Evaluation of Large Language Models for Natural Language Processing: A Survey

www.surveyx.cn

Abstract

The evaluation of large language models (LLMs) is pivotal for advancing natural language processing (NLP) applications and ensuring model reliability, particularly in high-stakes domains like healthcare and software engineering. This survey paper systematically examines the methodologies employed to assess LLM performance, focusing on metrics and benchmarks that evaluate accuracy, efficiency, contextual understanding, and ethical considerations. The study highlights the transformative impact of LLMs across various sectors, including healthcare, finance, and academia, while identifying persistent challenges such as biases, data contamination, and limitations in current evaluation frameworks. By integrating insights from diverse benchmarks, this survey underscores the necessity for comprehensive evaluation strategies that address ethical and safety concerns, ensuring alignment with societal norms. Future directions emphasize the expansion and diversification of benchmarks, integration with external knowledge, and the development of advanced evaluation metrics to enhance LLM adaptability and performance. The findings advocate for a systematic evaluation approach to guide the refinement and deployment of LLMs, ensuring their effective application in real-world scenarios and fostering responsible AI development.

1 Introduction

1.1 Purpose and Significance of LLM Evaluation

Evaluating large language models (LLMs) is essential for advancing natural language processing (NLP) applications and ensuring their reliability, especially in high-stakes domains like healthcare, where risks such as factual inaccuracies, biases, and toxicity are prevalent [1, 2, 3]. Rigorous evaluation is necessary to identify and mitigate these issues, particularly as LLMs are increasingly used to enhance clinical decision-making and improve healthcare delivery through the processing of electronic health records (EHRs).

In software engineering, LLM evaluation is crucial for applications such as automated code generation and unit testing, which improve model performance and reliability [4]. The ethical implications of LLM deployment, especially in conversational agents like ChatGPT, underscore the need for comprehensive evaluation to align with societal norms and user expectations, including addressing gender biases in educational contexts related to STEM fields [5, 6].

A systematic evaluation framework is vital for identifying LLM competencies—reasoning, knowledge, reliability, and safety—while addressing challenges posed by traditional evaluation methods that struggle to keep pace with diverse real-world applications. Comprehensive assessment frameworks and benchmarks provide insights into LLM performance across various tasks, essential for mitigating risks and enhancing practical utility in both academic and industrial settings [7, 8, 9]. By illuminating their capabilities and limitations, evaluations guide researchers and practitioners in refining and deploying LLMs effectively.

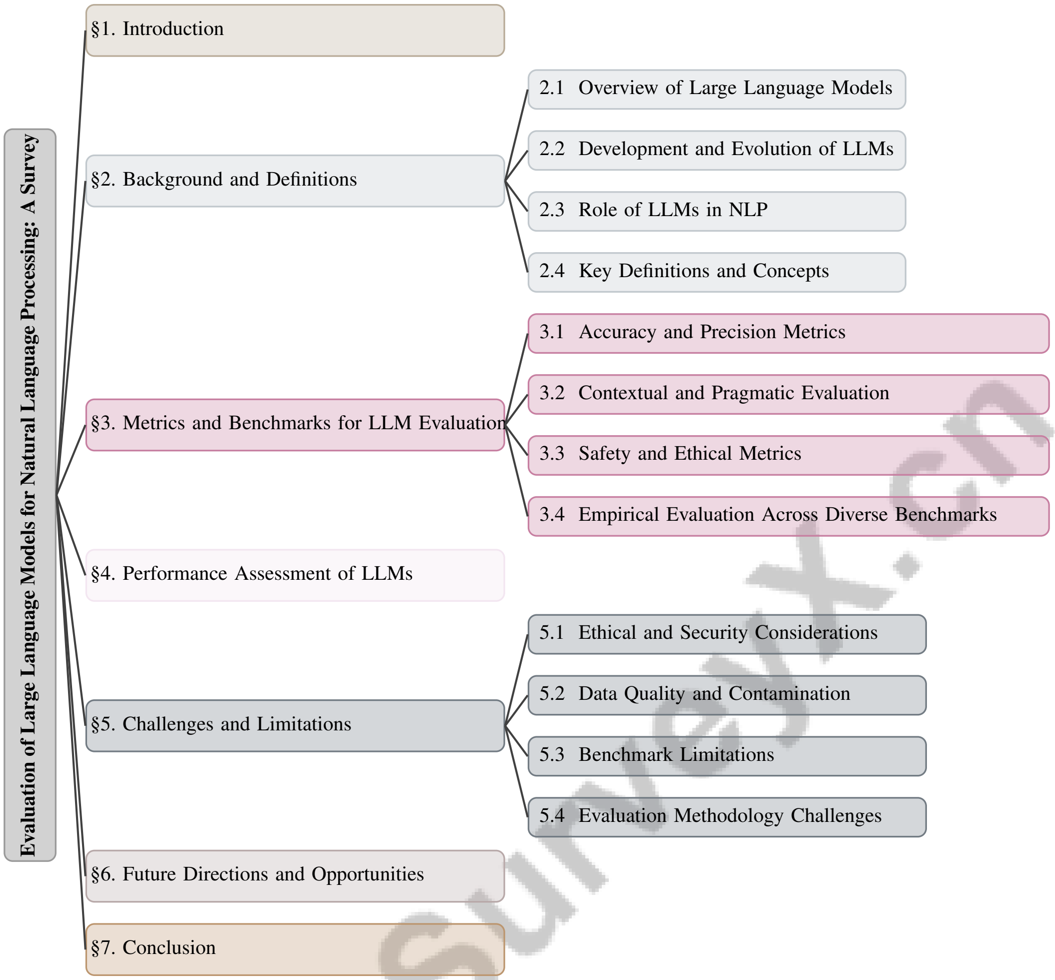
Evaluation of Large Language Models for Natural Language Processing: A Survey
§1. Introduction
§2. Background and Definitions
2.1 Overview of Large Language Models
2.2 Development and Evolution of LLMs
2.3 Role of LLMs in NLP
2.4 Key Definitions and Concepts
§3. Metrics and Benchmarks for LLM Evaluation
3.1 Accuracy and Precision Metrics
3.2 Contextual and Pragmatic Evaluation
3.3 Safety and Ethical Metrics
3.4 Empirical Evaluation Across Diverse Benchmarks
§4. Performance Assessment of LLMs
§5. Challenges and Limitations
5.1 Ethical and Security Considerations
5.2 Data Quality and Contamination
5.3 Benchmark Limitations
5.4 Evaluation Methodology Challenges
§6. Future Directions and Opportunities
§7. Conclusion

Figure 1: chapter structure

1.2 Growing Importance of LLMs in AI

The rise of large language models (LLMs) like ChatGPT represents a significant evolution in artificial intelligence (AI), profoundly impacting various sectors. These models have driven advancements in natural language processing (NLP), highlighting their critical role in AI applications [5]. In higher education, LLMs are transforming student engagement with digital writing and assessments, reflecting their growing influence in academic contexts.

In healthcare, the integration of LLMs into clinical diagnosis is a promising area that enhances medical decision-making processes [2]. Their application in health-related question-answering for neglected tropical diseases further illustrates their potential to address global health challenges.

In financial technology, LLMs are employed for sentiment analysis and named entity recognition, where accuracy and reliability are paramount. In software development, open-source LLMs are increasingly adopted for their benefits in data privacy and performance, facilitating customized code generation and troubleshooting. However, even advanced models like GPT-4 can produce code with high rates of API misuse, necessitating careful evaluation to ensure reliability and robustness in real-world applications [10, 8, 9].

The rapid expansion of biomedical literature also calls for efficient automated information extraction methods, where LLMs contribute significantly to BioNLP. Their transformative impact on AI-driven business decision-making is evident in critical areas like mergers and acquisitions and stock market investments, where advanced reasoning and knowledge integration enhance strategic analysis and forecasting [11, 8, 9].

