A Survey of Large Language Models and Their Interactions with User Interfaces and Datasets

www.surveyx.cn

Abstract

This survey paper explores the transformative role of Large Language Models (LLMs) within artificial intelligence and computer science, emphasizing their integration with user interfaces, datasets, usability testing, human-computer interaction, and natural language processing. LLMs have significantly advanced AI capabilities, improving accessibility and applicability across various domains, including healthcare, education, and legal sectors. The paper highlights the importance of methodological integration frameworks that combine LLMs with user interfaces and datasets, enhancing AI systems' performance and user experience. It underscores the necessity of well-structured datasets and comprehensive usability testing to ensure reliable AI applications. The survey also discusses the challenges and limitations of LLMs, such as computational demands, data quality, and ethical considerations, advocating for ongoing research to address these issues. Future directions include refining LLM integration techniques, enhancing personalization, and expanding datasets to improve AI systems' robustness and adaptability. By advancing research on LLMs and related methodologies, the field can achieve greater innovation and practical applications, ultimately enhancing AI systems' efficacy and reliability in addressing complex real-world challenges.

1 Introduction

1.1 Significance of Large Language Models

Large Language Models (LLMs) have become pivotal in artificial intelligence, significantly enhancing natural language processing by converting complex human inputs into structured outputs, thereby improving accessibility for non-experts and expanding AI's applicability [1]. In multilingual contexts, LLMs enhance performance across various languages, promoting inclusivity in AI applications [2]. Their capacity to generate human-like text is crucial for IT operations, particularly in managing extensive data sets [3].

In healthcare, LLMs improve diagnostic accuracy and support clinical decision-making, despite deployment challenges [3]. They facilitate automatic text summarization in clinical settings, reducing documentation burdens and potentially mitigating physician burnout [4]. Furthermore, LLMs enhance automated grading in education, improving feedback efficiency and consistency [5]. The educational sector recognizes both opportunities and challenges presented by LLMs, as perceived by students and educators [6].

LLMs also optimize web element localization for web-based test automation, enhancing the accuracy and efficiency of testing processes [7]. In the legal field, the need for domain-specific entity recognition highlights LLMs' role in navigating complex legal texts and ensuring accurate information retrieval [7]. Additionally, in urban planning, they foster communication and collaboration, improving participation and efficiency [8]. Their ability to generate novel scientific hypotheses, particularly in medical research, underscores their potential in advancing scientific inquiry [9].

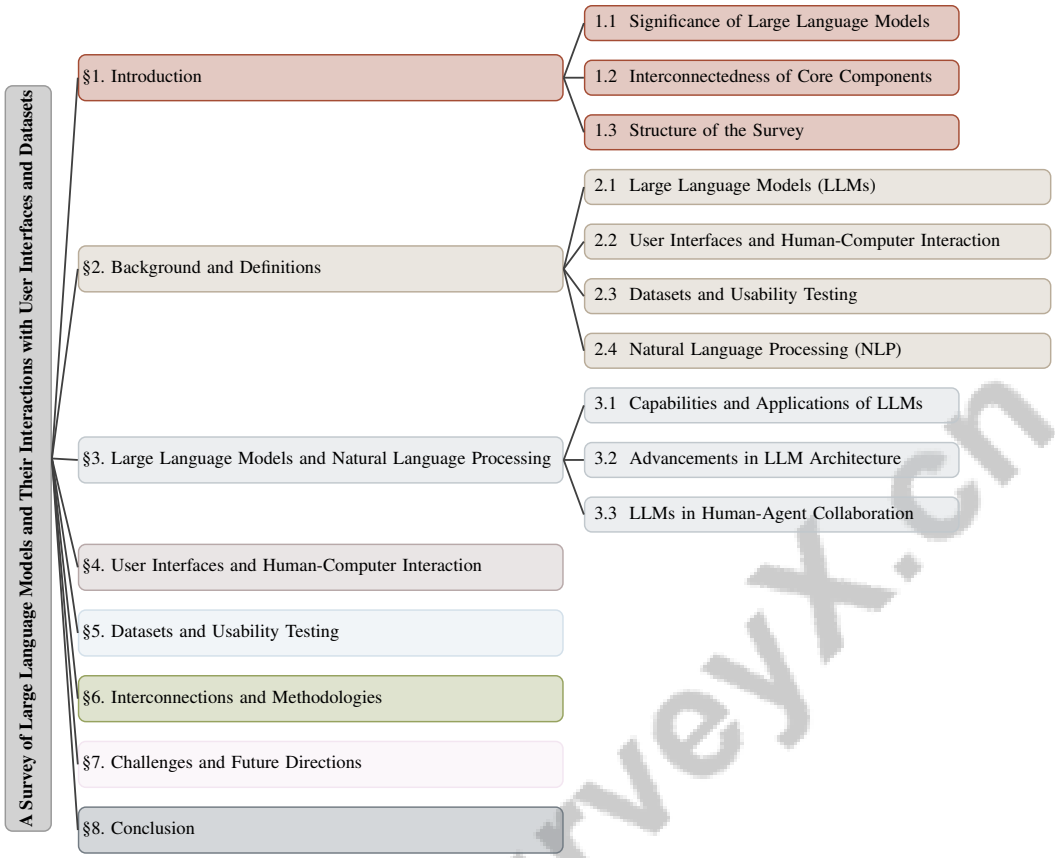


Figure 1: chapter structure

Despite their advantages, LLMs encounter challenges such as misalignment with human values and the generation of non-factual content, or hallucinations, necessitating ongoing research and refinement [10]. The integration of LLMs into user interfaces is increasingly common, particularly in web and mobile applications, highlighting the need for intelligent interfaces that adapt to real-time changes [11]. The rise of LLMs has transformed user interactions with knowledge-based systems, enabling chatbots to synthesize vast information and assist with complex tasks [8]. Moreover, leveraging LLMs for mental health interventions addresses the urgent need for effective solutions amidst rising mental health challenges exacerbated by the COVID-19 pandemic [12].

The automation of complex processes, such as visualization generation, which can be tedious for novice users, highlights LLMs' significance in modern AI applications [10]. As these models evolve, their impact on AI applications and research remains profound, driving innovation across multiple domains and enhancing the efficacy, accessibility, and ethical considerations of AI systems. However, the emergence of biases in LLMs, which can perpetuate stereotypes, necessitates the development of safer and more responsible models [13]. Understanding the complexity of LLM architectures and their emergent capabilities is essential for effectively harnessing their potential [14]. Integrating multi-modal data, such as imaging and non-imaging patient data, is crucial for accurate medical diagnoses, exemplifying LLMs' importance in healthcare [15].

1.2 Interconnectedness of Core Components

The integration of Large Language Models (LLMs) with user interfaces, datasets, usability testing, human-computer interaction (HCI), and natural language processing (NLP) forms a comprehensive framework essential for enhancing AI systems' functionality and user experience. LLMs act as a critical interface in natural language communication, facilitating complex interactions between users and data agents, thus promoting interdisciplinary collaboration [7]. In educational contexts, LLMs enrich learning experiences and develop competencies necessary for their effective use, demonstrating their interconnectedness with educational interfaces [6].

User interface design, grounded in HCI principles, is vital for fostering user trust and minimizing biases within AI systems. Integrating LLMs into visualization systems presents challenges that must be addressed to enhance intelligent analysis and user experience [7]. Ensuring that LLM outputs align with user-defined constraints is crucial for improving usability and seamless integration into workflows [16].

Usability testing is closely linked to LLMs and user interfaces, necessitating a comprehensive understanding of user needs and AI capabilities. Systems that engage users in reflective dialogues, such as Gradschool.chat, exemplify how LLMs can enhance user engagement and satisfaction [6]. The complexity of prompt evaluation across diverse datasets further emphasizes the interconnectedness of these components, necessitating thorough usability testing to ensure effective AI performance [17]. However, users often face challenges in initializing and refining prompts due to cognitive barriers and biased perceptions, which can hinder task completion [16].

In NLP, LLMs are indispensable for processing and generating human-like text, essential for applications ranging from chatbots to automated content analysis. The absence of a systematic analytical framework for incorporating nonverbal behaviors in LLM-voice assistant interactions limits understanding of user engagement, highlighting the need for more comprehensive frameworks [6]. Additionally, LLMs support early psychological interventions through their conversational capabilities, addressing the demand for effective mental health solutions [16].

The interconnectedness of LLM safety, user interfaces, datasets, usability testing, HCI, and NLP is vital, as LLMs pose risks in sensitive applications [17]. Leveraging these interconnected components enables researchers and developers to create more intuitive, efficient, and trustworthy AI applications that cater to diverse user needs and contexts. Integrating LLMs into conversational assistants for specific tasks illustrates the multifaceted interactions between LLMs, user interfaces, and human-computer interaction [7]. The necessity for a human-centered approach in designing effective explanations in explainable AI (XAI) systems further underscores the importance of interconnectedness [6].

