokens over others. Ovadia et al. (2024) compare two approaches for embedding factual knowledge into LLMs—unsupervised

fine-tuning and Retrieval-Augmented-Generation (RAG). Their findings suggest that RAG outperforms unsupervised fine-tuning. They also note that unsupervised training often exposes models to multiple variations of the same fact, complicating the accurate retention of factual information. These challenges emphasize the importance of exploring non-parametric methods for integrating external knowledge into LLMs (cf., Section 4.2).

4.1.3 Knowledge Editing. Incorporating new knowledge into the parameters of LLMs through fine-tuning is computationally expensive due to their vast scale and the billions of parameters they contain. This challenge has motivated the development of more efficient mechanisms for integrating external knowledge without the high computational cost of full model retraining. Knowledge-based model editing (KME), or knowledge editing, addresses this by modifying only a small and targeted subset of parameters responsible for encoding specific information, thereby updating the model's knowledge while preserving its pre-trained capabilities (Pinter & Elhadad, 2023; Wang et al., 2024c).

Unlike fine-tuning, which typically adjusts all or a broad range of parameters to optimize model performance for a downstream task, KME focuses on localized and semantically precise modifications. As highlighted by Ishigaki et al. (2024), knowledge editing generally proceeds in two stages: (a) localization, where the model identifies the neurons or attention heads associated with the target knowledge, and (b) modification, where only those parameters are updated to reflect new or corrected information. This targeted approach enables locality, ensuring that updates affect only the intended knowledge, and generality, allowing the edit to generalize across semantically related contexts (Wang et al., 2024d).

Beyond its application in LLMs, knowledge editing techniques have broader implications in machine learning, such as mitigating data biases and improving model robustness on downstream tasks (Ilharco et al., n.d). By maintaining the integrity of previously learned information while efficiently integrating new knowledge, KME provides a computationally efficient and semantically stable alternative to fine-tuning, making it especially valuable for dynamically updating LLMs in knowledge-intensive domains.

To better illustrate the conceptual and computational distinctions between pre-training, fine-tuning, and knowledge editing, Table 3 provides a structured comparison highlighting their objectives, data requirements, training scope, efficiency, and key representative techniques. This comparative overview clarifies the boundary between full-scale model adaptation and targeted knowledge modification, situating knowledge editing as a lightweight yet powerful alternative for dynamically updating LLMs in knowledge-intensive settings.

4.1.4 Steering and Styling LLMs. Steering an LLM involves guiding its output to align with specific stylistic, instructional, or knowledge-driven constraints (Lai et al., 2024). While it is commonly applied to text style transfer (TST)—for instance, adapting an LLM to generate Shakespearean-style language, steering also plays a crucial role in knowledge integration, ensuring that models correctly interpret and apply external information in their responses. There are multiple perspectives on steering and styling LLMs, many of which facilitate the integration of external knowledge. Steering techniques can be broadly classified into parametric and non-parametric approaches[9]. Parametric methods, such as neuron deactivation or fine-tuning, modify model parameters to influence responses. On the other hand, non-parametric methods, such as prompt-based steering, control model behavior without altering the underlying parameters. Lai et al. (2024) propose a novel parametric steering approach called sNeuron-TST for steering the style. This method identifies the neurons associated with source and target styles, deactivating the source-style neurons to enforce the target style in generated text. However, they observe that deactivating these neurons leads to performance degradation. To address this, they introduce a constructive decoding method that compensates for the removed neurons, improving output consistency. Beyond stylistic control, the similar parameter-level steering mechanisms of LLMs are also applicable to knowledge integration, as they direct the model's attention mechanisms to prioritize user-specified information (e.g., instructions or domain-specific knowledge). This is achieved by identifying subsets of attention heads and applying attention reweighting, which enables the model to effectively incorporate and process new information. Such steering methods have been shown to enhance performance on knowledge-intensive tasks (Lamb et al., 2025; Zhang et al., 2024c). Therefore, steering techniques ensure that models remain aligned with the external knowledge sources.

4.1.5 Other Methods and Limitations of Parametric Approaches. Other methods of parametric data augmentation to LLMs include embedding techniques. They transform external knowledge into vector representations that can be integrated into the LLM's latent space, thereby capturing the text's linguistic features. The NLP community has used embeddings for years to transform raw textual information into numerical representations that can be processed by AI algorithms (Zhang et al., 2024a). Additionally, methods such as graph embeddings (Jain & Lapata, 2024; Pan et al., 2024) and knowledge graph transformers (Liu et al., 2024) enable the model to learn richer representations of external knowledge, thereby improving its contextual understanding and reasoning abilities. Another prominent parametric technique is *knowledge distillation*, where a large teacher model transfers its knowledge to a smaller student model by training the

Table 3. Comparison of Pre-Training, Fine-Tuning, and Knowledge Editing in Large Language Models (LLMs).

Aspect	Pre-training (Sec. 4.1.1)	Fine-tuning (Sec. 4.1.2)	Knowledge Editing (Sec. 4.1.3)
Objective	Learn general linguistic and world knowledge from large unlabeled corpora.	Adapt pre-trained models for specific downstream tasks using labeled data.	Modify specific factual or conceptual knowledge without full retraining.
Data Type	Large-scale, unlabeled, diverse textual corpora.	Task-specific, labeled datasets.	Targeted factual or conceptual updates (e.g., correcting or adding specific facts).
Training Scope	Full model training from scratch (all parameters).	Updates all or a small subset of parameters or adds trainable modules while keeping the rest frozen (e.g., PEFT, adapters, LoRA).	Modifies only a small, localized subset of parameters or neurons (Wang et al., 2024c).
Supervision Type	Self-supervised learning (e.g., MLM, CLM).	Supervised or semi-supervised learning.	Usually unsupervised or weakly supervised; guided by target knowledge specification (Wang et al., 2024d).
Computational Cost	Extremely high (training from scratch with massive resources).	Moderate to high (depending on the number of parameters updated).	Low (localized parameter modification, no full retraining) (Pinter & Elhadad, 2023).
Parameter Efficiency	Low, all parameters trained.	Moderate to high, PEFT methods improve efficiency.	Very high, edits only a few targeted parameters.
Knowledge Integration Mechanism	Implicitly learns knowledge from text corpora.	Injects task- or domain-specific knowledge via fine-tuning or adapters.	Identifies and modifies neurons encoding specific knowledge (localization + modification) (Wang et al., 2024d).
Key Techniques / Examples	BERT (Devlin et al., 2019), GPT (Radford et al., 2018), DAPT (Chalkidis et al., 2020; Peng et al., 2021), TAPT (Gururangan et al., 2020).	Full fine-tuning (Devlin et al., 2019), LoRA (Hu et al., 2022), Adapters (Poth et al., 2023), Prefix/Prompt Tuning (Li & Liang, 2021; Petrov et al., 2023)	MEND (Mitchell et al., 2021), ROME (Andonian et al., 2022), MEMIT (Meng et al., 2022), EasyEdit (Wang et al., 2024c).
Advantages	Builds strong general-purpose representations.	Enables domain/task specialization and improved downstream performance.	Efficiently updates or corrects model knowledge while preserving prior learning (Ilharco et al., n.d).
Limitations	High computational cost and data requirements.	Risk of catastrophic forgetting and resource-intensive for large models.	Limited scope of edits; potential instability in compositional reasoning (Wang et al., 2024d).

student to approximate the teacher's outputs. This process embeds the teacher's knowledge directly into the student's parameters, allowing for model compression and efficiency gains while retaining much of the teacher's performance (Hinton et al., 2015; Sanh et al., 2019; Wang et al., 2020). In the context of LLMs, knowledge distillation is widely used to reduce computational overhead and memory requirements, thereby making LLMs easier to deploy and more scalable.

Parametric knowledge within LLMs often faces challenges, such as a lack of explainability, as it is stored within the model weights; that is, the knowledge is converted into numerical values, making its provenance difficult to trace. This might lead to security risks due to its opaque nature. Another challenge is that updating knowledge is computationally expensive and time-consuming, as parametric knowledge updates require some level of modification to model weights/layers/architecture. To address these limitations, external non-parametric knowledge emerges as a good option, offering enhanced transparency, flexibility, adaptability, and operational simplicity (Wang et al., 2024b). Table 4 summarizes how these methods differ from non-parametric strategies in terms of computation, explainability, flexibility, and update mechanisms.

Table 4. Comparison Between Parametric and Non-Parametric Methods.

Aspect	Parametric Methods	Non-Parametric Methods
Computational Requirements	Computationally expensive and time-consuming; requires high-end GPU clusters for fine-tuning.	Less computationally expensive; RAG reduces load by avoiding full parameter-based storage.
Knowledge Updates	Requires modification of model weights or layers, which carries the risk of catastrophic forgetting.	Knowledge is stored externally, allowing for dynamic updates via external sources.
Explainability	Low explainability; knowledge embedded in weights.	High transparency and provenance can be traced.
Knowledge Representation	Implicitly encoded in parameters, memorization may not ensure accurate recall.	Explicitly stored in some external system, such as knowledge bases, retrieval systems, or an external model.
Specific Challenges	Frequency shock, attention imbalance, and difficulty in reverse retrieval.	Retrieval quality affects generation; context window limits usable information.
Flexibility	Static knowledge; retraining required for updates.	Adaptable through external access; supports real-time updates.

4.2 Non-Parametric Knowledge

As previously discussed, non-parametric knowledge refers to information or knowledge provided to the LLM without altering its internal weights or architecture. This type of knowledge is stored in a separate system outside the model and is not encoded within its trainable parameters. The following are the most commonly used approaches for integrating external and up-to-date knowledge into LLMs in a non-parametric manner.