LLMs' integration across sectors underscores their growing significance in AI, driving innovation and revolutionizing industry practices. Their contributions to natural language processing tasks, domain-specific applications, and advanced techniques like Retrieval-Augmented Generation (RAG) necessitate robust evaluation frameworks to assess their competencies in reasoning, knowledge reliability, and safety [11, 12, 9, 13]. As the field evolves, LLMs' role in enhancing AI capabilities remains indispensable.

1.3 Applications of LLMs

Large language models (LLMs) exhibit remarkable versatility across various applications, significantly impacting multiple sectors. In mental health, LLMs analyze natural language inputs to comprehend and predict conditions like anxiety and depression, enhancing mental health assessment and intervention strategies [14].

In medicine, LLMs facilitate clinical summarization, improving the efficiency and accuracy of clinical documentation and decision-making [15]. They also play a pivotal role in evaluating clinical diagnosis capabilities, as demonstrated by the CLIBENCH framework, which comprehensively assesses LLM performance in clinical settings [2].

In software engineering, LLMs enhance code generation, summarization, and bug detection processes, thereby boosting productivity [4]. Their adaptability is further exemplified in robotics, where they synthesize robot policy code from natural language commands, showcasing their potential in automating complex tasks [16].

In academia, LLMs like ChatGPT assist researchers in navigating vast information landscapes and generating scholarly content, thereby contributing to research productivity and citation practices [5]. In finance, domain-specific LLMs such as BloombergGPT cater to unique financial analysis demands, providing insights that inform strategic market decisions [17]. Moreover, generative AI applications in business process management leverage LLMs to automate repetitive tasks and enhance operational efficiency [18].

The extensive applications of LLMs across sectors such as mental health, medicine, software engineering, academia, and finance underscore their transformative influence on operational efficiency, decision-making processes, and innovation. Their remarkable performance in tasks like disorder detection in mental health and integration into various evaluation frameworks illustrates their potential across multiple industries [8, 9, 19]. As LLM technology evolves, its applications are likely to expand further, driving advancements across diverse fields.

1.4 Structure of the Survey

This survey is meticulously organized to provide a comprehensive evaluation of large language models (LLMs) within the natural language processing (NLP) domain. It begins with an **Introduction** that details the purpose and significance of LLM evaluations, followed by an exploration of their growing importance in AI and diverse applications across sectors such as healthcare, software engineering, and finance. The introduction concludes with this subsection, outlining the survey's structure.

The second section, **Background and Definitions**, offers an in-depth overview of LLMs, tracing their development, evolution, and pivotal role in NLP. Key definitions and concepts critical to understanding LLMs and their evaluation are clarified.

In **Metrics and Benchmarks for LLM Evaluation**, various metrics and benchmarks are discussed, focusing on their role in assessing LLM accuracy, efficiency, contextual understanding, and ethical considerations. This section underscores the necessity of diverse benchmarks for comprehensive evaluation.

The fourth section, **Performance Assessment of LLMs**, delves into empirical analyses of LLM performance across different NLP tasks, highlighting strengths, weaknesses, and comparisons with traditional models. It also examines the impact of model architecture and fine-tuning on performance.

The section titled **Challenges and Limitations** provides a comprehensive analysis of the difficulties encountered in evaluating LLMs, addressing key issues such as ethical considerations, data contamination, and the inadequacies of current evaluation frameworks. Rapid advancements in LLM performance complicate the evaluation process, as traditional NLP tasks become insufficient. The diverse applications of LLMs in real-world scenarios outpace existing evaluation tasks, emphasizing the need for robust benchmarks and a structured approach to assess core competencies like reasoning, knowledge, reliability, and safety, while also addressing societal implications and potential risks associated with LLM deployment [7, 8, 9].

The survey explores **Future Directions and Opportunities**, identifying potential areas for advancing LLM evaluation methods, developing new benchmarks, and addressing ethical issues. The significance of ongoing evaluation in adapting to rapid technological advancements, particularly concerning LLMs, is emphasized due to the challenges posed by traditional evaluation methods that fail to keep pace with exceptional performance and diverse applications. This necessitates developing new benchmarks and evaluation frameworks that effectively assess key competencies such as reasoning, knowledge, reliability, and safety, ensuring that evaluation practices evolve alongside these technologies [20, 21, 22, 5, 9].

Finally, the **Conclusion** synthesizes key findings and insights, reinforcing the critical role of systematic evaluation in advancing LLM technology and its applications in NLP. Each section is designed to build upon the previous, providing a coherent and thorough examination of LLM evaluation.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Overview of Large Language Models

Large Language Models (LLMs) represent a significant leap in artificial intelligence, excelling in complex NLP tasks like text generation, translation, and summarization. Models such as GPT-3.5 and GPT-4 leverage extensive datasets and sophisticated deep learning architectures to produce human-like text [23]. As these models scale, they exhibit emergent capabilities that enhance task performance beyond what smaller models can achieve [24].

LLM architectures, including encoder-only, encoder-decoder, and decoder-only models, have advanced contextual comprehension and reasoning, critical for generating coherent outputs [4]. Despite these advancements, LLMs still struggle with multi-step logic and complex reasoning tasks, as shown by benchmarks like BIG-bench [25]. In healthcare, models like GatorTron, trained on vast clinical datasets, underscore the importance of standardized evaluation frameworks for reliable clinical applications [2]. The Tree of Thoughts framework further enhances inference capabilities by exploring coherent text units, vital for intricate reasoning tasks [23].

Specialized benchmarks, such as the LSAT Logic Games and README++, evaluate logical reasoning and multilingual readability, respectively, facilitating comprehensive model assessments across domains [26, 2]. In software engineering, LLMs excel in generating and repairing code, synthesizing functions from natural language descriptions, and performing code repairs without prior training on security fixes [10]. Benchmarks like CodeT5 standardize program understanding and generation evaluations, highlighting LLM applications in software testing [26].

The ability of LLMs to learn from vast datasets and perform a wide range of tasks with high accuracy is transformative. The evolution of AI technologies, particularly generative models like ChatGPT, enhances applications across various sectors, revolutionizing industries through improved decision-making and problem-solving capabilities [27, 24, 5, 22].

2.2 Development and Evolution of LLMs

The evolution of LLMs has been driven by scalable Transformer-based architectures, advancing NLP [28]. This progress is evident in software testing, where closed-source models have historically limited benchmark diversity [29]. LLMs have evolved from systems reliant on extensive labeled datasets to sophisticated models incorporating strategic planning and exploration, addressing inefficiencies in complex problem-solving [30].

Innovations like the SELF-INSTRUCT framework facilitate the semi-automatic generation of instructions and input-output pairs, overcoming limitations of existing instruction data [31]. This approach exemplifies efforts to enhance LLM adaptability and learning efficiency for diverse applications.

Incorporating programming language characteristics into LLM design, as seen with CODET's dual execution agreement, evaluates outputs against generated test cases, improving model performance in code generation tasks [32, 33]. The relationship between model scale and emergent capabilities underscores the importance of model size in achieving performance breakthroughs [34]. However, challenges in maintaining efficiency, especially in mathematics, code generation, and multilingual understanding, necessitate ongoing innovation and evaluation [35].

In multilingual processing, the lack of comprehensive benchmarks for languages like Traditional Chinese highlights the need for new benchmarks [36]. Datasets like README++ exemplify efforts to provide robust resources for evaluating model performance across languages, underscoring the importance of diverse benchmarks [26].

2.3 Role of LLMs in NLP

LLMs have significantly influenced NLP, advancing language understanding and generation. They enable direct generation of control code from natural language instructions in robotics, enhancing operational efficiency [16]. In NLI, LLMs have improved ethical reasoning, allowing nuanced decision-making [37].

LLMs' integration into NLP tasks highlights their potential in areas requiring precise semantic understanding. Generative LLMs outperform traditional encoder-based models in Semantic Textual Similarity tasks, especially in domain-specific contexts [38]. However, vulnerabilities such as selection bias in MCQs emphasize the need for robust evaluation frameworks [39].