Current models primarily excel in static environments and simple domains, such as web and mobile interfaces, highlighting the need for benchmarks that can evaluate dynamic GUI interactions [7]. These benchmarks aim to facilitate model comparisons in understanding computer UIs as states, enabling the development of automated systems that interpret user actions and automate workflows [16]. LLM-based chatbots often struggle to provide personalized support, particularly when users initiate vague queries or lack sufficient contextual information [6]. A novel approach that treats recommendation as instruction following allows users to express their preferences in natural language, enhancing personalization and user satisfaction [7]. Deep-learning techniques for generating UI mockups from natural language phrases further illustrate the interconnectedness of LLMs, user interfaces, and usability testing [16]. The survey categorizes different modes of human-LLM interactions, addressing knowledge gaps regarding effective interaction strategies and their applications [17]. Existing models often struggle with accurately identifying judicial entities, critical for various applications in legal information retrieval and analysis [7]. The proposed idea, GPTDroid, formulates mobile GUI testing as a QA task, allowing the LLM to interact with mobile apps to generate and execute testing scripts iteratively [16]. Effectively adapting user interfaces to meet changing user preferences and contexts is essential for improving user experience [6]. Understanding user interactions with LLMs, including user intents, satisfaction, and concerns, is crucial for effective human-AI collaboration [17].

1.3 Structure of the Survey

This survey is structured to comprehensively explore the role and impact of Large Language Models (LLMs) within the broader context of artificial intelligence and computer science. The **Introduction** establishes the significance of LLMs, emphasizing their transformative effects across various domains and their interconnectedness with user interfaces, datasets, usability testing, human-computer interaction (HCI), and natural language processing (NLP). This section sets the stage for subsequent detailed discussions.

Section 2: Background and Definitions delves into fundamental concepts, offering precise definitions of LLMs, user interfaces, datasets, usability testing, HCI, and NLP. This section elucidates the

interrelationships among these components, providing a foundational understanding necessary for appreciating their integration in AI systems.

Section 3: Large Language Models and Natural Language Processing focuses on the capabilities and applications of LLMs in NLP, reviewing significant advancements in LLM architecture and their implications for NLP tasks. This section also explores the role of LLMs in facilitating human-agent collaboration, a critical aspect of modern AI interactions.

Section 4: User Interfaces and Human-Computer Interaction examines the design and functionality of user interfaces as pivotal points of human-computer interaction. The discussion centers on foundational principles guiding the creation of intuitive and effective user interfaces, emphasizing accessibility and AI integration to enhance user experience. It draws on innovative concepts such as "Collage" from avant-garde literature to propose new design directions for AI writing tools and presents 18 validated design guidelines for human-AI interaction in light of recent AI advancements [8, 18].

Section 5: Datasets and Usability Testing highlights the critical role of datasets in training AI models and the significance of usability testing in assessing AI systems' ease of use and user experience. This section underscores the necessity of well-structured datasets for reliable usability testing outcomes and explores iterative prototyping and design models utilized in this process.

Section 6: Interconnections and Methodologies analyzes the interconnectedness of LLMs, user interfaces, datasets, usability testing, HCI, and NLP. It explores advanced methodologies that integrate various components, including fine-tuned LLMs and innovative retrieval frameworks, to enhance AI systems' performance and user experience. Case studies and evaluation frameworks illustrate successful applications, such as automating Systematic Literature Reviews (SLRs) and improving multi-document question answering through Inter-chunk Interactions. It emphasizes the importance of user interface design in AI writing tools, advocating for a collage-based approach that reflects historical literary innovations. The findings highlight the potential of these methodologies to streamline complex processes, improve retrieval accuracy, and foster effective collaboration in qualitative analysis, ultimately setting new standards for AI-driven research and writing practices [19, 20, 8, 21].

Section 7: Challenges and Future Directions identifies current challenges in integrating and applying the core components and methodologies discussed. The study addresses challenges in automating systematic literature reviews and enhancing question-answering capabilities using LLMs while outlining promising future research directions. These include refining methodologies for AI integration in academic research, improving retrieval frameworks for multi-document interactions, and enhancing lay language generation through retrieval-augmented techniques. By advocating for updates to PRISMA reporting guidelines and exploring external knowledge sources, the research aims to advance methodological transparency, improve accessibility of complex academic content, and streamline literature review processes across disciplines [19, 20, 22].

Finally, the **Conclusion** synthesizes the survey's key findings, reflecting on the significance of interconnected components and methodologies in advancing AI and computer science. The findings highlight implications for future research and practical applications, illustrating the transformative role of fine-tuned LLMs in automating SLRs and enhancing academic research methodologies. This ongoing evolution underscores the necessity for rigorous scrutiny and methodological updates, such as revising PRISMA reporting guidelines to incorporate AI-driven processes, improving reliability and transparency in scholarly work. Moreover, the critical examination of current LLM research reveals trends necessitating increased ethical considerations and reproducibility, emphasizing the need for a systematic approach to harnessing LLMs effectively across diverse research domains [23, 20].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Large Language Models (LLMs)

Large Language Models (LLMs) represent a significant advancement in artificial intelligence, excelling in processing and generating human-like text through pre-trained neural networks. They are pivotal in various natural language processing (NLP) applications, including text generation, translation, summarization, and question answering, underscoring their transformative role in AI

research and applications [24, 14]. LLMs demonstrate emergent behaviors as they scale, enhancing their capabilities and reinforcing their status as indispensable AI tools.

A notable feature of LLMs is their ability to discern user intent, categorized into diverse types such as 'Ask for Advice', 'Information Retrieval', 'Leisure', 'Seek Creativity', 'Solve Problems', 'Text Assistant', and 'Use through API' [6]. However, they encounter challenges in accessing real-time, accurate data from relational databases, limiting their practical utility [15]. In healthcare, LLMs hold promise for improving clinical practices through transfer learning and domain-specific fine-tuning, while also addressing ethical concerns [25].

In automated summarization, LLMs are crucial for converting doctor-patient dialogues into clinical notes, essential for effective communication and continuity of care [26]. They also serve as therapeutic counselors, particularly in cognitive behavioral therapy (CBT), with effectiveness comparable to human counselors [17]. In corporate environments, LLMs are assessed for their ability to predict legislative bills' relevance to companies and draft persuasive communications [13].

Despite their advanced functionalities, LLMs face limitations in managing complex reasoning tasks due to restricted memory retention, affecting their ability to utilize context from prior interactions [24]. This issue is compounded by hallucinations, where generated text may contradict or lack verification against source material [26]. Strategies to mitigate these issues without extensive retraining are needed to enhance LLM reliability [5].

As LLM architectures evolve, ongoing research is crucial to address inherent limitations and explore innovative applications leveraging their strengths. Initiatives like Legilimens, which extracts conceptual features from chat-oriented LLMs to identify unsafe content, exemplify efforts to enhance safety and moderation capabilities [16]. Developing benchmarks for evaluating LLM performance in process mining tasks is essential for accurate model comparisons, facilitating continuous improvement in LLM development.

2.2 User Interfaces and Human-Computer Interaction

User interfaces (UIs) are essential for interaction between humans and computer systems, incorporating visual elements and interactive features to facilitate communication and task execution. UI design, guided by human-computer interaction (HCI) principles, aims to optimize user experience by ensuring interfaces are intuitive, accessible, and responsive to user needs [27]. This focus on usability is crucial in AI writing tools, where fragmented text displays necessitate designs accommodating diverse user interactions [8].

Understanding user interactions, including active applications and contextual factors, is vital for creating systems that adapt to dynamic user environments, as seen in Intelligent User Interfaces (IUIs) utilizing fuzzy logic and AI to enhance user interaction [11]. HCI principles are evident in systems like LIDA, offering conversational interfaces for user interaction and visualization refinement [28], and the CARE system, which facilitates iterative query refinement in collaborative contexts [29].

The increasing reliance on web technologies necessitates continuous evaluation and enhancement of UI quality to meet user expectations and improve functionality [30]. This is particularly relevant in scenarios where non-experts program interactions with machines, such as ambient contexts designed to assist the elderly [31]. Creating user interface mock-ups from high-level text descriptions exemplifies HCI principles in generating intuitive designs [32].

As HCI evolves, integrating AI into user interfaces remains a critical focus, driving innovation and improving usability across various domains. Emphasizing adaptive and intelligent interface design enhances AI systems' effectiveness in supporting diverse tasks. This focus is underscored by 18 validated design guidelines aimed at improving usability and interpretability in Explainable AI (XAI) interfaces, where user experience is vital for effective AI explanations and support. Ongoing research and innovation in interface design are essential to bridge knowledge gaps and optimize user engagement with AI systems [33, 18].

2.3 Datasets and Usability Testing

Datasets and usability testing are fundamental to the development and evaluation of artificial intelligence (AI) systems, providing essential components for model training and user experience

optimization. Datasets, consisting of structured data collections, are critical for training AI models, enabling them to learn patterns and make predictions across various tasks. The quality and diversity of datasets directly impact model performance, robustness, and generalization capabilities [34]. The MedEval dataset, for instance, contains 2,680 medical consultation dialogues annotated based on LLM-specific Mini-CEX criteria, illustrating the significance of domain-specific data in enhancing healthcare model accuracy [35]. Similarly, the InLegalNER dataset, derived from real-world case law documents, encompasses a range of entities pertinent to the legal domain, emphasizing the necessity of specialized datasets for effective information retrieval [36].

