A Survey of Internal Consistency, Self-Feedback, and Reliability in Large Language Models

www.surveyx.cn

Abstract

The rapid advancement of large language models (LLMs) has significantly transformed natural language processing, offering capabilities that closely mimic human-like text generation and understanding. This survey explores critical dimensions such as internal consistency, self-feedback, and AI reliability, highlighting their importance for the safe and reliable deployment of LLMs across various sectors. Internal consistency ensures coherent outputs, particularly crucial in high-stakes environments like healthcare. Self-feedback mechanisms facilitate iterative refinement, enhancing model performance by addressing issues such as reward overoptimization. AI reliability underscores the need for trustworthy outputs, especially in applications involving high-stakes decision-making. The survey also examines innovative evaluation frameworks and metrics essential for assessing LLM effectiveness. Challenges such as hallucinations and biases in training data underscore the need for robust evaluation methods. Future directions include enhancing feedback mechanisms, optimizing hyperparameters, and improving calibration techniques. By addressing these challenges, the survey aims to contribute to the ongoing evolution of LLMs, ensuring they meet user requirements while maintaining safety and reliability. The integration of these elements is crucial for the successful application of LLMs in diverse domains, emphasizing their far-reaching implications for the future development of AI systems.

1 Introduction

1.1 Scope and Significance

The rapid advancement of large language models (LLMs) has significantly transformed the landscape of natural language processing, providing capabilities that closely mimic human-like text generation and understanding [1]. This survey delves into the critical dimensions of internal consistency, self-feedback, and reliability within LLMs, highlighting their significance for the safe and reliable deployment of these models across various sectors. The importance of maintaining internal consistency in LLMs is underscored by their application in high-stakes environments such as healthcare, where precision and coherence are paramount .

Addressing the challenges of validating outputs from LLMs is crucial, as internal consistency and reliability are pivotal for ensuring safe deployment [2]. Furthermore, the issue of unreliability in LLMs must be addressed to facilitate their effective deployment in knowledge-intensive tasks [3]. Enhancing reasoning capabilities in LLMs is also critical for their application in complex tasks, such as mathematical reasoning, which necessitates robust and reliable model outputs [4].

The survey also considers the significance of training-time optimization methods for compound AI systems, which are integral to improving the overall performance and reliability of LLMs [5]. By exploring these dimensions, the survey aims to contribute to the ongoing evolution and deployment of LLMs across diverse fields, ensuring these models meet user requirements while maintaining safety and reliability in their applications .

