Large Language Models and Emergent Abilities in Natural Language Processing: A Survey

www.surveyx.cn

Abstract

This survey examines the transformative impact of Large Language Models (LLMs) on natural language processing (NLP), emphasizing their emergent abilities in language understanding, generation, and manipulation. LLMs, such as Llama 3, showcase superior performance in language style imitation, particularly when guided by advanced prompting methods like the Tree-of-Thoughts (ToT). The rapid growth in LLM research, including models like ChatGPT, reflects their potential across diverse domains, despite significant ethical concerns necessitating careful consideration. The integration of knowledge graphs into LLMs is identified as a promising approach to enhance knowledge modeling capabilities, addressing limitations in factual reasoning. Key takeaways include the importance of model scaling, training data quality, and fine-tuning strategies in enhancing LLM performance. The survey highlights the necessity for a new regulatory category for LLMs, emphasizing continuous monitoring and patient involvement in regulatory decisions. Additionally, LLM-based data augmentation methods generate high-quality synthetic data, surpassing traditional approaches, while tailored learning approaches enhance reasoning abilities in smaller language models. Practical considerations for LLM deployment emphasize task decomposition, retrieval-augmented generation, and structured tool management for successful agent deployment. In sentiment analysis, dynamic adaptive optimization modules improve accuracy and mean squared error metrics. In the medical field, LLMs are valuable for improving patient care and education, yet significant limitations remain before clinical application. Overall, this survey underscores LLMs' significance in advancing NLP, highlighting their emergent abilities and the challenges to fully realize their potential across diverse applications.

1 Introduction

1.1 Significance of LLMs in AI and NLP

Large Language Models (LLMs) have become pivotal in advancing artificial intelligence (AI) and natural language processing (NLP), outperforming traditional machine learning frameworks and significantly enhancing NLP technologies. Their ability to process vast textual datasets enables them to execute complex language tasks with high precision, as evidenced by their successful applications in automated systematic reviews, table-to-text generation, and extractive summarization. Notably, models like GPT-4 have demonstrated exceptional performance in data extraction and coherent summary generation, revolutionizing the efficiency of NLP applications in real-world contexts [1, 2, 3, 4].

In domains such as disaster response, LLMs provide actionable intelligence that enhances decision-making and response strategies [5]. Their impact extends to automatic speech recognition (ASR), where they improve performance by refining the understanding and prediction of linguistic patterns [6]. Furthermore, LLMs have showcased their versatility across various NLP tasks, emphasizing their transformative influence [2].

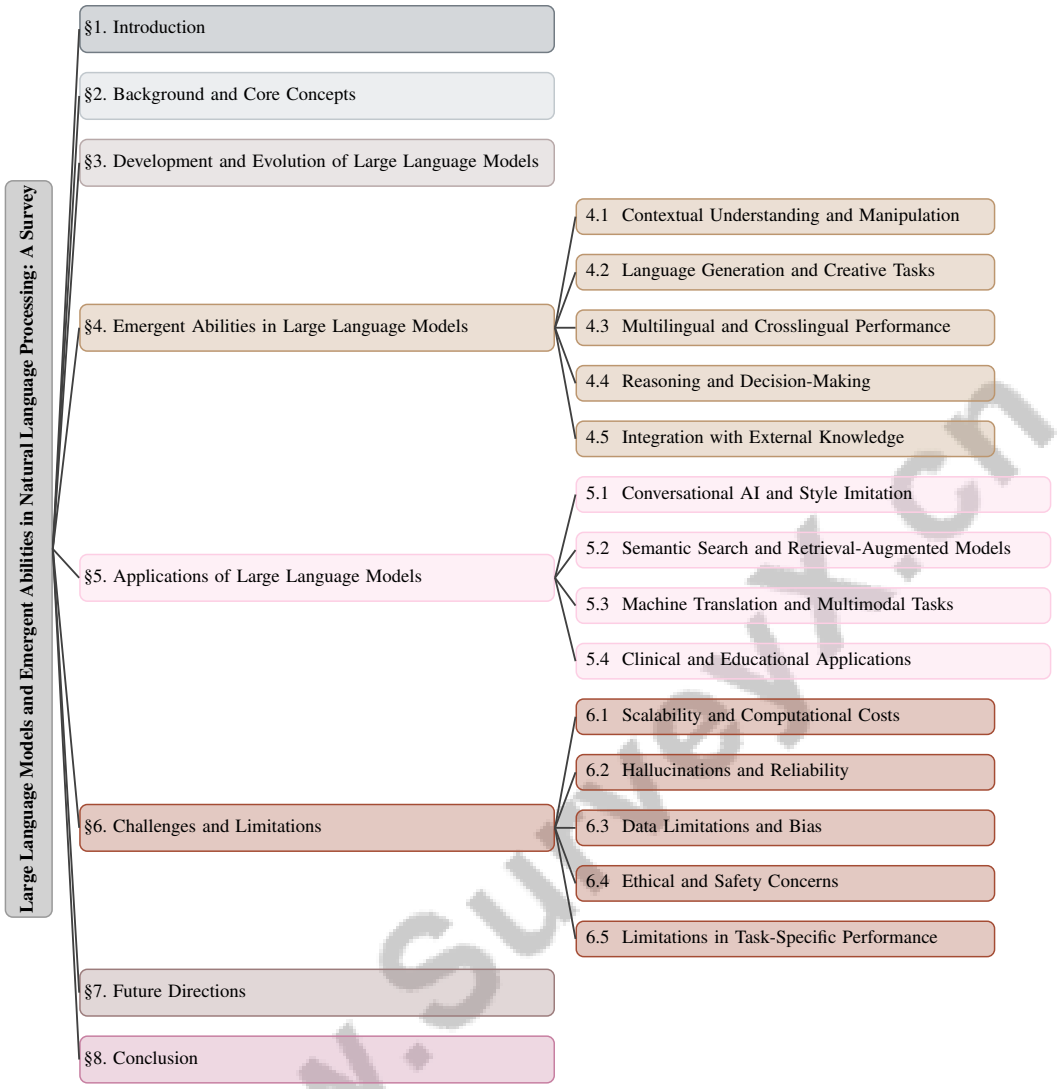


Figure 1: chapter structure

Despite their advantages, LLMs pose challenges, particularly regarding the potential for misuse in generating disinformation, as illustrated by their ability to create false news articles [7]. This underscores the need for a theoretical understanding of LLM behavior to accurately predict and explain their capabilities [8]. Additionally, LLMs have made significant contributions to computational social science, especially in data annotation, highlighting their broad applicability [9].

The integration of LLMs across various fields not only boosts workflow efficiency and adaptability but also addresses linguistic and accessibility gaps, thereby humanizing technology and promoting societal benefits [10]. Evaluating LLMs necessitates a multidisciplinary approach, incorporating insights from user experience research and human behavioral psychology to enhance the reliability of experimental designs and results [11]. As AI and NLP evolve, LLMs remain at the forefront, driving innovation and expanding the capabilities of these domains, paving the way for future advancements in AI-driven environments.

1.2 Emergent Abilities of LLMs

LLMs exhibit emergent abilities that significantly enhance their language understanding and generation capabilities, surpassing traditional models by effectively leveraging extensive datasets and computational resources [12]. These abilities are particularly apparent in complex tasks such as in-context learning, where LLMs dynamically adapt to new information without explicit retraining,

thereby refining their reasoning and decision-making processes [13]. Moreover, LLMs excel in chain-of-thought reasoning and instruction-following tasks, which are essential for tackling sophisticated linguistic challenges.

The emergent intelligence properties of LLMs are assessed through both quantitative and qualitative evaluations, revealing insights into their capabilities and limitations [14]. For instance, LLMs have shown proficiency in aligning auditory and textual information within ASR systems, addressing previous methodological limitations [6]. Additionally, models like Llama2-7B, GPT-3.5, and Falcon-7B have advanced keyword extraction methods by capturing contextual and semantic nuances, outperforming traditional techniques [15].

In text processing, frameworks such as EYEGLAXS utilize LLMs for extractive summarization, showcasing their emergent abilities in managing lengthy documents [2]. Despite their strengths, LLMs occasionally struggle with complex examples, prompting the development of innovative methods to enhance their performance [16]. Furthermore, integrating LLMs with human annotations improves statistical inference in research, demonstrating their capacity to yield valuable insights from complex datasets [9].

These emergent abilities underscore the transformative impact of LLMs on NLP, enabling more nuanced and sophisticated language tasks. As research progresses, a thorough examination of LLM capabilities is critical for unlocking their full potential across diverse applications, particularly in long document summarization, table-to-text generation, and predictive analytics. Innovative approaches like Extract-then-Evaluate improve summary evaluation by reducing computational costs and enhancing correlation with human assessments. LLMs also effectively translate expert knowledge into quantifiable features, thereby improving predictive analytics and decision-making accuracy. These advancements highlight the importance of ongoing exploration to optimize LLM utilization across various contexts [1, 17, 18].