The HELPERT dataset, consisting of 27 simulated CBT-based sessions, mirrors standard therapy dialogue formats, underscoring the role of datasets in training models for therapeutic applications [12]. In corporate settings, datasets with ground-truth labels indicating the relevance of Congressional bills to companies are sourced from official summaries and SEC filings, highlighting the critical role of datasets in predictive analytics [7]. Moreover, datasets relevant to manufacturing, which include structured and unstructured text data, facilitate knowledge sharing and operational efficiency [25].

Usability testing complements datasets by systematically evaluating the ease of use and user experience of AI systems. This process involves assessing user interactions with AI tools, identifying improvement areas, and ensuring effective performance in real-world scenarios. The significance of usability testing is evident in the CogBench dataset, which includes responses from various LLMs to multiple cognitive tasks, enabling diverse analyses of behaviors and interactions [37]. Additionally, structured literature reviews focusing on benchmarks for evaluating LLMs, especially those with multimodal capabilities, highlight the importance of usability testing in refining model performance and aligning with user needs [38].

Integrating high-quality datasets with thorough usability testing is crucial for advancing AI systems, enhancing performance and user-centricity while adhering to ethical standards. Recent research emphasizes user-driven value alignment, where users engage in correcting biased outputs from AI companions, fostering a collaborative relationship that aligns AI behavior with user values. Tools like CoAIcoder illustrate how AI can facilitate human collaboration in qualitative analysis, highlighting the need for AI system design to consider user independence and its impact on outcomes, such as code diversity. This approach underscores the necessity of combining robust data practices with user engagement to create effective, ethical, and inclusive AI solutions [39, 21]. By addressing dataset quality and usability testing challenges, researchers and developers can enhance AI applications' performance and user experience across diverse contexts, leading to more reliable and trustworthy systems.

2.4 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a vital subfield of artificial intelligence dedicated to enabling computers to understand, interpret, and generate human language meaningfully. This capability is increasingly integrated into systems such as user interfaces, dialogue systems, and automated customer service platforms, facilitating natural communication with users. However, NLP implementation often faces challenges related to processing speed, particularly in real-time applications where delays can impede user experience. Recent advancements, including syntactic parsers utilizing machine learning and statistical models, aim to address speed issues through techniques like Compressed POS Set and Syntactic Patterns Pruning. Additionally, fine-tuned Large Language Models (LLMs) are transforming academic research methodologies by automating systematic literature reviews, enhancing efficiency and accuracy. Efforts to improve the accessibility of complex biomedical literature through lay language generation further demonstrate the potential of automated methods to make expert content interpretable for diverse audiences [40, 20, 22]. NLP encompasses various tasks such as language translation, sentiment analysis, and information retrieval, essential for creating systems that interact effectively with human users. The integration of LLMs into NLP has significantly improved these capabilities, enhancing fluency, coherence, and the handling of complex linguistic constructs.

LLMs have been instrumental in advancing NLP tasks by leveraging large datasets and sophisticated architectures to enhance language understanding and generation. Their adaptability to linguistic challenges is demonstrated by their ability to optimize performance through advanced algorithms, improving efficiency in processing natural language. Nonetheless, challenges persist, including the

generation of harmful content, privacy leaks, and adversarial attacks that threaten the integrity and reliability of LLM outputs [10].

Innovative approaches like knowledge graph-enhanced pre-trained language models (KGPLMs) have been explored to enrich LLM capabilities by integrating structured knowledge, thereby improving factual accuracy and contextual understanding [26]. Moreover, methodologies such as Legilimens, which analyze features from both the first and last tokens generated by LLMs, have been introduced to enhance moderation capabilities across various tasks without increasing computational complexity [16].

In multilingual contexts, LLMs serve as evaluators that assess and calibrate outputs against a dataset of 20,000 human judgments across multiple languages and text generation tasks. This approach addresses traditional evaluation limitations that often depend on human annotators or existing benchmarks, enhancing reproducibility and transparency in NLP research. By incorporating LLMs into the evaluation process, researchers can mitigate biases inherent in single-model assessments and improve evaluation reliability, particularly for low-resource languages and non-Latin scripts [20, 41, 42, 43]. This capability is essential for addressing representation and adaptability issues in complex human behaviors within social systems. As research advances, the synergy between NLP and LLMs is expected to further enhance AI systems' capabilities and applications, driving innovation and improving human-computer interaction efficacy.

In recent years, the evolution of Large Language Models (LLMs) has significantly influenced the field of Natural Language Processing (NLP). These models have not only advanced in their architectural complexity but have also expanded their applications across various domains. To better understand this evolution, we can refer to Figure 2, which illustrates the hierarchical structure of LLMs. This figure categorizes their capabilities and applications, emphasizing key advancements in architecture and their role in human-agent collaboration. Notably, it highlights critical areas such as conversational agents, educational tools, workflow automation, safe model designs, modular frameworks, enhanced memory systems, interaction frameworks, social system modeling, and user experience optimization. By examining these components, we can gain deeper insights into how LLMs are shaping the landscape of NLP and their implications for future research and application.

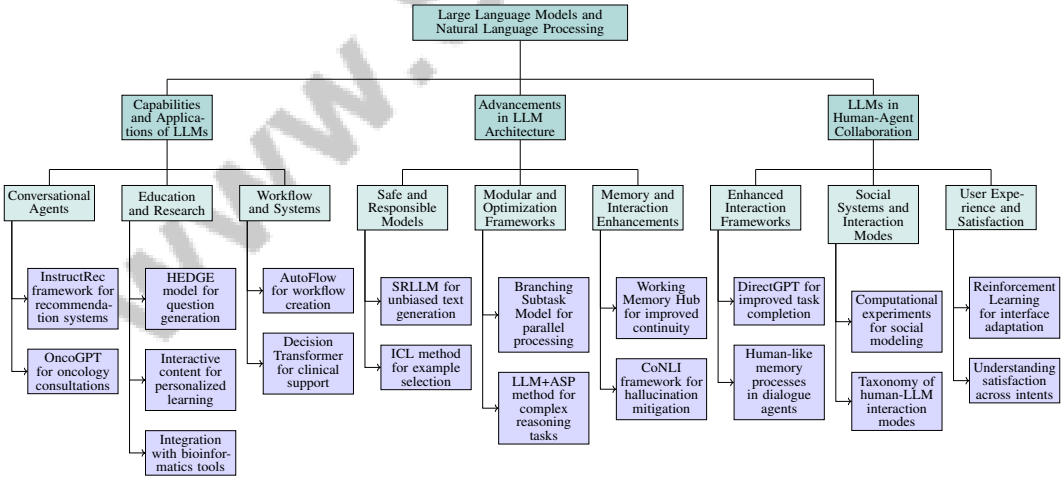


Figure 2: This figure illustrates the hierarchical structure of Large Language Models (LLMs) in Natural Language Processing, categorizing their capabilities and applications, advancements in architecture, and their role in human-agent collaboration. It highlights key areas such as conversational agents, educational tools, workflow automation, safe model designs, modular frameworks, enhanced memory systems, interaction frameworks, social system modeling, and user experience optimization.

3 Large Language Models and Natural Language Processing

3.1 Capabilities and Applications of LLMs

Large Language Models (LLMs) have revolutionized natural language processing (NLP), enhancing applications across diverse fields. Their adeptness at generating human-like text underpins the development of conversational agents that enrich user interactions. For example, the InstructRec framework leverages natural language instructions to refine recommendation systems, boosting user satisfaction [44]. In healthcare, models like OncoGPT excel in oncology consultations, demonstrating their reliability for medical inquiries [45].

In education, LLMs facilitate content creation and personalized learning. The HEDGE model automates the generation of question stems and distractors, optimizing instructional material development [46]. Additionally, they foster student engagement through interactive content and personalized learning experiences [6]. LLMs also integrate with bioinformatics tools in scientific research, enhancing data analysis and hypothesis generation [47].

As illustrated in Figure 3, the diverse capabilities and applications of LLMs span various domains, highlighting their significant impact on conversational agents, educational tools, and automation frameworks. Their adaptability is evident in frameworks like AutoFlow, which automates workflow creation using natural language, enabling task specification without extensive design knowledge [48]. In multi-agent systems, improved contextual memory management enhances collaborative capabilities [49]. In counseling, Decision Transformer models aid topic recommendations, enhancing automated clinical support systems [50].

LLMs are effective in GUI-centric applications, such as AutoWebGLM, which simplifies HTML content representation for better comprehension [51]. Models like Polymetis demonstrate LLMs' utility in handling large datasets across multiple materials domains [52]. Furthermore, frameworks like LDB enhance debugging processes by systematically analyzing generated programs [53].

A significant innovation is LLMs' ability to integrate and interpret multi-source data, including text, audio, and images, for comprehensive diagnostics in fields like dentistry [54]. Tools like CoAIcoder, which assist in collaborative qualitative coding, exemplify how LLMs enhance coding efficiency and coder agreement [21].