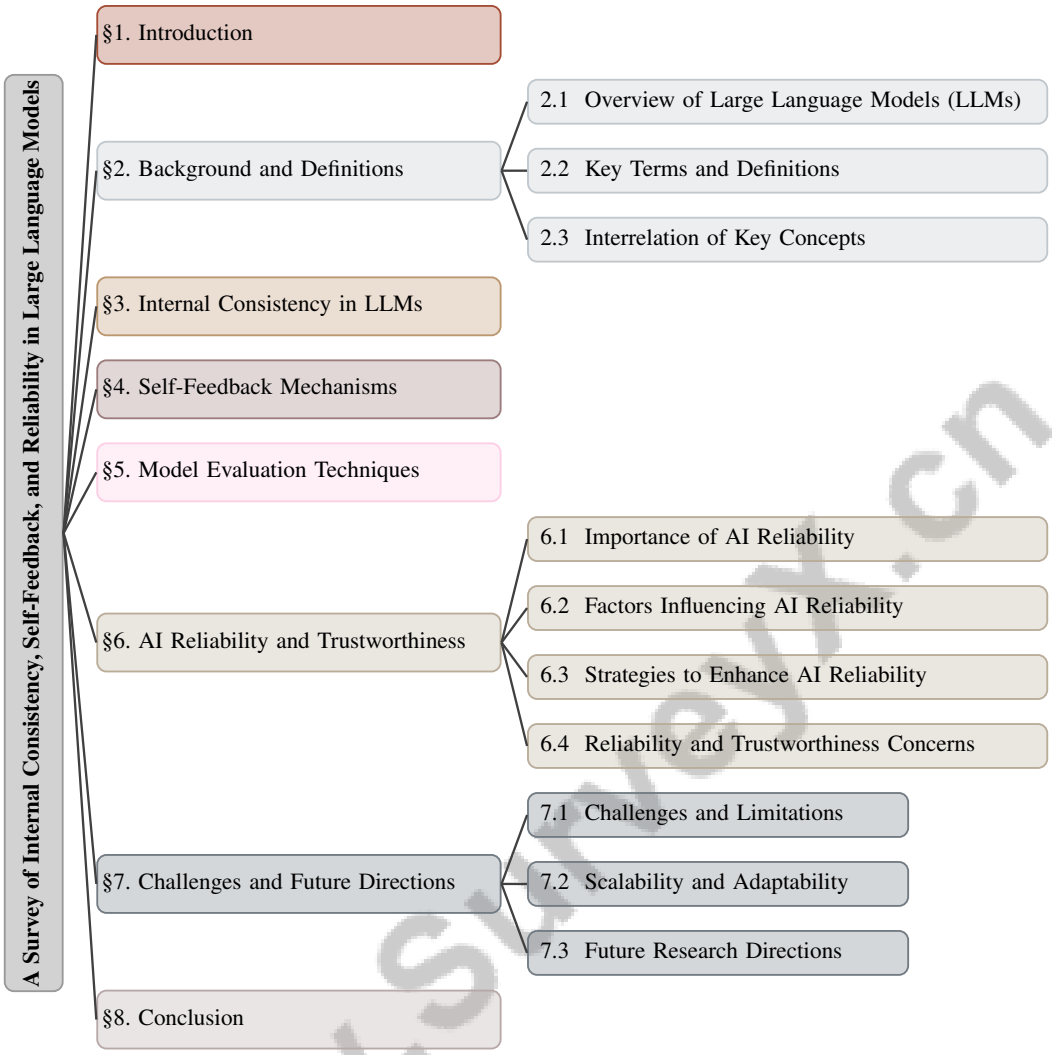

Figure 1: chapter structure

1.2 Role in Natural Language Processing and AI Model Evaluation

Large language models (LLMs) have revolutionized the field of natural language processing (NLP) by significantly enhancing the capabilities of machines to understand and generate human-like text. Internal consistency and reliability are crucial components in ensuring that these models adhere to user instructions, impacting their performance and trustworthiness [6]. The ability of LLMs to maintain internal consistency is vital for producing outputs that are not only syntactically correct but also semantically meaningful, which is essential in applications such as dialogue systems and translation tasks [3]. These attributes are fundamental for ensuring the reliability of the model and its evaluation, particularly in high-stakes environments like healthcare [7].

Self-feedback mechanisms are integral to the iterative refinement and adaptive learning processes in LLMs. These mechanisms enable models to continuously improve their performance by addressing issues such as reward overoptimization and ensuring alignment with desired performance metrics [8]. This iterative process is especially important in dynamic environments where continuous learning is necessary, and accurate predictions are paramount, such as in safety-sensitive applications [5].

The reliability of LLMs is a cornerstone for trustworthy AI model evaluation. Reliable models are essential for applications involving high-stakes decision-making and planning, where inaccuracies could lead to significant consequences [3]. The challenge of hallucinations, where models generate factually incorrect content, highlights the need for robust evaluation frameworks to maintain factual integrity and reliability in LLM outputs [9]. Furthermore, the lack of a clear understanding of LLM

capabilities hampers the ability to predict their performance across various tasks and to develop effective evaluation benchmarks [10].

Understanding user experiences and satisfaction when interacting with LLMs is another critical aspect, particularly in terms of their effectiveness as collaborative tools [11]. The trustworthiness of LLM outputs, especially in sensitive applications such as psychological consultation, is critical. The Psy-LLM framework serves as an example of how LLMs can bridge gaps in service availability while maintaining trust and reliability in their outputs.

Internal consistency, self-feedback, and reliability are integral to the successful application of LLMs in NLP. These elements not only enhance the models' ability to generate coherent and contextually appropriate outputs but also influence the methodologies used for AI model evaluation, ensuring that LLMs are reliable and trustworthy tools in various domains. The challenges associated with harmful, biased, or untruthful content generation, privacy leakage, and system vulnerabilities further emphasize the need for robust internal consistency and self-feedback mechanisms to mitigate risks and enhance the overall reliability of LLMs [3].

1.3 Structure of the Survey

This survey is meticulously structured to provide a comprehensive analysis of the critical dimensions influencing the performance and reliability of large language models (LLMs). The paper begins with an **Introduction**, which sets the stage by discussing the scope and significance of internal consistency, self-feedback, and reliability in LLMs, and their pivotal roles in natural language processing and AI model evaluation. Following this, the **Background and Definitions** section offers an overview of LLMs, elucidating key terms and their interrelations, establishing a foundational understanding for the subsequent discussions.

