A Survey of Large Language Models, Retrieval-Augmented Generation, Natural Language to SQL, Intelligent Systems, Plasma Donation, and Natural Language Processing

www.surveyx.cn

Abstract

This survey paper provides a comprehensive examination of advanced technologies, including Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), natural language to SQL (nl2sql), intelligent systems, plasma donation, and natural language processing (NLP). These technologies have significantly impacted various sectors by enhancing information retrieval, data interaction, and human-machine interactions. LLMs have revolutionized text generation and information access, particularly in e-commerce and healthcare, despite challenges related to technical complexity and costs. RAG enhances LLM capabilities by integrating external data, improving information accuracy and contextual relevance. The nl2sql conversion facilitates intuitive data analysis by translating human language queries into SQL, while intelligent systems offer adaptive solutions across industries. In medical applications, particularly plasma donation, these technologies optimize efficiency and safety. The survey highlights LLMs' role in biomedical sciences for evidence synthesis, emphasizing their interdisciplinary applications. Motivations include addressing AI adoption challenges and enhancing translation quality in diverse contexts. The survey underscores the interconnectedness of these technologies and their transformative potential in advancing scientific understanding and practical applications. Future research directions focus on optimizing RAG, expanding LLM applicability, and exploring hybrid approaches to enhance retrieval quality and system performance across domains.

1 Introduction

1.1 Scope and Significance

This survey on advanced technologies, encompassing Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), natural language to SQL (nl2sql), intelligent systems, plasma donation, and natural language processing (NLP), underscores their substantial impact across various sectors. RAG-enhanced LLMs improve the accuracy and contextual relevance of responses by integrating external data, thereby transforming user interactions with structured and unstructured knowledge. The automation of literature reviews through NLP techniques demonstrates LLMs' capability to synthesize extensive academic literature, addressing challenges posed by the increasing volume of research output. These advancements not only enhance the reliability of generative AI systems but also streamline complex tasks across domains, including academic research and intelligent systems applications [1, 2, 3, 4]. LLMs have revolutionized information access and text generation, significantly benefiting fields like e-commerce and healthcare by facilitating tasks such as product translation and health-related inquiries. However, the deployment of LLMs presents challenges

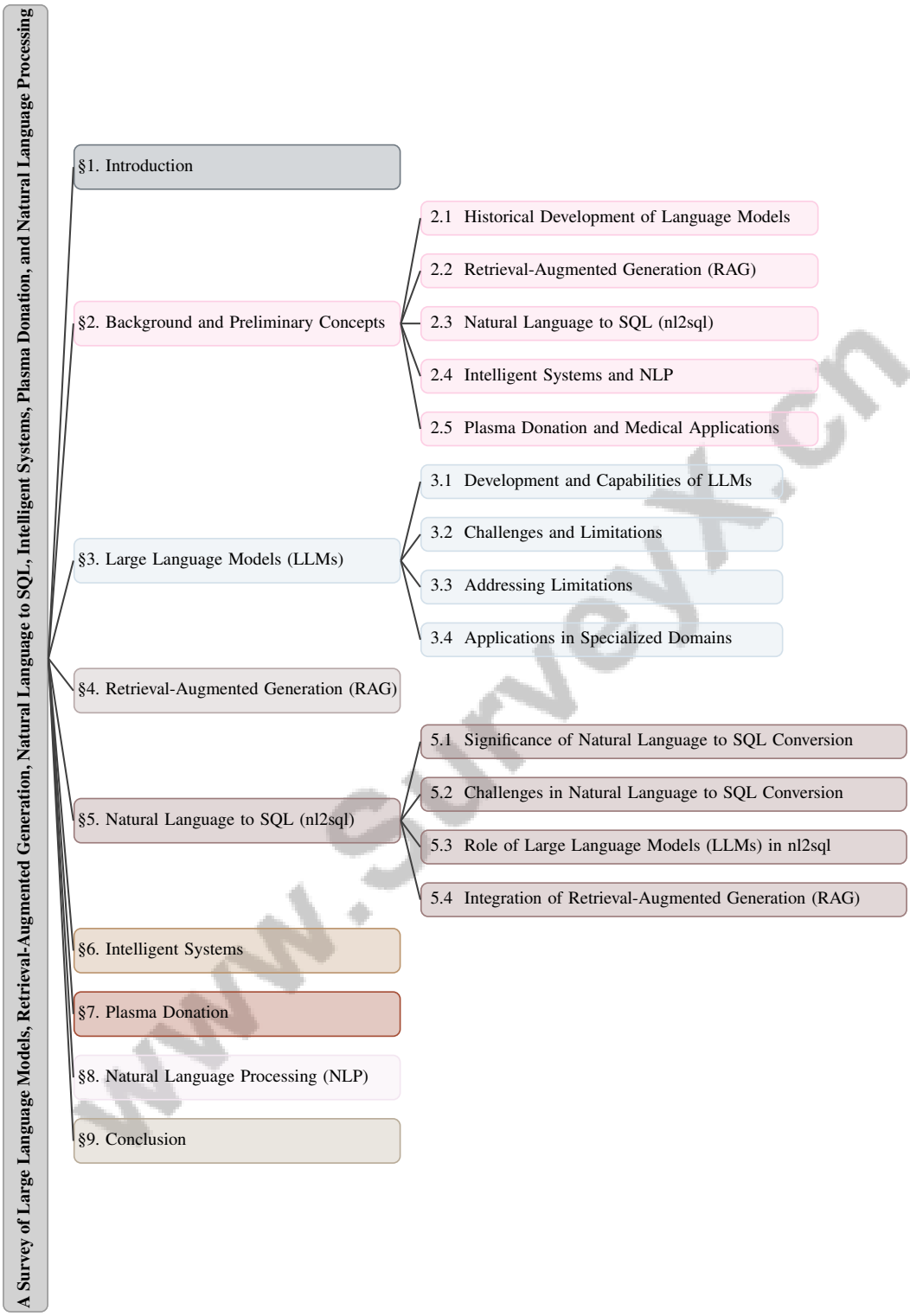


Figure 1: chapter structure

related to technical complexity and implementation costs that organizations must navigate to fully harness their potential.

RAG enhances LLMs by integrating external data, improving the relevance and accuracy of information retrieval and generation tasks, and addressing the limitations of traditional LLMs. This

integration is vital for enhancing user interactions across various complexities, enabling LLMs to better understand and retrieve relevant information, manage complex queries, and generate accurate evaluations, thereby expanding their applicability in fields such as academic literature review, medical summarization, and data insight generation [5, 6, 7, 3, 8].

The nl2sql conversion represents a significant advancement in data interaction, allowing for more intuitive and efficient data analysis. Intelligent systems are increasingly essential for facilitating seamless human-machine interactions [9]. In the medical domain, plasma donation exemplifies how intelligent systems and NLP can enhance efficiency and safety, showcasing the interdisciplinary applications of these technologies.

Furthermore, LLMs play a crucial role in biomedical sciences for evidence synthesis and knowledge extraction, highlighting their significance in advancing scientific research [10]. This survey aims to provide a comprehensive overview of these technologies, emphasizing their interconnectedness and the overarching importance of their integration in advancing both scientific understanding and practical applications.

1.2 Motivation for the Survey

The motivation for this survey stems from the need to understand and integrate advanced technologies, such as LLMs and RAG, to address critical challenges and opportunities across diverse domains. A primary focus is exploring how LLMs and RAG can enhance translation quality in e-commerce, yielding more accurate and contextually relevant outputs [11]. The survey also aims to tackle the complexities and inconsistencies in existing frameworks for LLM-based agents, which are vital for optimizing their deployment in various paradigms [12].

A significant aspect of this survey is addressing the practical challenges of AI adoption in enterprises, particularly regarding the implementation and management of large-scale AI models like LLMs [13]. Understanding effective deployment strategies is crucial, especially given the economic implications and sustainability concerns associated with LLMs [14]. Additionally, the survey seeks to examine LLMs' role in improving the reliability and accuracy of health-related information retrieval, assessing their performance against traditional search engines [15].

The survey also aims to leverage advances in machine translation technology to facilitate seamless communication between American Sign Language (ASL) and Indian Sign Language (ISL) users, thereby bridging communication gaps [16]. Ethical concerns related to sensitive information in LLMs, such as fairness in high-stakes applications like recruitment and education, are critical issues addressed by this survey.

By examining how data-augmented LLMs can effectively utilize external data, the survey aims to enhance performance, reduce hallucinations, and improve controllability and interpretability across various applications [4]. Furthermore, it provides a comprehensive overview of Retrieval-Augmented Language Models (RALMs), filling existing literature gaps regarding their paradigm, evolution, taxonomy, and applications [17]. Addressing the information scarcity faced by generative AI models, particularly LLMs, is essential for advancing their adaptability to new data and enhancing their overall efficacy [18].

1.3 Relevance in the Technological Landscape

The survey's exploration of advanced technologies such as LLMs, RAG, and nl2sql is highly relevant in today's rapidly evolving technological landscape. Integrating external data sources and advanced frameworks is essential to overcome the inherent limitations of LLMs, thereby enhancing their applicability and reliability in real-world scenarios [19]. This is particularly significant given the increasing reliance on LLMs across various sectors, including e-commerce and healthcare, where precise and contextually relevant information retrieval is critical.

Moreover, challenges posed by existing explanation methods often fail to provide clear and intuitive insights, exacerbating issues such as object hallucinations and a lack of fidelity in explanations [9]. Addressing these challenges is crucial for improving the transparency and trustworthiness of intelligent systems, which are integral to facilitating seamless human-machine interactions.

The focus on nl2sql highlights the growing importance of intuitive data interaction and analysis, enabling users to engage with complex databases more effectively. The integration of intelligent systems and NLP in the medical field, particularly in critical procedures such as plasma donation, illustrates the interdisciplinary potential of these technologies to significantly improve operational efficiency and safety. Recent advancements in LLMs have shown promise in automating complex tasks like evidence synthesis and data extraction, while also addressing challenges related to contextual understanding and the generation of accurate information. Employing techniques like RAG can enhance the relevance and reliability of these models in clinical settings, ultimately facilitating better decision-making and patient care. This underscores the necessity of continuous expert oversight in implementing these technologies to ensure their effectiveness in real-world applications [20, 10].

1.4 Structure of the Survey

The survey is meticulously structured to provide a comprehensive exploration of advanced technologies, including LLMs, RAG, nl2sql, intelligent systems, plasma donation, and NLP. The paper is organized into several key sections, each focusing on distinct yet interconnected aspects of these technologies.

The introductory section delineates the scope, significance, motivation, and relevance of the survey in the current technological landscape, highlighting the transformative impact of these technologies across various domains and setting the context for subsequent sections.