LLMs also improve web element localization, as seen in the VON Similo LLM method, which enhances web element identification accuracy [55]. In the legal domain, LLMs streamline entity recognition in complex legal texts, facilitating information retrieval and analysis [36].

The ongoing evolution of LLMs ensures their adaptability to diverse user needs across various domains. Their NLP capabilities not only enhance AI systems' functionality but also open new possibilities for human-computer interaction and digital communication. As LLMs advance, research must address ethical considerations to ensure responsible and effective use in tackling complex challenges, such as mental health support [56].

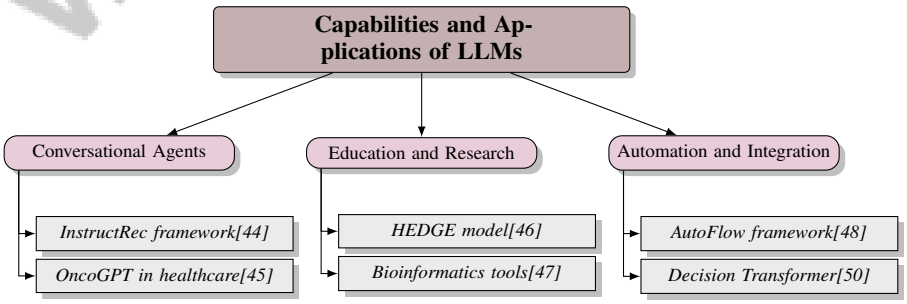


Figure 3: This figure illustrates the diverse capabilities and applications of Large Language Models (LLMs) across various domains, highlighting their impact on conversational agents, educational tools, and automation frameworks.

3.2 Advancements in LLM Architecture

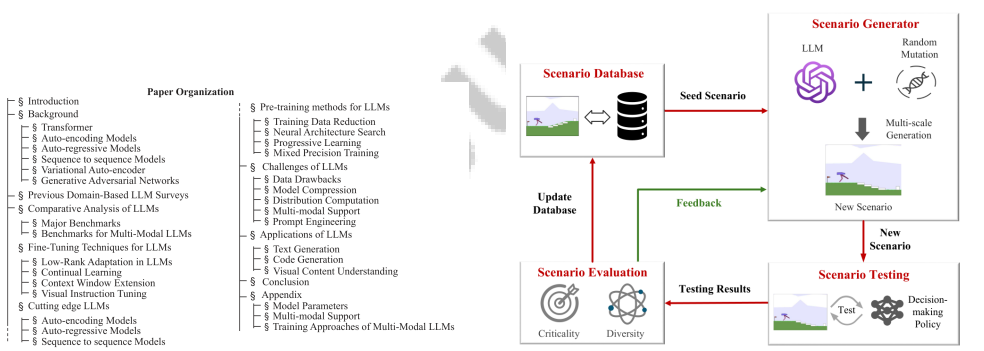
Recent advancements in Large Language Model (LLM) architecture have significantly bolstered their capabilities in NLP tasks. The Safe and Responsible Large Language Model (SRLLM) employs a dual-layered approach, combining instruction fine-tuning atop a pre-trained safe model, enhancing unbiased, contextually rich text generation while addressing ethical deployment concerns [13]. The In-Context Learning (ICL) method, which is tuning-free and label-free, uses a novel scoring function to select high-quality examples, improving LLM adaptability and efficiency [24].

Modular LLM designs, such as the Branching Subtask Model (BSM), facilitate parallel processing of sub-tasks, enhancing evaluation and generation capabilities. The RETA-LLM framework supports information retrieval and processing stages, including request rewriting and fact-checking, increasing efficiency and accuracy [57].

Integrating LLMs with optimization algorithms has emerged as a focal point, allowing these models to function as both search operators and generators of optimization strategies. The LLM+ASP method illustrates LLMs' ability to parse linguistic variability into structured logical forms, interpretable by Answer Set Programming (ASP), providing a robust framework for complex reasoning tasks [58].

To mitigate ungrounded hallucinations in LLM outputs, the CoNLI framework employs a two-stage process that detects and mitigates hallucinations through systematic rewriting, enhancing content reliability and factual accuracy. AutoFlow exemplifies adaptability in LLM frameworks by enabling workflow generation through fine-tuning and in-context learning [48].

Integrating a Working Memory Hub with an Episodic Buffer in LLMs allows for persistent memory links and improved continuity, enhancing complex interactions management [49]. These architectural advancements not only improve NLP task capabilities but also drive innovation across various domains, ensuring AI systems remain robust, efficient, and adaptable to evolving user needs and technological challenges.



(a) Paper Organization[59]

(b) Scenario Generation and Evaluation System[60]

Figure 4: Examples of Advancements in LLM Architecture

As shown in Figure 4, advancements in LLM architecture significantly influence NLP, exemplified by paper organization and scenario generation systems. The paper organization chart provides a structured overview of essential LLM architecture components, emphasizing the foundational Transformer model and pre-training methods, focusing on topics like the Transformer framework and auto-regressive models. Conversely, the scenario generation and evaluation system illustrates LLMs' dynamic application in generating and testing scenarios, using a scenario database and generator to create new scenarios evaluated through criticality and diversity metrics. The feedback loop ensures continuous improvement of generated scenarios, showcasing LLMs' practical utility in real-world applications [59, 60].

3.3 LLMs in Human-Agent Collaboration

Large Language Models (LLMs) are pivotal in enhancing collaboration between humans and AI agents, facilitating more effective interactions and problem-solving. Their integration into human-

agent collaboration frameworks addresses limitations of traditional AI systems, particularly in dynamic environments where understanding human needs is critical [61]. Incorporating human-like memory processes into LLM-based dialogue agents enables autonomous recall and utilization of memories during interactions, enhancing contextual understanding and response accuracy [62].

The development of DirectGPT illustrates advancements in LLM-based interfaces, offering faster task completion and improved clarity by reducing verbosity in prompts [63]. This aligns with the need for user-centered AI explanations that consider specific tasks and cognitive processes, enhancing understanding and trust [64]. However, designing effective human-AI interactions remains challenging due to unpredictable AI behavior, which can lead to user confusion and mistrust [18].

Integrating LLMs into computational experiments enhances modeling of social systems, providing a more accurate representation of human-like behaviors and facilitating nuanced human-agent interactions [65]. A structured taxonomy of human-LLM interaction modes identifies four key phases—planning, facilitating, iterating, and testing—that optimize these interactions [66].

To further enhance collaboration, a Reinforcement Learning-based framework incorporating physiological data can inform user interface adaptations, enabling real-time responses to user needs and improving interaction experiences [27]. Understanding user satisfaction across different intents is crucial, as it varies significantly; for instance, 'Text Assistant' interactions yield high satisfaction, while 'Seek Creativity' interactions often result in dissatisfaction, reflecting diverse user expectations [1].

LLMs significantly enhance human-agent collaboration by offering personalized, context-aware interactions and enabling AI systems to better understand and respond to human needs. As AI research progresses, fine-tuned LLMs are expected to play an increasingly critical role in developing sophisticated collaborative AI systems. They have already demonstrated potential in automating complex processes such as systematic literature reviews, enhancing question-answering capabilities through improved retrieval frameworks, and facilitating human collaboration in qualitative analysis. With ongoing advancements, LLMs are poised to streamline labor-intensive research tasks, improve the quality and efficiency of academic outputs, and support interdisciplinary ideation by simulating expert personas, addressing challenges posed by the growing volume of academic literature and the scarcity of domain experts [19, 67, 21, 68, 20].

4 User Interfaces and Human-Computer Interaction

4.1 User Interface Elements and Accessibility

Effective UI design is crucial for enhancing human-computer interaction, focusing on usability and accessibility. Direct manipulation principles, as seen in DirectGPT, promote intuitive interfaces that reduce cognitive load and support seamless interaction [63]. Accessibility is essential, with systems like MICEFrame enabling users without programming skills to control their environments, emphasizing the need for inclusive design [31]. AI integration further personalizes UIs, as demonstrated by MindGuide, which uses contextual data to enhance user satisfaction [56]. CoAIcoder exemplifies AI-driven efficiency in collaborative qualitative analysis [21].

Hybrid frameworks that combine user evaluations with automated metrics ensure UIs meet user expectations and technical standards [30]. Grounding UI elements in natural language instructions facilitates intuitive interaction [69]. Thoughtful AI writing tool design, considering historical practices, underscores UI design's role in user engagement [8].

Prioritizing usability, accessibility, and user involvement in UI development builds trust and enhances user experience. By integrating features like text fragmentation and natural language descriptions, developers can create inclusive UIs that meet diverse user needs, improve accessibility, and empower users to align with AI systems [70, 8, 39].

4.2 Conversational Interfaces and Human-AI Interaction

Conversational interfaces are key to improving human-AI interactions by enabling natural dialogues. They leverage advanced LLMs for structured, dynamic responses, enhancing engagement and satisfaction. The Interactive Explanatory Tool (IET) and Conceptual Model Interpreter illustrate how

conversational interfaces transform information and streamline processes [71, 72]. Design impacts the credibility and trust of AI content, with users employing strategies to mitigate AI biases [39].