1.3 Structure of the Survey

This survey is systematically organized to provide a comprehensive exploration of LLMs and their emergent abilities in NLP. It begins with an introduction that establishes the significance of LLMs in AI and NLP, emphasizing their transformative impact and emergent capabilities. Following the introduction, Section 2 provides an in-depth examination of the foundational principles of transformer models and essential concepts in NLP, highlighting the mechanisms that enable LLMs to perform a wide array of tasks. This section also discusses recent advancements in model architectures, training strategies, and the implications of data quality, which are critical for understanding LLM functionality and their practical applications across various NLP domains [17, 4, 19, 3, 20].

Section 3 traces the historical development and evolution of LLMs, highlighting pivotal milestones and architectural innovations that have shaped their advancement. This section also examines the emergence of specialized and multimodal LLMs, broadening the scope of their applications. Section 4 focuses on the specific emergent abilities of LLMs, such as contextual understanding, language generation, multilingual performance, reasoning, and integration with external knowledge, providing a detailed analysis of how these capabilities enhance language tasks.

The survey delves into the practical applications of LLMs in Section 5, thoroughly examining their roles in various domains such as conversational AI, where they enhance user interactions; semantic search, improving information retrieval accuracy; machine translation, facilitating cross-linguistic communication; and specialized fields like clinical settings, where they assist in patient data analysis, and educational environments, where they support personalized learning experiences [3, 4]. Section 6 addresses the challenges and limitations associated with LLMs, such as scalability, computational costs, and ethical considerations, offering insights into the complexities of deploying these models.

In Section 7, the survey presents potential future directions for LLM research and development, emphasizing strategies for enhancing emergent abilities, improving model efficiency, and expanding applications while addressing ethical and societal challenges. The concluding section summarizes the key points discussed, reinforcing the significance of LLMs and their emergent abilities in transforming NLP and setting the stage for future advancements in the field.The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Transformer Models and Their Role in LLMs

Transformer models form the foundation of Large Language Models (LLMs), utilizing self-attention mechanisms and feed-forward neural networks to significantly enhance natural language processing (NLP) capabilities. These models efficiently process large datasets, performing complex linguistic tasks with high accuracy [2]. A pivotal component of transformers is the Multi-Head Attention (MHA) mechanism, which enables parallel processing of input sequences, improving the model's ability to discern intricate word relationships.

Transformers are categorized into encoder-only, decoder-only, and encoder-decoder architectures. Encoder-only models, like BERT, excel in tasks requiring deep comprehension, such as text classification and semantic analysis, by extracting structured insights from unstructured data. Decoder-only models, exemplified by Llama 2, are optimized for generating coherent text, making them suitable for narrative construction and text generation. These models, with extensive parameter counts, have achieved state-of-the-art results in various NLP applications, including Rhetorical Structure Theory (RST) discourse parsing. Techniques like fine-tuning with QLoRA enhance Llama 2's adaptability across diverse parsing strategies, maintaining high performance across multiple benchmark datasets [21, 22]. Encoder-decoder models, such as T5, combine comprehension and generation capabilities, proving versatile for tasks like machine translation and summarization.

Recent transformer advancements address computational complexity and memory constraints. Innovations like Long-token-first Byte Pair Encoding (LBPE) prioritize long tokens during encoding, alleviating learning imbalances. The document-wise memory architecture (DWM) enhances LLMs by mapping document representations to specific memory entries, improving recall of document-specific information during text generation. This architecture uses a mechanism that softly masks irrelevant memories, prioritizing pertinent content retrieval. Experimental results show DWM significantly improves the accuracy of document-related content retrieval, leading to more coherent and contextually appropriate text generation [17, 23, 2, 24, 1].

Despite their advantages, transformer models face challenges related to quadratic memory and computational demands, limiting their ability to process extended contexts. Optimization strategies, such as grammar masking, ensure LLM outputs adhere to specified context-free grammar, enhancing text generation reliability. Research into quantized LLMs highlights their potential to optimize resource consumption while maintaining performance. Benchmarks assessing emergent abilities like in-context learning and chain-of-thought reasoning across varying quantization levels suggest that while 4-bit quantization retains these capabilities, 2-bit quantization significantly impairs performance, underscoring the importance of targeted strategies like model fine-tuning to mitigate accuracy loss in low-bit models [25, 26].

The continuous enhancement of transformer architectures is essential for expanding the functional scope of LLMs, particularly in efficiently processing and integrating multimodal data. Recent advancements in summarization frameworks, such as EYEGLAXS, leverage cutting-edge techniques to improve performance and adaptability in handling lengthy text documents [2, 20]. These developments underscore the pivotal role of transformer models in the evolution of AI and NLP technologies.

2.2 Foundational Concepts in Natural Language Processing

Natural Language Processing (NLP) includes foundational concepts critical to the functionality and advancement of Large Language Models (LLMs). Distributional semantics, which posits that words appearing in similar contexts tend to have similar meanings, enables LLMs to derive semantic understanding from extensive corpora, enhancing their proficiency in complex language tasks [27].

Prompt engineering is a vital technique that customizes LLM outputs to meet specific stylistic and contextual needs, improving adaptability and precision across various linguistic tasks [28]. Addressing the challenge of unseen intents in Knowledge-Based Question Answering (KBQA) systems emphasizes the necessity for LLMs to generalize across diverse representations and novel intents, enhancing flexibility and efficiency [29].

Named Entity Recognition (NER) is fundamental in NLP, enabling LLMs to identify and categorize entities within text, essential for extracting structured information from unstructured data [30]. In specific contexts, such as Chinese text processing, LLMs are utilized for error detection and correction, addressing challenges in Chinese Grammatical Error Correction (CGEC) and Chinese Spelling Check (CSC) [31].

Crosslingual capabilities are crucial for evaluating LLM performance in multilingual contexts, particularly in translating low-resource languages to high-resource languages, extending NLP technologies beyond predominantly English datasets, as evidenced in translations involving Southern Quechua and Spanish [32]. The comprehension of metaphors involves complex cognitive tasks requiring analogy and pragmatic reasoning, showcasing the advanced cognitive abilities of LLMs [33].

Analyzing public affairs documents is vital for promoting transparency, accountability, and informed decision-making, highlighting the societal impact of NLP technologies [34]. Traditional data augmentation methods have proven inadequate for specialized tasks, necessitating innovative approaches like LLM2LLM to enhance dataset quality and quantity [16]. Understanding biases and inaccuracies in LLM outputs is essential for refining performance and ensuring reliable results [9].

The foundational concepts of NLP are crucial for the ongoing advancement of LLM technologies, underpinning their ability to perform intricate language tasks with heightened precision and efficiency across diverse applications, including automated systematic reviews, table-to-text generation, and various knowledge-intensive tasks [1, 3, 4].

3 Development and Evolution of Large Language Models

3.1 Historical Milestones and Breakthroughs

The development of Large Language Models (LLMs) has been marked by key breakthroughs that have significantly advanced their capabilities in natural language processing (NLP). One major advancement is their integration into automatic speech recognition (ASR) systems, which has surpassed the limitations of n-grams and traditional neural network models, enhancing ASR through improved contextual understanding and language generation [6]. LLMs have also revolutionized extractive summarization, as demonstrated by the EYEGLAXS framework, which efficiently processes and distills long documents [2]. This showcases their potential in information retrieval and knowledge extraction.

Addressing the challenge of fine-tuning in low-data settings, the LLM2LLM framework offers innovative methods to enhance model performance with limited data, illustrating LLMs' adaptability to varied data environments [16]. These milestones underscore LLMs' transformative impact on NLP, exemplified by innovations in ASR, summarization, and adaptation techniques, which set new benchmarks in diverse tasks. Models like GPT-4 demonstrate versatility in table-to-text generation, further integrating LLMs into complex linguistic tasks and enhancing NLP application efficiency and accuracy [17, 4, 2, 3, 1]. As research advances, LLMs are expected to play a crucial role in the future of artificial intelligence.