The second section, "Background and Preliminary Concepts," delves into the core concepts and historical evolution of the technologies under review, providing foundational knowledge necessary for understanding the advancements and applications discussed later.

Each of the main technologies is explored in dedicated sections. The LLMs section examines their development, capabilities, challenges, and applications in specialized domains, offering insights into their role in generating human-like text. The RAG section discusses its integration with LLMs, methodologies for enhancing retrieval quality, and the impact on LLM performance.

The nl2sql section addresses the conversion process of natural language queries into SQL database queries, highlighting its significance, challenges, and the role of LLMs and RAG in improving this conversion. The intelligent systems section focuses on their adaptive and autonomous capabilities and explores their applications across various industries.

Plasma donation is examined in a dedicated section that provides an overview of the medical procedure, discusses challenges, and explores the role of intelligent systems and NLP in improving efficiency and safety. The NLP section investigates the role of LLMs and RAG in advancing NLP technologies and their applications in domain-specific contexts.

Finally, the conclusion synthesizes key findings and insights from the survey, discussing emerging trends, future directions, and research opportunities. This structured approach ensures a coherent narrative that aligns with the survey's objectives while enhancing comprehension of the interconnected technologies by integrating advanced methodologies for extracting and enriching tabular data from complex PDF documents, thereby improving the accuracy and effectiveness of information retrieval in professional knowledge-based question answering systems [21, 22].The following sections are organized as shown in Figure 1.

2 Background and Preliminary Concepts

2.1 Historical Development of Language Models

Language models have evolved from basic statistical approaches to sophisticated Large Language Models (LLMs), marking significant advancements in natural language processing. Early models relied on n-gram techniques, which were limited in contextual understanding, while the advent of neural networks and deep learning significantly enhanced the ability to capture complex language semantics, improving tasks like question answering and summarization [23]. LLMs have shown particular promise in biomedicine for tasks requiring precise language comprehension, such as evidence synthesis and knowledge extraction [10]. However, challenges like outdated knowledge retention and inaccuracies in mathematical computations persist, necessitating domain-specific data integration to enhance performance, particularly in healthcare [24, 25].

Despite their potential, LLMs often struggle with incorporating current, guideline-based knowledge critical for clinical applications [26]. Issues such as accuracy limitations, bias, and the inability to access real-time external data highlight the need for continuous updates and fine-tuning [27]. The vast biomedical literature and existing method limitations complicate the generation of explainable biomedical hypotheses [28]. Additionally, deploying LLMs in practical applications presents computational and memory challenges [5], yet their ability to automate systematic and scoping reviews across fields underscores their transformative impact [23]. In power engineering, LLMs have been applied for data analysis, forecasting, and risk assessment, demonstrating versatility across domains [29].

Developing domain-specific instruction datasets is crucial for fine-tuning LLMs in specialized areas [5]. The manual creation of competency questions for ontology learning can be streamlined with LLMs [30]. Benchmarks for lengthy, legally nuanced documents reflect increasing demands on LLMs for processing and generating specialized text [23]. Emerging benchmarks like VITALITY 2 and REASONS address the need for LLMs capable of navigating extensive information corpora and generating accurate citations, enhancing reliability in scholarly contexts [3, 23].

The historical trajectory of language models reflects efforts to enhance capabilities and address limitations, culminating in more robust and adaptable LLMs. Recent advancements, including system-level performance optimization, Retrieval-Augmented Generation (RAG), and domain-specific updates, demonstrate a concerted effort to improve LLM effectiveness across diverse applications, tackling challenges related to accuracy, efficiency, and sustainability [14, 31, 32, 33, 34].

2.2 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) represents a significant advancement in NLP by integrating external knowledge sources to enhance LLM capabilities. By leveraging LLMs' in-context learning abilities, RAG produces more accurate and contextually relevant responses through the incorporation of relevant passages retrieved from an Information Retrieval (IR) system [35]. RAG addresses the limitations of traditional LLMs that rely on static datasets, which can lead to inaccuracies due to outdated or incomplete information [36]. Grounding LLM responses in reliable knowledge bases enhances accuracy, especially in specialized domains like healthcare [26].

The integration of RAG techniques in LLMs opens new applications, such as hypothesis-driven investigations in biomedical research, where explainable and actionable predictions are crucial [28]. However, RAG systems face challenges like retrieval latency and errors in document selection, necessitating efficient retrieval mechanisms and architectures to optimize performance [37]. RAG systems have shown potential in enhancing generation quality across benchmarks, crucial for domains requiring diverse content integration and multi-aspect query handling, augmenting LLM capabilities [35]. RAG's significance in advancing intelligent systems and NLP applications underscores its indispensable role in the current technological landscape.

2.3 Natural Language to SQL (nl2sql)

The conversion of natural language queries (NLQs) into Structured Query Language (SQL) queries, or nl2sql, represents a crucial advancement in data interaction, particularly in complex domains like pharmacovigilance. Translating NLQs into SQL commands allows for more intuitive and efficient querying of databases, enhancing data accessibility [38]. LLM-based text-to-SQL systems have democratized data access by simplifying querying for non-experts [39], improving SQL query generation accuracy and efficiency. Querying real-world data (RWD) databases to answer epidemiological questions highlights the importance of precise SQL queries, requiring sophisticated nl2sql systems capable of navigating RWD database intricacies [40].

Evaluating LLMs within the RAG context further highlights their potential to enhance text-to-SQL systems. By leveraging datasets containing biomedical questions and documents, LLMs can be assessed for generating coherent, contextually relevant SQL queries [41]. Simplifying access to complex databases like the Land Matrix Initiative (LMI) emphasizes nl2sql systems' role in reducing technical barriers to data querying, facilitating data-driven decision-making [42].

Developing and implementing nl2sql systems, particularly through paradigms like Table-Augmented Generation (TAG) and RAG advancements, is essential for improving data accessibility and user

interaction with complex databases. These systems leverage LLM reasoning and knowledge capabilities to enable diverse natural language queries, transcending traditional Text2SQL limitations focused on relational algebra. By facilitating meaningful insights extraction from structured and unstructured data, these systems enhance query accuracy, addressing challenges like computational efficiency, model robustness, and data privacy, paving the way for effective data management and analysis [43, 44, 39, 22].

2.4 Intelligent Systems and NLP

The evolution of intelligent systems and NLP has been marked by advancements that revolutionize automated reasoning and human-computer interaction. Intelligent systems, characterized by adaptability and learning capabilities, increasingly integrate advanced NLP techniques to enhance functionality across domains. The incorporation of LLMs has transformed intelligent systems, enabling them to process and interpret complex human language with improved accuracy and contextual relevance. Studies show that LLMs, particularly those based on the GPT architecture, have significantly enhanced automation in systematic reviews, achieving high precision and recall rates in data extraction tasks. Advanced models like GPT-4o introduce multi-modality capabilities, enhancing human-computer interaction and enabling effective data scraping from unstructured text, streamlining the review process and paving the way for future applications across fields [14, 31, 32, 33, 34].

Notably, LLMs are applied in domain-specific tasks like medical education, where challenges like hallucinations and effective summarization methods are prevalent [45]. LLM-based systems' ability to answer clinical questions effectively is critical, with benchmarks developed to evaluate their performance in generating relevant, evidence-based responses [46]. This capability is crucial for intelligent systems to efficiently manage vast data, especially in specialized fields like healthcare.

In sign language translation, frameworks using LLMs to convert American Sign Language (ASL) gestures into Indian Sign Language (ISL) gestures demonstrate intelligent systems' potential to bridge communication gaps through advanced gesture recognition and NLP techniques [16]. This integration highlights intelligent systems' transformative potential in facilitating seamless human-machine interactions across diverse linguistic contexts.

The emergence of LLMs significantly impacts NLP by addressing token-level Named Entity Recognition (NER) in clinical texts, especially for rare diseases. Existing methods often focus on document-level NER, limiting accurate recognition of medical entities with multiple synonyms and abbreviations. Software systems' complexity further complicates querying and understanding code, necessitating specialized knowledge and advanced NLP techniques to effectively manage vast information [47].

Despite advancements, challenges like bias in LLMs and data contamination persist, impacting evaluation integrity and intelligent systems' reliability. Bias in LLMs, particularly concerning culturally significant practices like traditional Chinese medicine, remains a critical issue [48]. Data contamination, categorized into guideline, raw text, and annotation contamination, poses varying implications for evaluation integrity, necessitating robust methodologies to ensure accurate and reliable NLP applications [49].

The integration of intelligent systems and NLP significantly enhances automated decision-making and reasoning capabilities, as evidenced by advancements like automated literature review systems using various NLP techniques, including retrieval-augmented generation with LLMs like GPT-3.5-turbo. Innovations like VITALITY 2 allow researchers to efficiently identify semantically relevant literature through advanced text embeddings and user-friendly interfaces, streamlining the literature review process. Improved PDF structure recognition in systems like ChatDOC demonstrates how enhanced data retrieval methods lead to more accurate responses in professional knowledge-based applications, highlighting these technologies' transformative potential across industries [2, 3, 21]. The ongoing evolution of these technologies promises to enhance adaptability and effectiveness in addressing complex real-world problems, underscoring intelligent systems and NLP's indispensable role in the technological landscape.

2.5 Plasma Donation and Medical Applications

Plasma donation is a critical medical procedure involving the extraction of plasma, the liquid component of blood, from donors. This process is essential for producing therapeutic products to treat

conditions like immune deficiencies, hemophilia, and other blood disorders. Advanced technologies, particularly intelligent systems and NLP, have significantly enhanced plasma donation procedures' efficiency and safety by automating complex tasks like evidence synthesis and data extraction, improving clinical decision-making through tailored prompts, and ensuring reliable information retrieval using techniques like Retrieval-Augmented Generation (RAG). These advancements facilitate adherence to clinical guidelines, allowing real-time adjustments in response to evolving medical knowledge, leading to safer, more effective donation practices [20, 10, 21].

LLMs in the medical domain show promise in generating evidence-based responses and enhancing clinical decision-making. However, traditional RAG methods, including standard RAG and GraphRAG, face limitations in addressing medical queries' complexities [50]. Specialized benchmarks, such as those evaluating LLMs in simulated tertiary care settings, underscore the necessity of assessing their performance in clinical decision-making and their potential as autonomous agents [20].