Incorporating verbal and nonverbal cues enriches interactions, as seen in LLM-based voice assistants. The Adaptive Explanation System (AES) aligns user needs with XAI capabilities, improving transparency and trust [64]. Conversational interfaces address challenges like information overload, with tools like RELIC enabling users to verify LLM-generated responses [73].

Developing conversational interfaces that emphasize engagement, adaptability, and transparency is vital for trust and effectiveness. Users perceive AI content as clearer and more engaging, highlighting the need for systems that refine prompts and manage expectations. As AI companions are increasingly seen as friends, fostering user-driven value alignment is essential to mitigate biases and ensure systems resonate with user values [1, 74, 75, 39, 76].

4.3 Integration of AI and User Interfaces

AI integration into UIs enhances functionality and user experience by streamlining design and improving interaction quality. The DIDUP framework allows users to iteratively develop UI designs through LLM-generated code, supporting responsive and adaptive designs [77]. In education, AI-driven interfaces like HEDGE facilitate instructional material development, ensuring content accuracy and relevance [46].

Deep-learning models augment UI design by creating rapid mock-ups for prototyping, aligning with user expectations [32]. Automating design processes through AI reduces development time, enhancing user satisfaction. AI integration into UIs promotes intuitive, adaptive, and efficient technology, aligning with user needs and preferences. User-driven value alignment strategies engage users in correcting biased AI outputs, while concepts like "Collage" facilitate dynamic AI interactions [33, 8, 39, 21]. These advancements foster effective human-computer interactions, enhancing emotional connections and overall user experience.

5 Datasets and Usability Testing

Datasets are fundamental to the development and evaluation of AI systems, serving as the cornerstone for training models and conducting usability tests. The quality, structure, and diversity of datasets are critical in shaping the performance and applicability of AI models across various fields. This section explores the integral role of datasets in AI model training, highlighting their significance in ensuring reliable usability testing methodologies. We will discuss how well-structured datasets contribute to the robustness of AI models, setting the foundation for further examination of their necessity in subsequent subsections.

5.1 Role of Datasets in AI Model Training

Datasets are crucial for AI model development, providing the essential data needed for learning and adapting to diverse applications. The robustness and generalizability of AI models are heavily influenced by dataset quality, diversity, and structure. For example, a dataset of 329,411 instructions for grounding tasks underscores the importance of structured data in enhancing AI reasoning capabilities [69]. In specialized domains, such as manufacturing and healthcare, domain-specific datasets are vital for achieving high performance, as demonstrated by datasets derived from factory manuals and issue analysis reports [25] and multi-modal data for Alzheimer's diagnosis [15].

Challenges in dataset quality and evaluation are prevalent, with practitioners often relying on personal intuition and lacking standardized frameworks for data assessment [34]. The demand for larger datasets and comprehensive testing across various LLM architectures further complicates effective model validation [16]. Real-world datasets, such as those from road traffic and hospital management, highlight the necessity of authentic data for training models in practical applications [78].

In user interface design, datasets reflecting real-time changes are essential for evaluating intelligent user interfaces in dynamic environments [11]. The absence of standardized evaluation metrics complicates model performance comparisons, emphasizing the need for comprehensive datasets to ensure reliable usability testing [38]. Access to high-quality and diverse datasets is crucial for

enhancing AI system reliability, accuracy, and usability, promoting methodological transparency and trustworthy outputs across research domains [34, 20].

5.2 Importance of Well-Structured Datasets

Well-structured datasets are vital for effective usability testing in AI systems, offering a framework for evaluating model performance comprehensively. By categorizing datasets for specific purposes—such as safety, value alignment, and bias—they play a critical role in aligning AI models with ethical standards and user expectations [17]. Systematic assessments using structured datasets ensure AI system robustness, accuracy, and user-friendliness, capturing metrics like user interaction logs and task performance.

Current evaluation methods often focus on specific usability aspects or rely on subjective feedback, leading to incomplete assessments [30]. Structured datasets, such as those used in PubMedQA evaluations, highlight their importance in achieving reliable outcomes in specialized domains like medical applications [3]. Incorporating datasets with numerical preference scores, rankings, and textual critiques across multiple aspects provides a nuanced understanding of AI performance and user satisfaction, especially for evaluating multimodal capabilities [5].

The complexity of integrating physiological data into the adaptation process presents challenges, necessitating advanced analytical techniques [27]. Well-structured datasets are essential for accommodating intricate data types, providing a reliable basis for usability testing.

5.3 Iterative Prototyping and Design Models

Iterative prototyping and design models are crucial methodologies in usability testing, enabling continuous refinement of user interfaces and AI systems. These approaches focus on creating and evaluating prototypes, gathering user feedback from diverse sources like social media and product forums. This feedback helps identify defects, inspire enhancements, and inform new features, significantly enhancing user experience through user-centered design [79, 68, 20, 41].

A unified hub orchestrating data flows between components enhances contextual retention and decision-making capabilities, supporting the development of responsive systems [49]. Iterative prototyping allows exploration of various design alternatives, enabling early detection of usability issues. By managing large volumes of feedback through techniques like data mining and natural language processing, teams can minimize costly redesigns and enhance alignment with user expectations, leading to a more satisfactory user experience [80, 81, 79, 20, 82].

Integrating iterative prototyping into design processes ensures AI systems prioritize user-friendliness and efficiency, allowing continuous refinement based on feedback and real-time adjustments. This approach enhances user interactions and facilitates a dynamic and responsive design process, ultimately leading to AI solutions that better meet user needs and expectations [68, 8, 77, 21].

6 Interconnections and Methodologies

Category	Feature	Method
Methodological Integration Frameworks	User-Centric Design	IMAUI[11]
Enhancement Techniques in AI Systems	Domain-Specific Adjustments	OGPT[45], GatorTronGPT[4]
	Adaptive Learning Strategies	LDB[53], AWGLM[51], VSL[55]
	Data Fusion Techniques	MMLLMAIS[54]
Case Studies and Evaluation Frameworks	Human-Centric Evaluation	HEDGE[46]
	Data Integration Techniques	MMSEAN[15]

Table 1: This table provides a comprehensive overview of the methodologies applied in AI systems, categorized into three main areas: Methodological Integration Frameworks, Enhancement Techniques in AI Systems, and Case Studies and Evaluation Frameworks. Each category lists specific features and methods, highlighting the integration of user-centric design, adaptive strategies, and evaluation techniques that contribute to the advancement of AI technology across various domains.

This section explores the complex interrelations between methodologies that drive advancements in AI systems, focusing on integration frameworks that combine Large Language Models (LLMs), user

interfaces, datasets, and usability testing. Table 4 provides a detailed examination of the methodologies employed in AI systems, showcasing the diverse approaches that underpin the development and enhancement of user-centric and technically advanced solutions. These frameworks are pivotal in creating user-centric AI solutions that enhance both performance and user experience. The following subsections delve into specific methodologies within these frameworks, highlighting their role in optimizing AI systems across various sectors.

6.1 Methodological Integration Frameworks

Methodological integration frameworks are essential for advancing AI systems by merging LLMs, user interfaces (UIs), datasets, and usability testing, ensuring both technical proficiency and user-centric design. For example, integrating LLMs for knowledge embedding and multi-modal alignment significantly enhances medical diagnosis by enabling precise information retrieval [15]. In manufacturing, structured benchmarking of LLMs addresses evaluation gaps, promoting knowledge sharing and operational efficiency [25].

The Intelligent Multi-Agent User Interface (IMAUI) employs multi-agent systems, fuzzy logic, and situational control to improve user-AI interaction through adaptive interfaces [11]. In education, LLMs support personalized learning and content generation, transforming educational practices by enhancing student engagement [6]. Novel prompting strategies for LLMs in process mining enhance efficiency and accuracy [78], while frameworks categorizing dataset practitioners highlight their role fluidity, crucial for effective AI development [34].

These frameworks are vital for creating technically advanced, user-focused AI systems. Fine-tuned LLMs automate Systematic Literature Reviews (SLRs) with high factual accuracy, while retrieval-augmented LLMs, like the RETA-LLM toolkit, enhance domain-specific responses through external information retrieval, setting a new standard in academic research methodologies [57, 20].

6.2 Enhancement Techniques in AI Systems

Method Name	Integration Methodologies	Domain-Specific Applications	Adaptability and Feedback
OGPT[45]	-	Oncology-related Queries	-
AWGLM[51]	Reinforcement Learning Techniques	Web Navigation	Adaptive Learning
LDB[53]	-	-	Runtime Execution Information
MMLLMAIS[54]	Multi-source Data	Dental Diagnosis	User Feedback
VSL[55]	Traditional Algorithms	Web Element Localization	Adaptive Strategies
GatorTronGPT[4]	Prompt Tuning	Clinical Settings	Reinforcement Learning

Table 2: Overview of various AI enhancement methods, detailing their integration methodologies, domain-specific applications, and adaptability mechanisms. The table highlights the diverse approaches employed by different models, including OncoGPT, AutoWebGLM, and GatorTronGPT, to enhance functionality across medical, web navigation, and clinical domains. It underscores the importance of adaptive learning and feedback in advancing AI systems.