The survey then delves into **Internal Consistency in LLMs**, exploring its significance, the challenges in achieving it, and the methods and metrics employed for its assessment. The subsequent section provides a comprehensive analysis of , detailing their role in enhancing internal consistency among large language models (LLMs), facilitating iterative refinement of outputs, and improving overall model learning and error correction. Specifically, these mechanisms, which include frameworks like Self-Evaluation and Self-Update, enable LLMs to assess and adjust their responses based on internal signals, thereby addressing common issues such as reasoning deficiencies and hallucinations. By leveraging these self-assessment processes, models can refine their outputs iteratively without the need for external supervision, ultimately leading to more accurate and reliable performance across various tasks. [12, 13, 14, 15, 16]

The paper provides a comprehensive examination of , focusing on the latest evaluation metrics and innovative frameworks while critically assessing their strengths and weaknesses. It highlights the importance of meta-evaluation and benchmarking in enhancing the robustness of evaluations for large language models (LLMs). Additionally, the paper proposes future directions for model evaluation, emphasizing the need for standardized benchmarks and the integration of peer-review-inspired frameworks to address the limitations of current methodologies. [17, 18, 19, 20]. The focus then shifts to **AI Reliability and Trustworthiness**, analyzing the importance of reliability, factors influencing it, and strategies to enhance the trustworthiness of AI outputs.

The survey concludes with a discussion on **Challenges and Future Directions**, identifying current limitations, scalability issues, and proposing future research avenues to advance the development and application of LLMs. Each section is crafted to build upon the previous, ensuring a coherent and logical progression of ideas, ultimately providing a holistic view of the complexities and advancements in the field of large language models.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Overview of Large Language Models (LLMs)

Large language models (LLMs) are transformative in natural language processing (NLP), excelling in tasks such as language translation, dialogue generation, and complex reasoning [4]. In healthcare, LLMs enhance clinical decision-making by automating processes like cancer staging [7], though their adaptation for sensitive medical applications requires specialized approaches [21]. In education,

they automate content assessment, providing critical feedback for student development, particularly in computational fields [22]. In business, LLMs improve recommendation systems by generating natural language profiles [5].

Despite their capabilities, LLMs face challenges such as generating biased or erroneous content, necessitating robust evaluation methods to ensure reliability. The issue of 'hallucinations', where models produce incorrect information, is particularly concerning in applications like medical question-answering [23]. Additionally, LLMs struggle with long inputs, potentially losing critical historical information [24].

To address these issues, strategies like Confidence-Probability Alignment align an LLM's internal confidence with its outputs [8]. Optimizing compound AI systems using LLMs is crucial for enhancing performance across domains [5]. Fine-tuning LLMs in low-data regimes addresses existing data augmentation limitations [25]. LLMs are also specialized for task-oriented dialogue agents, especially where traditional data collection is costly or impractical [26]. In computational social science, LLMs aid data annotation, though challenges in statistical inference remain [27].

Ongoing research is essential to mitigate risks and enhance the robustness of LLMs across sectors. Establishing universal analysis frameworks is vital for evaluating LLM quality and trustworthiness, ensuring safety and reliability in various applications [3].

2.2 Key Terms and Definitions

Understanding key terms is crucial for comprehending LLM operational frameworks and evaluation methodologies. Internal consistency refers to LLMs generating logically coherent and contextually aligned outputs, vital in precision-critical applications like medical diagnostics and dialogue systems [7]. Self-feedback involves mechanisms for iterative output refinement through internal assessments and external feedback, enhancing performance over time. This includes self-reflection, where models engage in internal dialogues to generate training data, facilitating continuous learning [4]. Intrinsic self-correction allows models to autonomously rectify errors, improving translation quality and model refinement.

Model evaluation includes methodologies and metrics assessing LLM output quality, focusing on accuracy, coherence, and relevance [7]. Evaluation criteria influence consistency and reliability of assessments. Structured solution planning and verification through visible tests are crucial for evaluating code generation accuracy in LLMs.

AI reliability concerns the dependability of AI systems in producing consistent and accurate results, emphasizing factually correct, contextually appropriate outputs. The challenge of hallucinations, particularly in medical contexts, underscores the need for robust evaluation methods [4, 7].

Additional terms include LLM annotations, statistical estimation, and demographic biases, crucial for understanding challenges in adapting to varying data availability and task requirements [27]. The sensitivity of LLMs to prompts and the complexity of optimizing interdependent parameters highlight model tuning intricacies.