In plasma donation, LLMs and RAG systems play a crucial role in optimizing information retrieval and ensuring accurate medical responses. Comprehensive benchmarks like MedExpQA and MIRAGE significantly enhance LLMs' evaluation in medical question-answering tasks by providing structured frameworks assessing performance across languages and reasoning capabilities. MedExpQA addresses multilingual medical QA gaps by incorporating reference gold explanations authored by medical professionals, enabling accurate LLM prediction assessment. MIRAGE systematically evaluates RAG techniques, using diverse medical questions to identify optimal configurations for improving LLM accuracy. These benchmarks highlight ongoing LLM challenges, such as hallucinations and outdated knowledge, while paving the way for future advancements in medical AI applications [51, 52]. These benchmarks are essential for understanding LLM capabilities and limitations in handling complex medical information and providing accurate, contextually relevant responses.

LLM deployment in plasma donation exemplifies these technologies' interdisciplinary potential. By enhancing medical response accuracy and optimizing information retrieval, LLMs contribute to plasma donation safety and efficiency, ultimately improving patient outcomes. Continuous multilingual benchmark development supports LLM advancement in the medical domain, enabling effective cross-language comparisons and research [53].

Furthermore, HealthPariksha, a benchmark for assessing LLM performance in healthcare queries across multiple Indic languages, facilitates fair comparisons among models in real-world health chatbot settings [54]. This benchmark highlights the importance of evaluating LLMs in diverse linguistic contexts to ensure applicability and reliability in global health applications.

Integrating advanced language models, particularly RAG systems, in plasma donation and medical applications signifies transformative healthcare technology advancement. By customizing domain knowledge, these systems enhance medical response accuracy and speed, as demonstrated in studies where RAG significantly improved language model performance in tasks like preoperative medicine and medication consultations. For example, an LLM-RAG model achieved a response time of 15-20 seconds compared to the 10 minutes typically required by human practitioners, while also enhancing accuracy in extracting structured clinical data from unstructured reports. This technological synergy streamlines healthcare delivery and fosters improved patient care through more reliable, evidence-based information retrieval [55, 45, 56, 57, 26].

3 Large Language Models (LLMs)

The exploration of Large Language Models (LLMs) encompasses not only their remarkable development and capabilities but also the multifaceted challenges they face. Understanding these challenges is crucial for appreciating the limitations that may impede their effectiveness in real-world applications. As we transition to the first subsection, we will delve into the specific challenges and limitations that LLMs encounter, particularly in high-stakes environments where accuracy and reliability are paramount.

3.1 Development and Capabilities of LLMs

The development of Large Language Models (LLMs) has significantly advanced the field of natural language processing, enabling the generation of human-like text across various domains. These

models have become pivotal in transforming diverse fields, such as business process management (BPM), where their strong reasoning abilities facilitate the generation of structured outputs from unstructured textual data [58]. This capability is further exemplified in the realm of academic literature, where the VITALITY 2 framework leverages LLMs to overcome traditional literature review limitations, enabling more sophisticated textual interactions [3].

In software engineering, LLMs face challenges related to context length limitations, which restrict their ability to process extensive source code. The RAG method has been proposed to address these issues by integrating retrieval techniques with language models, thereby enhancing the generation of commit messages and improving software inquiries . The integration of large language models (LLMs) with advanced frameworks, such as the Missing Information Guided Retrieve-Extraction-Solving (MIGRES) paradigm and iterative in-context learning techniques, highlights their capacity to generate contextually relevant and precise outputs even in complex and technically demanding environments. These frameworks not only enhance the LLMs' ability to identify and retrieve necessary information step-by-step but also improve their evaluation capabilities in diverse applications, ranging from academic literature reviews to technical writing assistance. By addressing challenges like outdated knowledge and reasoning errors, these innovations demonstrate the significant potential of LLMs to operate effectively in intricate scenarios. [5, 3, 7, 6]

The adaptability of LLMs is further demonstrated in the field of automated literature review, where models like GPT-3.5-TURBO-0125 have achieved superior performance, as evidenced by high ROUGE-1 scores, surpassing traditional models such as T5 and spaCy [2]. This highlights the capabilities of LLMs in synthesizing and summarizing complex information efficiently.

Moreover, the exploration of LLM capabilities extends to enhancing commit message generation through frameworks like REACT, which integrates retrieval techniques to improve the quality and relevance of the generated messages [59]. These advancements reflect the transformative potential of LLMs in various applications, underscoring their role in driving innovation and efficiency across diverse sectors.

The ongoing advancement of Large Language Models (LLMs) significantly enhances their role in text generation and comprehension, leading to transformative developments across diverse fields such as software engineering, academic research, and information retrieval. For instance, LLMs like GPT-4 have been shown to automate systematic literature reviews, effectively summarize large datasets, and facilitate complex queries through innovative tools like VITALITY 2. These capabilities not only streamline the research process but also improve user efficiency in extracting insights from tabular data, thereby demonstrating the versatility and practical applicability of LLMs in tackling intricate challenges in real-world scenarios. [3, 33, 8]

3.2 Challenges and Limitations

Large Language Models (LLMs) have transformed natural language processing, yet they encounter significant challenges that constrain their effectiveness across various applications. A prominent issue is the occurrence of hallucinations, where LLMs produce outputs that seem plausible but are factually incorrect or misleading. This issue is particularly critical in high-stakes domains such as healthcare, where accuracy is vital for patient safety . The reliance on pre-trained knowledge without grounding in institutional guidelines exacerbates this problem, leading to potential inaccuracies [26].

The integration of domain-specific knowledge remains a critical challenge, as current models often struggle to incorporate specialized information effectively, impacting their performance in niche applications such as biomedical Named Entity Recognition (NER) [24]. This is compounded by the scarcity of domain-specific data, which limits the ability of LLMs to provide accurate and contextually relevant responses. Furthermore, LLMs face limitations in reasoning depth, particularly when utilizing Retrieval-Augmented Generation (RAG), which restricts their ability to perform deeper reasoning tasks effectively [60].

The real-time retrieval process in RAG introduces latency and errors, complicating the integration of retrieval and generation components [37]. This challenge is further complicated by the inability of existing RAG models to reflect real-time data after configuration, leading to inaccurate responses, especially for complex questions [36]. Additionally, LLMs may lead to hallucinatory responses, making RAG crucial for achieving accurate information [28].

Bias in LLMs, such as against culturally specific practices, poses a multifaceted problem, as models trained on extensive internet text corpora are likely to reflect the biases inherent in those sources. The excessive presence of irrelevant content increases token counts and introduces noise, complicating LLM processing and reducing the quality of generated outputs. Moreover, the challenge of generating accurate citations requires LLMs to analyze multiple references and determine precisely when to incorporate citations, a task at which current methods struggle [36].

Addressing these challenges is crucial for enhancing the reliability, accuracy, and applicability of LLMs across diverse domains. By enhancing the integration of domain-specific knowledge through advanced techniques such as Retrieval-Augmented Generation (RAG) and Sequential Fusion methods, Large Language Models (LLMs) can significantly improve their factual accuracy and contextual relevance. These methods not only address the issue of hallucinations—where LLMs generate incorrect or nonsensical information—but also help mitigate biases that may arise from limited or skewed datasets. As a result, LLMs can produce outputs that are more meaningful and aligned with real-world applications, thereby reducing the risks associated with their deployment in critical fields such as healthcare and finance. This approach leverages curated datasets and sophisticated reasoning techniques to ensure that LLMs can effectively handle complex, domain-specific queries, ultimately enhancing their reliability in knowledge-intensive tasks. [6, 61, 34]

3.3 Addressing Limitations

Addressing the limitations of Large Language Models (LLMs) necessitates the implementation of advanced strategies and methodologies that enhance their robustness, accuracy, and contextual understanding across diverse applications. One innovative approach involves the creation of structured vector databases, which facilitate accurate data retrieval and improve the contextual comprehension capabilities of LLMs [62]. This structured approach ensures that LLMs can access and utilize relevant information more effectively, thereby reducing the occurrence of hallucinations and improving the reliability of generated outputs.

The development of a Knowledge Boundary Model (KBM) represents another significant advancement in optimizing retrieval processes. By categorizing questions based on their necessity, KBMs enhance retrieval efficiency and ensure that LLMs are equipped with the most pertinent information for generating accurate responses [63]. This approach is particularly beneficial in high-stakes environments where the precision of information is critical.

SubgraphRAG, which integrates a lightweight multilayer perceptron with a parallel triple-scoring mechanism, offers a flexible and efficient solution for subgraph retrieval. This method enhances retrieval effectiveness through directional structural distances, allowing LLMs to access and process information more efficiently [64]. The integration of such innovative retrieval techniques is crucial for overcoming the limitations associated with static datasets and improving the dynamic adaptability of LLMs.

In the realm of business process management (BPM), LLM-based BPM Task Processing (LLM-BPM) utilizes pre-trained LLMs to analyze textual descriptions and generate outputs relevant to BPM tasks [58]. This method underscores the potential of LLMs to transform complex workflows by providing contextually relevant insights and enhancing decision-making processes.

DRAGIN, a dynamic retrieval-augmented approach, innovates by employing Real-time Information Needs Detection (RIND) and Query Formulation based on Self-attention (QFS) to optimize retrieval processes [65]. This dual approach moves beyond static methods, enabling LLMs to dynamically adjust their retrieval strategies based on real-time information needs, thereby enhancing the accuracy and relevance of generated outputs.

DPrompt tuning is another promising strategy designed to leverage document information effectively while minimizing the computational overhead associated with processing multiple transformer layers [60]. By optimizing the use of document data, DPrompt tuning enhances the reasoning capabilities of LLMs and reduces the risk of generating erroneous or misleading outputs.

Finally, the development of benchmarks inspired by the need to mitigate hallucinations in LLM outputs plays a crucial role in enhancing model performance [57]. By augmenting LLMs with relevant, domain-specific documents during inference, these benchmarks ensure that the models are grounded in reliable information, thereby improving the accuracy and reliability of their outputs.

These strategies and methodologies effectively address the inherent limitations of large language models (LLMs) by enhancing their ability to generate accurate, contextually relevant, and reliable outputs across various applications. For instance, the Missing Information Guided Retrieve-Extraction-Solving (MIGRES) paradigm improves retrieval-augmented generation (RAG) systems by enabling LLMs to identify and target missing information during the reasoning process, which leads to more effective document retrieval and content generation. Additionally, LLM-Ref enhances reference handling in technical writing by directly retrieving and generating content from text paragraphs, significantly improving the accuracy and contextual relevance of outputs, as evidenced by a notable increase in Ragas scores compared to traditional RAG systems. These advancements collectively bolster the efficacy of LLMs in diverse applications, ensuring they produce high-quality, relevant results. [5, 6]. By focusing on robust knowledge integration, dynamic data updating, and comprehensive evaluation frameworks, researchers can significantly improve the performance and applicability of LLMs in diverse fields.