Table 2 provides a comprehensive summary of current enhancement techniques in AI systems, illustrating the integration of LLM strengths with traditional machine learning to improve domain-specific functionalities and adaptability. Enhancement techniques in AI systems utilize integrated methodologies that combine LLM strengths with traditional machine learning to improve functionality. Galactica's curated dataset and multi-modal capabilities exemplify a robust framework for scientific knowledge organization [83]. In medicine, fine-tuning the LLaMA-7B model with oncology dialogues enhances conversational capabilities, as evidenced by OncoGPT's performance in medical consultations [45].

AutoWebGLM's integration of multiple learning strategies improves web navigation tasks, demonstrating LLM adaptability [51]. Similarly, the LDB framework provides runtime feedback, focusing LLMs on specific code segments to enhance debugging [53]. In dentistry, multi-modal data integration enhances diagnostic accuracy [54], while combining traditional algorithms with LLM capabilities improves web element localization [55].

The GatorTronGPT model shows that prompt tuning can outperform traditional fine-tuning, highlighting the effectiveness of enhancement techniques [4]. LLM-based evaluators provide a scalable framework for multilingual task assessment, offering a comprehensive approach to AI system evaluation [43]. In AI writing tools, the 'Collage' framework integrates fragmented text management,

enhancing content creation [8]. These techniques underscore the importance of integrated methodologies in advancing adaptable, user-centric AI systems.

6.3 Case Studies and Evaluation Frameworks

Benchmark	Size	Domain	Task Format	Metric
InLegalNER[36]	1,000,000	Judicial Entity Recognition	Entity Recognition	Precision, Recall
LLM-Lobbyist[7]	485	Law	Relevance Prediction And Letter Drafting	Accuracy
LLM-PM[78]	561,470	Process Mining	Event Log Analysis	Accuracy, F1-score
ASAG-LLM[84]	1,770	Code Comprehension	Semantic Similarity Assessment	Pearson, Spearman
ChatGPT-Text-Detection[85]	10,000	Text Classification	Text Classification	Accuracy, MCC
GUIDE[86]	1,000,000	Robotic Process Automation	Next Action Prediction	Accuracy, Grounding Accuracy
PIP[87]	3,360	Copyright	Text Generation	Rouge-L
ROBUSTAPI[88]	1,208	Java Api Usage	Code Generation	API Misuse Rate, Compilation Rate

Table 3: Table ef presents a comprehensive overview of various benchmarks utilized across different domains to evaluate the performance of AI systems. The table includes details on the size, domain, task format, and specific metrics used for assessment, providing a clear framework for understanding AI capabilities in judicial, law, process mining, code comprehension, and other areas. This structured comparison facilitates the identification of key performance indicators and the applicability of AI models in diverse contexts.

Case studies and structured evaluation frameworks provide critical insights into the effectiveness of integrated AI systems across diverse domains. In medical applications, integrating multi-modal data sources enhances diagnostic accuracy and decision-making [15]. In education, LLMs for automated content creation and personalized learning exemplify AI's transformative potential in improving educational outcomes, as demonstrated by the HEDGE model [46].

In legal domains, the InLegalNER dataset showcases AI's ability to analyze complex legal entities within judicial texts, enhancing legal information retrieval [36]. In corporate environments, AI systems for predictive analytics support corporate strategy, as seen in datasets assessing legislative bills' relevance to companies [7]. Evaluation frameworks for process mining tasks provide structured methodologies for assessing AI performance, ensuring robustness and alignment with user needs [78].

Table 3 provides a detailed overview of representative benchmarks that illustrate the diverse applications and evaluation metrics of AI systems across multiple domains. These case studies and frameworks reveal essential methodologies for developing and assessing AI systems, including fine-tuned LLMs for automating systematic literature reviews and a novel retrieval framework that enhances question-answering tasks by refining context and reasoning chains [19, 20]. By leveraging these insights, researchers and developers can enhance AI solutions' effectiveness and applicability across various domains, leading to more sophisticated and impactful applications.

Feature	Methodological Integration Frameworks	Enhancement Techniques in AI Systems	Case Studies and Evaluation Frameworks
Integration Focus	Llms And Uis	Llms And ML	Multi-modal Data
Application Domain	Medical And Manufacturing	Medicine And Web	Medical And Legal
Key Benefit	User-centric Design	Improved Functionality	Enhanced Accuracy

Table 4: This table presents a comparative analysis of various methodologies employed in AI systems, highlighting their integration focus, application domains, and key benefits. It categorizes the methodologies into three main areas: Methodological Integration Frameworks, Enhancement Techniques in AI Systems, and Case Studies and Evaluation Frameworks. The comparison underscores the significance of user-centric design, improved functionality, and enhanced accuracy across different sectors such as medical, manufacturing, and legal domains.

7 Challenges and Future Directions

7.1 Challenges and Limitations

The deployment of Large Language Models (LLMs) across various domains presents significant challenges impacting their efficacy and applicability. A major issue is the high computational demand and environmental impact of training these models, which require vast datasets and resources. Current LLM benchmarks often inadequately evaluate model performance in complex scenarios [38]. The quality of input data is crucial, as poor-quality event logs can degrade LLM performance, highlighting the need for high-quality datasets [78]. Implementing Intelligent User Interfaces (IUIs) also poses challenges due to the requisite extensive training data and complex algorithms [11].

LLMs' lack of explainability is problematic, especially in fields requiring deep understanding of human behaviors. Rigorous evaluation methods and ethical considerations are crucial, particularly in educational contexts where nuanced student assessments depend on transparent AI systems [23, 84]. The black-box nature of neural networks complicates testing, validation, and the identification of critical scenarios, necessitating standardized metrics to enhance transparency.

User interaction research often overlooks LLMs as independent entities, neglecting user experience and satisfaction dimensions. The unpredictable nature of user interactions with LLM agents calls for user-centric research to better understand intents and experiences. Methodologies like user surveys and interaction log analysis can develop effective LLM systems that resonate with human interaction complexities and support personalized problem-solving [29, 68, 1].

In educational applications, LLMs struggle to consistently generate valid misconceptions and distractors, leading to inaccuracies in assessments. This inconsistency challenges their ability to differentiate correct from incorrect information. Current research stresses rigorous evaluation, reproducibility, and ethical considerations in educational contexts [23, 84]. Manual labeling for discourse data analysis limits scalability, posing obstacles to analyzing large datasets. AI reliance in coding may reduce diversity as coders converge on AI-generated suggestions.

Multilingual contexts expose further limitations, especially for low-resource languages often overlooked by existing frameworks. This gap underscores the need for inclusive AI frameworks capable of addressing diverse linguistic contexts in biomedical literature and academic research. Making expert-authored content accessible to the public and providing high-quality metadata for expanding research datasets require enhanced AI systems that simplify and summarize information while offering essential background explanations. Advanced techniques like Retrieval-Augmented Lay Language (RALL) generation and fine-tuned LLMs can improve tasks like systematic literature reviews and metadata annotation, broadening knowledge access across disciplines [89, 20, 22].

Ongoing research and innovation are crucial for enhancing LLM transparency, reliability, and adaptability. Addressing biases and inaccuracies in LLM-generated responses can improve AI effectiveness and usability across fields. Such advancements automate complex tasks like systematic literature reviews and enhance Explainable AI accessibility to non-experts. Promoting user-driven value alignment, where users correct biased outputs, ensures AI applications align with societal values, contributing to robust, reliable, and ethically sound AI applications [90, 20, 39].

7.2 Ethical Considerations and Safety

Deploying Large Language Models (LLMs) necessitates addressing ethical and safety concerns due to their societal impact. A key ethical issue is biases in LLM-generated content, which can perpetuate stereotypes and misalign with human values, affecting democratic processes [7]. The lack of interpretability in LLM outputs complicates these issues, requiring robust frameworks for transparency and accountability [6].

In therapeutic contexts, LLMs raise ethical questions about the reliability and safety of AI-driven interventions in areas like mental health. The potential for biased recommendations necessitates stringent oversight and evaluation to safeguard user well-being [12]. The risk of hallucinations in LLM-generated content requires human oversight to ensure accuracy and reliability [61].

Ethical implications extend to LLM training and operational methodologies. The substantial computational resources required for training raise environmental concerns, emphasizing the need for sustainable AI practices [13]. The dependency on high-quality datasets and prompt engineering

challenges bring ethical considerations about the manual effort and expertise needed for system development [9].

Future research must refine training techniques, enhance alignment strategies, and explore LLM implications across applications [9]. Developing governance frameworks that promote interpretability and facilitate interdisciplinary collaboration addresses LLM safety concerns [13]. Prioritizing ethical considerations and safety measures ensures LLMs positively contribute to society while mitigating risks.

7.3 Future Directions for Research and Development

Future LLM development presents numerous research avenues for enhancing capabilities and addressing challenges. Optimizing LLMs with traditional methods, developing locally installed models to reduce API costs and latency, and automating reasoning library construction for frameworks like ARALLM offer opportunities to expand LLM utility in complex scenarios [55, 91].