These interconnected terms enhance understanding and application of LLMs, ensuring effective deployment across sectors. Methodologies like METACRITIQUE provide frameworks for evaluating model-generated critiques, improving alignment and refinement. The ALLURE approach audits LLM evaluations to rectify reasoning errors through iterative learning, reducing dependence on human annotators. The Guide-Align framework integrates safety guidelines to mitigate risks from biased content generation, resulting in safer outputs. These strategies underscore the necessity of continuous LLM improvement for enhanced capabilities and trustworthiness [28, 29, 19].

2.3 Interrelation of Key Concepts

The interplay between internal consistency, self-feedback, model evaluation, and AI reliability is crucial for advancing LLM performance and evaluation. These components ensure LLMs generate coherent, contextually appropriate outputs across applications. Internal consistency maintains information integrity in applications like sentiment analysis and question answering, though LLM architecture complexity necessitates flexible evaluation methods [30].

Self-feedback enhances internal consistency by enabling iterative output refinement through internal assessments and external feedback, improving LLM performance and adaptability. The interplay between self-critique and external verification is crucial for LLM performance assessment and capability enhancement. METACRITIQUE evaluates critique quality through Atomic Information Units (AIUs), improving alignment. Critique models in a two-player framework enhance reasoning efficiency, while self-contrast techniques address biases in LLM self-evaluations, leading to more stable reflections [29, 15, 16].

Model evaluation techniques assess LLM effectiveness and reliability. Integrating LLM annotations with human annotations enhances statistical estimation and addresses biases [27]. Critiquing empirical literature on LLM evaluation emphasizes innovative frameworks to capture LLM output complexities [30].

AI reliability ensures trustworthy LLM outputs, especially in high-stakes applications. It is influenced by alignment with human values, safety-trained models, and comprehensive evaluation frameworks assessing trustworthiness across fairness, robustness, and explainability dimensions. Addressing these aspects enhances LLM output safety and quality, fostering confidence in sensitive contexts [28, 31, 32, 33]. Aligning LLM capabilities with diverse user intents and expectations requires addressing training data biases for reliable outputs. Understanding these interdependencies is crucial for advancing LLM capabilities and trustworthiness across sectors, ensuring effective deployment while maintaining safety and reliability. Balancing context length, accuracy, and performance underscores the complexity of these interrelated factors in LLM development and application.

3 Internal Consistency in LLMs

Exploring the significance of internal consistency within large language models (LLMs) reveals its crucial role in ensuring coherent and reliable outputs across diverse applications. This section examines the importance of maintaining internal consistency, particularly in high-stakes environments, to appreciate its impact on LLM functionality and broader applicability.

3.1 Significance of Internal Consistency

Internal consistency is fundamental to the reliability of LLMs, ensuring coherent and contextually appropriate outputs, especially in critical areas like healthcare where accuracy is vital for decision-making and patient safety [24]. As illustrated in Figure 2, the significance of internal consistency in LLMs is particularly pronounced in healthcare, where various frameworks and methods are employed to enhance this consistency. Maintaining internal consistency enhances practical utility and reduces the risk of misleading or erroneous content [7]. The LLM2LLM framework exemplifies how targeted data generation can refine outputs and minimize inconsistencies [25]. The SCM framework improves contextual recall by processing ultra-long inputs without modifying LLMs [24]. Furthermore, interactive self-reflection methods systematically mitigate hallucinations, enhancing LLM accuracy and reliability [23]. Self-feedback mechanisms allow models to refine outputs through structured self-assessment and external feedback, which is crucial for task-oriented dialogue tasks [26]. Philosophical analyses of LLM credences provide insights into factors influencing coherence and reliability, emphasizing robust evaluation frameworks [30].

3.2 Challenges in Achieving Internal Consistency

Achieving internal consistency in LLMs involves overcoming challenges related to model architecture and evaluation frameworks. Reward hacking, where models exploit reward signals, degrades reasoning performance and consistency [4]. Variability in reasoning paths leads to inconsistent predictions for identical inputs [34]. Poorly calibrated confidence scores affect reliability and sampling decisions [27]. The inability to self-verify reasoning outputs accumulates errors [35]. Human feedback, while valuable, is costly and lacks real-time applicability, complicating automated feedback mechanisms [36]. Noisy annotations impact modular AI system performance [37]. Benchmarks often suffer from saturation and contamination, complicating accurate performance assessment across languages and tasks. Current studies' narrow focus impedes generalization across applications [14]. Managing LLMs' slow and fast thinking processes, resulting in fluctuating gradient norms, poses a significant hurdle [38]. Constraints like model size and data quality challenge reasoning tasks [39]. Addressing