4.2.1 Prompting Methods. Prompting refers to the process of providing textual instructions to an LLM to elicit desired task-specific behavior. Common prompting strategies include zero-shot prompting, where only the instruction is provided without examples; few-shot prompting, which supplements the instruction with a few input–output examples; and chain-of-thought (CoT) prompting, which guides the model by decomposing the reasoning process into intermediate steps (Brown et al., 2020; Sun et al., 2024). As a non-parametric approach, prompting does not alter the model's internal parameters or weights. Instead, it conditions the model externally via contextual cues by embedding relevant information directly into the prompt, within the LLM's context limit. While techniques such as CoT prompting (Bosma et al., 2022) improve reasoning tasks like retrieval and classification, they remain insufficient for more complex operations, such as inverse search, where models must infer relationships beyond explicitly stated data, as explained in Section 4.1.1. In such cases, advanced methodologies like RAG and reversal training are essential, as only a limited amount of context can be included in the prompt due to the LLMs 's context limits. Lack of context results in issues like hallucination and inappropriate response generation. To solve this problem, a new field of research emerged called context engineering (Piñeiro-Martín et al., 2025; Rasmussen et al., 2025), in which relevant context is provided to the LLM to support knowledge-intensive NLU tasks. RAG and knowledge graph (KG) integration, discussed in detail in the following sub-sections, can indeed be conceptualized as forms of context engineering. These findings highlight the importance of knowledge integration in LLMs, bridging the gap between limited contextual understanding and the effective application of knowledge in real-world, knowledge-intensive tasks (Allen-Zhu & Li, 2025a, 2025b)

4.2.2 Knowledge-Based Methods: Knowledge Graphs and Ontologies.

4.2.2.1 Knowledge Graphs and their role in LLMs:. Knowledge graphs, such as DBpedia (Auer et al., 2007), or Wikidata (Vrandečić & Krötzsch, 2014), are external knowledge sources that store knowledge in the form of nodes and edges (Fensel et al., 2020). By linking LLMs with these graphs, it is possible to enhance the models' ability to reason about entities and their relationships. In recent years, significant research has focused on integrating knowledge graphs and LLMs to reduce hallucinations and improve factual accuracy. LLMs and KGs complement each other's capabilities. Merging both helps overcome each other's limitations. LLMs gain access to factual knowledge from KGs, which improves accuracy and trustworthiness, as in fact-checking. KGs, on the other hand, benefit from LLMs in language processing and language understanding tasks like: synthesis of user responses, question-answering, automated KG construction, etc. (Choudhary & Reddy, 2023; Luo et al., 2024a; Pan et al., 2024). There are various ways to integrate KGs and LLMs. First, *KG-enhanced LLMs* leverage KGs as external knowledge sources to provide domain-specific context to LLMs, thereby improving the

factual accuracy of their outputs. Second, *LLM-augmented KGs*, where LLMs are used for populating the triples in the knowledge graph, assist in KG-related tasks, and incorporate ontologies to ensure that the generated triples conform to domain-specific rules and constraints. Third is the *synergized LLMs+KGs*, a unified framework that aims to enhance each other's capabilities through knowledge representation and reasoning (Pan et al., 2024).

KG-enhanced LLMs, LLM-augmented KGs, and synergized KG+LLMs are methods for integrating KGs and LLMs. However, the focus of the work is KG-enhanced LLMs, where KG serves as the external source. For instance, Ye et al. (2024) introduce a method to integrate KGs with Large Language Models (LLMs) to enhance factual accuracy. Their approach employs deep reinforcement learning, which identifies relevant inference paths within the KG based on user input and incorporates this information into the LLM prompt, thereby providing context that yields more domain-specific and accurate results. The work by Wang et al. (2025) discusses the challenges of handling evolving knowledge in the medical domain, emphasizing the need for continuous model updates and the integration of external knowledge. They propose a 3-step framework to develop an LLM-powered AI application for the medical domain: (i) modeling (which breaks down a complex task into simpler sub-tasks), (ii) optimization (enhancing the model response generation relevancy and accuracy by integrating external knowledge), and (iii) system engineering. They also highlight that more research is needed on system optimization, specifically on augmenting external knowledge with LLMs. Chen et al. (2024) handle out-of-date information issues in LLMs by integrating knowledge graphs to facilitate accurate fact identification and logical reasoning. They propose a method called Graph Memory-based Editing for Large Language Models (GMeLLo), which combines the strengths of both KGs and LLMs to perform multi-hop question answering in dynamic environments, that is, those with frequently updated external knowledge. Such advancements not only enhance the performance of LLMs in NLU tasks but also foster greater trust and adoption of these models across critical fields such as healthcare, science, and education. Building on these integration strategies, researchers have explored other practical applications that leverage the synergy between KGs and LLMs to address specific tasks, such as question answering and improving explainability. Feng et al. (2023), integrate KGs with LLMs to develop a multiple options question answering system. They call their methodology a knowledge solver, which allows them to search for relevant facts in the integrated knowledge graph. The approach also increases the explainability of LLMs' reasoning processes by providing complete retrieval paths, as demonstrated through experiments on the datasets: CommonsenseQA (Talmor et al., 2021), OpenbookQA (Mihaylov et al., 2018), and MedQA-USMLE (Jin et al., 2021). There are other works in the same research area. For instance, Jiang et al. (2024) highlight the importance of connecting domain-specific KGs to LLMs for the task of domain-related question answering. They propose a subgraph retrieval method based on the CoT and PageRank, which returns the paths most likely to contain the answer, thereby improving the efficiency of the given NLU task (domain-dependent question answering). Additionally, Luo et al. (2024a) integrate KGs with LLMs to enable reasoning abilities for the task of knowledge-graph question answering (KGQA). They propose planning-retrieval-reasoning: LLMs first create a reasoning plan and then perform reasoning, which involves fetching reasoning paths from the KG using the generated plan. Hallucinations happen if the reasoning plan is incorrect. The aim is to distill the knowledge from KGs into LLMs to generate faithful relation paths as plans. Research studies such as (Ji et al., 2024; Ma et al., 2025; Wang et al., 2023a) explore similar approaches, focusing on using LLMs for reasoning and question answering by incorporating knowledge graphs. However, when developing methods that combine KGs and LLMs, scalability must be a key consideration. Without effective scalability, integrating extensive knowledge bases or graphs can demand substantial resources, posing significant challenges for large-scale deployment (Zhang et al., 2024b).

4.2.2.2 Constrained-Decoding in LLMs:. Recent advancements in natural language generation (NLG) have introduced constrained-decoding methods for LLMs (Chen et al., 2022; Hokamp & Liu, 2017; Post & Vilar, 2018), which apply ontological rules, among other constraints, to ensure that LLM outputs maintain logical consistency and conform to domain-specific structures (Pan et al., 2024). It can also be seen as a way of integrating external knowledge with LLMs. The purpose of this integration is to provide a response adhering to the provided input syntax or semantics. Traditional constrained decoding with LLMs was applied at the syntax level; a typical example is the introduction of JSON mode (Tam et al., 2024). Other frameworks for syntax-level constrained generation include Outline (Outline, 2026) and Guidance (Guidance, 2026). They can output formats like JSON, XML, python scripts, and SQL, amongst others. However, the concept can also be extended to the level of semantics. A foundational work in this area is presented by Hokamp and Liu (2017), where they utilize lexicons to control the generation process of LLMs. An evolving research area under the umbrella of semantic-based constrained decoding concerns interactions and potential synergies between graph-based knowledge systems and LLMs. Ontologies, the foundational framework for structuring and organizing knowledge graphs, formally represent the domain by serving as the Terminology Box (T-Box) that defines classes and their relationships (Luo et al., 2024b). These structured frameworks are instrumental in guiding the extraction and construction of domain-specific knowledge graphs and in constraining and refining LLM outputs to align with established schemas. Constrained-decoding

leverages entities from the KG or the schema and structure of ontologies to regulate LLM responses, ensuring relevance and adherence to domain-specific logic. Technically, constrained decoding encompasses methods for controlling the output tokens of LLMs using external knowledge sources, such as controlled vocabularies, taxonomies, structured ontologies, or domain-specific knowledge graphs. To summarize, ontologies and knowledge graphs provide the necessary structure and knowledge to anchor LLM-generated outputs, enabling the generation of domain-relevant language that aligns with the ontology's semantics. Although constrained decoding yields precise and structured responses, it is limited by increased latency, resulting in longer response generation times compared to unconstrained generation (Geng et al., 2023).

4.2.3 Memory-Based Systems.

4.2.3.1 Retrieval-Augmented Generation (RAG):. It is a kind of memory-oriented framework for LLMs, since it facilitates the external storage (*in the vector-database*) and dynamic retrieval of information. The work by Akbar et al. (2025) discusses the use of vector databases and RAG frameworks to manage memory for conversational AI systems. The RAG approach integrates dense vector search with LLM-based text generation, thereby improving response quality in knowledge-intensive applications. As its name implies, RAG retrieves contextually relevant information for a query using dense vector retrieval techniques. The retrieved information is then embedded into the LLM's prompt during text generation, improving both the factual accuracy and contextual relevance of the output. Rather than training or fine-tuning the LLM with vast amounts of knowledge, which consumes substantial resources and time, external knowledge can be dynamically retrieved using methods such as RAG. This reduces computational load and enables LLMs to scale efficiently while still benefiting from external knowledge (Li et al., 2025c). It is most useful in scenarios where LLMs lag behind task-specific architectures. This approach enables the model to access and incorporate knowledge in real time, eliminating the need to store the entire knowledge base within its parameters. This significantly reduces computational overhead and enhances the model's decision provenance (Lewis et al., 2020b).