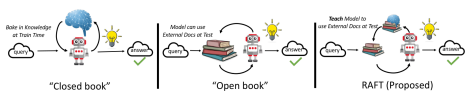
3.4 Applications in Specialized Domains

The application of Large Language Models (LLMs) in specialized domains has demonstrated their transformative potential across various industries. In the field of business process management (BPM), LLMs such as GPT-4 have shown significant promise. The ability of these models to perform BPM tasks effectively, often surpassing traditional methods, underscores their potential to streamline and optimize complex business workflows [58]. This capability is indicative of the broader applicability of LLMs in specialized domains, where they can provide valuable insights and enhance decision-making processes.

In the realm of problem-solving and decision-making, the integration of retrieval mechanisms such as ARM-RAG has further enhanced the capabilities of LLMs. Experiments have demonstrated that ARM-RAG significantly improves the problem-solving abilities of LLMs, outperforming baseline systems that lack retrieval augmentation [66]. This enhancement is particularly crucial in domains where the complexity of tasks requires access to external knowledge bases, enabling LLMs to generate more accurate and contextually relevant responses.

Despite the advancements in LLM capabilities, challenges remain in accurately capturing their true potential through existing benchmarks. Studies have revealed inadequacies in current benchmarks, such as those involving state-of-the-art models like GPT-4 and Gemini, highlighting the need for more comprehensive evaluation frameworks that can better assess the full range of LLM functionalities [67]. These insights emphasize the importance of developing robust benchmarks that can effectively measure the performance of LLMs in specialized applications.

The deployment of Large Language Models (LLMs) in specialized domains is significantly enhancing innovation and efficiency by addressing unique challenges such as Few-Shot Domain-Expert Reasoning, improving knowledge retrieval through advanced frameworks like Missing Information Guided Retrieval-Augmented Generation, and facilitating effective table-to-text generation across various industries. These advancements not only demonstrate the transformative potential of LLMs but also highlight their ability to achieve substantial accuracy gains in specific applications, such as medical and economic datasets, thereby driving progress in real-world information seeking and decision-making processes. [6, 8, 34]. By leveraging advanced retrieval mechanisms and addressing benchmark limitations, LLMs can further enhance their applicability and impact in specialized fields, paving the way for more effective and intelligent solutions.



(a) Closed Book vs. Open Book: A Comparative Study on Knowledge Transfer and Retrieval[68]

	$\text{Acc}_{\text{retrieval\_db}}$
With Open-Domain Queries	0.53
Without Open-Domain Queries	0

(b) Accuracy of Retrieval in Database with and without Open-Domain Queries[69]

Figure 2: Examples of Applications in Specialized Domains

As shown in Figure 2, Large Language Models (LLMs) have demonstrated significant potential in various specialized domains, providing innovative solutions and enhancing the efficiency of

knowledge transfer and retrieval processes. The examples illustrated in Figure 2 highlight two distinct applications of LLMs in specialized domains. The first example, "Closed Book vs. Open Book: A Comparative Study on Knowledge Transfer and Retrieval," explores the efficacy of two approaches to knowledge management. The "Closed book" approach relies on pre-stored knowledge to answer queries, whereas the "Open book" approach enhances model performance by allowing access to external documents during testing. This comparison underscores the potential of LLMs to adaptively utilize external information sources to improve accuracy and relevance in knowledge retrieval tasks. The second example, "Accuracy of Retrieval in Database with and without Open-Domain Queries," examines how the inclusion of open-domain queries influences retrieval accuracy. The data suggests that incorporating open-domain queries can significantly impact the precision of information retrieval from databases, showcasing the transformative impact of LLMs in optimizing data-driven decision-making processes. These examples collectively demonstrate the versatility and adaptability of LLMs in addressing complex challenges across specialized domains, paving the way for more dynamic and informed applications. [**?**]zhang2024raft,li2024grammargroundedmodularmethodology)

In recent years, the integration of Retrieval-Augmented Generation (RAG) has emerged as a pivotal advancement in the realm of Large Language Models (LLMs). This approach not only enhances the capabilities of LLMs but also addresses critical challenges related to factual accuracy and retrieval quality. As illustrated in Figure 3, the hierarchical structure of RAG delineates its various components, encompassing the concepts, methodologies, and impacts associated with its implementation. This figure effectively highlights how structured optimization techniques and advanced technological integrations contribute to the overall performance improvements seen in LLMs, thereby reinforcing the significance of RAG in contemporary natural language processing. By understanding this framework, researchers can better appreciate the multifaceted benefits that RAG brings to the field, paving the way for future innovations in language model development.

4 Retrieval-Augmented Generation (RAG)

4.1 Concept and Integration of RAG with LLMs

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) by integrating external knowledge sources, thus addressing issues like hallucinations and outdated information. This integration improves the factual accuracy and relevance of generated outputs by retrieving data from extensive databases. RAG combines LLMs with retrieval systems to ensure responses are reliable and contextually relevant [26]. The Query-Based Retrieval Augmented Generation (QB-RAG) method optimizes retrieval by matching queries through vector search, enhancing content accuracy [35]. DiRAG improves biomedical Named Entity Recognition (NER) by incorporating medical knowledge, enhancing information extraction accuracy. RUGGED synthesizes biomedical knowledge from curated databases, supporting hypothesis generation [26].

Graph technology further enhances RAG systems, as demonstrated by Jeong et al., who improved response accuracy through diverse data synthesis. Cache-augmented generation (CAG) preloads resources, enhancing RAG efficiency by eliminating real-time retrieval [35]. RAG's integration with LLMs optimizes retrieval through techniques like query rewriting and chunking, addressing common LLM challenges and leading to more reliable responses across domains [70, 6, 4, 1, 71]. RAG not only enhances LLM reliability but also expands their potential in complex environments, paving the way for intelligent solutions in natural language processing.

4.2 Methodologies for Enhancing Retrieval Quality

Enhancing retrieval quality in RAG systems is crucial for producing accurate and contextually relevant LLM outputs. Various methodologies optimize retrieval processes, focusing on precision, recall, and integration of retrieved information. The RAG framework improves LLMs by retrieving relevant text data, integrating specialized knowledge [26]. Cache-Augmented Generation (CAG) enhances efficiency by preloading context, eliminating real-time retrieval [37]. DiRAG integrates external biomedical knowledge, improving NER precision and relevance [24].

Structured methodologies are key in enhancing retrieval quality. Wang et al. developed a benchmark to optimize RAG workflows, focusing on query classification, retrieval, reranking, and summarization [35]. Enhancements like multi-query Query Rewriter, Knowledge Filter, and Memory Knowledge

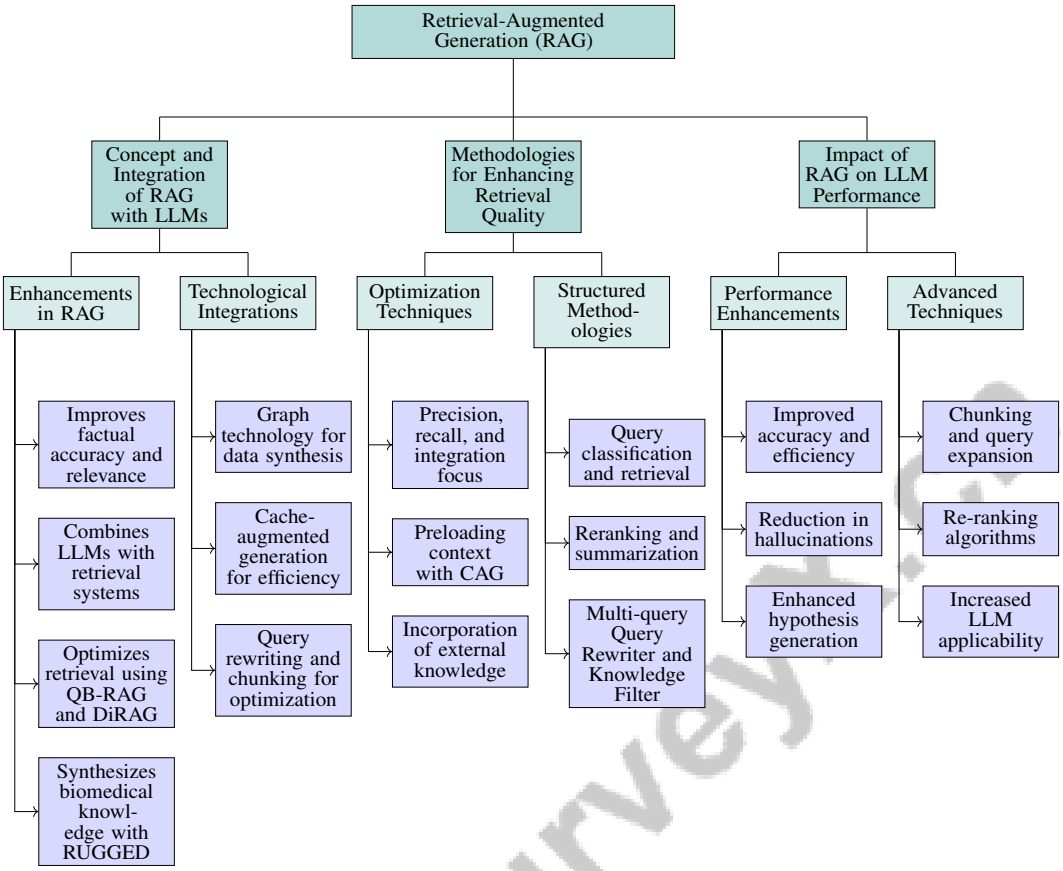
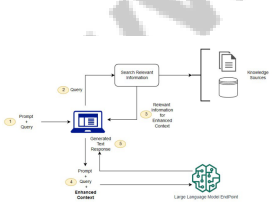
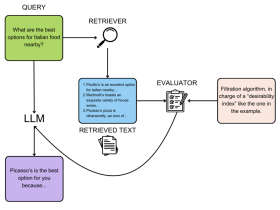


Figure 3: This figure illustrates the hierarchical structure of Retrieval-Augmented Generation (RAG) in enhancing Large Language Models (LLMs). It categorizes the concepts, methodologies, and impacts of RAG, highlighting enhancements in factual accuracy, retrieval quality, and LLM performance through structured optimization techniques and advanced technological integrations.

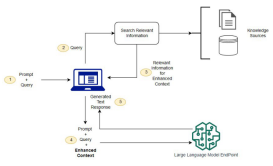
Reservoir significantly improve the ability to generate accurate outputs. Empirical validation across datasets supports these advancements, refining response quality and efficiency in generating information from complex data sources [71, 22].