The unpredictability of emergent abilities in LLMs poses design challenges, as these capabilities can introduce unforeseen risks [34]. Comprehensive benchmarks, like the LSAT Logic Games, are necessary to assess reasoning capabilities [40]. Training on task-relevant data during pretraining raises concerns about the validity of model comparisons and claims of emergent capabilities [41].

LLM integration in autonomous systems addresses the limitations of traditional machine learning methods, which rely on historical data that may not encompass unprecedented scenarios, such as those in autonomous driving [42]. This integration underscores LLMs' transformative potential in enhancing decision-making across various fields. LLMs have also improved performance in software engineering tasks, with decoder-only architectures demonstrating notable effectiveness [4].

The benchmark developed by Wang et al. evaluates LLMs based on factual correctness, logical reasoning, non-toxicity, and fairness, which are vital for safe deployment [1]. This comprehensive framework is essential for the reliable integration of LLMs into high-stakes applications without compromising ethical standards.

LLMs' influence in NLP is profound, with their integration into diverse tasks highlighting their significance in ongoing research. As LLM technology advances, it promises to enhance NLP by broadening language comprehension and processing capabilities. This evolution aims to improve traditional evaluation benchmarks and foster innovation across sectors, including specialized domains like law, where LLMs facilitate complex reasoning and knowledge retrieval [11, 9].

2.4 Key Definitions and Concepts

Evaluating LLMs requires understanding critical terms and concepts integral to their development and assessment. 'Readability assessment' evaluates text comprehension across different audiences, crucial for multilingual models processing diverse linguistic contexts [43]. Robust evaluation frameworks are essential for effective task handling across languages.

'Pragmatic competence' refers to an LLM's ability to grasp context-dependent meanings and conversational implicatures, assessed through benchmarks focusing on specific languages like Korean to ensure appropriate interactions [44]. In software engineering, 'functional correctness' pertains to the accuracy of generated code, verified through unit tests, emphasizing the importance of evaluating code generation tasks [45].

Ethical considerations in AI deployment highlight the responsible use of LLMs, particularly in research and educational settings, where maintaining 'academic integrity' and ethical AI integration are crucial for enhancing learning experiences. 'Human evaluation' is vital in LLM assessments, especially in standardized methods and healthcare applications, ensuring model outputs align with intended goals [21].

’Self-knowledge evaluation’ assesses LLMs’ ability to generate and verify their outputs, providing insights into their capabilities and limitations [46]. Adaptive sampling techniques enhance model assessment efficiency by dynamically selecting optimal methods based on benchmark characteristics [47]. The FLASK benchmark addresses the challenge of evaluating LLMs on their ability to follow diverse user instructions requiring multiple skills, highlighting the insufficiency of traditional single-metric evaluations [48].

The benchmark by Dominguez-Olmedo et al. aims to provide a fair evaluation framework by adjusting for training effects on test tasks, thereby improving performance assessment reliability [41]. In multilingual contexts, efforts to evaluate language models in tasks like contextual question answering and summarization have addressed the lack of comprehensive benchmarks for languages like Traditional Chinese [36].

The faithfulness of summaries generated by LLMs is critical, focusing on whether the summary accurately reflects the source document [49]. This aspect is essential for ensuring the reliability of model-generated content across applications.

The key definitions and concepts outlined in the literature establish a comprehensive framework for understanding and evaluating LLMs. This framework clarifies core competencies necessary for effective assessment—including reasoning, knowledge, reliability, and safety—informing the development of tailored evaluation strategies that address the diverse applications and challenges presented by LLMs in academic and real-world contexts [8, 50, 51, 9].

3 Metrics and Benchmarks for LLM Evaluation

A structured framework for evaluating large language models (LLMs) is vital for a comprehensive assessment of their performance through various metrics and benchmarks. This section examines critical evaluation criteria, emphasizing their roles in measuring accuracy, precision, context, and ethical considerations. By categorizing these metrics, we enhance our understanding of LLM performance and its implications across diverse applications. Figure 2 illustrates the hierarchical structure of metrics and benchmarks used in evaluating LLMs, categorizing them into accuracy and precision metrics, contextual and pragmatic evaluation, safety and ethical metrics, and empirical evaluation across diverse benchmarks. Each category is further divided into specific metrics, benchmarks, and evaluation frameworks, highlighting the comprehensive approach needed to assess LLM performance effectively. The following subsection focuses on accuracy and precision metrics, foundational elements in assessing LLM capabilities.

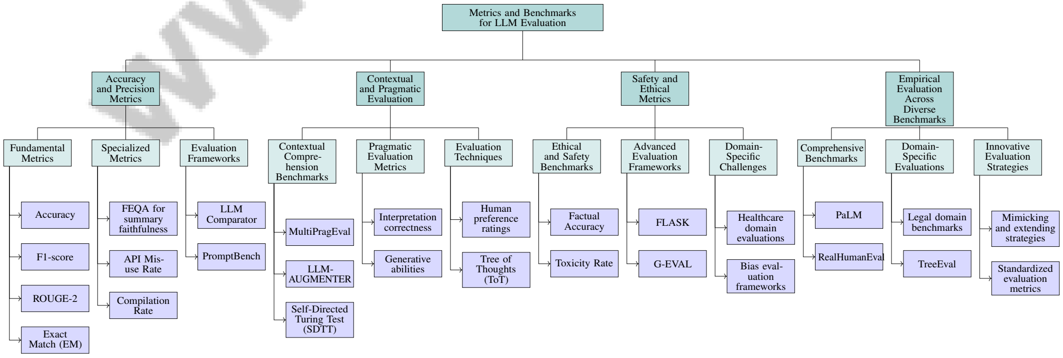
Metrics and Benchmarks for LLM Evaluation
Accuracy and Precision Metrics
Contextual and Pragmatic Evaluation
Safety and Ethical Metrics
Empirical Evaluation Across Diverse Benchmarks
Fundamental Metrics
Specialized Metrics
Evaluation Frameworks
Contextual Compre-hension Benchmarks
Pragmatic Evaluation Metrics
Evaluation Techniques
Ethical and Safety Benchmarks
Advanced Evaluation Frameworks
Domain-Specific Challenges
Comprehensive Benchmarks
Domain-Specific Evaluations
Innovative Evaluation Strategies
Accuracy
F1-score
ROUGE-2
Exact Match (EM)
FEQA for summary faithfulness
API Mis-use Rate
Compilation Rate
LLM Comparator
PromptBench
MultiPragEval
LLM-AUGMENTER
Self-Directed Turing Test (SDTT)
Interpretation correctness
Generative abilities
Human preference ratings
Tree of Thoughts (ToT)
Factual Accuracy
Toxicity Rate
FLASK
G-EVAL
Healthcare domain evaluations
Bias eval-uation frameworks
PaLM
RealHumanEval
Legal domain benchmarks
TreeEval
Mimicking and extending strategies
Standardized evaluation metrics

Figure 2: This figure illustrates the hierarchical structure of metrics and benchmarks used in evaluating large language models (LLMs), categorizing them into accuracy and precision metrics, contextual and pragmatic evaluation, safety and ethical metrics, and empirical evaluation across diverse benchmarks. Each category is further divided into specific metrics, benchmarks, and evaluation frameworks, highlighting the comprehensive approach needed to assess LLM performance effectively.

3.1 Accuracy and Precision Metrics

LLM evaluation relies on diverse metrics assessing accuracy and precision across tasks. Metrics like Accuracy and F1-score are fundamental for evaluating correctness and robustness, particularly in code synthesis tasks, where correctness is measured against test cases [26]. These metrics are crucial for high-stakes applications, providing insights into the model's ability to generate accurate outputs.

In text generation tasks, ROUGE-2 and Exact Match (EM) metrics evaluate the quality and coherence of generated text, ensuring outputs are contextually relevant and faithful to source material. FEQA, for instance, assesses summary faithfulness, underscoring the need for models to maintain factual accuracy [49]. This highlights the importance of precision in text generation, where information integrity is paramount.