As highlighted above, RAG involves storing data in vector databases, such as Elasticsearch (Elastic, 2025), ChromaDB (Chroma, 2025), or others. Vector databases store vector representations of text, which are multidimensional arrays of numbers. They serve as the external memory. Several frameworks are available online that facilitate the seamless integration of RAG with LLMs. Examples include Hugging Face's RAG implementation, Haystack (deepset, 2025), and LangChain (Mavroudis, 2024), among others. These frameworks provide tools and libraries to streamline the development of RAG pipelines, enabling efficient retrieval and context-enhanced text generation.

External knowledge integration into LLMs via RAG can be viewed as both parametric and non-parametric. Naive RAG implementations include using a pre-trained dense vector retriever, for example, instruct-embeddings (Su et al., 2023) and connecting it with a pre-trained LLM. In this scenario, neither of the two models—the retrievers nor the LLMs—is further fine-tuned, that is, following the non-parametric approach. The naive RAG approach focuses solely on the inference stage. In contrast, advanced RAG implementations offer two methods. The first method fine-tunes the dense-vector retriever for the specific task or domain. This is a non-parametric approach because the LLM's parameters remain unchanged. The second method fine-tunes both the retriever and the LLM for the specific task or domain. This is a parametric approach, in which the LLM's parameters and weights are adjusted to suit the use case. Some progress has also been made in pre-training RAG systems (Gao et al., 2023).

Further research on RAG implementations includes the following: RAG task categorization (Zhao et al., 2024a), where queries are classified into levels based on the type of external data required: explicit fact queries, implicit fact queries, interpretable rationale queries, and hidden rationale queries. This is done to ensure that the correct and most appropriate data is fetched for the given task and provided to the prompt as context. Zhao et al. (2024a) propose three methods of ingesting data into LLMs, that is, context (providing the retrieved context directly to the LLM), small model (adding a small model trained on the domain to help external data integration with LLMs), and fine-tuning (fine-tuning the LLM). They also highlight the challenges of deploying data-augmented LLMs and believe that there is no one-size-fits-all solution for it. Anjum et al. (2025) propose a framework called HALO for mitigating hallucinations in the medical question-answering systems to enhance reliability and accuracy. They use the RAG technique to integrate domain-specific information with the LLMs. The results show an increase in LLM accuracies from 44% to 65% for Llama-3.1 and from 56% to 70% for ChatGPT. Additional work in a similar line of research includes Hwang et al. (2025), who introduce Reliability RAG, an extension of traditional RAG systems designed to handle multiple data sources. Their approach focuses on estimating the reliability of various sources within the database. Similarly, Khan et al. (2024) developed a PDF-based, LLM-powered RAG system, showcasing its application in processing and retrieving information from PDF documents.

Traditional RAG techniques have several limitations that can negatively impact the quality of generated responses. In particular, the retrieval component often returns irrelevant or low-quality results, which directly affects the final output. Key issues in retrieval include the lack of pre-retrieval query processing, the absence of post-retrieval enhancements such

as reranking, and text chunking that disregards semantic boundaries. The generation component also faces constraints, including the context window limitation of LLMs and the inherent performance differences between smaller and larger models. To mitigate retrieval-related shortcomings, optimized approaches such as Advanced RAG and Modular RAG have been proposed, with a primary focus on enhancing retrieval quality. For a detailed discussion of these methods, see Abo El-Enen et al. (2025). To address generation-related issues, such as context-window limitations and model capacity constraints, prior work has proposed methods, including positional interpolation, to extend context windows (Chen et al., 2023a) and architectural remedies for generation breakdowns in long-context settings (Hosseini et al., 2025).

RAG research is ongoing and rapidly evolving, with emerging approaches, such as GraphRAG[10] (Peng et al., 2025) and other optimization methods, being proposed. Therefore, this work can be further extended to incorporate additional state-of-the-art research in this direction.

4.2.3.2 Other Memory-Based Methods:. Knowledge graphs and ontologies, as discussed in Section 4.2.2, serve as a form of semantic memory for LLMs (Akbar et al., 2025, p. 12). Unlike the plain text-based embeddings typically used in RAG systems, they organize information into structured graph representations. This graphical organization enables graph-based retrieval, which goes beyond simple semantic-similarity search. Such capabilities form the key motivation behind the development of GraphRAG (Peng et al., 2025). In a similar manner, Rasmussen et al. (2025) adopt temporal knowledge graphs, that is, graph structures that evolve over time and retain historical information, as a mechanism for episodic memory *(facts as episodes, changing over time)* in the implementation of GraphRAG. Beyond retrieval-centric approaches, MemGPT (Packer et al., 2023) introduces a complementary memory paradigm by simulating operating-system-like virtual memory management. Whereas RAG primarily augments LLMs with external embedding-based or semantic knowledge retrieved from document corpora, MemGPT equips them with working memory by storing and recalling prior interactions from external storage, thereby sustaining dialogue continuity and overcoming fixed context window limitations. Instead of modifying the model's parameters, MemGPT externalizes memory management: the LLM maintains a limited "active context" (analogous to RAM or short-term memory) while offloading less immediately relevant information into external storage (analogous to disk or long-term memory). This external memory may include prior conversation history, task states, or knowledge chunks, all stored outside the model's internal weights. When needed, MemGPT dynamically recalls and reinserts this information back into the context window, effectively integrating relevant working memory. In this way, MemGPT functions as another non-parametric memory-based method for LLMs. Similarly, Li et al. (2025d) investigate memory integration methods motivated by operating systems for LLMs. Refer to the survey (Zhang et al., 2025) for more literature related to memory-based systems for LLMs.

4.2.4 Tool Usage and Function Calling. The core of agentic AI systems lies in their ability to use external tools or functions. Tool usage refers to the ability of LLM-based systems to interact with external resources or services to perform tasks beyond their internal capabilities or training data. Function calling is a specific type of tool usage in which an LLM invokes a predefined function or API. A simple example of a tool is a calculator. Since LLMs do not perform actual calculations but instead predict the next word based on previous text, they might correctly answer simple questions like 2×2 because they have likely seen it in their training data. However, for less common numbers such as 2.356×0.627171111, the model may produce an incorrect result unless it is connected to a calculator. This illustrates the concept of tool usage or function calling. Similarly, other tools could include a calendar or any external API, such as a weather service, or even a simple standalone Python function/method. For instance, early versions of ChatGPT often provided incorrect answers to questions like "What is today's date?", a problem that tool integration can address. The paper by Cancedda et al. (2023), titled *Toolformer*, discusses how LLMs can learn to use external tools autonomously. Using such tools or calling APIs enables LLMs to access external information that is not stored in the model's internal weights or parameters. Frameworks like LangGraph (LangChain-AI, 2025) and LangChain (LangChain-AI, 2025) serve as orchestration engines for integrating various tools and functions with LLMs. There are methods proposed in the literature for enhancing the tool-calling capabilities of LLMs, for instance, Li et al. (2025a) propose MeCo, an adaptive decision-making strategy for external tool use. Similarly, Abdelaziz et al. (2024) introduce *granite-20B-functioncalling*[11], a specialized model trained through a multi-task learning framework covering seven core function-calling tasks, including nested function calling, function chaining, next-based functions, parallel functions, and others. In the context of knowledge-intensive tasks, Hertling and Sack (2024) integrate knowledge graphs with LLMs through function calling. For more research related to tool usage, consult the following papers (Cui et al., 2025; Li et al., 2024b; Qin et al., 2025).

4.2.5 Reinforcement Learning. Reinforcement learning (RL) is a subfield of machine learning in which an agent improves its performance by interacting with an environment and receiving feedback in the form of rewards or penalties (Sutton & Barto, 1998). In the context of Large Language Models (LLMs), RL can be employed to optimize model behavior across

different tasks. Here, the model acts as the agent, while the environment is more broadly defined as the task setup and feedback loop; within this loop, the reward function or reward model, that is, another machine learning system trained to evaluate the outputs of the LLMs, serves as an approximation of the environment's feedback. Importantly, Reinforcement Learning from Human Feedback (RLHF) extends this paradigm by training the environment's reward signal using human feedback, thereby aligning the model's outputs more closely with human expectations. Instruction-following is one such task in which the reward model guides the agent toward producing responses aligned with human preferences. A prominent example is ChatGPT, which was tuned using RLHF. In this case, the reward model was trained on human-labeled data that distinguished between high-quality and low-quality responses, as described in Agarwal et al. (2022). Following the similar area of research, Yan et al. (2025), propose a variant of RLHF called Reinforcement Learning from Knowledge Graph Feedback (RLKGF), replacing human feedback as they are time-consuming and costly to accumulate, with a knowledge graph. They assume that the human reasoning process could be substituted by the reasoning paths in a knowledge graph. However, they also note that the knowledge in the KG represents human thinking, thereby aligning the model with human preferences. Their results show that RLKGF outperforms Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022; Lee et al., 2024), another variant of RLHF in which the human feedback component is replaced by AI-generated feedback. Any discussion of RL in the context of LLMs cannot overlook the InstructGPT paper (Agarwal et al., 2022), although classifying it *and its related variants* as parametric or non-parametric is not straightforward. The reward model is trained externally, making that component non-parametric, while the LLMs policy is updated internally, meaning its weights are adjusted based on signals from the reward model, which is parametric. Taken together, this constitutes a hybrid approach. However, there are other RL-based methods that are purely non-parametric. One such method is *RL for prompt optimization*. For instance, Deng et al. (2022) introduce RLPrompt, a discrete prompt-optimization method that harnesses reinforcement learning to generate optimal prompts that are transferable across models. Extending the work of RLPrompt, Kwon et al. (2024) introduce StablePrompt, an automatic prompt tuning method based on reinforcement learning. In this framework, an agent model generates candidate prompts, while an anchor model is incorporated to stabilize policy updates. The target LLM, evaluated against the dataset, supplies reward signals that reflect the quality of the generated prompts and drive the optimization process.