(a) A diagram illustrating the process of generating text responses from large language models[72]



(b) A flowchart illustrating the process of a language model (LLM) generating a response to a query about Italian food options nearby.[73]



(c) A diagram illustrating the process of generating text responses from large language models[70]

Figure 4: Examples of Methodologies for Enhancing Retrieval Quality

As illustrated in Figure 4, RAG is crucial in enhancing LLM response quality. This methodology integrates retrieval systems with generative models to improve the relevance and accuracy of generated text. The diagrams exemplify this process, demonstrating the systematic approach of generating

responses by utilizing enhanced context from relevant information to produce accurate outputs [72, 73, 70].

4.3 Impact of RAG on LLM Performance

RAG integration significantly enhances LLM performance by improving accuracy, efficiency, and contextual relevance. RAG configurations, including monoT5, monoBERT, and Llama-2-7b, demonstrate substantial improvements in response accuracy and efficiency [35]. Wang et al.'s benchmark evaluates RAG's impact on LLM performance, covering general and specialized capabilities [35].

In biomedical research, RUGGED minimizes hallucinations and provides actionable insights, enhancing hypothesis generation reliability [28]. The LLM-RAG model generates complex preoperative instructions with human-comparable accuracy, showcasing RAG's practical benefits in medical applications [26]. Cache-Augmented Generation (CAG) outperforms traditional RAG systems in efficiency and accuracy by preloading context, optimizing retrieval speed and reliability [37]. LLMs with RAG exhibit improved accuracy and reduced hallucination rates, illustrating RAG's positive impact on performance [57].

RAG integration enhances LLM performance by improving accuracy, relevance, and efficiency of responses. This improvement results from RAG's ability to source pertinent information, addressing outdated knowledge and inaccuracies. Advanced techniques like chunking, query expansion, and reranking algorithms refine the RAG process, enhancing retrieval quality and elevating LLM reliability across specialized fields [60, 70, 6, 1, 4]. These advancements expand LLM applicability in complex environments, paving the way for reliable solutions in natural language processing.

5 Natural Language to SQL (nl2sql)

5.1 Significance of Natural Language to SQL Conversion

Transforming natural language queries (NLQs) into Structured Query Language (SQL) marks a pivotal advancement in data interaction, significantly boosting information retrieval efficiency and accuracy across various fields. This process is essential in domains like pharmacovigilance, where accurate translation of NLQs into SQL commands facilitates intuitive database querying, thus improving access to vital data [38]. In e-commerce, precise translation of product titles highlights the importance of context in effective data interaction [11].

Large Language Models (LLMs) play a crucial role in this conversion, generating correct SQL queries from natural language input and underscoring the necessity of context and precision in data interaction [74]. Surveys on text-to-SQL systems have underscored benchmarks, evaluation methods, and knowledge graphs' role in enhancing system accuracy and efficiency [39].

Integrating retrieval-augmented generation (RAG) with text-to-SQL generation enhances the ability to answer complex queries using electronic health records (EHR) and claims data [40]. This integration is crucial for effective data extraction and analysis, especially in specialized fields where database complexity presents challenges. Moreover, querying unstructured text data with SQL is complicated by poorly defined schemas, highlighting the need for sophisticated nl2sql systems capable of managing natural language input intricacies [44].

NLQ to SQL conversion represents a transformative step in data interaction, driven by LLMs and RAG systems. This evolution enables users to extract insights from extensive text corpora with improved ease and accuracy. Unlike traditional methods that often struggled with schema alignment, modern LLM-based text-to-SQL systems facilitate hybrid querying, integrating structured and unstructured data. This capability broadens accessible data and addresses challenges such as computational efficiency and model robustness, empowering users to derive precise insights from their queries [75, 44, 39].

5.2 Challenges in Natural Language to SQL Conversion

Converting natural language queries (NLQs) to Structured Query Language (SQL) poses several challenges impacting data retrieval efficiency and accuracy. A primary issue is the mismatch between user query formulation in natural language and the complex structures of databases, leading to high

rates of query errors [38]. This is further complicated by differences in entity descriptions between natural language and their database representations, hindering accurate SQL query generation [74].

Current data extraction methods are often time-consuming, error-prone, and reliant on manually prepared static pipelines, limiting their effectiveness [44]. The complexity of linguistic expressions and the need to understand diverse database schemas necessitate advanced reasoning capabilities to handle a wide range of user queries [43]. Additionally, SQL query equivalence checking is undecidable, with existing methods relying on heuristic rules requiring significant domain expertise and engineering effort [75].

These challenges are intensified by the need for thorough exploration and validation of complex research hypotheses across multiple datasets, which current LLM and RAG methods struggle to address effectively. Benchmarks, especially those related to financial questions, highlight the difficulty of ensuring factual correctness and avoiding hallucinations in generated responses [76]. Addressing these challenges is essential for improving data interaction accuracy and efficiency, enabling more intuitive engagement with complex databases.

5.3 Role of Large Language Models (LLMs) in nl2sql

Large Language Models (LLMs) significantly enhance the natural language to SQL (nl2sql) conversion by leveraging advanced contextual understanding and retrieval mechanisms. Their integration into nl2sql systems facilitates accurate translation of complex natural language queries into structured SQL commands, improving data accessibility across various domains. Qin et al.'s framework employs a memory architecture enabling LLMs to select relevant databases and generate SQL queries, emphasizing contextual knowledge's importance in optimizing query generation [74].

The FACTS framework, as discussed by Akkiraju et al., illustrates how optimizing Retrieval-Augmented Generation (RAG) pipelines can enhance nl2sql processes' accuracy and efficiency, underscoring LLMs' potential to streamline data retrieval and query formulation by incorporating relevant contextual information [77]. Additionally, the RATSG methodology combines text-to-SQL generation with RAG to effectively query Electronic Health Records (EHR) and claims data, showcasing LLMs' applicability in specialized fields like epidemiology [40].

Biswal et al.'s benchmark evaluates various models' performance in answering complex natural language questions over databases, focusing on queries requiring semantic reasoning and world knowledge. This evaluation demonstrates LLMs' capability to handle intricate queries demanding a deep understanding of database structures and contextual nuances [43]. Zhao et al. propose using LLMs to assess SQL query equivalence through techniques like Miniature Mull for semantic equivalence and Explain Compare for relaxed equivalence, further demonstrating LLMs' versatility in enhancing nl2sql processes [75].

Moreover, Mohammadjafari et al.'s survey categorizes text-to-SQL methods into rule-based, deep learning, and LLM approaches, highlighting LLMs' evolution and capabilities in this domain [39]. The MLPrompt framework, tested on text-to-MIP and text-to-SQL tasks, exemplifies LLMs' adaptability in improving data retrieval precision and expressiveness [78]. Zhao et al. emphasize that while RAG is effective for explicit fact queries, more complex queries require iterative RAG and specialized techniques like text-to-SQL for data linkage [4].

In facilitating user interaction with databases, Kbir et al. propose using LLMs to generate REST and GraphQL queries, enabling more intuitive and user-friendly access to complex datasets like the LMI database [42]. This approach illustrates LLMs' versatility in bridging the gap between natural language input and structured database queries.

LLMs significantly advance nl2sql processes by enhancing SQL query generation's accuracy, efficiency, and contextual relevance. By integrating advanced retrieval techniques and utilizing contextual knowledge, LLMs improve user interaction with complex databases, enabling precise data extraction and analysis across multiple fields. This is achieved through methods like the Missing Information Guided Retrieve-Extraction-Solving (MIGRES) framework, which enhances information retrieval by identifying knowledge gaps, and the contextual enrichment of tabular data, improving complex query accuracy. These innovative approaches facilitate a more intuitive engagement with data, leading to more meaningful insights and informed decision-making [4, 34, 6, 22].

5.4 Integration of Retrieval-Augmented Generation (RAG)

The integration of Retrieval-Augmented Generation (RAG) into natural language to SQL (nl2sql) systems is crucial for enhancing the accuracy and efficiency of converting natural language queries into SQL commands. RAG combines retrieval mechanisms with generation capabilities, facilitating complex database queries by incorporating relevant external information into the query generation process. This integration improves semantic understanding and ensures generated SQL statements accurately reflect user intent [39].

The LLM-SQL-Solver exemplifies RAG's application in determining SQL queries' semantic and relaxed equivalence. By utilizing LLMs, this method enhances SQL query evaluation and generation, improving data retrieval precision and reliability [75]. Balancing fine-tuning with retrieval-augmented generation is an emerging trend, optimizing nl2sql systems' performance by ensuring LLMs access the most relevant, up-to-date information during query processing [39].

The Galois prototype further demonstrates RAG's potential in nl2sql applications by executing SQL queries over data stored in pre-trained LLMs. This DB-first approach utilizes a logical query plan to simplify complex tasks, leveraging SQL's structured nature to enhance data retrieval accuracy [44]. The effectiveness of this method is supported by LLMs' inherent ability to store factual information, which, when combined with SQL's structured framework, facilitates precise and efficient data interaction [44].

Integrating RAG into nl2sql systems significantly enhances their capacity to produce accurate and contextually relevant SQL queries through techniques like knowledge retrieval and context enrichment. This evolution from traditional rule-based models to sophisticated LLM approaches not only improves query translation accuracy but also addresses challenges related to computational efficiency, model robustness, and data privacy. Incorporating modules like the Query Rewriter and Knowledge Filter further refines the query generation process, ensuring systems can effectively handle complex queries and setting a new standard in the field [39, 71, 22]. By combining advanced retrieval techniques with LLMs' generative power, RAG enhances the interaction between natural language inputs and structured database queries, enabling more effective and intuitive data extraction and analysis across diverse domains.

6 Intelligent Systems

6.1 Adaptive and Autonomous Capabilities

Intelligent systems' adaptive and autonomous features have revolutionized automated processes and human-computer interaction by enabling decision-making with minimal human input. These systems, exemplified by optimized Retrieval-Augmented Generation (RAG) pipelines, dynamically adjust retrieval strategies to enhance user interaction [77]. Adaptive systems like VITALITY 2 facilitate seamless navigation of large datasets, streamlining literature reviews and enhancing data-driven decision-making [3].

Autonomous capabilities allow these systems to function independently, learning and adapting in real-time, which is essential for applications requiring immediate data processing and decision-making. By leveraging RAG systems, they integrate extensive database information, refreshing their knowledge to maintain efficiency across domains [3, 79]. Through advanced algorithms and machine learning, these systems manage complex tasks autonomously, reducing manual intervention and boosting operational efficiency.

The continuous evolution of adaptive and autonomous capabilities marks a significant advancement in artificial intelligence, fostering intuitive technology interactions. Technologies like RAG and LLMs provide innovative solutions that enhance information retrieval and comprehension in academia, engineering, and environmental compliance. VITALITY 2 aids in identifying semantically relevant literature, while RAG-Fusion improves customer service response accuracy. Additionally, advancements in PDF structure recognition support knowledge-based question-answering systems, demonstrating their transformative potential across industries [80, 3, 81, 21].