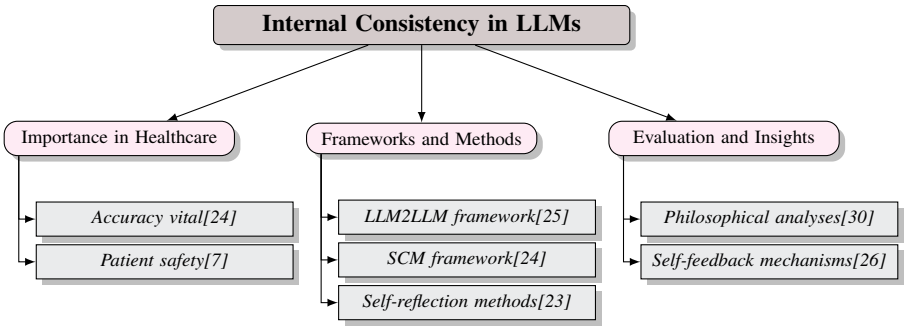

Figure 2: This figure illustrates the significance of internal consistency in LLMs, particularly in healthcare, the frameworks and methods employed to enhance consistency, and the evaluation insights gained from philosophical analyses and feedback mechanisms.

these challenges is essential for enhancing LLM capabilities and ensuring effective implementation in applications like software requirements [18, 40].

3.3 Methods and Metrics for Assessing Internal Consistency

Evaluating internal consistency in LLMs involves diverse methods and metrics ensuring coherence and reliability. The LLM2LLM framework enhances consistency through targeted data augmentation [25]. Reward Refinement Techniques (RRT) align reward signals with performance metrics, maintaining reasoning consistency [4]. Metrics like recall, precision, and F1-score evaluate model-generated content quality, ensuring outputs meet expected standards. Systematic evaluation metrics like accuracy and F1-score assess reasoning task consistency, with proprietary models outperforming open-source counterparts [41, 42, 43, 14, 33]. Integrating classical machine learning models with LLMs through adaptive weighting improves classification accuracy, reducing reliance on human annotations [44, 27, 45]. Confidence-driven inference prioritizes human annotations, enhancing data collection efficiency and statistical estimates. Multicalibration enhances confidence score reliability across datasets, aligning LLM confidence with certainty and minimizing overfitting [8, 46].

4 Self-Feedback Mechanisms

This section explores the role of self-feedback mechanisms in enhancing the internal consistency of large language models (LLMs). These mechanisms are crucial for refining model outputs and ensuring coherence in generated responses, providing a foundation for examining various methodologies and frameworks that facilitate these enhancements.

4.1 Improving Internal Consistency through Self-Feedback

Self-feedback mechanisms are essential for enhancing LLMs' internal consistency through iterative refinement and self-assessment. They enable LLMs to identify and correct output inconsistencies, thereby improving coherence and reliability. Interactive self-reflection, as proposed by [23], allows LLMs to iteratively generate, evaluate, and refine knowledge and answers, enhancing factual accuracy and consistency. The LLM2LLM framework [25] illustrates self-feedback's role in targeting data augmentation to reduce inconsistencies. 'Self-talk', where LLMs simulate dialogues in different roles, generates data that enhances internal consistency [26]. Furthermore, refining reward signals [4] mitigates reward hacking, aligning model outputs with desired performance metrics. These mechanisms collectively underscore the importance of self-feedback in maintaining reliable outputs, enhancing LLMs' trustworthiness across sectors [12, 13, 47, 14, 16].

4.2 Iterative Refinement and Self-Assessment

Iterative refinement and self-assessment are crucial for enhancing LLM performance and reliability. The RASC framework [48] exemplifies optimizing sampling through dynamic reasoning assessment, improving reliability. Addressing self-bias, where LLMs optimize incorrect outputs, requires

balancing self-assessment with external validation. The USC method [49] uses multiple response sampling to enhance output quality. Alternating feedback and refinement steps [13] systematically address inconsistencies, improving task understanding. Using weak LLMs for preference feedback [50] reduces human annotation reliance. Frameworks like TEaR and SELF-REFINE demonstrate significant output improvements through systematic self-correction, enhancing performance across applications [47, 13, 29, 51, 19]. These techniques ensure coherent, contextually appropriate outputs.