3.2 Architectural Innovations and Model Evolution

Method Name	Architectural Focus	Integration Techniques	Application Enhancements
HoT[35]	Medical Reasoning Enhancement	Structured Thinking Process	Comprehensive Medical Responses
MRKL[36]	Flexible Architecture	Neuro-symbolic Architecture	Memory Capability Advancements
PORTLLM[37]	Efficiency Scalability Reasoning	Lightweight Model Patches	Fine-tuning Advancements
Path-LLaMA[38]	Reasoning Capabilities	Locally Trained LIms	Fine-tuned Tasks
AO[39]	Resource Utilization	Attention Offloading	Cost Efficiency
CDI[9]	Validity, Accuracy	Confidence-driven Inference	Annotation Process Optimization
LARM[40]	Reasoning Task-specific Performance	Neuro-symbolic Architectures	Embodied AI Autoregressive

Table 1: An overview of recent architectural innovations and integration techniques in large language models (LLMs), highlighting their focus areas and application enhancements. The table summarizes key methods, detailing their architectural focus, integration techniques, and the specific application enhancements they contribute to the field of LLMs.

The evolution of LLM architectures has focused on improving efficiency, scalability, and task-specific performance. Current models struggle with generating accurate medical responses due to limited

reasoning capabilities, prompting the development of architectures that incorporate comprehensive reasoning [35]. The MRKL method, integrating neuro-symbolic architectures with discrete knowledge modules, exemplifies advancements in processing and reasoning [36].

As illustrated in Figure 2, the key architectural innovations and model evolution in large language models highlight critical areas such as efficiency and scalability, reasoning and processing, and task-specific performance improvements. Each category showcases specific methods and frameworks contributing to advancements in LLM capabilities. Table 1 provides a comprehensive summary of the architectural innovations and model evolution in large language models, illustrating the various methods, their architectural focus, integration techniques, and the resulting application enhancements.

Fine-tuning pretrained models for specific tasks remains inefficient, especially with frequent model updates [37]. Local LLMs for fine-tuning generative instructions offer solutions for complex, multi-label tasks [38]. Additionally, separating memory-intensive from computation-intensive tasks enhances resource utilization and cost efficiency compared to homogeneous systems [39].

Advancements in memory capabilities draw parallels with human memory [41]. The use of geometric features for toxicity detection surpasses direct prompt engineering [42]. The CONFIDENCE-DRIVEN INFERENCE method exemplifies innovations in data annotation accuracy [9].

The Large-scale Autoregressive Model (LARM) predicts subsequent skills from textual and visual inputs, expanding embodied AI capabilities [40]. Collectively, these architectural innovations enhance LLM potential across applications, improving complex task precision. As research continues, these models are poised to integrate advanced reasoning and memory capabilities, solidifying their future role in artificial intelligence.

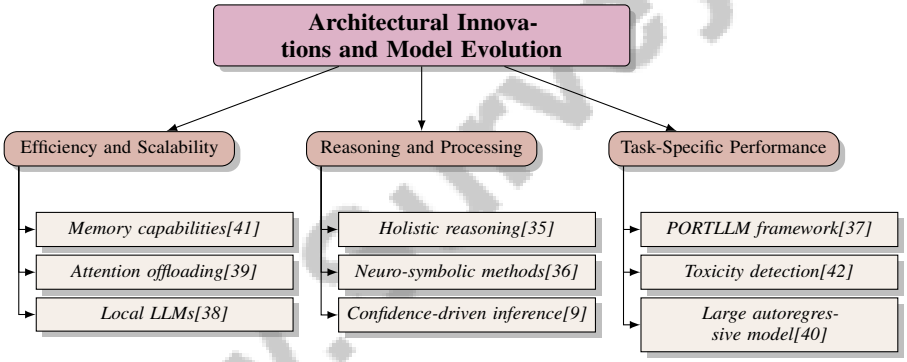


Figure 2: This figure illustrates the key architectural innovations and model evolution in large language models, highlighting efficiency and scalability, reasoning and processing, and task-specific performance improvements. Each category showcases specific methods and frameworks contributing to advancements in LLM capabilities.

3.3 Innovations in Specialized and Multimodal LLMs

Recent advancements in specialized and multimodal LLMs have significantly expanded their capabilities, enabling them to tackle complex tasks across various domains. The Tree-of-Thoughts (ToT) framework enhances language style imitation, allowing LLMs to replicate intricate linguistic patterns more effectively [43]. In translation, LLMs with a 2048-token context window demonstrate proficiency in handling complex linguistic nuances across languages [44].

Multimodal LLM innovations, such as the SEED-LLaMA model, integrate visual and textual data, improving performance by merging inputs effectively [21]. The AdaptVision model optimizes visual data processing through dynamic image partitioning [45]. The MindMerger framework enhances linguistic proficiency by embedding external language understanding capabilities [46]. The LLaMA-Rider method promotes adaptive learning through exploration and feedback [47].

In dental diagnostics, multimodal capabilities integrate diverse data sources, enhancing data synthesis and interpretation [48]. The late-fusion approach facilitates robust multimodal system development without full pretraining [49]. Evaluation benchmarks, like those from the Math Stack Exchange dataset, illustrate LLM advancements in addressing complex mathematical queries and enhancing

problem-solving skills [50]. Novel vocabulary substitution methods, such as Unigram and BPE tokenization, demonstrate ongoing refinement of tokenization techniques [51].

Advancements in specialized and multimodal LLMs showcase substantial progress in enhancing functionality and versatility. These innovations redefine roles across sectors, particularly in healthcare, integrating diverse data types for improved clinical decision-making and patient engagement while addressing data limitations and ethical considerations. As these models evolve, they promise to refine capabilities and broaden applications, transforming artificial intelligence utilization in diverse fields [52, 4, 2, 53].

4 Emergent Abilities in Large Language Models

The capabilities of Large Language Models (LLMs) are shaped by both their architectural design and their adeptness at navigating complex contextual frameworks. This section delves into the emergent abilities of LLMs, focusing on their proficiency in contextual understanding, language generation, multilingual performance, reasoning, and integration with external knowledge, which collectively advance language generation, reasoning, and decision-making.

4.1 Contextual Understanding and Manipulation

LLMs exhibit sophisticated contextual understanding and manipulation, essential for executing complex language tasks across various domains. Through advanced in-context learning mechanisms, they surpass traditional methods in tasks like named entity recognition and keyword extraction [15], with their ability to recall learned information based on input cues [41]. Practical applications, such as ChartGPT, demonstrate contextual understanding by selecting relevant columns based on user input and filtering rows according to intent [11]. The LARM model enhances decision-making by generating single tokens for actions, reducing ambiguity [40]. However, challenges in functional competence remain, indicating limitations in real-world contextual understanding [54]. Tool learning integration shows promise in expanding LLM capabilities and accuracy [55].

In automatic speech recognition (ASR), LLMs align speech and text, leveraging both modalities to enhance recognition accuracy [6]. Spanish LLMs, combined with classifiers, effectively manage multi-label datasets, showcasing enhanced contextual manipulation [34]. The transformative potential of LLMs in contextual information manipulation is evident in tasks like table-to-text generation and query-based generation, underscoring their adaptability in real-world information-seeking scenarios [56, 8, 3, 57, 1].

4.2 Language Generation and Creative Tasks

LLMs excel in language generation and creative tasks, using advanced prompting techniques to emulate specific linguistic styles and produce contextually relevant content [43]. Despite advancements, LLMs often struggle with genuine novelty, producing outputs that may lack unexpected creativity [58]. Models like GPT-4 demonstrate superior narrative generation capabilities, effectively utilizing contextual cues to create coherent narratives [1]. Techniques such as EVAPORATE-CODE+ enhance creative language tasks by generating multiple candidate functions and employing weak supervision for output ensembling [59]. In machine translation, refined editing techniques have reduced language mismatch and repetition errors, enhancing quality [60]. Segmentation methods improve outcomes in long-form spoken language by effectively segmenting transcripts [61].

LLMs produce summaries rivaling those by experts in fields like medicine, highlighting their potential in accuracy-critical domains [62]. In dentistry, they generate personalized treatment plans based on multimodal data analysis [48]. The Ophtha-LLaMA2 model exemplifies multimodal data integration with high diagnostic accuracy [63]. ChartGPT enhances visualization accuracy through a structured reasoning pipeline [11]. These innovations illustrate the expansive potential of LLMs in language generation and creative tasks, paving the way for further advancements.

4.3 Multilingual and Crosslingual Performance

LLMs show varying proficiency in multilingual and crosslingual tasks, facing challenges in low-resource languages and non-Latin scripts [64]. Despite these challenges, LLMs trained on parallel

data exhibit competitive machine translation performance, particularly with larger vocabulary sizes [44]. In the medical domain, new evaluation datasets for multilingual tasks emphasize the need for tailored benchmarks [65]. Many modern LLMs demonstrate surface-level crosslingual capabilities, struggling with deeper knowledge transfer [66]. The adaptMLLM method showcases advancements in multilingual capabilities, focusing on low-resource language pairs like English-Irish and English-Marathi [67]. Imbalanced training data and insufficient multilingual capabilities hinder effectiveness in low-resource languages [68].