5 Future Research Directions

As discussed in the sections above, there are limitations with LLMs which require further research. Additionally, the wide range of applications of LLMs in NLU tasks presents numerous opportunities for future investigation. The following section outlines potential research directions for enhancing NLU by leveraging LLMs and incorporating external knowledge to create highly efficient systems or models. These research directions emerged as key areas for further study, based on challenges highlighted in the surveyed literature and LLM optimization papers.

5.0.0.1 Explainability in Knowledge-Augmented LLMs:. As LLMs integrate more external knowledge, ensuring their explainability becomes increasingly essential. Future research should focus on developing techniques to understand how knowledge is incorporated into the model's decision-making process. Additionally, improving the trustworthiness of applications built with LLMs will require methods to recover accurate citations for LLM-generated answers and effectively identify potential biases in the information they rely on Rorseth et al. (2024). For instance, improving the transparency of how knowledge graphs are integrated with LLMs and clarifying their reasoning processes (Chen et al., 2023b) can enhance user trust.

5.0.0.2 Optimizing LLM Finetuning:. Future research can also address the issue of Frequency Shock (cf. Section 4.1.2) (Kazemi et al., 2023). Other areas of improvement related to fine-tuning include mitigation strategies for overcoming attention imbalance (Ghosal et al., 2024). Attention imbalance, as the name suggests, is a phenomenon observed in LLMs in which the attention mechanism disproportionately focuses on certain tokens. This can lead to the suppression of important information, such as factual knowledge stored in the model's parameters, from being used in generating responses.

5.0.0.3 Scalability and Efficiency Improvements in Knowledge Integration:. A promising direction is improving the scalability and efficiency of knowledge-augmented LLMs. It is a critical area of research given the increasing demands for LLMs. Numerous strategies have been proposed to optimize LLMs, focusing on reducing computational cost and improving inference speed without compromising the model performance. These techniques include model compression, architectural advancements, and data efficiency methods, among others. Each of these techniques has its advantages and disadvantages and therefore requires further investigation and exploration (Huang et al., 2024a; Li et al., 2025b).

5.0.0.4 Agentic AI and Knowledge Integration:. Future research could explore the integration of Agentic AI with LLMs, representing a significant advancement in artificial intelligence by combining the strengths of symbolic paradigms (symbolic representation and logic) with those of connectionist paradigms (neural networks). This integration has the potential to enhance knowledge integration and decision-making capabilities, enabling the development of autonomous agents that can navigate complex environments, reason, and learn from experiences. The synergy between these paradigms is crucial for enhancing the adaptability and reasoning capabilities of AI systems (Sharma, 2024; Xiong et al., 2024).

5.0.0.5 Interactive Knowledge Retrieval and Integration:. Interactive NLP systems (iNLP), where knowledge is dynamically retrieved and integrated during inference, offer a promising avenue for improving LLM-based NLU systems. For example, developing dynamically evolving knowledge graphs (Chen et al., 2024). These systems engage in back-and-forth dialogues with external knowledge sources, refining their understanding of the task and accessing up-to-date information (Wang et al., 2023b). Combining LLM, Semantic Web (KGs and ontologies), and iNLP can also be a promising future research area.

6 Discussion

This paper explores LLM-based NLU, particularly for knowledge-intensive tasks such as IE and KGC in domain-specific contexts. It surveys methods for integrating domain-specific knowledge into LLMs to enhance accuracy and address common limitations of their out-of-the-box usage. It provides an overview of LLM-based NLU systems, discussing the technical details of how NLU tasks benefit from the use of LLMs since the advent of transformers. As LLMs have limitations, such as hallucinations, limited contextual understanding in specialized domains, and incomplete/outdated data, these limitations directly affect the performance of NLU systems. As part of the survey, the challenges of LLMs are discussed, and approaches to mitigate them are investigated. We acknowledge that while LLMs have demonstrated exceptional performance across a wide range of NLU tasks, integrating external knowledge sources, such as ontologies and knowledge graphs, among others, is a critical pathway to address issues of factual accuracy, domain-specific reasoning, and ambiguity resolution. The discussion emphasizes the importance of integrating external knowledge to enhance factual accuracy, minimize hallucinations, and optimize performance on domain-specific tasks. Two primary approaches for incorporating knowledge into LLMs were explored, namely parametric and non-parametric methods. Parametric methods encode knowledge within a model's parameters by updating its weights through techniques such as pre-training, fine-tuning, steering, knowledge editing, embedding-based approaches, and knowledge distillation. These approaches offer advantages such as higher accuracy in knowledge extraction and manipulation. However, they are computationally intensive and prone to challenges such as catastrophic forgetting, knowledge conflicts, and limited explainability. In contrast, non-parametric approaches, such as prompting strategies, knowledge-based techniques, memory-augmented systems (e.g., RAG), and tool-usage or function-calling methods, operate without extensive model training. These methods offer greater explainability, improved hardware efficiency, and increased flexibility, making them well-suited for scenarios where adaptability and explainability are crucial. Table 4 provides a comparison of these approaches across other relevant dimensions.

The choice of integration technique ultimately depends on the specific use case and the nature and volume of the data. Furthermore, the method's efficiency can only be determined through experimentation and thorough evaluation, as with other AI methods and algorithms. As research into improving knowledge-intensive NLU systems with LLMs is ongoing, there is ample room for experimentation and further investigation into developing autonomous methods to enhance the explainability of LLM-generated outputs and to explore the integration of Agentic AI with LLMs. Combining fields such as the Semantic Web, NLP, and interactive NLP (iNLP) to create systems that incorporate real-time human feedback, thereby adding a human-in-the-loop dimension, represents a significant opportunity to advance LLM-based NLU systems.

7 Conclusion

The survey investigates techniques of integrating external knowledge with LLMs. The taxonomy and the Python script[12] used for searching and crawling the seed papers can be found at Yadav (2025). The study seeks to provide insights to enhance the effectiveness of the out-of-the-box LLMs in handling domain-specific and knowledge-intensive tasks. Its emphasis lies in examining research on LLM-driven NLU systems, while highlighting the inherent challenges of LLMs that constrain the performance of such NLU applications. Approaches for overcoming the common limitations of LLMs, including hallucinations, lack of contextual understanding, and outdated data, are discussed in detail. These approaches suggest integrating external knowledge to achieve better factual accuracy with LLMs for knowledge-intensive tasks, such as IE and KGC. The investigation highlights that while parametric methods for integrating knowledge into LLMs have shown a positive impact on performance, they face challenges, including a lack of explainability and the

extensive use of hardware resources during model training. Non-parametric approaches, which rely on external storage systems for knowledge retention, offer improved explainability and allow the provenance of results to be traced. Moreover, implementing non-parametric approaches is less computationally expensive. However, the decision of which approach to pick depends on the particular problem. Trade-offs must be made between performance and hardware requirements. Lastly, future research directions for improving LLM-based NLU systems are discussed.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iDs

Itisha Yadav https://orcid.org/0009-0002-7957-9600
Sirko Schindler https://orcid.org/0000-0002-0964-4457
Diana Peters https://orcid.org/0000-0002-5855-2989
Roman Klinger https://orcid.org/0000-0002-2014-6619

Notes

1 Other challenges associated with LLMs that are not directly mitigated by incorporating external knowledge include scalability limitations (Zhang et al., 2024b) and response inconsistency (Saxena et al., 2024).
2 https://scholar.google.com/
3 https://www.semanticscholar.org/
4 https://dblp.org/
5 https://aclanthology.org/
6 https://ieeexplore.ieee.org/
7 In the context of this survey, we consider semi-structured data as documents like datasheets comprised of both tables with complex structures and floating text.
8 Due to data privacy constraints, the original domain ontology concepts and datasheet contents are omitted from the example prompt.
9 Approaches discussed in the context of knowledge-intensive tasks are mostly parametric and are thus discussed as parametric knowledge integration.
10 Refer to the following Github repository for detailed research related to GraphRAG: https://github.com/pengboci/GraphRAG-Survey
11 https://huggingface.co/ibm-granite/granite-20b-functioncalling
12 The script can be reused in the future for extending this work with the latest developments.

References

Abdelaziz, I., Basu, K., Agarwal, M., Kumaravel, S., Stallone, M., Panda, R., Rizk, Y., Bhargav, G. P. S., Crouse, M., Gunasekara, C., Ikbal, S., Joshi, S., Karanam, H., Kumar, V., Munawar, A., Neelam, S., Raghu, D., Sharma, U., Soria, A. M., … Kapanipathi, P. (2024). Granite-function calling model: Introducing function calling abilities via multi-task learning of granular tasks. In F. Dernoncourt, D. Preoţiuc-Pietro, & A. Shimorina (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing: industry track* (pp. 1131–1139). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-industry.85

Abo El-Enen, M., Saad, S., & Nazmy, T. (2025). A survey on retrieval-augmentation generation (RAG) models for healthcare applications. *Neural Computing and Applications*, *37*(33), 28191–28267. https://doi.org/10.1007/s00521-025-11666-9

Agarwal, S., Almeida, D., Askell, A., Christiano, P., Hilton, J., Jiang, X., Kelton, F., Leike, J., Lowe, R., Miller, L., Mishkin, P., Ouyang, L., Ray, A., Schulman, J., Simens, M., Slama, K., Wainwright, C., Welinder, P., Wu, J., & Zhang, C. (2022). Training language models to follow instructions with human feedback. In *Advances in neural information processing systems 35*, *NeurIPS 2022* (Vol. 35, pp. 27730–27744). Neural Information Processing Systems Foundation, Inc. (NeurIPS). https://doi.org/10.52202/068431-2011

Akbar, N. A., Dembani, R., Lenzitti, B., & Tegolo, D. (2025). Rag-driven memory architectures in conversational llms—A literature review with insights into emerging agriculture data sharing. *IEEE access*, *13*, 123855. https://doi.org/10.1109/access.2025.3589241

Albtosh, L. (2024). *Challenges and limitations of using LLMs in software security, Pp. 439–464*. IGI Global.