4.3 Self-Generated Data and Learning

Self-generated data is pivotal for enhancing LLM learning and performance. The SELF framework [52] allows LLMs to autonomously generate and refine responses, promoting continuous improvement. SELF-REFINE methodology [13] enhances outputs in creative and precise tasks. Iterative refinement, involving generating initial responses and refining them through proxy metrics [53], achieves high-quality outputs. The SDG framework [54] uses LLMs to propose and verify sub-goals, supporting adaptive learning. The IWSI framework [55] filters high DSE samples, yielding robust learning outcomes without heavy external supervision. The USC method [49] offers flexible evaluation for diverse tasks. However, challenges remain in preventing error amplification, as LLMs may fail to correct mistakes [56]. Addressing these challenges is crucial for maximizing self-generated data benefits.

4.4 Self-Correction and Error Mitigation

Self-correction mechanisms enhance LLM accuracy by autonomously refining predictions to reduce error propagation. The Self-Contrast approach [16] improves reflection capabilities, addressing overconfidence and inconsistency in self-feedback, leading to precise self-correction. Despite advancements, self-correction relies on external feedback for effective error mitigation [12], necessitating robust validation processes. Memory recall mechanisms [39] enhance accuracy by providing contextual information, informing predictions and reducing errors. The SELF-REFINE methodology [13] leverages self-generated feedback for continuous improvement, enhancing prediction accuracy and reliability.

5 Model Evaluation Techniques

Category	Feature	Method
Evaluation Metrics and Techniques	Prediction Reliability	VERITAS[3], CDI[27], SCM[24]
	Dialogue Evaluation	STDG[26]
	Performance Enhancement	RRT[4]
Innovative Evaluation Frameworks	Structured Assessment Methods	CAIRDD[57]
	Model Evaluation Techniques	SLR-AF[51]
	Adaptive Evaluation Techniques	ALLURE[19]
Meta-Evaluation and Benchmarking	Collaborative Evaluation	RGV[58], PRD[59]
Future Directions in Model Evaluation	Adaptive Evaluation	TALEC[60]

Table 1: This table provides a comprehensive overview of the various evaluation metrics, techniques, and frameworks utilized in assessing large language models (LLMs). It categorizes the methods into evaluation metrics and techniques, innovative evaluation frameworks, meta-evaluation and benchmarking, and future directions in model evaluation. Each category highlights specific features and methods, demonstrating the breadth and depth of current and emerging evaluation strategies in the field.

Evaluating large language models (LLMs) requires a systematic approach that integrates various methodologies and metrics tailored to their unique characteristics. This section explores diverse evaluation metrics and techniques foundational for assessing LLM performance across applications. As illustrated in **??**, the hierarchical structure of model evaluation techniques for LLMs highlights core and advanced evaluation metrics, innovative frameworks, meta-evaluation and benchmarking strategies, and future directions in model evaluation. Table 1 presents a detailed summary of the methods and techniques employed for evaluating large language models, categorizing them into key areas of focus. Table 5 presents a comprehensive comparison of different evaluation methods for large language models, detailing the core metrics, advanced techniques, and frameworks for bias mitigation employed in assessing their performance. By examining both traditional and advanced

metrics, we aim to elucidate how these strategies contribute to understanding model reliability and effectiveness. The subsequent subsection provides an in-depth exploration of specific evaluation metrics and techniques employed in this context.

5.1 Evaluation Metrics and Techniques

Method Name	Core Metrics	Advanced Techniques	Task-Specific Metrics
VERITAS[3]	Accuracy	Ece	Exact Match
SCM[24]	Answer Accuracy	Confidence Calibration Metrics	Memory Retrieval Recall
CDI[27]	Effective Sample Size	Confidence Scores	Coverage Metrics
STDG[26]	Dialogue Diversity	Confidence Calibration Metrics	Subgoal Completion
RRT[4]	Greedy Decoding Accuracy	Confidence Calibration Metrics	Pass@16 Score

Table 2: This table presents a comparative analysis of various methods used in the evaluation of large language models (LLMs), detailing their core metrics, advanced techniques, and task-specific metrics. The methods include VERITAS, SCM, CDI, STDG, and RRT, each employing unique metrics to address specific challenges in LLM performance assessment.

Table 2 provides a detailed overview of the different evaluation methods applied to large language models (LLMs), highlighting the core metrics, advanced techniques, and task-specific metrics utilized by each method. A comprehensive evaluation of LLMs involves a variety of metrics and techniques to accurately assess their performance. Core metrics such as accuracy, precision, recall, and F1-score are essential for evaluating the reliability and coherence of LLM outputs, ensuring adherence to task instructions and high linguistic acceptability [3]. Advanced techniques address LLM-specific challenges, such as confidence calibration metrics like Expected Calibration Error (ECE) and AUROC, which offer insights into prediction confidence and reliability [30]. Performance assessments also include metrics for answer accuracy, memory retrieval recall, and summarization coherence, crucial for evaluating language understanding and generation [24].