The LLM Comparator method enhances evaluation by visually connecting aggregated metrics with individual examples, offering deeper insights into model behavior [52]. This approach enables targeted improvements in model design and deployment. Figure 3 illustrates the hierarchical classification of LLM evaluation metrics, segmented into accuracy, text generation, and code evaluation categories. Each category includes specific metrics used to assess the performance and reliability of large language models in various tasks.

In software engineering, metrics like API Misuse Rate and Compilation Rate are vital for assessing the reliability of generated code. Studies indicate a significant portion of code generated by LLMs contains API misuses, emphasizing the need for comprehensive evaluation frameworks assessing practical implications of code generation [10, 53]. These metrics highlight the importance of evaluating LLM outputs in terms of applicability and precision.

The PromptBench framework incorporates metrics like Accuracy and F1-score to evaluate model performance against adversarial prompts, ensuring robustness against various challenges [26]. Selecting metrics tailored to application domains ensures a comprehensive assessment of core competencies—reasoning, knowledge, reliability, and safety—while addressing limitations of traditional evaluation methods [51, 9]. This diverse array of metrics provides a robust framework for effective LLM deployment.

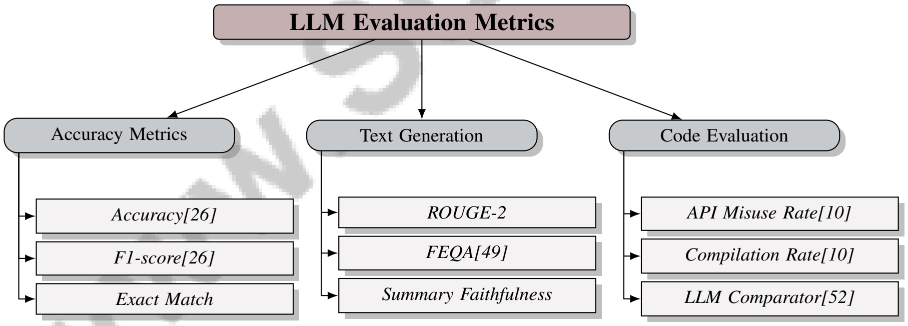

Figure 3: This figure illustrates the hierarchical classification of LLM evaluation metrics, segmented into accuracy, text generation, and code evaluation categories. Each category includes specific metrics used to assess the performance and reliability of large language models in various tasks.

3.2 Contextual and Pragmatic Evaluation

Evaluating LLMs based on contextual understanding and pragmatic use involves assessing their ability to comprehend and generate contextually appropriate language. Benchmarks like MultiPragEval assess pragmatic comprehension across multiple languages, emphasizing the importance of evaluating LLMs in diverse linguistic contexts [54]. This evaluation is crucial for tasks requiring nuanced language understanding.

The LLM-AUGMENTER framework enhances contextual evaluation by integrating external knowledge, reducing hallucinations and improving output accuracy [55]. The Self-Directed Turing Test (SDTT) provides dynamic evaluation through dialogues mimicking human interactions, allowing comprehensive performance assessments in contextually rich scenarios [56].

Metrics for pragmatic evaluation focus on model interpretation correctness and generative abilities [44]. These metrics reveal how well LLMs interpret and generate language aligning with human expectations. Chain-of-thought prompting significantly enhances reasoning abilities, demonstrating the importance of structured prompting for contextual understanding [57].

Human preference ratings, as seen in evaluations of models like InstructGPT, provide insights into LLMs' contextual understanding by comparing outputs with human expectations [58]. Frameworks like the Tree of Thoughts (ToT) integrate search algorithms with language model capabilities to enhance problem-solving efficiency [30].

Evaluating LLMs in pragmatic contexts also includes assessing their ability to generate ethically sound explanations, comparing generated outputs against baseline methods using metrics for logical validity and moral foundation classification accuracy [37]. This emphasizes the need for models to produce contextually appropriate and ethically aligned outputs.

Evaluating LLMs based on contextual understanding and pragmatic use is integral for their effectiveness across applications. Utilizing diverse benchmarks and evaluation methodologies enables researchers to gain a nuanced understanding of LLM performance in rich contexts. This comprehensive approach addresses inadequacies of traditional NLP evaluation tasks and aligns with evolving LLM applications in real-world settings, identifying four core competencies—reasoning, knowledge, reliability, and safety—defined with specific benchmarks and metrics [7, 9].

3.3 Safety and Ethical Metrics

Evaluating LLMs from a safety and ethical perspective is crucial given their extensive deployment and potential to generate harmful or biased outputs. A comprehensive framework incorporating diverse metrics is necessary to assess ethical alignment and safety. Metrics like Factual Accuracy and Toxicity Rate measure factual information accuracy and toxic content prevalence in LLM outputs, reflecting a dual focus on factual correctness and ethical considerations [1].

The FLASK benchmark introduces metrics like Factuality and Logical Correctness, assessing both the correctness of model responses and the interpretability of reasoning processes [48]. These metrics ensure models produce responses that are both correct and ethically sound.

Evaluating LLMs in sensitive domains, such as healthcare, presents unique challenges due to the complexity of medical texts and current limitations in evaluation metrics. The necessity for expert human evaluation in these contexts is resource-intensive, highlighting the need for efficient evaluation methods [15]. The lack of standardized bias evaluation approaches can lead to inconsistent results, underscoring the importance of robust frameworks for assessing representational biases [6].

The G-EVAL framework demonstrates superior performance in evaluating natural language generation outputs, achieving higher correlation with human judgments than existing evaluators [59]. This highlights the potential for advanced evaluation methods to enhance accuracy and reliability in ethical and safety contexts.

Innovative benchmarks combining semi-synthetic and human-crafted datasets rigorously test LLMs across multiple safety dimensions and interaction formats [60]. These benchmarks are essential for evaluating LLMs' robustness in handling diverse and sensitive interactions, ensuring adherence to ethical standards.

Evaluating LLMs from a safety and ethical perspective is multifaceted, requiring diverse metrics to ensure reliability, fairness, and alignment with societal norms. By incorporating various evaluation metrics, researchers can gain deeper insights into the risks associated with deploying LLMs. This multifaceted approach enhances understanding of LLM performance across applications, such as reasoning, ethics, and healthcare, while aiding in identifying and mitigating safety concerns. Ultimately, this rigorous evaluation framework supports the advancement of AI systems that are safer and more ethically aligned with societal values [8, 9, 21].

3.4 Empirical Evaluation Across Diverse Benchmarks

Empirical evaluation of LLMs necessitates diverse benchmarks to assess performance across a wide array of tasks. Table 1 presents a comprehensive overview of diverse benchmarks used to empirically evaluate large language models (LLMs) across multiple domains and task formats. The inclusion of

Benchmark	Size	Domain	Task Format	Metric
RHE[61]	888	Programming	Coding Tasks	Acceptance Rate, Task Completion Time
LJP[11]	100,000	Legal Judgment Prediction	Classification	Accuracy, F1-score
NEW-BM[62]	10,000	Data Analysis	Classification	Accuracy, F1-score
NLBench[63]	500,000	Text Classification	Sentiment Analysis	F1-score
AU-Benchmark[64]	10,000	Educational Assessment	Question Answering	Accuracy, F1-score
KG-LLM[12]	10,000	Knowledge Graph Construction	Entity And Relation Extraction	F1-Score, Accuracy
InterCode[65]	168	Programming	Example-based Code Generation	Pass@K
MPE[54]	1,200	Pragmatics	Multiple-choice Questions	Accuracy, F1-score

Table 1: This table provides an overview of various benchmarks used for evaluating large language models (LLMs) across different domains. It details the benchmark name, dataset size, domain of application, task format, and evaluation metrics employed. The table highlights the diversity and specificity of tasks used to assess LLM performance, facilitating a comprehensive understanding of their capabilities and limitations.

over 150 tasks in benchmarks like PaLM allows for robust empirical evaluation across various aspects of language processing, providing insights into LLM capabilities and limitations [66]. This extensive range ensures models are assessed on diverse linguistic and cognitive challenges, facilitating a holistic understanding of their performance.