Allen-Zhu, Z., & Li, Y. (2025a). Physics of language models: Part 3.1, knowledge storage and extraction. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5250633

Allen-Zhu, Z., & Li, Y. (2025b). Physics of language models: Part 3.2, knowledge manipulation. https://doi.org/10.2139/ssrn.5250621.

Andonian, A., Bau, D., Belinkov, Y., & Meng, K. (2022). Locating and editing factual associations in gpt. In *Advances in neural information processing systems 35*, *NeurIPS 2022*, (Vol. 35, pp. 17359–17372). Neural Information Processing Systems Foundation, Inc. (NeurIPS). https://doi.org/10.52202/068431-1262.

Anjali, M. K., & Babu, A. P. (2014). Ambiguities in natural language processing. *International Journal of Innovative Research in Computer and Communication Engineering*, *2*(5), 392–394.

Anjum, S., Zhang, H., Zhou, W., Paek, E. J., Zhao, X., & Feng, Y. (2025). Halo: Hallucination analysis and learning optimization to empower llms with retrieval-augmented context for guided clinical decision making. In *Proceedings of the ACM/IEEE international conference on connected health: Applications, systems and engineering technologies*, CHASE '25 (pp. 187–198). ACM. https://doi.org/10.1145/3721201.3721385

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). *DBpedia: A nucleus for a web of open data, Pp. 722–735*. Springer Berlin Heidelberg.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* https://doi.org/10.48550/arxiv.2212.08073.

Ben Zaken, E., Goldberg, Y., & Ravfogel, S. (2022). Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th annual meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 1–9). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.1.

Bosma, M., Chi, E., Ichter, B., Le, Q. V., Schuurmans, D., Wang, X., Wei, J., Xia, F., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems 35* (pp. 24824–24837). NeurIPS 2022. Neural Information Processing Systems Foundation, Inc. (NeurIPS). https://doi.org/10.52202/068431-1800

Bouhoun, Z., Allali, A., Cocci, R., Assaad, M. A., Plancon, A., Godest, F., Kondratenko, K., Rodriguez, J., Vitillo, F., Malhomme, O., Bechet, L. B., & Plana, R. (2024). Curielm: Enhancing large language models for nuclear domain applications. *EPJ Web of Conferences*, *302*, 17006. https://doi.org/10.1051/epjconf/202430217006

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cancedda, N., Dessi, R., Dwivedi-Yu, J., Hambro, E., Lomeli, M., Raileanu, R., Schick, T., Scialom, T., & Zettlemoyer, L. (2023). Toolformer: Language models can teach themselves to use tools. In *Advances in neural information processing systems 36*, *NeurIPS 2023* (Vol. 36, pp. 68539–68551). Neural Information Processing Systems Foundation, Inc. (NeurIPS). https://doi.org/10.52202/075280-2997

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2898–2904). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.261

Chen, R., Jiang, W., Qin, C., Rawal, I. S., Tan, C., Choi, D., Xiong, B., & Ai, B. (2024). Llm-based multi-hop question answering with knowledge graph integration in evolving environments. In Y. Al-Onaizan, M. Bansal, & Y. N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 14438–14451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.844

Chen, S., Wong, S., Chen, L., & Tian, Y. (2023a). Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*. https://doi.org/10.48550/arxiv.2306.15595

Chen, X., Wan, X., & Yang, Z. (2022). Relation-constrained decoding for text generation. In *Advances in neural information processing systems 35*, *NeurIPS 2022* (Vol. 35, pp. 26804–26819). Neural Information Processing Systems Foundation, Inc. (NeurIPS). https://doi.org/10.52202/068431-1944

Chen, Z., Chen, J., Chen, Y., Yu, H., Singh, A. K., & Sra, M. (2023b). *Lmexplainer: Grounding knowledge and explaining language models*. https://doi.org/10.48550/arXiv.2303.16537.

Choudhary, N., & Reddy, C. K. (2023). *Complex logical reasoning over knowledge graphs using large language models*. https://doi.org/10.48550/arxiv.2305.01157

Chroma (2025) *Chroma: Open source embedding database*. https://github.com/chroma-core/chroma.

Cleti, M., & Jano, P. (2024). *Hallucinations in LLMs: Types, causes, and approaches for enhanced reliability*. https://doi.org/10.31219/osf.io/tj93u

Cui, S., He, A., Xu, S., Zhang, H., Wang, Y., Zhang, Q., Wang, Y., & Xu, B. (2025). Self-guided function calling in large language models via stepwise experience recall. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 10842–10854). Association for Computational Linguistics. ISBN 979-8-89176-335-7. https://doi.org/10.18653/v1/2025.findings-emnlp.574

da Costa, L. Y., & Oliveira e Souza Filho, J. B. D. (2024). Adapting llms to new domains: A comparative study of fine-tuning and rag strategies for portuguese qa tasks. In *Anais do XV Simpósio brasileiro de tecnologia da informação e da linguagem humana (STIL 2024)*, STIL 2024 (pp. 267–277). Sociedade Brasileira de Computação. https://doi.org/10.5753/stil.2024.245443

deepset (2025) *Haystack: An end-to-end framework for building nlp applications*. from https://haystack.deepset.ai/.

Deng, M., Wang, J., Hsieh, C. P., Wang, Y., Guo, H., Shu, T., Song, M., Xing, E., & Hu, Z. (2022). Rlprompt: Optimizing discrete text prompts with reinforcement learning. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 3369–3391). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.222

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north american chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423

Dong, G., Li, R., Wang, S., Zhang, Y., Xian, Y., & Xu, W. (2023). Bridging the kb-text gap: Leveraging structured knowledge-aware pre-training for kbqa. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, CIKM'23 (pp. 3854–3859). Association for Computing Machinery. ISBN 9798400701245. https://doi.org/10.1145/3583780.3615150

Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., Jin, G., Qi, Y., Hu, J., Meng, J., Bensalem, S., & Huang, X. (2025). Safeguarding large language models: A survey. *Artificial Intelligence Review*, *58*(12), 382. https://doi.org/10.1007/s10462-025-11389-2

Elastic (2025) *Elastic: Search, observe, protect*. https://www.elastic.co/.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T. S., & Li, Q. (2024). A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, KDD '24 (pp. 6491–6501). Association for Computing Machinery. ISBN 9798400704901. https://doi.org/10.1145/3637528.3671470

Feng, C., Zhang, X., & Fei, Z. (2023). *Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs*. https://doi.org/10.48550/arxiv.2309.03118

Fensel, D., Simsek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I., Umbrich, J., & Wahler, A. (2020). *Introduction: What is a knowledge graph? (Pp. 1–10)*. Springer International Publishing.

Fields, J., Chovanec, K., & Madiraju, P. (2024). A survey of text classification with transformers: How wide? How large? How long? How accurate? How expensive? How safe?. *IEEE Access*, *12*, 6518–6531. https://doi.org/10.1109/access.2024.3349952

Gao, D., Chen, K., Chen, B., Dai, H., Jin, L., Jiang, W., Ning, W., Yu, S., Xuan, Q., Cai, X., Yang, L., & Wang, Z. (2024). LLMs-based machine translation for e-commerce. *Expert Systems With Applications*, *258*, 125087. https://doi.org/10.1016/j.eswa.2024.125087

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, *2*(1), 32. https://doi.org/10.48550/arXiv.2312.10997

Geng, S., Josifoski, M., Peyrard, M., & West, R. (2023). Grammar-constrained decoding for structured nlp tasks without finetuning. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 10932–10952). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.674

Ghosal, G. R., Hashimoto, T., & Raghunathan, A. (2024). *Understanding finetuning for factual knowledge extraction*. https://doi.org/10.48550/arxiv.2406.14785

Guidance (2026) *Guidance, an efficient programming paradigm for steering language models*. https://github.com/guidance-ai/guidance.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 8342–8360). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.740

Hertling, S., & Sack, H. (2024). Towards large language models interacting with knowledge graphs via function calling. In *KBC-LM/LM-KBC@ ISWC*.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* https://doi.org/10.48550/arxiv.1503.02531

Hokamp, C., & Liu, Q. (2017). Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/p17-1141

Hosseini, P., Castro, I., Ghinassi, I., & Purver, M. (2025). Efficient solutions for an intriguing failure of LLMs: Long context window does not mean LLMs can analyze long sequences flawlessly. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 1880–1891). Association for Computational Linguistics.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (v Volume 1: long papers)* (pp. 328–339). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1031

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LORA: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3. https://doi.org/10.48550/arXiv.2106.09685

Hu, L., Liu, Z., Zhao, Z., Hou, L., Nie, L., & Li, J. (2023). A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*, *36*(4), 1413–1430. https://doi.org/10.1109/tkde.2023.3310002

Hu, L., Liu, Z., Zhao, Z., Hou, L., Nie, L., & Li, J. (2024). A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*, *36*(4), 1413–1430. https://doi.org/10.1109/tkde.2023.3310002

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55. https://doi.org/10.1145/3703155

Huang, S., Yang, K., Qi, S., & Wang, R. (2024a). When large language model meets optimization. *Swarm and Evolutionary Computation*, *90*, 101663. https://doi.org/10.1016/j.swevo.2024.101663

Huang, Y., Tang, K., & Chen, M. (2024b). Leveraging large language models for enhanced nlp task performance through knowledge distillation and optimized training strategies. https://doi.org/10.48550/arxiv.2402.09282

Hwang, J., Park, J., Park, H., Kim, D., Park, S., & Ok, J. (2025). Retrieval-augmented generation with estimation of source reliability. In *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 34267–34291). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1738

Ilharco, G., Ribeiro, M. T., Wortsman, M., Schmidt, L., Hajishirzi, H., & Farhadi, A. (n.d). Editing models with task arithmetic. In *The eleventh international conference on learning representations*.