In question answering and reasoning tasks, metrics like exact match, QA-F1, and macro precision evaluate model effectiveness, as demonstrated by the EnsReas method [27]. Discrepancies in benchmarks necessitate innovative frameworks like CONFIDENCE-DRIVEN INFERENCE (CDI), which combines LLM and human annotations based on confidence scores for statistically valid estimates. Task-oriented dialogue systems use metrics such as dialogue diversity, subgoal completion, and character consistency, complemented by human ratings [26]. In programming and code generation, metrics like greedy decoding accuracy and Pass@16 score measure performance, setting new field benchmarks [4].

5.2 Innovative Evaluation Frameworks

Method Name	Evaluation Techniques	Performance Analysis	Bias Mitigation
RGV[58]	Voting Mechanism	Specific Dimensions	Reducing Biases
TALEC[60]	Correlations With Human	Specific Evaluation Criteria	Flexible Evaluation Framework
ALLURE[19]	F1 Score	Specific Dimensions	Iterative Learning

Table 3: Summary of evaluation methods for large language models, detailing their evaluation techniques, performance analysis, and bias mitigation strategies. The table highlights the unique approaches of RGV, TALEC, and ALLURE in advancing the accuracy and fairness of LLM evaluations.

Innovative frameworks have advanced LLM evaluation, offering nuanced performance insights. Table 3 presents a comparative analysis of innovative evaluation frameworks for large language models, emphasizing their methodologies, performance metrics, and strategies for bias mitigation. The Reference-Guided Verdict Method (RGV) enhances evaluation accuracy by using multiple LLMs as judges, reducing biases and improving robustness [58]. Factored evaluation mechanisms dissect chatbot responses into specific dimensions, providing granular performance analysis and targeted improvements [61]. TALEC combines zero-shot and few-shot learning to enhance judge models' evaluation capabilities, improving adaptability and precision [60]. Exploring self-bias in LLM outputs through statistical bias and distance skewness metrics offers insights into inherent evaluation biases, contributing to more equitable methods [56]. These frameworks collectively refine LLM evaluations, reducing human annotator dependence and enhancing performance across applications [19].

5.3 Meta-Evaluation and Benchmarking

Benchmark	Size	Domain	Task Format	Metric
LIMBS[62]	1,000	Conversational AI	Language Style Imitation	Human Evaluation, Automatic Evaluation
TrustScore[33]	1,000	Question Answering	Open-ended Question Answering	TrustBC, TrustF C
ViLLM-Eval[63]	32,296	Education	Multiple-choice Questions	Perplexity, Accuracy
SELFEXP[64]	100	Sentiment Analysis	Sentiment Classification	Comprehensiveness, Sufficiency
AES-LLM[42]	20	Education	Essay Scoring	Spearman's r, ICC
LLM-B[65]	23	Legal Reasoning	Multiple Choice Questions	Accuracy, F1-score
LLM-BWL[66]	87	Behavioral Weight Loss	Message Evaluation	Helpfulness Rating
PVQ[67]	50,000	Psychology	Value Stability Assessment	Rank-Order Stability, Ipsative Stability

Table 4: This table presents a comprehensive overview of various benchmarks used for evaluating large language models (LLMs) across different domains and task formats. It includes details on the size of each benchmark, the specific domain it pertains to, the task format employed, and the metrics used for evaluation. These benchmarks serve as a foundation for assessing the effectiveness and reliability of LLMs in diverse applications.

Meta-evaluation and benchmarking are critical for assessing LLMs, providing insights into evaluation methodologies' effectiveness and reliability. Peer Rank (PR) and Peer Discussion (PD) frameworks leverage collective model intelligence for accurate evaluations [59]. The Reference-Guided Verdict Method (RGV) employs majority voting among annotators and LLM judges, using kappa statistics for agreement assessment [58]. Table 4 provides a detailed enumeration of representative benchmarks essential for the meta-evaluation and benchmarking of large language models, highlighting the diversity in domain applications and evaluation metrics. Benchmarking establishes standardized metrics and datasets, identifying model strengths and weaknesses for targeted improvements. Meta-evaluation techniques in benchmarking analyze performance across dimensions, addressing language-specific challenges and biases, promoting equitable evaluations across linguistic and cultural settings [68, 20, 18, 69, 19].