To illustrate the hierarchical structure of multilingual and crosslingual performance in LLMs, Figure 3 categorizes the challenges and solutions, model advancements, and domain-specific applications. The challenges include low-resource language support, knowledge transfer issues, and tokenization strategies. Model advancements highlight innovations such as Dynamic Learning Strategies, the PLUME model, and the SEED Tokenizer. Additionally, domain-specific applications focus on Medical mT5, RoLlama for Romanian, and adaptMLLM in crisis situations, showcasing the diverse applications of LLMs across different fields.

Challenges in knowledge transfer complicate multilingual applications, leading to issues like catastrophic forgetting and insufficient domain adaptation [68]. Specialized models like RoLlama have shown improved performance in Romanian tasks, generating more contextually relevant responses than general LLMs [69]. Benchmarks for vocabulary adaptation, such as those for Russian tasks, provide insights into optimizing LLM performance in specific linguistic contexts [51]. The SEED tokenizer represents a significant advancement in bridging gaps in multimodal LLMs, enhancing their ability to process diverse data effectively [21]. Addressing these challenges is crucial for maximizing the multilingual potential of LLMs across a broader range of languages and domains.

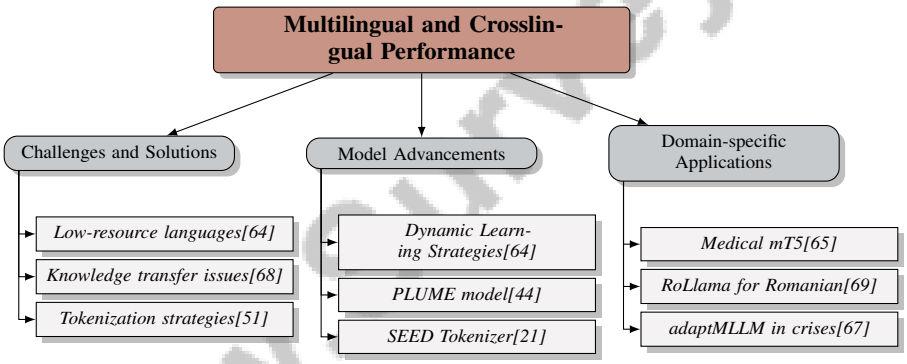


Figure 3: This figure illustrates the hierarchical structure of multilingual and crosslingual performance in large language models (LLMs), categorizing the challenges and solutions, model advancements, and domain-specific applications. The challenges include low-resource language support, knowledge transfer issues, and tokenization strategies. Model advancements highlight innovations like Dynamic Learning Strategies, the PLUME model, and the SEED Tokenizer. Domain-specific applications focus on Medical mT5, RoLlama for Romanian, and adaptMLLM in crisis situations, showcasing the diverse applications of LLMs across different fields.

4.4 Reasoning and Decision-Making

LLMs demonstrate substantial capabilities in reasoning and decision-making tasks, yet face challenges necessitating ongoing research and refinement. The LLM2LLM framework highlights advancements in reasoning by focusing on challenging examples, enhancing decision-making accuracy through targeted learning strategies [16]. This approach shows the potential for LLMs to improve reasoning abilities through strategic learning. LLMs leverage human annotations to enhance reasoning processes, selecting annotations based on confidence scores to improve output validity [9]. Despite advancements, LLMs struggle with generating accurate logical forms, especially in understanding formal languages, posing obstacles in decision-making processes. While state-of-the-art LLMs can comprehend formal languages comparably to humans, their ability to produce precise logical representations remains inconsistent [4, 3, 17, 70]. Improved methodologies are needed to enhance logical reasoning capabilities.

Advancements in LLM reasoning and decision-making are vital for effective deployment across fields like scientific research automation, predictive analytics, and education. These improvements enhance their ability to process complex information, integrate domain-specific knowledge, and deliver accurate insights [71, 72, 4, 18, 1]. As research progresses, refining these capabilities will ensure that LLMs effectively support informed decision-making across diverse contexts.

4.5 Integration with External Knowledge

LLMs enhance performance in various applications through the integration of external knowledge, improving their handling of knowledge-intensive tasks. This integration is achieved through methodologies like Knowledge Graph-based Pre-trained Language Models (KGPLMs), enhancing language understanding and generating factually accurate content. The incorporation of multimodal data in Multimodal Large Language Models (MLLMs) exemplifies this potential, transforming healthcare diagnostics and personalized treatment through diverse data synthesis [73].

To optimize LLM performance, relevance metrics and robustness validation identify suitable proxy tasks during early evaluations, enhancing prediction accuracy. Empirical evidence suggests that LLMs experience sudden performance improvements at critical model size and training compute thresholds, underscoring the importance of scaling [41]. Explainable AI developments address LLM opacity, facilitating external knowledge integration and enhancing transparency. Experiments indicate significant performance variations based on quantization strategies and calibration data, highlighting the need for tailored approaches [42]. Domain-specific training methods, particularly in local LLMs, enhance understanding and classification of complex language in medical contexts [74].

Anecdotal evidence illustrates how LLMs can overcome mechanizing bottlenecks faced by various user groups, emphasizing adaptability and the role of external knowledge in enhancing user experiences. The LUNA framework provides a structured approach to evaluate LLMs from multiple quality perspectives, offering deeper insights into performance and integration strategies [75].

Integration of external knowledge into LLMs significantly enhances functionality and reliability across fields. This integration improves models' ability to generate contextually relevant and accurate outputs, expanding applicability in real-world scenarios like table-to-text generation and systematic reviews. Advanced LLMs like GPT-4 excel in tasks requiring data insight generation and evaluation, while methodologies like Retrieval-Augmented Generation (RAG) and Knowledge Graph integration bolster performance by providing factual context and addressing domain-specific challenges. Such advancements ensure that LLMs meet user expectations and adapt effectively to diverse information-seeking needs [71, 17, 4, 3, 1]. Continued exploration of these integration strategies is essential for maximizing LLM potential across various applications.

5 Applications of Large Language Models

The evolution of Large Language Models (LLMs) is notably transformative in conversational AI and style imitation, significantly enhancing user interaction and engagement. This section examines how LLMs revolutionize the generation of contextually and stylistically appropriate responses, optimizing conversational dynamics through innovative frameworks. Figure 4 illustrates the diverse applications of LLMs across multiple domains, including Conversational AI, Semantic Search, Machine Translation, and Clinical and Educational settings. This figure highlights key frameworks, techniques, and innovations that enhance user interaction, optimize computational efficiency, and improve decision-making processes, thereby providing a comprehensive overview of the impact of LLMs in various fields.

5.1 Conversational AI and Style Imitation

LLMs have revolutionized conversational AI by generating contextually relevant and stylistically diverse responses. Integrating models like LLMCRS has enhanced recommendation quality and user interaction, showcasing LLMs' potential for more engaging conversations [76]. AutoFlow exemplifies the automation of workflow generation, facilitating complex style imitation and task execution [77].

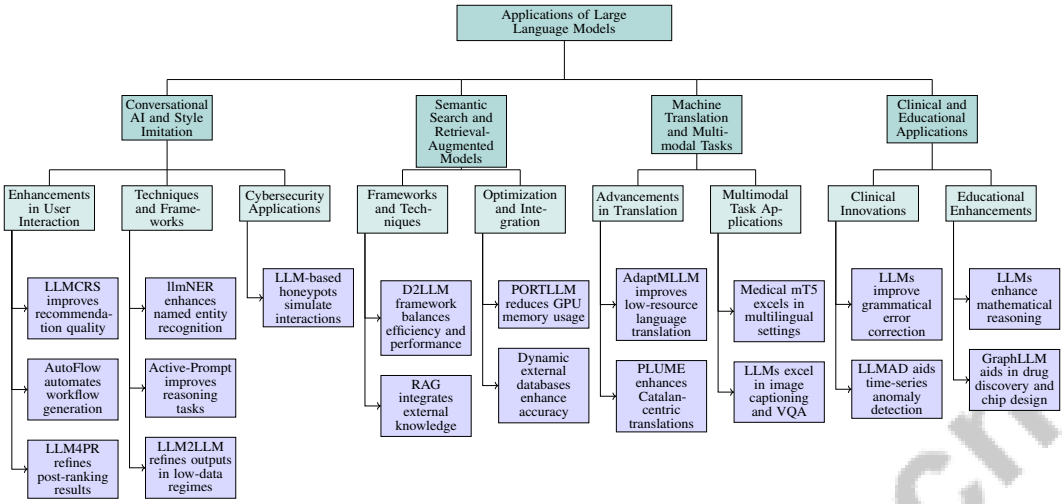


Figure 4: This figure illustrates the diverse applications of Large Language Models (LLMs) across multiple domains, including Conversational AI, Semantic Search, Machine Translation, and Clinical and Educational settings. It highlights key frameworks, techniques, and innovations that enhance user interaction, optimize computational efficiency, and improve decision-making processes.