Ishigaki, R., Suzuki, J., Shuzo, M., & Maeda, E. (2024). Knowledge editing of large language models unconstrained by word order. In X. Fu & E. Fleisig (Eds.), *Proceedings of the 62nd annual meeting of the Association for Computational linguistics (Volume 4: Student Research Workshop)* (pp. 159–169). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-srw.23

Jain, P., & Lapata, M. (2024). *Integrating large language models with graph-based reasoning for conversational question answering*. https://doi.org/10.48550/arxiv.2407.09506

Ji, Y., Wu, K., Li, J., Chen, W., Zhong, M., Jia, X., & Zhang, M. (2024). Retrieval and reasoning on kgs: Integrate knowledge graphs into large language models for complex question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 7598–7610). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.446

Jiang, B., Wang, Y., Luo, Y., He, D., Cheng, P., & Gao, L. (2024). Reasoning on efficient knowledge paths: Knowledge graph guides large language model for domain question answering. In *2024 IEEE international conference on knowledge graph (ICKG)* (pp. 142–149). IEEE. https://doi.org/10.1109/ickg63256.2024.00026

Jin, D., Pan, E., Oufattole, N., Weng, W. H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, *11*(14), 6421. https://doi.org/10.3390/app11146421

Kazemi, M., Mittal, S., & Ramachandran, D. (2023). *Understanding finetuning for factual knowledge extraction from language models*. https://doi.org/10.48550/arxiv.2301.11293

Keluskar, A., Bhattacharjee, A., & Liu, H. (2024). Do llms understand ambiguity in text? A case study in open-world question answering. In *2024 IEEE international conference on big data (bigdata)* (pp. 7485–7490). IEEE. https://doi.org/10.1109/bigdata62323.2024.10825265

Kerner, T. (2024). *Domain-specific pretraining of language models: A comparative study in the medical field*. https://doi.org/10.48550/arxiv.2407.14076

Khan, A., Hasan, M. T., Kemell, K. K., Rasku, J., & Abrahamsson, P. (2024). Developing retrieval augmented generation (rag) based LLM systems from PDFs: An experience report. https://doi.org/10.48550/arxiv.2410.15944

Krugmann, J. O., & Hartmann, J. (2024). Sentiment analysis in the age of generative AI. *Customer Needs and Solutions*, *11*(1). https://doi.org/10.1007/s40547-024-00143-4

Kukreja, S., Kumar, T., Purohit, A., Dasgupta, A., & Guha, D. (2024). A literature survey on open source large language models. In *Proceedings of the 2024 7th international conference on computers in management and business*, ICCMB 2024 (pp. 133–143). ACM. https://doi.org/10.1145/3647782.3647803

Kulkarni, C. S. (2023). The evolution of large language models in natural language understanding. *Journal of Artificial Intelligence, Machine Learning and Data Science*, *1*(4), 49–53. https://doi.org/10.51219/jaimld/chinmay-shripad-kulkarni/28

Kumar, P. (2024). Large language models (LLMs): Survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, *57*(10), 260. https://doi.org/10.1007/s10462-024-10888-y

Kwon, M., Kim, G., Kim, J., Lee, H., & Kim, J. (2024). Stableprompt: Automatic prompt tuning using reinforcement learning for large language model. In Y. Al-Onaizan, M. Bansal, & Y. N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 9868–9884). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.551

Lai, W., Hangya, V., & Fraser, A. (2024). Style-specific neurons for steering llms in text style transfer. In Y. Al-Onaizan, M. Bansal, & Y. N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 13427–13443). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.745

Lamb, T. A., Davies, A., Paren, A., Torr, P., & Pinto, F. (2025). Focus on this, not that! steering llms with adaptive feature specification. In A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, & J. Zhu (Eds.), *Forty-second international conference on machine learning, ICML 2025, July 13–19, 2025*, Proceedings of Machine Learning Research. PMLR/OpenReview.net.

LangChain-AI. (2025). *langgraph: Build resilient language agents as graphs*. https://github.com/langchain-ai/langgraph.

LangChain-AI. (2025). *Langchain: Build context-aware reasoning applications*. https://github.com/langchain-ai/langchain.

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., & Prakash, S. (2024). Rlaif vs. rlhf: scaling reinforcement learning from human feedback with ai feedback. In *Proceedings of the 41st international conference on machine learning*, ICML'24. JMLR.org.

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

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020a). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 7871–7880). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, Wt., Rocktäschel, T., Riedel, S., & Kiela, D. (2020b). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://dl.acm.org/doi/abs/10.5555/3495724.3496517

Li, W., Li, D., Dong, K., Zhang, C., Zhang, H., Liu, W., Wang, Y., Tang, R., & Liu, Y. (2025a). Adaptive tool use in large language models with meta-cognition trigger. In *Proceedings of the 63rd annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 13346–13370). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.655

Li, X., Ma, Y., Huang, Y., Wang, X., Lin, Y., & Zhang, C. (2025b). *Synergized data efficiency and compression (sec) optimization for large language models*. https://doi.org/10.20944/preprints202409.0662.v3

Li, X., Mei, S., Liu, Z., Yan, Y., Wang, S., Yu, S., Zeng, Z., Chen, H., Yu, G., Liu, Z., Sun, M., & Xiong, C. (2025c). RAG-DDR: Optimizing retrieval-augmented generation using differentiable data rewards. In *The Thirteenth international conference on learning representations*.

Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. 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)* (pp. 4582–4597). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.353

Li, Y., Zhao, H., Jiang, H., Pan, Y., Liu, Z., Wu, Z., Shu, P., Tian, J., Yang, T., Xu, S., Lyu, Y., Blenk, P., Pence, J., Rupram, J., Banu, E., Liu, N., Wang, S., Song, W., Zhai, X., … Liu, T. (2024a). Large language models for manufacturing. https://doi.org/10.48550/arxiv.2410.21418

Li, Z., Chen, Z., Ross, M., Huber, P., Moon, S., Lin, Z., Dong, X., Sagar, A., Yan, X., & Crook, P. (2024b). Large language models as zero-shot dialogue state tracker through function calling. In L. W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8688–8704). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.471

Li, Z., Song, S., Xi, C., Wang, H., Tang, C., Niu, S., Chen, D., Yang, J., Li, C., Yu, Q., Zhao, J., Wang, Y., Liu, P., Lin, Z., Wang, P., Huo, J., Chen, T., … Xiong, F. (2025d). Memos: A memory os for ai system. *arXiv preprint arXiv:2507.03724* https://doi.org/10.48550/arxiv.2507.03724

Liu, J., Mao, Q., Jiang, W., & Li, J. (2024). Knowformer: Revisiting transformers for knowledge graph reasoning. In *Proceedings of the 41st international conference on machine learning* (pp. 31669–31690).

Liu, Y., Yao, Y., Ton, J. F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., & Li, H. (2023). *Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment*. https://doi.org/10.48550/arxiv.2308.05374

Luo, L., Li, Y., Haffari, G., & Pan, S. (2024a). Reasoning on graphs: Faithful and interpretable large language model reasoning. In *The Twelfth international conference on learning representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. OpenReview.net.

Luo, Z., Wang, Y., Ke, W., Qi, R., Guo, Y., & Wang, P. (2024b). Boosting LLMs with ontology-aware prompt for ner data augmentation. In *ICASSP 2024—2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 12361–12365). IEEE. https://doi.org/10.1109/icassp48485.2024.10446860

Luo, Z., Yang, Y., Qi, R., Fang, Z., Guo, Y., & Wang, Y. (2023). Incorporating large language models into named entity recognition: Opportunities and challenges. In *2023 4th international conference on computer, big data and artificial intelligence (ICCBD+AI)* (pp. 429–433). IEEE. https://doi.org/10.1109/iccbd-ai62252.2023.00079

Ma, J., Gao, Z., Chai, Q., Sun, W., Wang, P., Pei, H., Tao, J., Song, L., Liu, J., Zhang, C., & Cui, L. (2025). Debate on graph: A flexible and reliable reasoning framework for large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(23), 24768–24776. https://doi.org/10.1609/aaai.v39i23.34658

Maes, S. (2025). *Fixing reference hallucinations of LLMs*. https://doi.org/10.31219/osf.io/u38w4_v2

Mavroudis, V. (2024). *Langchain*. https://doi.org/10.20944/preprints202411.0566.v1

McIntosh, T. R., Susnjak, T., Liu, T., Watters, P., & Halgamuge, M. N. (2024). The inadequacy of reinforcement learning from human feedback—Radicalizing large language models via semantic vulnerabilities. *IEEE Transactions on Cognitive and Developmental Systems*, *16*(4), 1561–1574. https://doi.org/10.1109/tcds.2024.3377445

McTear, M., Varghese Marokkie, S., & Bi, Y. (2023). *A comparative study of chatbot response generation: Traditional approaches versus large language models*. Springer Nature Switzerland.