6.2 Applications Across Industries

Intelligent systems have significantly enhanced operational efficiency and decision-making across industries. In claim verification, LLMs expedite fact-checking, demonstrating their impact on information validation [82]. These systems efficiently analyze claims using advanced natural language processing, reducing manual fact-checking time and resources.

In healthcare, LLMs and RAG techniques improve clinical decision-making and patient care by integrating comprehensive medical knowledge bases. They provide healthcare professionals with precise, contextually relevant information, enhancing clinical documentation and supporting evidence-based medicine [83, 10].

In finance, intelligent systems enhance data analysis and risk assessment with RAG techniques, improving information retrieval accuracy from financial documents. This optimization supports informed decision-making and risk management through refined text chunking, query expansion, and metadata annotations [84, 70, 72]. LLMs provide insights into market trends and consumer behavior, aiding strategic decisions and enhancing fraud detection and credit scoring.

In manufacturing, intelligent systems optimize production and supply chain management by forecasting equipment failures and optimizing maintenance schedules. These predictive capabilities reduce downtime and improve operational efficiency, allowing proactive resource management [79, 6, 71, 2, 3].

The widespread implementation of RAG models and LLMs across industries underscores their transformative potential in fostering innovation and operational efficiency. Innovations like the Golden-Retriever enhance document retrieval by interpreting domain-specific jargon, while VITALITY 2 streamlines literature reviews using semantic embeddings. These advancements demonstrate significant progress in information management and operational capabilities in enterprise settings [84, 18, 71, 2, 3]. Harnessing LLMs and RAG capabilities continues to transform industry practices, offering new growth opportunities in an evolving technological landscape.

7 Plasma Donation

7.1 Overview of Plasma Donation

Plasma donation is a critical medical procedure that involves collecting plasma from volunteer donors, essential for treating clotting disorders, burn victims, and surgical patients [45, 85, 3, 10]. The process utilizes plasmapheresis, where blood is extracted, plasma is separated, and the remaining components are returned to the donor. Intelligent systems and NLP technologies, including RAG methods like MedGraphRAG, enhance the efficiency and safety of plasma donation by accurately processing donor information and improving medical response reliability [86]. These technologies optimize the extraction and synthesis of medical knowledge, facilitating evidence-based clinical workflows and providing real-time, contextually relevant information, thereby improving operational efficiencies and the quality of patient care in plasma donation [3, 46, 20, 10].

7.2 Challenges in Plasma Donation

Plasma donation faces several challenges, including donor recruitment, retention, and managing potential adverse reactions [85, 3, 10, 21]. Ensuring plasma quality and safety requires stringent screening and testing protocols. Effective recruitment and retention strategies are crucial to meet the demand for plasma-derived therapies [87, 21, 57, 85, 3]. Intelligent systems and NLP, particularly LLMs and RAG, address these challenges by automating data extraction and evidence synthesis, improving donor interactions, and streamlining literature reviews for medical guidelines [10, 31, 49, 2, 3]. These technologies support donor screening, eligibility assessments, and information management, enhancing donor experience and encouraging repeat donations. Despite technological advancements, logistical challenges such as specialized equipment and trained personnel remain critical for safety and efficiency [10, 45, 81, 21]. A multifaceted approach combining technological innovation with effective donor engagement and education strategies is essential to enhance safety protocols, operational efficiency, and donor satisfaction [2, 88, 3, 10].

7.3 Role of Intelligent Systems in Plasma Donation

Intelligent systems significantly enhance the efficiency and safety of plasma donation by leveraging technologies like MedGraphRAG, which facilitates structured information management and generates evidence-based responses [86]. These systems improve donor screening accuracy and eligibility assessments, reducing errors and enhancing safety. Through advanced algorithms and data processing techniques, intelligent systems optimize donor information management and scheduling, streamlining workflows and improving donor experiences with efficient communication and personalized interactions [6, 71, 2, 3, 77]. This optimization encourages donor retention by minimizing wait times and ensuring a seamless donation process. The ability to process large data volumes quickly allows for real-time monitoring and decision-making, crucial for maintaining plasma quality and safety. These systems exemplify transformative potential in healthcare, improving data management, operational efficiency, and decision-making, thereby enhancing the safety and effectiveness of plasma donation procedures [22, 3, 21].

7.4 Application of NLP in Plasma Donation

NLP technologies significantly improve the efficiency and safety of plasma donation by automating and optimizing various aspects, from donor recruitment to post-donation follow-ups. Advanced language models enable NLP systems to process extensive unstructured data, enhancing donor screening and eligibility assessments through RAG techniques for accurate clinical information extraction. Studies indicate that fine-tuned language models can achieve over 98

8 Natural Language Processing (NLP)

8.1 Advancements in NLP through LLMs and RAG

The integration of Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) has significantly advanced natural language processing (NLP) by improving accuracy, contextual understanding, and application across various domains. SubgraphRAG exemplifies this by efficiently retrieving relevant subgraphs with high reasoning accuracy and explainability, enhancing complex Knowledge Graph Question Answering (KGQA) tasks [64]. This approach underscores RAG's potential in leveraging structured data for improved information retrieval.

In automated literature reviews, Ali et al.'s benchmark highlights RAG's effectiveness in synthesizing complex information, crucial for streamlining research and enhancing scientific knowledge accessibility [2]. The synergy of LLMs and RAG also bolsters evidence synthesis, as Garcia et al. emphasize the need for unified benchmarks and improved contextual understanding [10]. Tailored RAG approaches, such as those by Gilson et al. in ophthalmology, demonstrate enhanced factuality and attribution by utilizing large, curated datasets [57].

Furthermore, RAG's adaptability is evident in diverse contexts, including source code, as evaluated by Kamiya et al. [89]. Ke et al. showcase the LLM-RAG model's ability to generate precise preoperative instructions, reducing hallucination rates and enhancing NLP reliability in specialized domains [26]. Recent advancements in retrieval processes, including sophisticated chunking techniques and metadata annotations, have improved response accuracy in complex scenarios, such as querying intricate tabular data in PDF documents [70, 22].

8.2 Applications in Domain-Specific NLP

NLP has extensive applications across various domains, leveraging its capability to process and interpret human language with precision. In healthcare, NLP technologies enhance clinical decision-making by extracting and synthesizing medical information from vast literature and electronic health records, facilitating evidence-based medicine and improving patient outcomes [10]. In the legal sector, NLP streamlines the analysis of complex legal documents, enhancing accessibility and research efficiency [23].

The financial industry utilizes NLP to analyze market trends and consumer sentiment, enabling strategic investment decisions and enhancing competitiveness [29]. In education, NLP supports intelligent tutoring systems and automated grading tools, providing personalized learning experiences

and timely feedback [30]. Additionally, NLP applications in customer service, such as chatbots, improve engagement by providing timely and accurate responses [65].

The diverse applications of NLP in specialized fields, including automated literature reviews and domain-specific model adaptation, underscore its transformative potential across industries by enhancing efficiency, accuracy, and knowledge integration [90, 68, 2, 3, 34]. By improving information processing and decision-making capabilities, NLP technologies continue to drive innovation and efficiency, offering new growth opportunities in the evolving technological landscape.

8.3 Challenges and Ethical Considerations

The development and application of NLP technologies face numerous challenges and ethical considerations crucial for ensuring reliability and integrity. RAG methods' reliance on execution trace collection tools may inadequately capture function calls from dynamically loaded modules, leading to incomplete data processing [89]. Additionally, data quality and representation for underrepresented groups often result in biased outcomes in RAG applications [27].

The dependence on well-curated knowledge bases necessitates ongoing updates to maintain accuracy, particularly in rapidly evolving fields like healthcare [91]. Noise in documents presents a persistent challenge, requiring additional reasoning depth to address effectively [60]. Current benchmarks for addressing LLM hallucinations and evidence attribution may not fully capture all related challenges, indicating a need for further refinement [57].

In the biomedical domain, reliance on resources like the UMLS knowledge base can limit performance, affecting NLP efficacy in specific datasets [24]. These challenges highlight the necessity for comprehensive evaluation frameworks and continuous updates to ensure the effectiveness and fairness of NLP systems.

Ethical considerations are paramount, particularly regarding fairness and transparency. The potential for misinformation and the nuances of human language necessitate comprehensive measures to understand the broader implications of LLM-generated content. Privacy risks associated with RAG systems underscore the need for robust measures to protect user data and ensure compliance with privacy regulations [92, 93].

Addressing these challenges and ethical considerations demands a multifaceted approach combining technological innovation with ethical oversight. By adopting comprehensive evaluation frameworks and establishing auditable fairness criteria, the NLP community can enhance the reliability, fairness, and applicability of NLP technologies across critical domains. Initiatives like ALLURE aim to improve LLM evaluation capabilities, reducing reliance on human annotators and ensuring equitable outcomes [88, 7].

8.4 Enhancements in Knowledge-Driven NLP

Knowledge-driven approaches have significantly enhanced NLP by integrating structured knowledge bases and advanced retrieval mechanisms to improve accuracy and contextual relevance. These methods leverage domain-specific knowledge to inform NLP systems, enabling precise and contextually appropriate responses. For example, in text-to-SQL systems, RAG integration enhances handling complex database queries by incorporating relevant external information [43].

Advanced benchmarks drive further enhancements in knowledge-driven NLP. By expanding these benchmarks to encompass a broader variety of query types, researchers can optimize Text-to-Action Generation (TAG) systems for greater efficiency [43]. This optimization is essential for improving scalability and adaptability in addressing diverse linguistic and contextual challenges.

Moreover, knowledge-driven approaches facilitate the integration of structured data, such as knowledge graphs and ontologies, into NLP systems, enhancing their capacity for complex reasoning tasks. Utilizing a RAG framework mitigates hallucinations while incorporating domain-specific insights and intermediate reasoning results, improving LLMs' overall reasoning capability. Advanced optimizations refine retrieval quality and numerical computation abilities, leading to substantial performance improvements and reduced error rates [3, 94, 95, 96]. Continuous refinement ensures NLP technologies remain robust and adaptable to evolving linguistic landscapes.

Advancements in knowledge-driven NLP underscore the transformative potential of integrating structured knowledge into language models. By enhancing accuracy, relevance, and efficiency, approaches like VITALITY 2 and automated literature review techniques streamline the literature review process, traditionally hindered by keyword limitations, enhancing users' ability to conduct complex queries and generate comprehensive reviews [2, 3].