In optimizing conversational outputs, LLM4PR refines post-ranking results through user and item representations, leveraging contextual understanding [78]. The llmNER library streamlines named entity recognition, enhancing conversational AI efficiency through improved prompting and parsing [30].

Techniques like Active-Prompt enhance LLM performance in reasoning tasks, demonstrating adaptability in generating stylistically varied responses [79]. The LLM2LLM framework improves performance in low-data regimes by refining stylistic and contextual outputs [16].

In cybersecurity, LLM-based honeypots simulate realistic interactions to engage attackers, offering insights into attacker behavior and data collection [80]. These advancements highlight LLMs' transformative potential in enhancing communication platforms across various natural language processing tasks [3, 4].

Figure 5 illustrates the advancements in conversational AI and style imitation, focusing on enhanced user interactions, optimized outputs, and security adaptations using large language models. This visual representation underscores the significant impact of these technologies on user engagement and the overall efficacy of conversational systems.

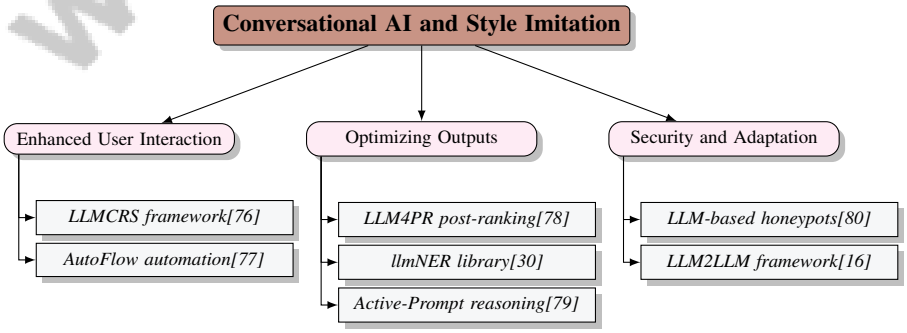


Figure 5: This figure illustrates the advancements in conversational AI and style imitation, focusing on enhanced user interactions, optimized outputs, and security adaptations using large language models.

5.2 Semantic Search and Retrieval-Augmented Models

LLMs have significantly enhanced semantic search by improving query-passages interaction, as shown by the D2LLM framework, which balances efficiency and performance [81]. Retrieval-Augmented Generation (RAG) addresses LLM limitations in accuracy and credibility, integrating external knowledge to enhance generative capabilities and reliability [24].

The PORTLLM framework optimizes computational efficiency, reducing GPU memory usage, crucial for deploying LLMs in retrieval-augmented models [37]. Integrating semantic search with retrieval-augmented models advances LLM applications, improving semantic understanding and information retrieval across domains. Dynamic external databases enhance accuracy and user experience by providing relevant information [75, 24].

5.3 Machine Translation and Multimodal Tasks

LLMs have advanced machine translation, particularly in low-resource languages, with methods like adaptMLLM improving translation quality in crisis contexts [67]. PLUME, a Catalan-centric multilingual LLM ensemble, demonstrates enhanced translation capabilities in specific linguistic contexts [44].

In the medical domain, the Medical mT5 model excels in multilingual settings, outperforming baselines in multi-task evaluations [65]. Fine-tuned LLMs leverage translation memories, optimizing translation workflows [82].

LLMs also excel in multimodal tasks like image captioning and VQA, integrating visual and textual data to enhance performance [45]. Ongoing research focuses on refining LLM applications in machine translation and multimodal tasks, enhancing communication across linguistic and modality barriers [53, 4].

5.4 Clinical and Educational Applications

LLMs impact clinical and educational settings by enhancing language processing, reasoning, and decision-making. In clinical environments, LLMs improve grammatical error correction in low-resource languages, aiding communication in healthcare [83]. Models like LLMAD aid time-series anomaly detection, crucial for patient care [84].

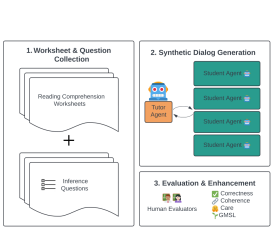
In education, LLMs enhance mathematical reasoning, with benchmarks revealing strengths and weaknesses [50]. Structured reasoning tasks make complex concepts accessible in resource-constrained settings [85].

Integrating expert intuition into predictive analytics enhances decision-making in clinical and educational domains [18]. GraphLLM's applications in drug discovery and chip design highlight graph reasoning's importance [86].

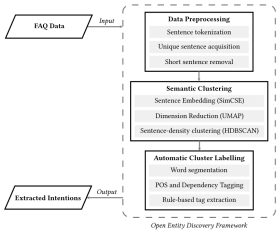
LLMs streamline legal analyses and compliance in regulatory contexts, improving administrative efficiency in clinical and educational settings [87]. In disaster response, LLMs provide timely, accurate information for emergency management [5].

These applications underscore LLMs' transformative potential in clinical and educational settings, offering innovative solutions for improved outcomes. As research progresses, LLMs are poised to enhance educational methodologies and clinical decision-making processes, playing a pivotal role in shaping future landscapes in these fields [53, 18, 4].

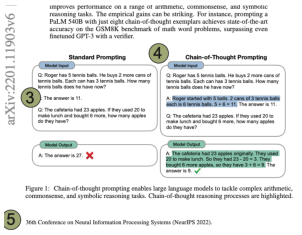
As shown in Figure 6, in artificial intelligence, LLMs demonstrate significant potential across various domains, particularly in clinical and educational settings. The "Synthetic Tutoring System for Reading Comprehension" exemplifies LLMs' role in personalized learning through structured approaches. The "Open Entity Discovery Framework" highlights innovative knowledge extraction for enhanced data retrieval. "Chain-of-Thought Prompting" showcases advanced techniques improving LLM performance in complex reasoning tasks. These examples illustrate LLMs' cutting-edge applications in creating intelligent, adaptive systems that revolutionize educational methodologies and clinical practices [88, 71, 89].



(a) Synthetic Tutoring System for Reading Comprehension[88]



(b) Open Entity Discovery Framework[71]



(c) Chain-of-Thought Prompting Enhances Large Language Models for Complex Arithmetic, Commonsense, and Symbolic Reasoning Tasks[89]

Figure 6: Examples of Clinical and Educational Applications

6 Challenges and Limitations

Large Language Models (LLMs) face several challenges that affect their deployment, performance, and accessibility. These challenges, including scalability, computational costs, hallucinations, data limitations, biases, ethical concerns, and task-specific performance, have significant implications for their real-world applications.

6.1 Scalability and Computational Costs

LLMs require significant computational resources, particularly for tasks like extractive summarization, which is challenging in low-data scenarios. Approaches such as LLM2LLM offer potential solutions [2, 16]. The high computational demands during training and inference, especially during token prediction, often exceed current technologies [39]. High inference costs and real-time validation needs further complicate scalability [90]. Model explosion, where task-specific fine-tuning degrades performance on other tasks, and the limitations of synthetic training datasets present additional challenges [36, 91]. Solutions like local LLMs and modular architectures could optimize resources, although their applicability is domain-specific [38]. Addressing these challenges is crucial for maximizing LLM potential through more efficient training strategies.

6.2 Hallucinations and Reliability

Hallucinations, or the generation of incorrect content, undermine LLM reliability due to training data biases and lack of updates [90]. Models like ChatGPT exhibit higher error rates than alternatives such as Bard [92]. Inconsistent functional competence requires specialized fine-tuning, especially in quantized models [54, 26]. Synthetic benchmarks may not fully capture real-world complexities, affecting optimization [93]. Biases in recommendation systems also pose reliability issues [94]. Mitigation efforts include frameworks like LUNA for distinguishing normal from abnormal behaviors [95], though LLMs often underperform compared to simpler methods [75]. Robust evaluation metrics and real-time updates are essential to enhance LLM accuracy and reliability.

6.3 Data Limitations and Bias

Data limitations and biases significantly impact LLM performance, particularly in low-resource languages and specialized domains [9, 51]. Traditional benchmarks focus on transformer models, limiting understanding of scalability and applicability in linear complexity models [34]. Biases in training data can perpetuate misinformation, complicating ethical alignment [42]. In clinical settings, data limitations pose challenges for Multimodal Large Language Models (MLLMs) [18]. Addressing these issues requires diversifying datasets, developing robust benchmarks, and implementing ethical guidelines. Strategies like iterative data augmentation can improve LLM accuracy and relevance [1, 16, 96, 17].