The RealHumanEval framework integrates human feedback into the evaluation process, offering a nuanced assessment of LLM capabilities compared to traditional benchmarks [61]. This highlights the importance of incorporating human judgment to capture the subtleties of model outputs, particularly in tasks requiring high contextual understanding.

In the legal domain, benchmarks designed for evaluating LLM competency in legal judgment prediction enable comparisons of models in real-world legal tasks [11]. This underscores the necessity of domain-specific evaluations to ensure LLMs navigate complex legal scenarios effectively.

The introduction of novel datasets and evaluation metrics enhances the representation of real-world tasks, ensuring the evaluation process remains relevant to real-world applications [62]. The TreeEval benchmark introduces a multi-task evaluation framework that allows simultaneous assessment of model performance across various natural language processing tasks [63]. This framework facilitates a comprehensive understanding of LLM capabilities, highlighting strengths and weaknesses in different contexts.

Challenges in categorizing generative AI models and establishing standardized evaluation metrics are addressed by Bandi et al., emphasizing the need for consistent frameworks to assess LLM performance effectively [24]. This standardization is crucial for reliable comparisons and advancements in model development.

Furthermore, the benchmark introduced by Ying et al. employs novel strategies such as mimicking and extending to generate samples at varying cognitive levels, enhancing evaluation reliability [64]. These strategies contribute to a robust evaluation process, ensuring LLMs are tested across a spectrum of challenges.

A comprehensive empirical evaluation of LLMs across diverse benchmarks is crucial for understanding their capabilities and limitations, as existing methods often fall short in addressing advancements and varied applications. This evaluation should encompass core competencies—reasoning, knowledge, reliability, and safety—while considering qualitative and quantitative performance aspects in real-world scenarios. By employing a wide array of evaluation frameworks, researchers can gain a deeper understanding of LLM performance, guiding the development and deployment of more effective and reliable AI systems [8, 7, 50, 9, 67].

4 Performance Assessment of LLMs

4.1 Comparison with Traditional Models

The evolution from traditional NLP models to large language models (LLMs) showcases significant advancements, particularly in specialized tasks. Transformer-based models, like the one achieving

83.8% accuracy on the MathQA-Python dataset, exemplify LLMs' proficiency in code synthesis [26]. In clinical NLP, models such as GatorTron surpass BioBERT and ClinicalBERT, highlighting LLMs' domain-specific strengths. However, models like ChatGPT and LLaMa display performance variability across benchmarks, especially in non-English contexts, indicating challenges in maintaining consistency across languages and tasks [1]. Neural abstractive summarization models evaluated with the FEQA metric demonstrate the critical role of precision in summarization [49].

Despite progress, LLMs face hurdles in complex domains like clinical decision-making, necessitating further refinement. In multimodal applications, models such as mPLUG-Owl excel in instruction understanding and visual reasoning through a two-stage training approach integrating visual knowledge and language capabilities. This innovation enhances performance in tasks like multi-turn dialogues and multi-image correlation, underscoring LLMs' potential in addressing real-world challenges, including vision-only document comprehension [25, 67].

Figure 4 illustrates the key advancements in large language models (LLMs), highlighting their performance in natural language processing tasks, multimodal applications, and the development of comprehensive evaluation frameworks. Addressing evaluation gaps requires comprehensive frameworks that tackle traditional NLP inadequacies and adapt to LLMs' diverse applications. Emphasizing core competencies—reasoning, knowledge, reliability, and safety—will aid in developing robust benchmarks to enhance LLM assessment and guide future innovations while mitigating deployment risks [8, 9].

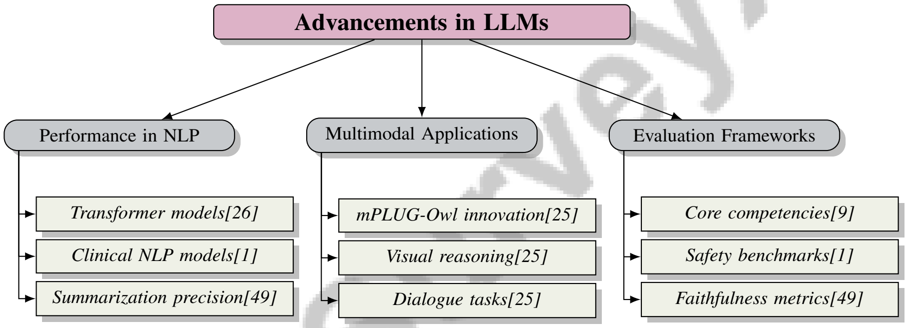

Figure 4: This figure illustrates the key advancements in large language models (LLMs), highlighting their performance in natural language processing tasks, multimodal applications, and the development of comprehensive evaluation frameworks.

4.2 Strengths in Specific Domains

LLMs excel in specific domains, notably in complex reasoning, code generation, and domain-specific knowledge tasks. Models like GPT-4 outperform fine-tuned models in problem-solving and logical deduction, with chain-of-thought prompting enhancing interpretability [12, 68]. In code generation, frameworks such as CODET and Mixtral demonstrate significant performance improvements, highlighting LLMs' potential in driving software development innovation [32, 35].

In multimodal tasks, advanced frameworks address hallucinations in Visual Language Models (VLLMs) through preference fine-tuning and modularized training, enhancing image-text integration. The POVID and mPLUG-Owl frameworks exemplify two-stage alignment strategies that improve visual and textual interaction, enhancing model accuracy and reliability [25, 19, 69, 50, 67].

LLMs also perform well in specific linguistic contexts, with models like IndicBERT surpassing multilingual models on various tasks [70]. However, performance variations across subfields highlight both strengths and weaknesses in knowledge capabilities [67]. LLMs' capabilities in legal judgment prediction and code generation demonstrate their potential to enhance advancements across fields. Developing comprehensive evaluation benchmarks is crucial for assessing LLM competencies in real-world scenarios, ensuring strengths in reasoning, knowledge application, reliability, and safety are effectively leveraged [11, 9]. Continuous research is necessary to address limitations and fully utilize LLMs' capabilities across diverse applications.

4.3 Challenges in Complex and Low-Resource Tasks

LLMs encounter significant challenges in complex and low-resource tasks, where traditional metrics often fail to capture nuanced performance variations. As task complexity increases, LLM performance declines, highlighting the limitations of static evaluation methods [71]. Complex tasks require evolving frameworks to accurately assess capabilities [72]. Models like GPT-4 struggle with coherent interactions in extended dialogues, especially in low-resource languages where performance lags behind supervised counterparts [56, 73].

Security vulnerabilities further complicate LLM outputs, with a significant portion of generated code identified as vulnerable, raising concerns about reliability in real-world applications [74]. The variability in quality of fixes for complex security issues indicates challenges in generating robust solutions [75]. LLMs also struggle with tasks requiring specialized knowledge, such as unit conversions and numeral transformations, highlighting the need for long-tail knowledge integration [76]. Poor performance in self-knowledge tasks reveals gaps in LLMs' self-assessment abilities [46].

Deterministic evaluations inadequately detect residual information leakage, suggesting probabilistic metrics may offer a more comprehensive framework [77]. These challenges underscore the need for innovative evaluation methodologies and enhanced training to address LLM limitations in complex and low-resource tasks, ensuring reliability across diverse applications.

4.4 Impact of Model Architecture and Fine-Tuning

The architecture and fine-tuning of LLMs are crucial in determining their performance across tasks and domains. Modular and extensible frameworks like FreeEval enhance adaptability and fairness by integrating various evaluation methods [78]. Fine-tuning strategies, such as those proposed by Ma et al., focus on preserving functional integrity during post-training, enabling efficient adaptation with minimal data [79]. Dynamic evaluation benchmarks reveal significant performance drops, emphasizing the need for evolving frameworks to capture architectural and fine-tuning impacts accurately [80].