9 Conclusion

9.1 Emerging Trends and Future Directions

The integration of Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs) is significantly advancing natural language processing (NLP) across multiple domains. Future research should focus on enhancing the robustness and retrieval quality of RAG systems to minimize computational overhead and broaden the applicability of Retrieval-Augmented Language Models (RALMs). Improving the interaction between Elasticsearch and LLMs, particularly through semantic understanding and context-awareness, offers a promising avenue for optimizing retrieval processes.

In the healthcare sector, models like Health-LLM demonstrate the potential for personalized applications, especially in disease prediction. The HealthPariksha benchmark emphasizes the need for tailored evaluation metrics and datasets that reflect real-world linguistic diversity. Future endeavors should aim at refining these benchmarks, integrating comprehensive clinical guidelines, and validating model performance in diverse healthcare environments.

Hybrid approaches remain vital, with RAG techniques proving effective in generating accurate responses for knowledge-based systems. In education, there is a need to explore alternative data sources and chunk sizes to improve answer accuracy and expand the application of these methodologies across various subjects. The potential of offline open-source LLMs to enhance flexibility and extend the applicability of LLM-Ref across research domains is another promising area for exploration.

In power systems, while LLMs show potential for improving operational efficiency, further refinement is needed to address their limitations in handling complex physical principles. Future studies could enhance hybrid models by incorporating additional contextual information or developing sophisticated scoring metrics to improve the retrieval of rare biomedical discoveries.

Exploring LLM applications in other BPM lifecycle tasks and refining prompting techniques to enhance output consistency and reliability are crucial for future research. The automotive sector also presents opportunities to expand system capabilities across sub-domains, optimize real-time performance, integrate multi-modal processing for visual elements, and address ethical considerations in data privacy.

These advancements underscore the necessity for ongoing exploration and optimization of RAG and LLM technologies, facilitating the development of more intelligent and effective solutions in the evolving technological landscape. Leveraging these insights will further enhance the capabilities and applicability of LLMs across diverse domains.

9.2 Future Directions and Research Opportunities

Future research in Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) presents numerous opportunities to enhance their capabilities and applicability across diverse domains. One area of focus is the expansion of datasets and the exploration of ensemble approaches to improve performance in legal decision support systems. This may involve developing robust LLM frameworks capable of managing knowledge conflicts, enhancing multilingual capabilities, and improving the interpretability of generated explanations.

In LLM inference, optimizing system-level solutions while addressing ethical implications and sustainability concerns is essential. Research should aim to refine frameworks for diverse model types and explore scalability in large-scale applications to ensure LLMs remain efficient and environmentally sustainable.

Investigating multi-perspective retrieval methodologies is another critical avenue for future research. Addressing limitations in these methodologies and expanding applications across knowledge-dense domains will enhance the adaptability of LLMs in complex information retrieval tasks.

Additionally, developing alternative learning paradigms, such as teacher-student or adversarial learning approaches, could facilitate collaboration between workflow generator and interpreter LLMs, leading to more efficient automated workflow generation systems.

In ontology learning tasks, future research should focus on enhancing LLMs through hybrid approaches and expanding evaluations to diverse knowledge domains, ensuring their versatility in adapting to various linguistic and contextual challenges.

Refining Knowledge Boundary Model (KBM) thresholds and exploring strategies to enhance adaptability across LLMs is another promising research direction, improving the accuracy and reliability of LLMs in managing domain-specific knowledge.

Moreover, strengthening the robustness of RAG-based unlearning frameworks against adversarial attacks and extending their applicability to multimodal LLMs and LLM-based agents is crucial to ensure the security and reliability of LLM applications.

These research opportunities highlight the importance of continuous exploration and optimization of LLM and RAG technologies, paving the way for more intelligent solutions in the evolving technological landscape. Leveraging these insights will further enhance the capabilities and applicability of LLMs across diverse domains.

References

[1] Ayman Asad Khan, Md Toufique Hasan, Kai Kristian Kemell, Jussi Rasku, and Pekka Abrahamsson. Developing retrieval augmented generation (rag) based llm systems from pdfs: An experience report, 2024.

[2] Nurshat Fateh Ali, Md. Mahdi Mohtasim, Shakil Mosharrof, and T. Gopi Krishna. Automated literature review using nlp techniques and llm-based retrieval-augmented generation, 2024.

[3] Hongye An, Arpit Narechania, Emily Wall, and Kai Xu. vitality 2: Reviewing academic literature using large language models, 2024.

[4] Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, and Lili Qiu. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely, 2024.

[5] Kazi Ahmed Asif Fuad and Lizhong Chen. Llm-ref: Enhancing reference handling in technical writing with large language models, 2024.

[6] Keheng Wang, Feiyu Duan, Peiguang Li, Sirui Wang, and Xunliang Cai. Llms know what they need: Leveraging a missing information guided framework to empower retrieval-augmented generation, 2024.

[7] Hosein Hasanbeig, Hiteshi Sharma, Leo Betthauser, Felipe Vieira Frujeri, and Ida Momennejad. Allure: Auditing and improving llm-based evaluation of text using iterative in-context-learning, 2023.

[8] Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. Investigating table-to-text generation capabilities of llms in real-world information seeking scenarios, 2023.

[9] Sule Tekkesinoglu and Lars Kunze. From feature importance to natural language explanations using llms with rag, 2024.

[10] Gabriel Lino Garcia, João Renato Ribeiro Manesco, Pedro Henrique Paiola, Lucas Miranda, Maria Paola de Salvo, and João Paulo Papa. A review on scientific knowledge extraction using large language models in biomedical sciences, 2024.

[11] Bryan Zhang, Taichi Nakatani, and Stephan Walter. Enhancing e-commerce product title translation with retrieval-augmented generation and large language models, 2024.

[12] Xinzhe Li. A review of prominent paradigms for llm-based agents: Tool use (including rag), planning, and feedback learning, 2024.

[13] Cheonsu Jeong. Beyond text: Implementing multimodal large language model-powered multi-agent systems using a no-code platform, 2025.

[14] Haiwei Dong and Shuang Xie. Large language models (llms): Deployment, tokenomics and sustainability, 2024.

[15] Marcos Fernández-Pichel, Juan C. Pichel, and David E. Losada. Search engines, llms or both? evaluating information seeking strategies for answering health questions, 2024.

[16] Malay Kumar, S. Sarvajit Visagan, Tanish Sarang Mahajan, and Anisha Natarajan. Enhanced sign language translation between american sign language (asl) and indian sign language (isl) using llms, 2024.

[17] Yucheng Hu and Yuxing Lu. Rag and rau: A survey on retrieval-augmented language model in natural language processing, 2024.

[18] Cheonsu Jeong. A study on the implementation of generative ai services using an enterprise data-based llm application architecture, 2023.

[19] Giorgio Roffo. Exploring advanced large language models with llmsuite, 2024.

[20] Akhil Vaid, Joshua Lampert, Juhee Lee, Ashwin Sawant, Donald Apakama, Ankit Sakhuja, Ali Soroush, Sarah Bick, Ethan Abbott, Hernando Gomez, Michael Hadley, Denise Lee, Isotta Landi, Son Q Duong, Nicole Bussola, Ismail Nabeel, Silke Muehlstedt, Silke Muehlstedt, Robert Freeman, Patricia Kovatch, Brendan Carr, Fei Wang, Benjamin Glicksberg, Edgar Argulian, Stamatios Lerakis, Rohan Khera, David L. Reich, Monica Kraft, Alexander Charney, and Girish Nadkarni. Natural language programming in medicine: Administering evidence based clinical workflows with autonomous agents powered by generative large language models, 2024.

[21] Demiao Lin. Revolutionizing retrieval-augmented generation with enhanced pdf structure recognition, 2024.

[22] Uday Allu, Biddwan Ahmed, and Vishesh Tripathi. Beyond extraction: Contextualising tabular data for efficient summarisation by language models, 2024.

[23] Cheonsu Jeong. Generative ai service implementation using llm application architecture: based on rag model and langchain framework. *Journal of Intelligence and Information Systems*, 29(4):129–164, 2023.

[24] Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Melika Emami, Fazlolah Mohaghegh, Mozhdeh Rouhsedaghat, and Kai-Wei Chang. Llms in biomedicine: A study on clinical named entity recognition, 2024.

[25] Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models, 2024.

[26] YuHe Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, and Daniel Shu Wei Ting. Development and testing of retrieval augmented generation in large language models – a case study report, 2024.

[27] Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. Retrieval-augmented generation for generative artificial intelligence in medicine, 2024.

[28] Alexander R. Pelletier, Joseph Ramirez, Irsyad Adam, Simha Sankar, Yu Yan, Ding Wang, Dylan Steinecke, Wei Wang, and Peipei Ping. Explainable biomedical hypothesis generation via retrieval augmented generation enabled large language models, 2024.

[29] Subir Majumder, Lin Dong, Fatemeh Doudi, Yuting Cai, Chao Tian, Dileep Kalathi, Kevin Ding, Anupam A. Thatte, Na Li, and Le Xie. Exploring the capabilities and limitations of large language models in the electric energy sector, 2024.

[30] Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. Llms4ol: Large language models for ontology learning, 2023.

[31] Aman Ahluwalia and Suhrud Wani. Leveraging large language models for web scraping, 2024.

[32] Baolin Li, Yankai Jiang, Vijay Gadepally, and Devesh Tiwari. Llm inference serving: Survey of recent advances and opportunities, 2024.

[33] Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A. Lenert. The emergence of large language models (llm) as a tool in literature reviews: an llm automated systematic review, 2024.

[34] Xin Zhang, Tianjie Ju, Huijia Liang, Ying Fu, and Qin Zhang. General llms as instructors for domain-specific llms: A sequential fusion method to integrate extraction and editing, 2024.

[35] Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. Searching for best practices in retrieval-augmented generation, 2024.

[36] Cheonsu Jeong. A study on the implementation method of an agent-based advanced rag system using graph, 2024.

[37] Brian J Chan, Chao-Ting Chen, Jui-Hung Cheng, and Hen-Hsen Huang. Don't do rag: When cache-augmented generation is all you need for knowledge tasks, 2025.

[38] Jeffery L. Painter, Venkateswara Rao Chalamalasetti, Raymond Kassekert, and Andrew Bate. Automating pharmacovigilance evidence generation: Using large language models to produce context-aware sql, 2024.

[39] Ali Mohammadjafari, Anthony S. Maida, and Raju Gottumukkala. From natural language to sql: Review of llm-based text-to-sql systems, 2025.

[40] Angelo Ziletti and Leonardo D'Ambrosi. Retrieval augmented text-to-sql generation for epidemiological question answering using electronic health records, 2024.