6.4 Ethical and Safety Concerns

LLM deployment raises ethical and safety concerns, including inconsistent safety filters and potential biases in decision-making [7, 10]. The opacity of LLM decisions complicates accountability, especially in sensitive areas like healthcare. Ethical implications extend to aligning LLM outputs with human values and managing job displacement risks [10]. In recommendation systems, transparency, fairness, and privacy are crucial. Addressing these concerns requires a multidisciplinary approach and robust ethical frameworks. Despite advancements, questions about LLM robustness and ethical alignment persist, necessitating tailored frameworks for accountability and transparency [72, 9].

6.5 Limitations in Task-Specific Performance

LLMs face task-specific performance limitations, often due to narrow study focuses and synthetic datasets that fail to capture real-world complexities [43, 88]. Synthetic data biases affect detector performance [97]. High-resolution document processing methods are limited by input resolution capabilities [45]. LLMs show low performance in specific tracks, such as Bio-ML, indicating a need for domain-specific models [98]. Overfitting and poor generalization complicate performance assurance on small datasets [82]. Benchmarks often overlook real-world task complexities, affecting applicability [99]. In legal tasks, data sufficiency for all concepts is a challenge [87]. Addressing these challenges requires ongoing research to optimize LLMs for specific tasks and improve their effectiveness across applications.

7 Future Directions

7.1 Enhancing Emergent Abilities

To enhance the emergent abilities of Large Language Models (LLMs), a comprehensive approach is essential, focusing on adaptability, reasoning, and external knowledge integration. Future research should aim to develop robust evaluation frameworks and improve tool accessibility to better capture user intent and advance multimodal tool learning [55]. Refining projector modules and exploring alternative training strategies can significantly enhance LLM capabilities in automatic speech recognition (ASR), thereby expanding their emergent abilities [6]. Incorporating sliding attention mechanisms and graph-based methods may improve the model's capacity for complex summarization tasks [2]. Additionally, optimizing hyperparameters and employing LLM2LLM techniques, such as prompt tuning, could enhance performance on larger datasets [16].

Addressing challenges like hallucinations and improving keyword extraction requires integrating human expertise into the LLM process, which can refine outputs and enhance reliability [15]. Calibrating LLM confidence scores is also crucial for methods like CONFIDENCE-DRIVEN INFERENCE, significantly improving decision-making processes [9]. A collaborative effort in integrating diverse strategies and ongoing research initiatives is vital for enhancing LLM capabilities across various applications and sectors, driving advancements in understanding, generating, and manipulating human language. By addressing computational efficiency, data biases, and refining evaluation methodologies, these strategies will ensure LLMs remain at the forefront of natural language processing innovations [1, 3, 17, 4].

7.2 Improving Model Efficiency

Enhancing the efficiency of Large Language Models (LLMs) is crucial for optimizing computational resources while maintaining or improving performance. Future research should focus on advancing parsing accuracy through more efficient algorithms and methodologies, thereby reducing the computational burden associated with training and inference [22]. Emerging trends emphasize developing better alignment techniques that enhance the model's ability to integrate seamlessly with external tools and databases, which is essential for improved performance across various applications [20]. Novel strategies to enhance model efficiency may include refining hyperparameter optimization, leveraging quantization techniques to reduce model size, and implementing modular architectures that facilitate targeted computation. By focusing on optimizing document length and employing effective sentence extraction techniques, researchers can address scalability challenges and high computational costs associated with LLMs. This approach enhances LLM efficiency in processing long documents,

addressing issues like the "Lost-in-the-Middle" problem, and ensuring practicality for deployment in resource-constrained environments [1, 3, 17]. As research progresses, these efforts will maximize the efficiency and applicability of LLMs across diverse domains, facilitating their integration into complex systems and expanding their utility.

7.3 Expanding Applications

The potential for expanding LLM applications is vast, driven by their proficiency in performing complex language tasks. In disaster response, LLMs can enhance decision-making by providing timely and actionable intelligence during emergencies [5], underscoring their transformative impact in critical situations. In healthcare, LLMs have demonstrated potential in improving diagnostic accuracy and personalizing treatment plans by integrating multimodal data, such as textual and visual inputs, for comprehensive analyses [48]. This innovative approach highlights LLMs' capacity to revolutionize medical diagnostics and patient care. Educational applications also significantly benefit from LLM advancements. By facilitating complex mathematical reasoning and providing tailored educational content, LLMs can enhance learning experiences across diverse educational settings [50]. Furthermore, integrating LLMs into legal and regulatory frameworks can streamline compliance checking and legal analyses, offering significant efficiency improvements in administrative functions [87]. This versatility illustrates LLMs' capability to handle specialized and technical domains. As LLM research advances, exploring diverse application domains—such as predictive analytics, literature reviews, and information extraction from tabular data—is essential to harness their capabilities, enhance decision-making accuracy, automate systematic reviews, and improve user efficiency in real-world scenarios [1, 18, 4]. Addressing current limitations and enhancing capabilities will position LLMs to drive innovation across various fields, from emergency management to education and beyond.

7.4 Addressing Ethical and Societal Challenges

The development and deployment of LLMs require a comprehensive approach to addressing ethical and societal challenges, crucial for ensuring responsible integration into various domains. A key focus area is integrating knowledge graphs (KGs) into LLMs, enhancing ethical robustness by providing structured and contextual information. However, significant research gaps remain in developing robust evaluation methodologies that encompass multilingual and multitask approaches [100]. To further enhance ethical robustness, future research should explore advanced techniques for modifying model weights, ensuring LLMs align with ethical standards and societal values [101]. Continuous ethical scrutiny is essential, and frameworks must adapt to keep pace with rapid technological advancements in LLMs, bridging identified gaps in ethics and ensuring responsible deployment [72]. Refining intervention techniques, such as the SARA framework, is crucial for guiding the moral compass of LLMs, especially in sensitive areas with profound ethical implications [102]. Additionally, developing tailored transparency approaches for diverse stakeholders can enhance communication methods and facilitate the exploration of regulatory frameworks that improve LLM transparency [103]. The long-term implications of LLM-generated disinformation pose significant societal risks, necessitating research focused on enhancing detection mechanisms to effectively counteract disinformation [7]. Addressing these ethical and societal challenges is paramount for maximizing the positive impact of LLMs while minimizing potential harms, ensuring deployment aligns with societal values and ethical standards.

7.5 Interdisciplinary Collaboration

Interdisciplinary collaboration is pivotal in advancing LLM technology, addressing the multifaceted challenges in safety, alignment, and ethical deployment. Integrating diverse expertise from computer science, linguistics, ethics, and social sciences is crucial for developing robust frameworks that ensure responsible LLM use. The complexities of LLM safety and alignment necessitate collaborative efforts to devise strategies that mitigate risks and enhance model reliability [104]. Fostering interdisciplinary partnerships allows researchers to integrate insights from ethics, predictive analytics, information extraction, and natural language processing to effectively tackle the unique foundational challenges posed by LLMs, including accountability, bias, and data interpretation complexities [105, 1, 72, 18]. This collaborative approach facilitates developing comprehensive evaluation metrics and ethical guidelines that align with societal values. Moreover, interdisciplinary collaboration enables exploring

innovative methodologies that enhance LLM capabilities, ensuring applicability across diverse contexts and domains. As LLM technology advances, the necessity for interdisciplinary collaboration intensifies to address the unique ethical challenges, performance gaps, and practical applications associated with LLMs. This collaboration is essential for tackling issues such as accountability, bias reduction, and effective LLM integration into fields like automated literature reviews and data interpretation, ensuring responsible development and deployment that aligns with ethical standards and societal needs [72, 17, 4, 3, 1]. By integrating diverse perspectives and expertise, researchers can drive innovation and ensure the ethical and effective deployment of LLMs, ultimately maximizing their potential to benefit society.

8 Conclusion

This survey highlights the profound impact of Large Language Models (LLMs) on natural language processing (NLP), emphasizing their advanced capabilities in understanding, generating, and manipulating language. Models like Llama 3 demonstrate proficiency in mimicking language styles, especially when using sophisticated prompting methods such as Tree-of-Thoughts (ToT). The growing research interest in LLMs, including ChatGPT, illustrates their potential across multiple fields, although ethical considerations remain a crucial area for discussion.

Integrating knowledge graphs (KGs) into LLMs presents a promising avenue for enhancing knowledge representation and overcoming current challenges in factual reasoning. The survey underscores the significance of scaling models, ensuring high-quality training data, and employing effective fine-tuning techniques to boost LLM performance. Additionally, there is a call for a new regulatory framework for LLMs, advocating for ongoing oversight and stakeholder engagement in regulatory processes.