Tools like the LLM Comparator provide insights into qualitative differences in model responses, aiding in identifying architectural and fine-tuning elements contributing to performance variations [52]. Strategic decisions in architecture and fine-tuning influence adaptability, efficiency, and effectiveness, with ongoing research crucial for addressing challenges like hallucination and limited controllability. Developing robust evaluation frameworks encompassing reasoning, knowledge, reliability, and safety, alongside retrieval-augmented generation techniques, will improve LLM accuracy and relevance, facilitating adaptation to diverse real-world demands while mitigating potential risks [8, 13, 7, 50, 9].

5 Challenges and Limitations

Evaluating large language models (LLMs) involves navigating a complex array of challenges, particularly in ethical and security domains. These aspects are vital in shaping LLM deployment and evaluation methodologies. The following subsections delve into the ethical implications and security challenges in LLM evaluation, underscoring the need to address these issues for responsible AI development.

5.1 Ethical and Security Considerations

LLM evaluation is fraught with ethical and security challenges, particularly in real-time applications where biases in model outputs can perpetuate stereotypes, such as gender disparities in STEM encouragement [6]. This highlights the need for rigorous frameworks to ensure fairness and inclusivity [5]. Security concerns are heightened by LLMs' potential to generate misinformation and inappropriate content, posing risks in enterprise settings where factuality, toxicity, hallucination, and bias are critical [60]. Adversarial attacks and the inadvertent extraction of sensitive information further complicate the landscape, necessitating robust security measures [81]. Challenges in deploying LLMs in autonomous systems, like vehicles, due to communication delays and inference times, limit their practical application in real-time scenarios [42]. Benchmarks often fail to capture programming language diversity or cultural nuances, limiting security evaluations [10]. The potential misuse of LLMs for harmful content generation underscores the need for defenses to align outputs with human values

AI-generated, for reference only.

[82]. The pressure to report state-of-the-art results can lead to questionable practices, complicating the ethical evaluation landscape [83]. Addressing these challenges requires comprehensive evaluation frameworks integrating diverse benchmarks, safety measures, and ethical guidelines, enhancing understanding of LLM performance and fostering more reliable AI systems [22, 8, 9, 37].

5.2 Data Quality and Contamination

LLM evaluation critically depends on the quality and diversity of training and testing datasets. Data contamination can distort perceptions of LLM capabilities and compromise evaluation reliability [80]. Curated datasets, while beneficial for targeted analyses, often lack the breadth required for generalizable findings [84]. Demographic biases among dataset contributors can further impact representativeness and model training [85]. In sensitive domains like mental health, data biases and complexity challenge benchmark applicability [14]. Generated test case quality and understanding of complex programming issues limit methodologies like CODET [32]. Privacy concerns arise from LLMs' potential to memorize and leak sensitive training data, raising ethical issues about data handling [81]. The exclusion of low-resource languages from benchmarks like README++ restricts evaluation frameworks for multilingual models [43], while reliance on translated datasets may introduce biases not present in original datasets [36]. Addressing data quality and contamination is vital for reliable LLM evaluation. Prioritizing diverse, high-quality datasets and comprehensive frameworks can enhance model performance assessment, identify risks related to data contamination, and support ethical AI development. Automating dataset updates and controlling difficulty can ensure fair evaluations, contributing to academic integrity in advanced AI [83, 64, 5, 22].

5.3 Benchmark Limitations

Current benchmarking practices significantly constrain LLM evaluation, often failing to provide comprehensive and accurate assessments. Many benchmarks are static, not accounting for varying test task training, leading to misleading evaluations [41]. This static nature also allows for leakage, inflating performance metrics [64]. Benchmarks often lack data diversity and fail to capture application nuances, complicating evaluations [62]. Inconsistent criteria and metrics hinder comparisons across studies [63]. Domain-specific focuses, such as clinical narratives, limit generalizability [3]. Many studies neglect necessary metrics for generative AI, creating gaps in understanding [24]. Benchmarks focusing on short programming tasks fail to capture real-world complexity [26]. FEQA metric accuracy depends on underlying QA model quality, introducing variability [49]. Current benchmarks lack comprehensiveness and flexibility, often missing nuances in adversarial contexts [23]. Assumptions like independence in win probabilities may not hold, leading to inaccuracies [86]. Addressing benchmark limitations requires dynamic, comprehensive frameworks that assess LLM capabilities across domains and languages, ensuring alignment with real-world applications [7, 51, 9].

5.4 Evaluation Methodology Challenges

Evaluating LLMs presents methodological challenges due to traditional tasks' inadequacy in capturing advanced capabilities and real-world application evolution. A multifaceted approach, incorporating qualitative and quantitative methods, addressing LLM-specific issues like hallucination and controllability, and integrating diverse metrics, is essential for understanding potential risks and effectiveness [7, 50, 8, 9]. Many benchmarks are static, failing to adapt to LLMs' evolving capabilities and suffering from data contamination. Human evaluations are resource-intensive and inconsistent, affecting outcomes [21]. Reliance on LLM calls for rationale clustering can introduce errors [52]. LLM safety assessment lacks comprehensive frameworks simulating various risks. The SEVAL benchmark uses structured prompts to simulate safety risks, providing robust assessments [87]. Evaluation methods often fail to mitigate selection bias, overlooking token bias influencing predictions [39]. In resource-constrained environments, advanced evaluation methods' computational demands limit applicability [30]. Developing innovative methodologies to address LLM evaluation complexities is crucial for accurate, reliable, and comprehensive assessments. Given NLP advancements and LLM applications, traditional tasks are insufficient. Proposed benchmarks focus on reasoning, knowledge, reliability, and safety competencies, creating structured systems for meaningful assessments that reflect LLMs' multifaceted abilities and adapt to evolving roles in academic and practical contexts [8, 9]. Overcoming these hurdles enables better understanding and improvement of LLM performance across applications.

AI-generated, for reference only.

6 Future Directions and Opportunities

6.1 Expansion and Diversification of Benchmarks

Expanding and diversifying benchmarks are crucial for assessing large language models (LLMs) across varied tasks and contexts. As LLMs evolve, establishing standardized frameworks that incorporate trends in generative AI, like hybrid models, is imperative [24]. Future efforts should enhance datasets and refine evaluation protocols with advanced prompt engineering to boost evaluation robustness [23]. Expanding benchmarks such as README++ is essential for capturing linguistic and cultural nuances, particularly in languages like Traditional Chinese, necessitating inclusive datasets and metrics [36]. Broadening benchmarks to encompass diverse programming languages and improving LLM-generated code reliability are also key areas for research [10]. In clinical settings, refining benchmarks is vital for high-stakes evaluations [2]. Integrating external knowledge sources and exploring cognitive levels for nuanced evaluations can create benchmarks reflecting real-world scenarios, guiding the development of effective models [64]. Ethical guidelines for AI use and innovative applications that uphold ethical standards must be prioritized [5]. Developing undesirable pattern analysis, rationale clustering, and LLM-based custom metrics are critical future work areas [52]. Overall, expanding and diversifying benchmarks is essential for comprehensive LLM evaluations, enhancing applicability across domains, and improving real-world deployment.

6.2 Integration with External Knowledge and Multimodal Tools

Integrating LLMs with external knowledge bases and multimodal tools offers promising enhancements for diverse applications. This integration can augment LLM performance by providing contextual and domain-specific information not inherently available within the models. Future research should explore multimodal tool integration to enhance complex task capabilities, such as code translation, through improved prompt designs and extraction methods [88, 53]. Incorporating external knowledge bases can enhance factual accuracy and reliability, crucial for tasks like legal judgment prediction, where information retrieval systems can improve LLM capabilities [11]. Exploring multilingual capabilities and pragmatic evaluation frameworks benefits from integrating external knowledge, fostering nuanced understanding across languages and contexts [44]. Developing probabilistic evaluation frameworks offers opportunities for integration, enhancing metrics for compliance verification [77]. Integrating LLMs with diverse informational resources and creating innovative evaluation frameworks addresses challenges in assessing LLMs, strengthening evaluation rigor, and fostering advancements in LLM technology [7, 9].