Meng, K., Sharma, A. S., Andonian, A., Belinkov, Y., & Bau, D. (2022). Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229* https://doi.org/10.48550/arxiv.2210.07229

Mihaylov, T., Clark, P., Khot, T., & Sabharwal, A. (2018). Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2381–2391). Association for Computational Linguistics. https://doi.org/10.18653/v1/d18-1260

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*. https://doi.org/10.48550/arxiv.2402.06196

Mitchell, E., Lin, C., Bosselut, A., Finn, C., & Manning, C. D. (2021). Fast model editing at scale. *arXiv preprint arXiv:2110.11309*. https://doi.org/10.48550/arxiv.2110.11309

Mukanova, A., Milosz, M., Dauletkaliyeva, A., Nazyrova, A., Yelibayeva, G., Kuzin, D., & Kussepova, L. (2024). LLM-powered natural language text processing for ontology enrichment. *Applied Sciences*, *14*(13), 5860. https://doi.org/10.3390/app14135860

Muralidharan, B., Beadles, H., Marzban, R., & Mupparaju, K. S. (2024). *Knowledge AI: Fine-tuning NLP models for facilitating scientific knowledge extraction and understanding*. https://doi.org/10.48550/arxiv.2408.04651

Nagar, A., Schlegel, V., Nguyen, T. T., Li, H., Wu, Y., Binici, K., & Winkler, S. (2025). LLMs are not zero-shot reasoners for biomedical information extraction. In *The sixth workshop on insights from negative results in NLP* (pp. 106–120). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.insights-1.11

of Standards, N. I., & AUTHOR(NIST), T. (2004). *Reuters corpora (rcv1, rcv2, trc2)*. https://trec.nist.gov/data/reuters/reuters.html.

Omeliyanenko, J., Zehe, A., Hotho, A., & Schlör, D. (2023). Capskg: Enabling continual knowledge integration in language models for automatic knowledge graph completion. In *The semantic web—ISWC 2023: 22nd international semantic web conference, Athens, Greece, November 6–10, 2023, Proceedings, Part I* (pp. 618–636). Springer-Verlag. https://doi.org/10.1007/978-3-031-47240-4_33

Outline (2026) *Outline for structured generation*. https://dottxt-ai.github.io/outlines/latest/.

Ovadia, O., Brief, M., Mishaeli, M., & Elisha, O. (2024). Fine-tuning or retrieval? Comparing knowledge injection in llms. In Y. Al-Onaizan, M. Bansal, & Y. N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 237–250). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.15

Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S. G., Stoica, I., & Gonzalez, J. E. (2023). *Memgpt: Towards LLMs as operating systems*. https://doi.org/10.48550/arxiv.2310.08560

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, *36*(7), 3580–3599. https://doi.org/10.1109/tkde.2024.3352100

Pantha, N., Ramasubramanian, M., Gurung, I., Maskey, M., & Ramachandran, R. (2024). *Challenges in guardrailing large language models for science*. https://doi.org/10.48550/arxiv.2411.08181

Peng, B., Chersoni, E., Hsu, Y. Y., & Huang, C. R. (2021). Is domain adaptation worth your investment? comparing bert and finbert on financial tasks. In U. Hahn, V. Hoste, & A. Stent (Eds.), *Proceedings of the third workshop on economics and natural language processing* (pp. 37–44). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.econlp-1.5

Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y., & Tang, S. (2025). Graph retrieval-augmented generation: A survey. *ACM Transactions on Information Systems*, *44*(2), 1–52. https://doi.org/10.1145/3777378

Petrov, A., Torr, P. H., & Bibi, A. (2023). When do prompting and prefix-tuning work? A theory of capabilities and limitations. *arXiv preprint arXiv:2310.19698*. https://doi.org/10.48550/arxiv.2310.19698

Piñeiro-Martín, A., Santos-Criado, F. J., García-Mateo, C., Docío-Fernández, L., & López-Pérez, MdC. (2025). Context is king: Large language models' interpretability in divergent knowledge scenarios. *Applied Sciences*, *15*(3), 1192. https://doi.org/10.3390/app15031192

Pinter, Y., & Elhadad, M. (2023). Emptying the ocean with a spoon: Should we edit models? In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 15164–15172). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.1012

Post, M., & Vilar, D. (2018). Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1314–1324). Association for Computational Linguistics. https://doi.org/10.18653/v1/n18-1119

Poth, C., Sterz, H., Paul, I., Purkayastha, S., Engländer, L., Imhof, T., Vulić, I., Ruder, S., Gurevych, I., & Pfeiffer, J. (2023). Adapters: A unified library for parameter-efficient and modular transfer learning. In Y. Feng & E. Lefever (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing: System demonstrations* (pp. 149–160). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-demo.13

Qiang, Z., Taylor, K., Wang, W., & Jiang, J. (2024). *OAEI-LLM: A benchmark dataset for understanding large language model hallucinations in ontology matching*. https://doi.org/10.48550/arxiv.2409.14038

Qin, S., Zhu, Y., Mu, L., Zhang, S., & Zhang, X. (2025). Meta-tool: Unleash open-world function calling capabilities of general-purpose large language models. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 30653–30677). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1481

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Rajabzadeh, H., Valipour, M., Zhu, T., Tahaei, M. S., Kwon, H. J., Ghodsi, A., Chen, B., & Rezagholizadeh, M. (2024). Qdylora: Quantized dynamic low-rank adaptation for efficient large language model tuning. In *Proceedings of the 2024 conference on empirical methods in natural language processing: Industry track* (pp. 712–718). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-industry.53

Rajasekharan, A., Zeng, Y., Padalkar, P., & Gupta, G. (2023). Reliable natural language understanding with large language models and answer set programming. *Electronic Proceedings in Theoretical Computer Science*, *385*, 274–287. https://doi.org/10.4204/eptcs.385.27

Rasmussen, P., Paliychuk, P., Beauvais, T., Ryan, J., & Chalef, D. (2025). Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*. https://doi.org/10.48550/arxiv.2501.13956

Rorseth, J., Godfrey, P., Golab, L., Srivastava, D., & Szlichta, J. (2024). Towards explainability in retrieval-augmented LLMs. In *2024 IEEE 40th international conference on data engineering (ICDE)* (pp. 5669–5670). IEEE. https://doi.org/10.1109/icde60146.2024.00466

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. https://doi.org/10.48550/arxiv.1910.01108

Sanu, E., Amudaa, T. K., Bhat, P., Dinesh, G., Kumar, A. U., & Chate, P. R. K. (2024). Limitations of large language models. In *2024 8th International conference on computational system and information technology for sustainable solutions (CSITSS)* (pp. 1–6). IEEE. https://doi.org/10.1109/csitss64042.2024.10817070

Saxena, Y., Chopra, S., & Tripathi, A. M. (2024). Evaluating consistency and reasoning capabilities of large language models. In *2024 Second international conference on data science and information system (ICDSIS)* (pp. 1–5). IEEE. https://doi.org/10.1109/icdsis61070.2024.10594233

Sedova, A., Litschko, R., Frassinelli, D., Roth, B., & Plank, B. (2024). To know or not to know? Analyzing self-consistency of large language models under ambiguity. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 17203–17217). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.1003

Sharma, A. (2024). Merging paradigms: The synergy of symbolic and connectionist ai in LLM-powered autonomous agents. *Journal of Artificial Intelligence General science (JAIGS)*, *6*(1), 138–150. https://doi.org/10.60087/jaigs.v6i1.237

Sharma, R. K., & Joshi, M. (2020). An analytical study and review of open source chatbot framework, rasa. *International Journal of Engineering Research and Technology*, *V9*(06), 1011–1014. https://doi.org/10.17577/ijertv9is060723

Shimizu, C., & Hitzler, P. (2025). Accelerating knowledge graph and ontology engineering with large language models. *Journal of Web Semantics*, *85*, 100862. https://doi.org/10.1016/j.websem.2025.100862

Singh, O. P., & Patil, D. M. E. (2024). Analysis of ambiguity, vagueness, fuzziness, uncertainty, possibility and probability in the natural language semantics with fuzzy logic. *International Research Journal on Advanced Engineering Hub (IRJAEH)*, *2*(05), 1478–1483. https://doi.org/10.47392/irjaeh.2024.0204

Singh, S., Zaidi, N., & Singh, A. (2018). Deep learning for natural language understanding: A review of recent advances. *International Journal of Applied Research*, *4*(10), 310–314. https://doi.org/10.22271/allresearch.2018.v4.i10d.11459

Some, L., Yang, W., Bain, M., & Kang, B. H. (2025). *A comprehensive survey on integrating large language models with knowledge-based methods*. https://doi.org/10.2139/ssrn.5111497

Soudani, H., Kanoulas, E., & Hasibi, F. (2024). Fine tuning vs. retrieval augmented generation for less popular knowledge. In *Proceedings of the 2024 annual international ACM SIGIR conference on research and development in information retrieval in the asia pacific region*, SIGIR-AP 2024 (pp. 12–22). Association for Computing Machinery. https://doi.org/10.1145/3673791.3698415

Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, Wt., Smith, N. A., Zettlemoyer, L., & Yu, T. (2023). One embedder, any task: Instruction-finetuned text embeddings. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 1102–1121). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.71

Sulaiman, N., & Hamzah, F. (2024). Evaluation of transfer learning and adaptability in large language models with the glue benchmark. *Authorea Preprints*. https://doi.org/10.36227/techrxiv.171077989.99407624/v1

Sun, S., Liu, Y., Wang, S., Iter, D., Zhu, C., & Iyyer, M. (2024). Pearl: Prompting large language models to plan and execute actions over long documents. In Y. Graham & M. Purver (Eds.), *Proceedings of the 18th conference of the European chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 469–486). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-long.29

Susnjak, T., Hwang, P., Reyes, N., Barczak, A. L., McIntosh, T., & Ranathunga, S. (2025). Automating research synthesis with domain-specific large language model fine-tuning. *ACM Transactions on Knowledge Discovery From Data*, *19*(3), 1–39. https://doi.org/10.1145/3715964