Refining LLM integration techniques and developing improved evaluation frameworks are essential for advancing applications in social simulations, where ethical implications must be considered [65]. Expanding human-LLM interaction modes to include image and video generation will enhance versatility [66]. Expanding datasets to include diverse legal documents and fine-tuning models for improved judicial applications is a critical research area [36].

Developing efficient, effective, and ethically aligned models is paramount. This involves exploring alignment and safety techniques in LLM outputs, ensuring ethical standards while maintaining performance [92]. Integrating advanced AI techniques into existing frameworks optimizes evaluation processes and enhances user experience, particularly in web-based interfaces [30].

Advanced prompt engineering and fine-tuning of open-source LLMs for improved GUI testing are critical for enhancing robustness and reliability [93]. Enhancing personalization in interactions, refining user intent models, and addressing user concerns about trust and capability are vital for future exploration, impacting user satisfaction and system effectiveness [1].

In healthcare, developing efficient models, addressing ethical considerations, and exploring cloud-based solutions to enhance LLM accessibility are key research priorities [3]. Focusing on these research directions allows AI systems to evolve, meeting complex user needs and leading to robust, efficient, and user-centric applications.

8 Conclusion

This survey has explored the profound influence of Large Language Models (LLMs) across multiple domains, underscoring their pivotal role in advancing natural language processing, user interfaces, datasets, usability testing, human-computer interaction, and AI systems. The synergy between LLMs and user interfaces, alongside datasets, has been instrumental in propelling AI applications forward, as evidenced by benchmarks that evaluate LLM performance in practical contexts. Methodological advancements, such as integrating LLMs with information retrieval systems, have significantly improved the accuracy of domain-specific question answering, highlighting the importance of structured frameworks in enhancing LLM outputs. These frameworks, designed to reduce inaccuracies and increase relevance, underscore the necessity of methodological rigor in refining LLM capabilities to ensure dependable AI outputs. Moreover, aligning user context with technical design in explainable AI systems has the potential to enhance human-AI collaboration, supporting a wide range of operational tasks. The interconnectedness of LLMs, user interfaces, and datasets remains crucial for developing AI systems that are both technically advanced and user-centered, adaptable to diverse contexts. Future research should focus on integrating external knowledge sources to improve verification processes and synthesize information from multiple LLM responses, paving the way for more robust and effective AI systems capable of addressing complex challenges across various domains.

References

[1] Jiayin Wang, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. Understanding user experience in large language model interactions, 2024.

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

[3] Mert Karabacak and Konstantinos Margetis. Embracing large language models for medical applications: opportunities and challenges. *Cureus*, 15(5), 2023.

[4] Mengxian Lyu, Cheng Peng, Xiaohan Li, Patrick Balian, Jiang Bian, and Yonghui Wu. Automatic summarization of doctor-patient encounter dialogues using large language model through prompt tuning, 2024.

[5] Ryan Mok, Faraaz Akhtar, Louis Clare, Christine Li, Jun Ida, Lewis Ross, and Mario Campanelli. Using ai large language models for grading in education: A hands-on test for physics, 2024.

[6] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[7] John J. Nay. Large language models as corporate lobbyists, 2023.

[8] Daniel Buschek. Collage is the new writing: Exploring the fragmentation of text and user interfaces in ai tools, 2024.

[9] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[10] Tianyu Cui, Yanling Wang, Chuanpu Fu, Yong Xiao, Sijia Li, Xinhao Deng, Yunpeng Liu, Qinglin Zhang, Ziyi Qiu, Peiyang Li, Zhixing Tan, Junwu Xiong, Xinyu Kong, Zujie Wen, Ke Xu, and Qi Li. Risk taxonomy, mitigation, and assessment benchmarks of large language model systems, 2024.

[11] Ben Khayut, Lina Fabri, and Maya Abukhana. Intelligent user interface in fuzzy environment, 2014.

[12] Zainab Iftikhar, Sean Ransom, Amy Xiao, and Jeff Huang. Therapy as an nlp task: Psychologists' comparison of llms and human peers in cbt, 2024.

[13] Shaina Raza, Oluwanifemi Bamgbose, Shardul Ghuge, Fatemeh Tavakol, Deepak John Reji, and Syed Raza Bashir. Developing safe and responsible large language model : Can we balance bias reduction and language understanding in large language models?, 2025.

[14] Ari Holtzman, Peter West, and Luke Zettlemoyer. Generative models as a complex systems science: How can we make sense of large language model behavior?, 2023.

[15] Yingjie Feng, Jun Wang, Xianfeng Gu, Xiaoyin Xu, and Min Zhang. Large language models improve alzheimer's disease diagnosis using multi-modality data, 2023.

[16] Jialin Wu, Jiangyi Deng, Shengyuan Pang, Yanjiao Chen, Jiayang Xu, Xinfeng Li, and Wenyuan Xu. Legilimens: Practical and unified content moderation for large language model services, 2024.

[17] Jaymari Chua, Yun Li, Shiyi Yang, Chen Wang, and Lina Yao. Ai safety in generative ai large language models: A survey, 2024.

[18] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[19] Tiezheng Guo, Chen Wang, Yanyi Liu, Jiawei Tang, Pan Li, Sai Xu, Qingwen Yang, Xianlin Gao, Zhi Li, and Yingyou Wen. Leveraging inter-chunk interactions for enhanced retrieval in large language model-based question answering, 2024.

[20] Teo Susnjak, Peter Hwang, Napoleon H. Reyes, Andre L. C. Barczak, Timothy R. McIntosh, and Surangika Ranathunga. Automating research synthesis with domain-specific large language model fine-tuning, 2024.

[21] Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. Coaicoder: Examining the effectiveness of ai-assisted human-to-human collaboration in qualitative analysis. *ACM Transactions on Computer-Human Interaction*, 31(1):1–38, 2023.

[22] Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. Retrieval augmentation of large language models for lay language generation, 2024.

[23] Adrian de Wynter. Awes, laws, and flaws from today's llm research, 2024.

[24] Yuncheng Hua, Lizhen Qu, and Gholamreza Haffari. Assistive large language model agents for socially-aware negotiation dialogues, 2025.

[25] Samuel Kernan Freire, Chaofan Wang, Mina Foosherian, Stefan Wellsandt, Santiago Ruiz-Arenas, and Evangelos Niforatos. Knowledge sharing in manufacturing using large language models: User evaluation and model benchmarking, 2024.

[26] Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling, 2024.

[27] Daniel Gaspar-Figueiredo. Learning from interaction: User interface adaptation using reinforcement learning, 2023.

[28] Victor Dibia. Lida: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models, 2023.

[29] Yingzhe Peng, Xiaoting Qin, Zhiyang Zhang, Jue Zhang, Qingwei Lin, Xu Yang, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. Navigating the unknown: A chat-based collaborative interface for personalized exploratory tasks, 2024.

[30] Ebenezer Agbozo. A hybrid data-driven web-based ui-ux assessment model, 2023.

[31] Céline Jost, Brigitte Le Pévédic, and Dominique Duhaut. Creating interaction scenarios with a new graphical user interface, 2012.

[32] Forrest Huang, Gang Li, Xin Zhou, John F. Canny, and Yang Li. Creating user interface mock-ups from high-level text descriptions with deep-learning models, 2021.

[33] Thu Nguyen, Alessandro Canossa, and Jichen Zhu. How human-centered explainable ai interface are designed and evaluated: A systematic survey, 2024.

[34] Crystal Qian, Emily Reif, and Minsuk Kahng. Understanding the dataset practitioners behind large language model development, 2024.

[35] Xiaoming Shi, Jie Xu, Jinru Ding, Jiali Pang, Sichen Liu, Shuqing Luo, Xingwei Peng, Lu Lu, Haihong Yang, Mingtao Hu, Tong Ruan, and Shaoting Zhang. Llm-mini-cex: Automatic evaluation of large language model for diagnostic conversation, 2023.

[36] Atin Sakkeer Hussain and Anu Thomas. Large language models for judicial entity extraction: A comparative study, 2024.

[37] Julian Coda-Forno, Marcel Binz, Jane X. Wang, and Eric Schulz. Cogbench: a large language model walks into a psychology lab, 2024.

[38] 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.

[39] Xianzhe Fan, Qing Xiao, Xuhui Zhou, Jiaxin Pei, Maarten Sap, Zhicong Lu, and Hong Shen. User-driven value alignment: Understanding users' perceptions and strategies for addressing biased and discriminatory statements in ai companions, 2025.

[40] Zhe Chen and Dunwei Wen. Accelerating and evaluation of syntactic parsing in natural language question answering systems, 2009.

[41] Zhumin Chu, Qingyao Ai, Yiteng Tu, Haitao Li, and Yiqun Liu. Pre: A peer review based large language model evaluator, 2024.

[42] Seungyoon Kim and Seungone Kim. Can language models evaluate human written text? case study on korean student writing for education, 2024.