6.3 Advanced Evaluation Metrics and Feedback Mechanisms

Developing advanced evaluation metrics and feedback mechanisms is vital for refining LLM performance across applications. Future research should mitigate biases in LLM-based evaluators and enhance existing frameworks for fairness and reliability [59]. In summarization tasks, improving QA models and exploring coherence and informativeness dimensions provide comprehensive assessments [49]. Integrating advanced feedback mechanisms optimizes LLM performance in complex environments, such as software engineering and robotics, by refining metrics and addressing data quality challenges [4, 16]. Generating test cases for programming tasks underscores the need for sophisticated evaluation frameworks capturing LLM performance intricacies [32]. Including ethical considerations in real-time decision-making emphasizes developing metrics incorporating ethical dimensions [42]. Future work should improve data quality and reduce biases in frameworks like SELF-INSTRUCT, enhancing LLM performance across architectures [31]. Advancing evaluation metrics and feedback mechanisms drives LLM technology innovations, improving effectiveness and applicability in real-world scenarios.

6.4 Innovations in Model Adaptability and Performance Enhancement

Innovations in LLM adaptability and performance enhancement are crucial for broadening applicability across domains. Table 2 provides a comparative analysis of various methods aimed at enhancing model adaptability and performance, emphasizing their application across different domains and the reliability of their evaluations. Future research should optimize computational efficiency in frameworks like AutoDAN and explore applicability to advanced LLMs, developing defenses against

Method Name	Model Adaptability	Performance Enhancement	Evaluation and Reliability
AD[82]	Various Domains	Improves Effectiveness	Attack Success Rate
LE[37]	Diverse Cultural Perspectives	Neuro-symbolic Approach	Logical Validity

Table 2: Comparison of innovative methods in model adaptability and performance enhancement, focusing on adaptability to diverse domains, performance improvement techniques, and evaluation reliability. The table highlights the adaptability of methods across various domains and cultural perspectives, their impact on performance enhancement, and the reliability of their evaluation metrics.

jailbreak attacks [82]. Enhancing inference capabilities in ethical dilemmas and cultural perspectives improves real-world applicability [37]. Exploring PriDe's domain adaptability and investigating selection bias causes can lead to robust models [39]. Research should focus on comprehensive memorization mitigation strategies, privacy risk assessments, and robust auditing methods [81]. Extending frameworks like Self-Reflection and investigating game and problem types influence on performance are vital for understanding LLM capabilities [40]. Enhancing capabilities in numeral and measurement tasks and expanding benchmarks to diverse datasets can improve performance in specialized domains [76]. Future research should ensure fair evaluations and consider training practices' implications on benchmark performance [41]. Implementing stricter evaluation reporting guidelines and fostering integrity over competition addresses questionable practices [83]. Expanding datasets, incorporating longer prompts, and exploring fine-tuning for safety concerns enhance robustness and reliability [60]. Future work could explore complex tasks and improve evaluation automation, advancing adaptability and performance [46]. Advances in adaptability and performance are essential for LLM technology progression, ensuring reliability and robustness in real-world scenarios and addressing evaluation challenges and risks like API misuse. Focusing on reasoning, knowledge, reliability, and safety develops benchmarks and metrics for comprehensive LLM evaluation, leading to dependable applications in fields like software development [10, 9].

7 Conclusion

The evaluation of large language models (LLMs) is paramount for advancing their utility and effectiveness in natural language processing (NLP) applications. This survey underscores the importance of systematic assessments in uncovering both the capabilities and limitations of LLMs, as evidenced by diverse metrics and benchmarks that scrutinize their performance across various linguistic and cognitive dimensions. The insights gained from these evaluations reveal the need for further development in pragmatic capabilities, as highlighted by recent studies.

Integrating sophisticated evaluation methodologies that incorporate external knowledge and multimodal tools is crucial for enhancing LLMs' contextual comprehension and pragmatic functionality. Such approaches are essential for equipping LLMs to adeptly handle complex, real-world scenarios, thereby broadening their applicability across diverse sectors. Furthermore, the development of novel evaluation metrics and feedback systems is critical for refining LLM performance, reducing biases, and ensuring ethical and consistent outputs.

As LLM technology evolves, future evaluation efforts must aim to expand and diversify benchmarks to fully capture the breadth of LLM capabilities. This includes addressing the challenges associated with complex and resource-scarce tasks and examining the impact of model architecture and fine-tuning on performance. By fostering a culture of rigorous and comprehensive evaluation, the research community can drive forward advancements in LLM technology, ultimately enhancing their efficacy and dependability in NLP applications.

References

[1] Wenxuan Wang. Testing and evaluation of large language models: Correctness, non-toxicity, and fairness, 2024.

[2] Mingyu Derek Ma, Chenchen Ye, Yu Yan, Xiaoxuan Wang, Peipei Ping, Timothy S Chang, and Wei Wang. Clibench: A multifaceted and multigranular evaluation of large language models for clinical decision making, 2024.

[3] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.

[4] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*, 33(8):1–79, 2024.

[5] Brady D Lund, Ting Wang, Nishith Reddy Mannuru, Bing Nie, Somipam Shimray, and Ziang Wang. Chatgpt and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5):570–581, 2023.

[6] Smilla Due, Sneha Das, Marianne Andersen, Berta Plandolit López, Sniff Andersen Nexø, and Line Clemmensen. Evaluation of large language models: Stem education and gender stereotypes, 2024.

[7] Yue Zhang, Ming Zhang, Haipeng Yuan, Shichun Liu, Yongyao Shi, Tao Gui, Qi Zhang, and Xuanjing Huang. Llmeval: A preliminary study on how to evaluate large language models, 2023.

[8] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models, 2023.

[9] Ziyu Zhuang, Qiguang Chen, Longxuan Ma, Mingda Li, Yi Han, Yushan Qian, Haopeng Bai, Zixian Feng, Weinan Zhang, and Ting Liu. Through the lens of core competency: Survey on evaluation of large language models, 2023.

[10] Li Zhong and Zilong Wang. Can llm replace stack overflow? a study on robustness and reliability of large language model code generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21841–21849, 2024.

[11] Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. A comprehensive evaluation of large language models on legal judgment prediction, 2023.

[12] Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities, 2024.

[13] 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.

[14] Mihael Arcan, David-Paul Niland, and Fionn Delahunty. An assessment on comprehending mental health through large language models, 2024.

[15] Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen K. Wong, Graham Wills, Elliot First, Frank J. Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. Evaluation of large language models for summarization tasks in the medical domain: A narrative review, 2024.

[16] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

[17] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[18] Amin Beheshti, Jian Yang, Quan Z Sheng, Boualem Benatallah, Fabio Casati, Schahram Dustdar, Hamid Reza Motahari Nezhad, Xuyun Zhang, and Shan Xue. Processgpt: transforming business process management with generative artificial intelligence. In *2023 IEEE International Conference on Web Services (ICWS)*, pages 731–739. IEEE, 2023.

[19] Abdelrahman Hanafi, Mohammed Saad, Noureldin Zahran, Radwa J. Hanafy, and Mohammed E. Fouda. A comprehensive evaluation of large language models on mental illnesses, 2024.

[20] Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators, 2023.

[21] Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V Stolyar, Katelyn Polanska, Karleigh R McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, Piyush Mathur, Giovanni E. Cacciamani, Cong Sun, Yifan Peng, and Yanshan Wang. A framework for human evaluation of large language models in healthcare derived from literature review, 2024.

[22] Mike Perkins. Academic integrity considerations of ai large language models in the post-pandemic era: Chatgpt and beyond. *Journal of University Teaching and Learning Practice*, 20(2), 2023.

[23] Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. Promptbench: A unified library for evaluation of large language models, 2024.

[24] Ajay Bandi, Pydi Venkata Satya Ramesh Adapa, and Yudu Eswar Vinay Pratap Kumar Kuchi. The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, 15(8):260, 2023.