Sutton, R., & Barto, A. (1998). Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, *9*(5), 1054–1054. https://doi.org/10.1109/tnn.1998.712192

Talmor, A., Yoran, O., Le Bras, R., Bhagavatula, C., Goldberg, Y., Choi, Y., & Berant, J. (2021). Commonsenseqa 2.0: Exposing the limits of ai through gamification. *Advances in Neural Information Processing Systems*. https://nips.cc/virtual/2021/22765

Tam, Z. R., Wu, C. K., Tsai, Y. L., Lin, C. Y., Lee, Hy, & Chen, Y. N. (2024). Let me speak freely? A study on the impact of format restrictions on large language model performance. In F. Dernoncourt, D. Preoţiuc-Pietro, & A. Shimorina (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing: Industry track* (pp. 1218–1236). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-industry.91

Tian, S., Luo, Y., Xu, T., Yuan, C., Jiang, H., Wei, C., & Wang, X. (2024). KG-adapter: Enabling knowledge graph integration in large language models through parameter-efficient fine-tuning. In L. W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics ACL 2024* (pp. 3813–3828). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.229

Vasantharajan, C., Tun, K. Z., Thi-Nga, H., Jain, S., Rong, T., & Siong, C. E. (2022). Medbert: A pre-trained language model for biomedical named entity recognition. In *2022 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC)* (pp. 1482–1488). IEEE. https://doi.org/10.23919/apsipaasc55919.2022.9980157

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. N., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need 30*.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85. https://doi.org/10.1145/2629489

Wang, F., Bao, R., Wang, S., Yu, W., Liu, Y., Cheng, W., & Chen, H. (2024a). InfuserKI: Enhancing large language models with knowledge graphs via infuser-guided knowledge integration. In Y. Al-Onaizan, M. Bansal, & Y. N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 3675–3688). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.209

Wang, H., & Li, Y. F. (2023). Large language model empowered by domain-specific knowledge base for industrial equipment operation and maintenance. In *2023 5th International conference on system reliability and safety engineering (SRSE)* (pp. 474–479). IEEE. https://doi.org/10.1109/srse59585.2023.10336112

Wang, K., Duan, F., Wang, S., Li, P., Xian, Y., Yin, C., Rong, W., & Xiong, Z. (2023a). *Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering*. https://doi.org/10.48550/arxiv.2308.13259

Wang, M., Yao, Y., Xu, Z., Qiao, S., Deng, S., Wang, P., Chen, X., Gu, J. C., Jiang, Y., Xie, P., Huang, F., Chen, H., & Zhang, N. (2024b). Knowledge mechanisms in large language models: A survey and perspective. In Y. Al-Onaizan, M. Bansal, & Y. N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 7097–7135). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.416

Wang, P., Zhang, N., Tian, B., Xi, Z., Yao, Y., Xu, Z., Wang, M., Mao, S., Wang, X., Cheng, S., Liu, K., Ni, Y., Zheng, G., & Chen, H. (2024c). EasyEdit: An easy-to-use knowledge editing framework for large language models. In Y. Cao, Y. Feng, & D. Xiong (Eds.), *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics (Volume 3: system demonstrations)* (pp. 82–93). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-demos.9

Wang, R., Tang, D., Duan, N., Wei, Z., Huang, X., Ji, J., Cao, G., Jiang, D., & Zhou, M. (2021). K-adapter: Infusing knowledge into pre-trained models with adapters. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 1405–1418). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.121

Wang, S., Zhu, Y., Liu, H., Zheng, Z., Chen, C., & Li, J. (2024d). Knowledge editing for large language models: A survey. *ACM Computing Surveys*, *57*(3), 1–37. https://doi.org/10.1145/3698590

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, *33*, 5776–5788. https://proceedings.neurips.cc/paper_files/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, Z., Chu, Z., Doan, T. V., Ni, S., Yang, M., & Zhang, W. (2024e). History, development, and principles of large language models: An introductory survey. *AI and Ethics*, *5*(3), 1955–1971. https://doi.org/10.1007/s43681-024-00583-7

Wang, Z., Wang, H., Danek, B., Li, Y., Mack, C., Arbuckle, L., Biswal, D., Poon, H., Wang, Y., Rajpurkar, P., Xiao, C., & Sun, J. (2025). A perspective for adapting generalist AI to specialized medical ai applications and their challenges. *NPJ Digital Medicine*, *8*(1), 429. https://doi.org/10.1038/s41746-025-01789-7

Wang, Z., Zhang, G., Yang, K., Shi, N., Zhou, W., Hao, S., Xiong, G., Li, Y., Sim, M. Y., Chen, X., Zhu, Q., Yang, Z., Nik, A., Liu, Q., Lin, C., Wang, S., Liu, R., Chen, W., Xu, K., … Fu, J. (2023b). *Interactive natural language processing*. https://doi.org/10.48550/arxiv.2305.13246

Xiong, H., Wang, Z., Li, X., Bian, J., Xie, Z., Mumtaz, S., Al-Dulaimi, A., & Barnes, L. E. (2024). *Converging paradigms: The synergy of symbolic and connectionist ai in llm-empowered autonomous agents*. https://doi.org/10.48550/arxiv.2407.08516

Yadav, I. (2025). *Script and taxonomy for external knowledge integration in large language models: A survey on methods, challenges, and future directions*. https://doi.org/10.5281/zenodo.15064852

Yan, L., Tang, C., Guan, Y., Wang, H., Wang, S., Liu, H., Yang, Y., & Jiang, J. (2025). Rlkgf: Reinforcement learning from knowledge graph feedback without human annotations. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 6619–6633). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.344

Yan, R., Sun, L., Wang, F., & Zhang, X. (2021). *K-xlnet: A general method for combining explicit knowledge with language model pretraining*. https://doi.org/10.48550/arxiv.2104.10649

Yan, X., Xiao, Y., & Jin, Y. (2024). Generative large language models explained [AI-explained]. *IEEE Computational Intelligence Magazine*, *19*(4), 45–46. https://doi.org/10.1109/mci.2024.3431454

Yang, S., Chen, F., Yang, Y., & Zhu, Z. (2023). A study on semantic understanding of large language models from the perspective of ambiguity resolution. In *Proceedings of the 2023 international joint conference on robotics and artificial intelligence*, JCRAI 2023 (pp. 165–170). ACM. https://doi.org/10.1145/3632971.3632973

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., Zhou, J., Chen, S., Gui, T., Zhang, Q., & Huang, X. (2023). A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. https://doi.org/10.48550/arxiv.2303.10420

Ye, W., Zhang, Q., Zhou, X., Hu, W., Tian, C., & Cheng, J. (2024). Correcting factual errors in llms via inference paths based on knowledge graph. In *2024 International conference on computational linguistics and natural language processing (CLNLP)* (pp. 12–16). IEEE. https://doi.org/10.1109/clnlp64123.2024.00011

Zait, F., & Zarour, N. (2018). Addressing lexical and semantic ambiguity in natural language requirements. In *2018 Fifth international symposium on innovation in information and communication technology (ISIICT)* (pp. 1–7). IEEE. https://doi.org/10.1109/isiict.2018.8613726

Zhang, C., Peng, B., Sun, X., Niu, Q., Liu, J., Chen, K., Li, M., Feng, P., Bi, Z., Liu, M., Zhang, Y., Cheng, F., Yin, C. H., Yan, L., & Wang, T. (2024a). *From word vectors to multimodal embeddings: Techniques, applications, and future directions for large language models*. https://doi.org/10.48550/arxiv.2411.05036

Zhang, N., Yao, Y., Tian, B., Wang, P., Deng, S., Wang, M., Xi, Z., Mao, S., Zhang, J., Ni, Y., Cheng, S., Xu, Z., Xu, X., Gu, J. C., Jiang, Y., Xie, P., Huang, F., Liang, L., & Zhang, Z., , & Chen, H. (2024b). *A comprehensive study of knowledge editing for large language models*. https://doi.org/10.48550/arxiv.2401.01286

Zhang, Q., Singh, C., Liu, L., Liu, X., Yu, B., Gao, J., & Zhao, T. (2024c). Tell your model where to attend: Post-hoc attention steering for LLMs. In *The Twelfth international conference on learning representations, ICLR 2024, May 7–11, 2024*.

Zhang, T., Wang, C., Hu, N., Qiu, M., Tang, C., He, X., & Huang, J. (2022). Dkplm: Decomposable knowledge-enhanced pre-trained language model for natural language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(10), 11703–11711. https://doi.org/10.1609/aaai.v36i10.21425

Zhang, T., Xu, R., Wang, C., Duan, Z., Chen, C., Qiu, M., Cheng, D., He, X., & Qian, W. (2023). Learning knowledge-enhanced contextual language representations for domain natural language understanding. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 15663–15676). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.969

Zhang, Z., Dai, Q., Bo, X., Ma, C., Li, R., Chen, X., Zhu, J., Dong, Z., & Wen, J. R. (2025). A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, *43*(6), 1–47. https://doi.org/10.1145/3748302

Zhao, S., Yang, Y., Wang, Z., He, Z., Qiu, L. K., & Qiu, L. (2024a). *Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely*. https://doi.org/10.48550/arxiv.2409.14924

Zhao, Y., Yan, L., Sun, W., Xing, G., Wang, S., Meng, C., Cheng, Z., Ren, Z., & Yin, D. (2024b). Improving the robustness of large language models via consistency alignment. In N. Calzolari, M. Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, LREC-COLING (pp. 8931–8941). ELRA. https://doi.org/10.63317/4p5t4qbdw6ca