The exploration of LLM-based data augmentation techniques reveals their superiority in generating high-quality synthetic data compared to traditional methods. Tailored learning strategies have been shown to notably improve the reasoning abilities of smaller language models, advancing both mathematical and commonsense reasoning tasks. For practical deployment, the survey outlines the benefits of task decomposition, retrieval-augmented generation, and structured tool management as essential components for successful LLM agent implementation.

In sentiment analysis, LLM adaptability has been significantly enhanced through dynamic adaptive optimization modules, leading to improved accuracy and reduced mean squared error (MSE) metrics over previous approaches. In the medical field, LLMs are recognized for their potential to elevate patient care and education, though substantial limitations must be addressed before they can be fully integrated into clinical practice.

Overall, this survey underscores the critical role of LLMs in advancing NLP, highlighting their emergent capabilities and the challenges that need to be addressed to fully leverage their potential in diverse applications. The incorporation of retrieval processes, such as Retrieval-Augmented Generation (RAG), notably enhances LLM functionalities, underscoring the ongoing need for comprehensive evaluation frameworks.

References

[1] 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.

[2] Léo Hemamou and Mehdi Debiane. Scaling up summarization: Leveraging large language models for long text extractive summarization, 2024.

[3] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond, 2023.

[4] 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.

[5] Rajat Rawat. Disasterqa: A benchmark for assessing the performance of llms in disaster response, 2024.

[6] Xuelong Geng, Tianyi Xu, Kun Wei, Bingshen Mu, Hongfei Xue, He Wang, Yangze Li, Pengcheng Guo, Yuhang Dai, Longhao Li, Mingchen Shao, and Lei Xie. Unveiling the potential of llm-based asr on chinese open-source datasets, 2024.

[7] Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. Disinformation capabilities of large language models, 2024.

[8] Ryan Burnell, Han Hao, Andrew R. A. Conway, and Jose Hernandez Orallo. Revealing the structure of language model capabilities, 2023.

[9] Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candès, and Dan Jurafsky. Can unconfident llm annotations be used for confident conclusions?, 2025.

[10] Qin Chen, Jinfeng Ge, Huaqing Xie, Xingcheng Xu, and Yanqing Yang. Large language models at work in china's labor market, 2023.

[11] Yuan Tian, Weiwei Cui, Dazhen Deng, Xinjing Yi, Yurun Yang, Haidong Zhang, and Yingcai Wu. Chartgpt: Leveraging llms to generate charts from abstract natural language, 2023.

[12] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[13] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642, 2024.

[14] Nils Körber, Silvan Wehrli, and Christopher Irrgang. How to measure the intelligence of large language models?, 2024.

[15] Sandeep Chataut, Tuyen Do, Bichar Dip Shrestha Gurung, Shiva Aryal, Anup Khanal, Carol Lushbough, and Etienne Gnimpieba. Comparative study of domain driven terms extraction using large language models, 2024.

[16] Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipalli, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. Llm2llm: Boosting llms with novel iterative data enhancement, 2024.

[17] Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. Less is more for long document summary evaluation by llms, 2024.

[18] Phoebe Jing, Yijing Gao, Yuanhang Zhang, and Xianlong Zeng. Translating expert intuition into quantifiable features: Encode investigator domain knowledge via llm for enhanced predictive analytics, 2024.

[19] Yue Xing, Xiaofeng Lin, Chenheng Xu, Namjoon Suh, Qifan Song, and Guang Cheng. Theoretical understanding of in-context learning in shallow transformers with unstructured data, 2024.

[20] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[21] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer, 2023.

[22] Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, and Manabu Okumura. Can we obtain significant success in rst discourse parsing by using large language models?, 2024.

[23] Bumjin Park and Jaesik Choi. Memorizing documents with guidance in large language models, 2024.

[24] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[25] Ruihao Gong, Yang Yong, Shiqiao Gu, Yushi Huang, Chengtao Lv, Yunchen Zhang, Xianglong Liu, and Dacheng Tao. Llmc: Benchmarking large language model quantization with a versatile compression toolkit, 2024.

[26] Peiyu Liu, Zikang Liu, Ze-Feng Gao, Dawei Gao, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. Do emergent abilities exist in quantized large language models: An empirical study, 2023.

[27] Jumbly Grindrod. Large language models and linguistic intentionality, 2024.

[28] Melissa Ailem, Katerina Marazopoulou, Charlotte Siska, and James Bono. Examining the robustness of llm evaluation to the distributional assumptions of benchmarks, 2024.

[29] Yan Zhao, Zhongyun Li, Yushan Pan, Jiaxing Wang, and Yihong Wang. Lb-kbqa: Large-language-model and bert based knowledge-based question and answering system, 2024.

[30] Fabián Villena, Luis Miranda, and Claudio Aracena. llmner: (zero|few)-shot named entity recognition, exploiting the power of large language models, 2024.

[31] Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. On the (in)effectiveness of large language models for chinese text correction, 2023.

[32] Sara Court and Micha Elsner. Shortcomings of llms for low-resource translation: Retrieval and understanding are both the problem, 2024.

[33] Nicholas Ichien, Dušan Stamenković, and Keith J. Holyoak. Large language model displays emergent ability to interpret novel literary metaphors, 2024.

[34] Alejandro Peña, Aythami Morales, Julian Fierrez, Ignacio Serna, Javier Ortega-Garcia, Iñigo Puente, Jorge Cordova, and Gonzalo Cordova. Leveraging large language models for topic classification in the domain of public affairs, 2023.

[35] Yixuan Weng, Bin Li, Fei Xia, Minjun Zhu, Bin Sun, Shizhu He, Kang Liu, and Jun Zhao. Large language models need holistically thought in medical conversational qa, 2023.

[36] Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, Dor Muhlgay, Noam Rozen, Erez Schwartz, Gal Shachaf, Shai Shalev-Shwartz, Amnon Shashua, and Moshe Tenenholtz. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning, 2022.

[37] Rana Muhammad Shahroz Khan, Pingzhi Li, Sukwon Yun, Zhenyu Wang, Shahriar Nirjon, Chau-Wai Wong, and Tianlong Chen. Portllm: Personalizing evolving large language models with training-free and portable model patches, 2024.

[38] V. K. Cody Bumgardner, Aaron Mullen, Sam Armstrong, Caylin Hickey, and Jeff Talbert. Local large language models for complex structured medical tasks, 2023.

[39] Shaoyuan Chen, Yutong Lin, Mingxing Zhang, and Yongwei Wu. Efficient and economic large language model inference with attention offloading, 2024.

[40] Zhuoling Li, Xiaogang Xu, Zhenhua Xu, SerNam Lim, and Hengshuang Zhao. Larm: Large auto-regressive model for long-horizon embodied intelligence, 2025.

[41] Wei Wang and Qing Li. Schrodinger's memory: Large language models, 2024.

[42] Randall Balestriero, Romain Cosentino, and Sarath Shekkizhar. Characterizing large language model geometry helps solve toxicity detection and generation, 2024.

[43] Ziyang Chen and Stylios Moscholios. Using prompts to guide large language models in imitating a real person's language style, 2024.

[44] Javier García Gilabert, Carlos Escolano, Aleix Sant Savall, Francesca De Luca Fornaciari, Audrey Mash, Xixian Liao, and Maite Melero. Investigating the translation capabilities of large language models trained on parallel data only, 2024.

[45] Yonghui Wang, Wengang Zhou, Hao Feng, and Houqiang Li. Adaptvision: Dynamic input scaling in mllms for versatile scene understanding, 2024.

[46] Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. Mindmerger: Efficient boosting llm reasoning in non-english languages, 2024.

[47] Yicheng Feng, Yuxuan Wang, Jiazheng Liu, Sipeng Zheng, and Zongqing Lu. Llama rider: Spurring large language models to explore the open world, 2023.

[48] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, and Bing Shi. Chatgpt for shaping the future of dentistry: The potential of multi-modal large language model, 2023.

[49] Maohao Shen, Shun Zhang, Jilong Wu, Zhiping Xiu, Ehab AlBadawy, Yiting Lu, Mike Seltzer, and Qing He. Get large language models ready to speak: A late-fusion approach for speech generation, 2024.

[50] Ankit Satpute, Noah Giessing, Andre Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. Can llms master math? investigating large language models on math stack exchange, 2024.

[51] Mikhail Tikhomirov and Daniil Chernyshev. Impact of tokenization on llama russian adaptation, 2023.

[52] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine, 2024.

[53] Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Junyu Liu, Benji Peng, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. From text to multimodality: Exploring the evolution and impact of large language models in medical practice, 2024.

[54] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 2024.

[55] Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. Tool learning with large language models: A survey, 2024.

[56] Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. When do llms need retrieval augmentation? mitigating llms' overconfidence helps retrieval augmentation, 2024.