[41] Samy Ateia and Udo Kruschwitz. Can open-source llms compete with commercial models? exploring the few-shot performance of current gpt models in biomedical tasks, 2024.

[42] Fatiha Ait Kbir, Jérémy Bourgoin, Rémy Decoupes, Marie Gradeler, and Roberto Interdonato. Adaptations of ai models for querying the landmatrix database in natural language, 2024.

[43] Asim Biswal, Liana Patel, Siddarth Jha, Amog Kamsetty, Shu Liu, Joseph E. Gonzalez, Carlos Guestrin, and Matei Zaharia. Text2sql is not enough: Unifying ai and databases with tag, 2024.

[44] Mohammed Saeed, Nicola De Cao, and Paolo Papotti. Querying large language models with sql, 2023.

[45] S. S. Manathunga and Y. A. Illangasekara. Retrieval augmented generation and representative vector summarization for large unstructured textual data in medical education, 2023.

[46] Yen Sia Low, Michael L. Jackson, Rebecca J. Hyde, Robert E. Brown, Neil M. Sanghavi, Julian D. Baldwin, C. William Pike, Jananee Muralidharan, Gavin Hui, Natasha Alexander, Hadeel Hassan, Rahul V. Nene, Morgan Pike, Courtney J. Pokrzywa, Shivam Vedak, Adam Paul Yan, Dong han Yao, Amy R. Zipursky, Christina Dinh, Philip Ballentine, Dan C. Derieg, Vladimir Polony, Rehan N. Chawdry, Jordan Davies, Brigham B. Hyde, Nigam H. Shah, and Saurabh Gombar. Answering real-world clinical questions using large language model based systems, 2024.

[47] Kareem Shaik, Dali Wang, Weijian Zheng, Qinglei Cao, Heng Fan, Peter Schwartz, and Yunhe Feng. S3llm: Large-scale scientific software understanding with llms using source, metadata, and document, 2024.

[48] Xingcan Su and Yang Gu. Implementing retrieval-augmented generation (rag) for large language models to build confidence in traditional chinese medicine. 2024.

[49] Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *arXiv preprint arXiv:2310.18018*, 2023.

[50] Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*, 2024.

[51] Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. Medexpqa: Multilingual benchmarking of large language models for medical question answering, 2024.

[52] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine, 2024.

[53] Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. Medical mt5: An open-source multilingual text-to-text llm for the medical domain, 2024.

[54] Varun Gumma, Anandhita Raghunath, Mohit Jain, and Sunayana Sitaram. Health-pariksha: Assessing rag models for health chatbots in real-world multilingual settings, 2024.

[55] Zhongzhen Huang, Kui Xue, Yongqi Fan, Linjie Mu, Ruoyu Liu, Tong Ruan, Shaoting Zhang, and Xiaofan Zhang. Tool calling: Enhancing medication consultation via retrieval-augmented large language models, 2024.

[56] Mohamed Sobhi Jabal, Pranav Warman, Jikai Zhang, Kartikeye Gupta, Ayush Jain, Maciej Mazurowski, Walter Wiggins, Kirti Magudia, and Evan Calabrese. Language models and retrieval augmented generation for automated structured data extraction from diagnostic reports, 2024.

[57] Aidan Gilson, Xuguang Ai, Thilaka Arunachalam, Ziyou Chen, Ki Xiong Cheong, Amisha Dave, Cameron Duic, Mercy Kibe, Annette Kaminaka, Minali Prasad, Fares Siddig, Maxwell Singer, Wendy Wong, Qiao Jin, Tiarnan D. L. Keenan, Xia Hu, Emily Y. Chew, Zhiyong Lu, Hua Xu, Ron A. Adelman, Yih-Chung Tham, and Qingyu Chen. Enhancing large language models with domain-specific retrieval augment generation: A case study on long-form consumer health question answering in ophthalmology, 2024.

[58] Michael Grohs, Luka Abb, Nourhan Elsayed, and Jana-Rebecca Rehse. Large language models can accomplish business process management tasks, 2023.

[59] Linghao Zhang, Hongyi Zhang, Chong Wang, and Peng Liang. Rag-enhanced commit message generation, 2024.

[60] Jingyu Liu, Jiaen Lin, and Yong Liu. How much can rag help the reasoning of llm?, 2024.

[61] Jiarui Li, Ye Yuan, and Zehua Zhang. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases, 2024.

[62] Hang Yang, Jing Guo, Jianchuan Qi, Jinliang Xie, Si Zhang, Siqi Yang, Nan Li, and Ming Xu. A method for parsing and vectorization of semi-structured data used in retrieval augmented generation, 2024.

[63] Zhen Zhang, Xinyu Wang, Yong Jiang, Zhuo Chen, Feiteng Mu, Mengting Hu, Pengjun Xie, and Fei Huang. Exploring knowledge boundaries in large language models for retrieval judgment, 2024.

[64] Mufei Li, Siqi Miao, and Pan Li. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation, 2025.

[65] Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. Dragin: Dynamic retrieval augmented generation based on the information needs of large language models, 2024.

[66] Eric Melz. Enhancing llm intelligence with arm-rag: Auxiliary rationale memory for retrieval augmented generation, 2023.

[67] Timothy R. McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N. Halgamuge. Inadequacies of large language model benchmarks in the era of generative artificial intelligence, 2024.

[68] Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. Raft: Adapting language model to domain specific rag. In *First Conference on Language Modeling*, 2024.

[69] Xinzhe Li, Ming Liu, and Shang Gao. Grammar: Grounded and modular methodology for assessment of closed-domain retrieval-augmented language model, 2024.

[70] Spurthi Setty, Harsh Thakkar, Alyssa Lee, Eden Chung, and Natan Vidra. Improving retrieval for rag based question answering models on financial documents, 2024.

[71] Yunxiao Shi, Xing Zi, Zijing Shi, Haimin Zhang, Qiang Wu, and Min Xu. Enhancing retrieval and managing retrieval: A four-module synergy for improved quality and efficiency in rag systems, 2024.

[72] Spurthi Setty, Harsh Thakkar, Alyssa Lee, Eden Chung, and Natan Vidra. Improving retrieval for rag based question answering models on financial documents. *arXiv preprint arXiv:2404.07221*, 2024.

[73] Joel Suro. Semantic tokens in retrieval augmented generation, 2024.

[74] Zongyue Qin, Chen Luo, Zhengyang Wang, Haoming Jiang, and Yizhou Sun. Relational database augmented large language model, 2024.

[75] Fuheng Zhao, Lawrence Lim, Ishtiyaque Ahmad, Divyakant Agrawal, and Amr El Abbadi. Llm-sql-solver: Can llms determine sql equivalence?, 2024.

[76] Liang Zhang, Katherine Jijo, Spurthi Setty, Eden Chung, Fatima Javid, Natan Vidra, and Tommy Clifford. Enhancing large language model performance to answer questions and extract information more accurately, 2024.

[77] Rama Akkiraju, Anbang Xu, Deepak Bora, Tan Yu, Lu An, Vishal Seth, Aaditya Shukla, Pritam Gundecha, Hridhay Mehta, Ashwin Jha, Prithvi Raj, Abhinav Balasubramanian, Murali Maram, Guru Muthusamy, Shivakesh Reddy Annepally, Sidney Knowles, Min Du, Nick Burnett, Sean Javiya, Ashok Marannan, Mamta Kumari, Surbhi Jha, Ethan Dereszenski, Anupam Chakraborty, Subhash Ranjan, Amina Terfai, Anoop Surya, Tracey Mercer, Vinodh Kumar Thanigachalam, Tamar Bar, Sanjana Krishnan, Samy Kilaru, Jasmine Jaksic, Nave Algarici, Jacob Liberman, Joey Conway, Sonu Nayyar, and Justin Boitano. Facts about building retrieval augmented generation-based chatbots, 2024.

[78] Teng Wang, Zhenqi He, Wing-Yin Yu, Xiaojin Fu, and Xiongwei Han. Large language models are good multi-lingual learners : When llms meet cross-lingual prompts, 2024.

[79] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems, 2024.

[80] Hung Phan, Anurag Acharya, Rounak Meyur, Sarthak Chaturvedi, Shivam Sharma, Mike Parker, Dan Nally, Ali Jannesari, Karl Pazdernik, Mahantesh Halappanavar, Sai Munikoti, and Sameera Horawalavithana. Examining long-context large language models for environmental review document comprehension, 2024.

[81] Rag-f usion: A n ew t ake on r e.

[82] Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. Claim verification in the age of large language models: A survey, 2025.

[83] Jinge Wu, Zhaolong Wu, Ruizhe Li, Abul Hasan, Yunsoo Kim, Jason P. Y. Cheung, Teng Zhang, and Honghan Wu. Integrating knowledge retrieval and large language models for clinical report correction, 2024.

[84] Zhiyu An, Xianzhong Ding, Yen-Chun Fu, Cheng-Chung Chu, Yan Li, and Wan Du. Golden-retriever: High-fidelity agentic retrieval augmented generation for industrial knowledge base, 2024.

[85] Julien Delile, Srayanta Mukherjee, Anton Van Pamel, and Leonid Zhukov. Graph-based retriever captures the long tail of biomedical knowledge, 2024.

[86] Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation, 2024.

[87] Deepa Tilwani, Yash Saxena, Ali Mohammadi, Edward Raff, Amit Sheth, Srinivasan Parthasarathy, and Manas Gaur. Reasons: A benchmark for retrieval and automated citations of scientific sentences using public and proprietary llms, 2024.

[88] Vincent Freiberger and Erik Buchmann. Fairness certification for natural language processing and large language models, 2024.

[89] Toshihiro Kamiya. A rag method for source code inquiry tailored to long-context llms, 2024.

[90] Dehai Min, Nan Hu, Rihui Jin, Nuo Lin, Jiaoyan Chen, Yongrui Chen, Yu Li, Guilin Qi, Yun Li, Nijun Li, and Qianren Wang. Exploring the impact of table-to-text methods on augmenting llm-based question answering with domain hybrid data, 2024.

[91] Eric Yang, Jonathan Amar, Jong Ha Lee, Bhawesh Kumar, and Yugang Jia. The geometry of queries: Query-based innovations in retrieval-augmented generation, 2024.

[92] Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, et al. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). *arXiv preprint arXiv:2402.16893*, 2024.

[93] Dattaraj Rao. Bayesian inference to improve quality of retrieval augmented generation, 2024.

[94] Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation, 2024.

[95] H ow m uch c an rag h elp the r.

[96] Ye Yuan, Chengwu Liu, Jingyang Yuan, Gongbo Sun, Siqi Li, and Ming Zhang. A hybrid rag system with comprehensive enhancement on complex reasoning. *arXiv preprint arXiv:2408.05141*, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.