[43] Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. Are large language model-based evaluators the solution to scaling up multilingual evaluation?, 2024.

[44] Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. Recommendation as instruction following: A large language model empowered recommendation approach, 2023.

[45] Fujian Jia, Xin Liu, Lixi Deng, Jiwen Gu, Chunchao Pu, Tunan Bai, Mengjiang Huang, Yuanzhi Lu, and Kang Liu. Oncogpt: A medical conversational model tailored with oncology domain expertise on a large language model meta-ai (llama), 2024.

[46] Jaewook Lee, Digory Smith, Simon Woodhead, and Andrew Lan. Math multiple choice question generation via human-large language model collaboration, 2024.

[47] Abbi Abdel-Rehim, Hector Zenil, Oghenejokpeme Orhobor, Marie Fisher, Ross J. Collins, Elizabeth Bourne, Gareth W. Fearnley, Emma Tate, Holly X. Smith, Larisa N. Soldatova, and Ross D. King. Scientific hypothesis generation by a large language model: Laboratory validation in breast cancer treatment, 2024.

[48] 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.

[49] Jing Guo, Nan Li, Jianchuan Qi, Hang Yang, Ruiqiao Li, Yuzhen Feng, Si Zhang, and Ming Xu. Empowering working memory for large language model agents, 2024.

[50] Aylin Gunal, Baihan Lin, and Djallel Bouneffouf. Conversational topic recommendation in counseling and psychotherapy with decision transformer and large language models, 2024.

[51] Hanyu Lai, Xiao Liu, Iat Long Iong, Shuntian Yao, Yuxuan Chen, Pengbo Shen, Hao Yu, Hanchen Zhang, Xiaohan Zhang, Yuxiao Dong, and Jie Tang. Autowebglm: A large language model-based web navigating agent, 2024.

[52] Chao Huang, Huichen Xiao, Chen Chen, Chunyan Chen, Yi Zhao, Shiyu Du, Yiming Zhang, He Sha, and Ruixin Gu. Polymetis:large language modeling for multiple material domains, 2024.

[53] Li Zhong, Zilong Wang, and Jingbo Shang. Debug like a human: A large language model debugger via verifying runtime execution step-by-step, 2024.

[54] 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.

[55] Michel Nass, Emil Alegroth, and Robert Feldt. Improving web element localization by using a large language model, 2023.

[56] Aditi Singh, Abul Ehtesham, Saifuddin Mahmud, and Jong-Hoon Kim. Revolutionizing mental health care through langchain: A journey with a large language model, 2024.

[57] Jiongnan Liu, Jiajie Jin, Zihan Wang, Jiehan Cheng, Zhicheng Dou, and Ji-Rong Wen. Reta-llm: A retrieval-augmented large language model toolkit, 2023.

[58] Zhun Yang, Adam Ishay, and Joohyung Lee. Coupling large language models with logic programming for robust and general reasoning from text, 2023.

[59] Minghao Shao, Abdul Basit, Ramesh Karri, and Muhammad Shafique. Survey of different large language model architectures: Trends, benchmarks, and challenges, 2024.

[60] Weichao Xu, Huaxin Pei, Jingxuan Yang, Yuchen Shi, Yi Zhang, and Qianchuan Zhao. Exploring critical testing scenarios for decision-making policies: An llm approach, 2024.

[61] Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. Large language model-based human-agent collaboration for complex task solving, 2024.

[62] Yuki Hou, Haruki Tamoto, and Homei Miyashita. "my agent understands me better": Integrating dynamic human-like memory recall and consolidation in llm-based agents, 2024.

[63] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. Directgpt: A direct manipulation interface to interact with large language models, 2024.

[64] Garrick Cabour, Andrés Morales, Élise Ledoux, and Samuel Bassetto. Towards an explanation space to align humans and explainable-ai teamwork, 2021.

[65] Qun Ma, Xiao Xue, Deyu Zhou, Xiangning Yu, Donghua Liu, Xuwen Zhang, Zihan Zhao, Yifan Shen, Peilin Ji, Juanjuan Li, Gang Wang, and Wanpeng Ma. Computational experiments meet large language model based agents: A survey and perspective, 2024.

[66] Jie Gao, Simret Araya Gebreegziabher, Kenny Tsu Wei Choo, Toby Jia-Jun Li, Simon Tangi Perrault, and Thomas W. Malone. A taxonomy for human-llm interaction modes: An initial exploration, 2024.

[67] Lishan Zhang, Han Wu, Xiaoshan Huang, Tengfei Duan, and Hanxiang Du. Automatic deductive coding in discourse analysis: an application of large language models in learning analytics, 2024.

[68] Yiren Liu, Pranav Sharma, Mehul Jitendra Oswal, Haijun Xia, and Yun Huang. Personaflow: Boosting research ideation with llm-simulated expert personas, 2024.

[69] Julia Rozanova, Deborah Ferreira, Krishna Dubba, Weiwei Cheng, Dell Zhang, and Andre Freitas. Grounding natural language instructions: Can large language models capture spatial information?, 2021.

[70] Yang Li, Gang Li, Luheng He, Jingjie Zheng, Hong Li, and Zhiwei Guan. Widget captioning: Generating natural language description for mobile user interface elements, 2020.

[71] Francesco Sovrano and Fabio Vitali. From philosophy to interfaces: an explanatory method and a tool inspired by achinstein's theory of explanation, 2021.

[72] Felix Härer. Conceptual model interpreter for large language models, 2023.

[73] Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. Relic: Investigating large language model responses using self-consistency, 2024.

[74] Martin Huschens, Martin Briesch, Dominik Sobania, and Franz Rothlauf. Do you trust chatgpt? – perceived credibility of human and ai-generated content, 2023.

[75] Ben Wang, Jiqun Liu, Jamshed Karimnazarov, and Nicolas Thompson. Task supportive and personalized human-large language model interaction: A user study, 2024.

[76] Sachin Pathiyan Cherumanal, Lin Tian, Futoon M. Abushaqra, Angel Felipe Magnossao de Paula, Kaixin Ji, Danula Hettiachchi, Johanne R. Trippas, Halil Ali, Falk Scholer, and Damiano Spina. Walert: Putting conversational search knowledge into action by building and evaluating a large language model-powered chatbot, 2024.

[77] Jenny Ma, Karthik Sreedhar, Vivian Liu, Sitong Wang, Pedro Alejandro Perez, and Lydia B. Chilton. Didup: Dynamic iterative development for ui prototyping, 2024.

[78] Alessandro Berti and Mahnaz Sadat Qafari. Leveraging large language models (llms) for process mining (technical report), 2023.

[79] Walid Maalej, Volodymyr Biryuk, Jialiang Wei, and Fabian Panse. On the automated processing of user feedback, 2024.

[80] Tajmilur Rahman and Yuecai Zhu. Automated user story generation with test case specification using large language model, 2024.

[81] Michael Xieyang Liu, Frederick Liu, Alexander J. Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J. Cai. "we need structured output": Towards user-centered constraints on large language model output, 2024.

[82] Kevin Moran, Ali Yachnes, George Purnell, Junayed Mahmud, Michele Tufano, Carlos Bernal-Cárdenas, Denys Poshyvanyk, and Zach H'Doubler. An empirical investigation into the use of image captioning for automated software documentation, 2023.

[83] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022.

[84] Priti Oli, Rabin Banjade, Jeevan Chapagain, and Vasile Rus. Automated assessment of students' code comprehension using llms, 2023.

[85] Niful Islam, Debopom Sutradhar, Humaira Noor, Jarin Tasnim Raya, Monowara Tabassum Maisha, and Dewan Md Farid. Distinguishing human generated text from chatgpt generated text using machine learning, 2023.

[86] Rajat Chawla, Adarsh Jha, Muskaan Kumar, Mukunda NS, and Ishaan Bhola. Guide: Graphical user interface data for execution, 2024.

[87] Weijie Zhao, Huajie Shao, Zhaozhuo Xu, Suzhen Duan, and Denghui Zhang. Measuring copyright risks of large language model via partial information probing, 2024.

[88] Li Zhong and Zilong Wang. Can chatgpt replace stackoverflow? a study on robustness and reliability of large language model code generation, 2024.

[89] Shiwei Zhang, Mingfang Wu, and Xiuzhen Zhang. Utilising a large language model to annotate subject metadata: A case study in an australian national research data catalogue, 2023.

[90] Philip Mavrepis, Georgios Makridis, Georgios Fatouros, Vasileios Koukos, Maria Margarita Separdani, and Dimosthenis Kyriazis. Xai for all: Can large language models simplify explainable ai?, 2024.

[91] Junjie Wang, Dan Yang, Binbin Hu, Yue Shen, Wen Zhang, and Jinjie Gu. Know your needs better: Towards structured understanding of marketer demands with analogical reasoning augmented llms, 2024.

[92] 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.

[93] Zhe Liu, Chunyang Chen, Junjie Wang, Mengzhuo Chen, Boyu Wu, Xing Che, Dandan Wang, and Qing Wang. Make llm a testing expert: Bringing human-like interaction to mobile gui testing via functionality-aware decisions, 2023.

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.