[57] Bo-Wen Zhang, Yan Yan, Boxiang Yang, Yifei Xue, and Guang Liu. Predictable emergent abilities of llms: Proxy tasks are all you need, 2024.

[58] Giorgio Franceschelli and Mirco Musolesi. On the creativity of large language models. *AI & SOCIETY*, pages 1–11, 2024.

[59] Simran Arora, Brandon Yang, Sabri Eyuboglu, Avanika Narayan, Andrew Hojel, Immanuel Trummer, and Christopher Ré. Language models enable simple systems for generating structured views of heterogeneous data lakes, 2023.

[60] Weichuan Wang, Zhaoyi Li, Defu Lian, Chen Ma, Linqi Song, and Ying Wei. Mitigating the language mismatch and repetition issues in llm-based machine translation via model editing, 2024.

[61] Arya D. McCarthy, Hao Zhang, Shankar Kumar, Felix Stahlberg, and Axel H. Ng. Improved long-form spoken language translation with large language models, 2022.

[62] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Adapted large language models can outperform medical experts in clinical text summarization, 2024.

[63] Huan Zhao, Qian Ling, Yi Pan, Tianyang Zhong, Jin-Yu Hu, Junjie Yao, Fengqian Xiao, Zhenxiang Xiao, Yutong Zhang, San-Hua Xu, Shi-Nan Wu, Min Kang, Zihao Wu, Zhengliang Liu, Xi Jiang, Tianming Liu, and Yi Shao. Ophtha-llama2: A large language model for ophthalmology, 2023.

[64] Somnath Kumar, Vaibhav Balloli, Mercy Ranjit, Kabir Ahuja, Tanuja Ganu, Sunayana Sitaram, Kalika Bali, and Akshay Nambi. Bridging the gap: Dynamic learning strategies for improving multilingual performance in llms, 2024.

[65] 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.

[66] Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. Crosslingual capabilities and knowledge barriers in multilingual large language models, 2024.

[67] Séamus Lankford and Andy Way. Leveraging llms for mt in crisis scenarios: a blueprint for low-resource languages, 2024.

[68] Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. A survey on large language models with multilingualism: Recent advances and new frontiers, 2025.

[69] Mihai Masala, Denis C. Ilie-Ablachim, Dragos Corlatescu, Miruna Zavelca, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. Openllm-ro – technical report on open-source romanian llms, 2024.

[70] Jinxin Liu, Shulin Cao, Jiaxin Shi, Tingjian Zhang, Lunyiu Nie, Linmei Hu, Lei Hou, and Juanzi Li. How proficient are large language models in formal languages? an in-depth insight for knowledge base question answering, 2024.

[71] Tuan Bui, Oanh Tran, Phuong Nguyen, Bao Ho, Long Nguyen, Thang Bui, and Tho Quan. Cross-data knowledge graph construction for llm-enabled educational question-answering system: A case study at hcmut, 2024.

[72] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. Navigating llm ethics: Advancements, challenges, and future directions, 2024.

[73] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

[74] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies, 2023.

[75] Hao Kang and Chenyan Xiong. Researcharena: Benchmarking large language models' ability to collect and organize information as research agents, 2025.

[76] Yue Feng, Shuchang Liu, Zhenghai Xue, Qingpeng Cai, Lantao Hu, Peng Jiang, Kun Gai, and Fei Sun. A large language model enhanced conversational recommender system, 2023.

[77] Zelong Li, Shuyuan Xu, Kai Mei, Wenyue Hua, Balaji Rama, Om Raheja, Hao Wang, He Zhu, and Yongfeng Zhang. Autoflow: Automated workflow generation for large language model agents, 2024.

[78] Yang Yan, Yihao Wang, Chi Zhang, Wenyuan Hou, Kang Pan, Xingkai Ren, Zelun Wu, Zhixin Zhai, Enyun Yu, Wenwu Ou, and Yang Song. Llm4pr: Improving post-ranking in search engine with large language models, 2024.

[79] Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. Active prompting with chain-of-thought for large language models, 2024.

[80] Hakan T. Otal and M. Abdullah Canbaz. Llm honeypot: Leveraging large language models as advanced interactive honeypot systems, 2024.

[81] Zihan Liao, Hang Yu, Jianguo Li, Jun Wang, and Wei Zhang. D2llm: Decomposed and distilled large language models for semantic search, 2024.

[82] Inacio Vieira, Will Allred, Séamus Lankford, Sheila Castilho, and Andy Way. How much data is enough data? fine-tuning large language models for in-house translation: Performance evaluation across multiple dataset sizes, 2024.

[83] Agnes Luhtaru, Taido Purason, Martin Vainikko, Maksym Del, and Mark Fishel. To err is human, but llamas can learn it too, 2024.

[84] Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Large language models can deliver accurate and interpretable time series anomaly detection, 2024.

[85] Bumjun Kim, Kunha Lee, Juyeon Kim, and Sangam Lee. Small language models are equation reasoners, 2024.

[86] Ziwei Chai, Tianjie Zhang, Liang Wu, Kaiqiao Han, Xiaohai Hu, Xuanwen Huang, and Yang Yang. Graphllm: Boosting graph reasoning ability of large language model, 2023.

[87] Shabnam Hassani. Enhancing legal compliance and regulation analysis with large language models, 2024.

[88] Menna Fateen and Tsunenori Mine. Developing a tutoring dialog dataset to optimize llms for educational use, 2024.

[89] Roma Shusterman, Allison C. Waters, Shannon O'Neill, Phan Luu, and Don M. Tucker. An active inference strategy for prompting reliable responses from large language models in medical practice, 2024.

[90] S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models, 2024.

[91] Shawn Gavin, Tuney Zheng, Jiaheng Liu, Quehry Que, Noah Wang, Jian Yang, Chenchen Zhang, Wenhao Huang, Wenhu Chen, and Ge Zhang. Longins: A challenging long-context instruction-based exam for llms, 2024.

[92] Yufan Chen, Arjun Arunasalam, and Z. Berkay Celik. Can large language models provide security privacy advice? measuring the ability of llms to refute misconceptions, 2023.

[93] Pei-Fu Guo, Ying-Hsuan Chen, Yun-Da Tsai, and Shou-De Lin. Towards optimizing with large language models, 2024.

[94] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. Recommender systems in the era of large language models (llms), 2024.

[95] Da Song, Xuan Xie, Jiayang Song, Derui Zhu, Yuheng Huang, Felix Juefei-Xu, and Lei Ma. Luna: A model-based universal analysis framework for large language models, 2024.

[96] Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. Data augmentation using large language models: Data perspectives, learning paradigms and challenges, 2024.

[97] Swapnaja Achintalwar, Adriana Alvarado Garcia, Ateret Anaby-Tavor, Ioana Baldini, Sara E. Berger, Bishwaranjan Bhattacharjee, Djallel Bouneffouf, Subhajit Chaudhury, Pin-Yu Chen, Lamogha Chiazor, Elizabeth M. Daly, Kirushikesh DB, Rogério Abreu de Paula, Pierre Dognin, Eitan Farchi, Soumya Ghosh, Michael Hind, Raya Horesh, George Kour, Ja Young Lee, Nishtha Madaan, Sameep Mehta, Erik Miehling, Keerthiram Murugesan, Manish Nagireddy, Inkit Padhi, David Piorkowski, Ambrish Rawat, Orna Raz, Prasanna Sattigeri, Hendrik Strobelt, Sarathkrishna Swaminathan, Christoph Tillmann, Aashka Trivedi, Kush R. Varshney, Dennis Wei, Shalisha Witherspooon, and Marcel Zalmanovici. Detectors for safe and reliable llms: Implementations, uses, and limitations, 2024.

[98] Hamed Babaei Giglou, Jennifer D'Souza, Felix Engel, and Sören Auer. Llms4om: Matching ontologies with large language models, 2024.

[99] Flavio Petruzzellis, Alberto Testolin, and Alessandro Sperduti. Assessing the emergent symbolic reasoning abilities of llama large language models, 2024.

[100] Ernests Lavrinovics, Russa Biswas, Johannes Bjerva, and Katja Hose. Knowledge graphs, large language models, and hallucinations: An nlp perspective, 2024.

[101] Saaketh Koundinya Gundavarapu, Shreya Agarwal, Arushi Arora, and Chandana Thimmalapura Jagadeeshaiah. Machine unlearning in large language models, 2024.

[102] Alejandro Tlaie. Exploring and steering the moral compass of large language models, 2024.

[103] Q. Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap, 2023.

[104] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwan, Yoshua Bengio, Danqi Chen, Philip H. S. Torr, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models, 2024.

[105] Mahsa Shamsabadi, Jennifer D'Souza, and Sören Auer. Large language models for scientific information extraction: An empirical study for virology, 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.