[25] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[26] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[27] Euclides Lourenco Chuma and Gabriel Gomes De Oliveira. Generative ai for business decision-making: A case of chatgpt. *Management Science and Business Decisions*, 3(1):5–11, 2023.

[28] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

[29] Lin Yang, Chen Yang, Shutao Gao, Weijing Wang, Bo Wang, Qihao Zhu, Xiao Chu, Jianyi Zhou, Guangtai Liang, Qianxiang Wang, and Junjie Chen. On the evaluation of large language models in unit test generation, 2024.

[30] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[31] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[32] Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. Codet: Code generation with generated tests. *arXiv preprint arXiv:2207.10397*, 2022.

[33] Yue Wang, Weishi Wang, Shafiq Joty, and Steven CH Hoi. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. *arXiv preprint arXiv:2109.00859*, 2021.

[34] 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.

[35] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[36] Chan-Jan Hsu, Chang-Le Liu, Feng-Ting Liao, Po-Chun Hsu, Yi-Chang Chen, and Da shan Shiu. Advancing the evaluation of traditional chinese language models: Towards a comprehensive benchmark suite, 2023.

[37] Xin Quan, Marco Valentino, Louise A. Dennis, and André Freitas. Enhancing ethical explanations of large language models through iterative symbolic refinement, 2024.

[38] Joseph Gatto, Omar Sharif, Parker Seegmiller, Philip Bohlman, and Sarah Masud Preum. Text encoders lack knowledge: Leveraging generative llms for domain-specific semantic textual similarity, 2023.

[39] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors, 2024.

[40] Saumya Malik. Lost in the logic: An evaluation of large language models' reasoning capabilities on lsat logic games, 2024.

[41] Ricardo Dominguez-Olmedo, Florian E. Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence, 2024.

[42] Kotaro Tanahashi, Yuichi Inoue, Yu Yamaguchi, Hidetatsu Yaginuma, Daiki Shiotsuka, Hiroyuki Shimatani, Kohei Iwamasa, Yoshiaki Inoue, Takafumi Yamaguchi, Koki Igari, Tsukasa Horinouchi, Kento Tokuhiro, Yugo Tokuchi, and Shunsuke Aoki. Evaluation of large language models for decision making in autonomous driving, 2023.

[43] Tarek Naous, Michael J. Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. Readme++: Benchmarking multilingual language models for multi-domain readability assessment, 2024.

[44] Dojun Park, Jiwoo Lee, Hyeyun Jeong, Seohyun Park, and Sungeun Lee. Pragmatic competence evaluation of large language models for the korean language, 2024.

[45] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[46] Zhiquan Tan, Lai Wei, Jindong Wang, Xing Xie, and Weiran Huang. Can i understand what i create? self-knowledge evaluation of large language models, 2024.

[47] Cong Xu, Gayathri Saranathan, Mahammad Parwez Alam, Arpit Shah, James Lim, Soon Yee Wong, Foltin Martin, and Suparna Bhattacharya. Data efficient evaluation of large language models and text-to-image models via adaptive sampling, 2024.

[48] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets, 2024.

[49] Esin Durmus, He He, and Mona Diab. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*, 2020.

[50] Tomoki Doi, Masaru Isonuma, and Hitomi Yanaka. Comprehensive evaluation of large language models for topic modeling, 2024.

[51] Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. Can multiple-choice questions really be useful in detecting the abilities of llms?, 2024.

[52] Minsuk Kahng, Ian Tenney, Mahima Pushkarna, Michael Xieyang Liu, James Wexler, Emily Reif, Krystal Kallarackal, Minsuk Chang, Michael Terry, and Lucas Dixon. Llm comparator: Visual analytics for side-by-side evaluation of large language models, 2024.

[53] Marcos Macedo, Yuan Tian, Filipe R. Cogo, and Bram Adams. Exploring the impact of the output format on the evaluation of large language models for code translation, 2024.

[54] Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, and Sungeun Lee. Multiprageval: Multilingual pragmatic evaluation of large language models, 2024.

[55] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

[56] Weiqi Wu, Hongqiu Wu, and Hai Zhao. Self-directed turing test for large language models, 2024.

[57] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

[58] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[59] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

[60] David Nadeau, Mike Kroutikov, Karen McNeil, and Simon Baribeau. Benchmarking llama2, mistral, gemma and gpt for factuality, toxicity, bias and propensity for hallucinations, 2024.

[61] Hussein Mozannar, Valerie Chen, Mohammed Alsobay, Subhro Das, Sebastian Zhao, Dennis Wei, Manish Nagireddy, Prasanna Sattigeri, Ameet Talwalkar, and David Sontag. The realhumaneval: Evaluating large language models' abilities to support programmers, 2024.

[62] Mingyue Cheng, Hao Zhang, Jiqian Yang, Qi Liu, Li Li, Xin Huang, Liwei Song, Zhi Li, Zhenya Huang, and Enhong Chen. Towards personalized evaluation of large language models with an anonymous crowd-sourcing platform, 2024.

[63] Xiang Li, Yunshi Lan, and Chao Yang. Treeeval: Benchmark-free evaluation of large language models through tree planning, 2024.

[64] Jiahao Ying, Yixin Cao, Yushi Bai, Qianru Sun, Bo Wang, Wei Tang, Zhaojun Ding, Yizhe Yang, Xuanjing Huang, and Shuicheng Yan. Automating dataset updates towards reliable and timely evaluation of large language models, 2024.

[65] Yingjie Fu, Bozhou Li, Linyi Li, Wentao Zhang, and Tao Xie. The first prompt counts the most! an evaluation of large language models on iterative example-based code generation, 2024.

[66] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[67] Patrik Puchert, Poonam Poonam, Christian van Onzenoodt, and Timo Ropinski. Llmmaps – a visual metaphor for stratified evaluation of large language models, 2023.

AI-generated, for reference only.

[68] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[69] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.

[70] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, 2020.

[71] Zhehao Zhang, Jiaao Chen, and Diyi Yang. Darg: Dynamic evaluation of large language models via adaptive reasoning graph, 2024.

[72] Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Dynamic evaluation of large language models for reasoning tasks, 2024.

[73] Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning, 2023.

[74] Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. Asleep at the keyboard? assessing the security of github copilot's code contributions. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 754–768. IEEE, 2022.

[75] Hammond Pearce, Benjamin Tan, Baleegh Ahmad, Ramesh Karri, and Brendan Dolan-Gavitt. Examining zero-shot vulnerability repair with large language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2339–2356. IEEE, 2023.

[76] Ancheng Xu, Minghuan Tan, Lei Wang, Min Yang, and Ruifeng Xu. Numcot: Numerals and units of measurement in chain-of-thought reasoning using large language models, 2024.

[77] Yan Scholten, Stephan Günnemann, and Leo Schwinn. A probabilistic perspective on unlearning and alignment for large language models, 2025.

[78] Zhuohao Yu, Chang Gao, Wenjin Yao, Yidong Wang, Zhengran Zeng, Wei Ye, Jindong Wang, Yue Zhang, and Shikun Zhang. Freeeval: A modular framework for trustworthy and efficient evaluation of large language models, 2024.

[79] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023.

[80] Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. Dynamic evaluation of large language models by meta probing agents, 2024.

[81] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[82] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*, 2023.

[83] Gavin Leech, Juan J. Vazquez, Niclas Kupper, Misha Yagudin, and Laurence Aitchison. Questionable practices in machine learning, 2024.

[84] 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. *arXiv preprint arXiv:2211.09085*, 2022.

[85] Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36:47669–47681, 2023.

[86] Mehmet S. Ismail. Performance rating in chess, tennis, and other contexts, 2023.

[87] Xiaohan Yuan, Jinfeng Li, Dongxia Wang, Yuefeng Chen, Xiaofeng Mao, Longtao Huang, Hui Xue, Wenhai Wang, Kui Ren, and Jingyi Wang. S-eval: Automatic and adaptive test generation for benchmarking safety evaluation of large language models, 2024.

[88] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.

AI-generated, for reference only.

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.