5.4 Future Directions in Model Evaluation

Future LLM evaluation requires dynamic, comprehensive frameworks addressing current benchmark inadequacies. Expanding evaluation methods to include diverse tasks and competencies is crucial for capturing LLM capabilities [65, 18]. Integrating low-resource languages ensures global inclusivity [69]. Refining methods to assess LLM cognitive capacities and ethical implications ensures technical proficiency and ethical responsibility [70]. Non-uniform weights in performance correlation suggest a need for nuanced frameworks [32]. Dynamic methods adapting to LLM capabilities, interdisciplinary approaches, and improved dataset quality are essential for refining benchmarks [71, 72]. Bridging gaps between LLM and human evaluators is critical [73]. Future research should enhance evaluation methods for LLM special capabilities, such as handling hallucinations and contextual memory [60]. Developing frameworks for continual self-improvement and integrating model editing for self-correction are promising [36]. Reliable methods for assessing LLM credences and refining peer review processes are essential [30, 20]. Addressing these directions advances LLM evaluation towards robust, inclusive, and ethically grounded methodologies.

Feature	Evaluation Metrics and Techniques	Innovative Evaluation Frameworks	Meta-Evaluation and Benchmarking
Core Metrics	Accuracy, F1-score	Not Specified	Standardized Benchmarks
Advanced Techniques	Confidence Calibration	Factored Evaluation	Peer Rank, Discussion
Bias Mitigation	Not Specified	Bias Reduction	Equitable Evaluations

Table 5: This table provides a comparative analysis of various evaluation methods for large language models (LLMs), categorizing them into core metrics, advanced techniques, and frameworks for bias mitigation. It highlights the integration of traditional and innovative evaluation strategies, emphasizing the importance of standardized benchmarks and equitable evaluations to enhance model reliability and effectiveness.

6 AI Reliability and Trustworthiness

The growing reliance on artificial intelligence (AI) systems across sectors underscores the critical importance of AI reliability and trustworthiness. This section examines how AI reliability shapes user trust and impacts the efficacy of AI technologies, particularly in high-stakes environments like healthcare and software development, where accuracy and consistency are paramount.

6.1 Importance of AI Reliability

AI reliability is crucial for ensuring the trustworthiness of outputs from large language models (LLMs), especially in critical applications such as healthcare and software development. The AntGLM-Med-10B model illustrates the importance of AI reliability in medical contexts, demonstrating competitive performance in medical question answering [7]. In software development, reliable AI improves code generation accuracy, as evidenced by studies on LLMs like ChatGPT and Bard, which transform natural language prompts into executable code, thereby reducing errors and manual coding [74, 75]. Techniques such as Multi-Perspective Self-Consistency (MPSC) enhance AI reliability by integrating diverse outputs, leading to improved coding benchmark performance [74]. The LPW method further underscores the necessity of trustworthy LLM outputs in programming tasks.

In healthcare, the EnsReas method enhances LLM prediction reliability, crucial for patient care, by addressing LLM uncertainty and facilitating confident clinical decision-making [1, 76, 19, 27]. Critique-based supervision in LLMs also enhances AI reliability, reinforcing trust in AI systems. The VERITAS model exemplifies a unified approach to enhancing AI reliability across applications, emphasizing the need for dependable AI systems to generate consistent outcomes [28, 51, 31, 77]. Structured guidelines and safety training, as seen in Guide-Align, help AI systems navigate risks and ensure high-quality outputs. Comprehensive trustworthiness evaluations across dimensions such as reliability and safety reinforce user confidence in AI applications.

Moreover, trustworthy evaluations aligned with user-defined standards are essential. Reliable evaluation processes ensure integrity in LLM outputs, particularly in critical domains where inaccuracies can have serious consequences. Studies show that LLMs can effectively assess software requirements quality, enhancing development quality and reducing costs [28, 40]. Structured guidelines through safety-trained models mitigate risks like biased content generation and privacy violations. The ASC framework illustrates AI reliability benefits by improving atomic fact recall and reducing hallucinations, producing comprehensive and accurate responses across applications [40, 18, 19, 78, 1]. The importance of AI reliability is further emphasized by the need for careful reward design in reinforcement learning to improve reasoning accuracy [4].

6.2 Factors Influencing AI Reliability

The reliability of LLMs is shaped by multiple factors affecting their performance and trustworthiness. Human evaluations and feedback can introduce biases into model training and evaluation processes [72]. Small and non-representative feedback pools complicate LLM reliability, leading to skewed behaviors and outputs [79]. This underscores the necessity for comprehensive and representative datasets to enhance LLM generalizability [80]. High costs and low adaptability of benchmarks also challenge AI reliability, as they may not account for LLM task dynamics and can introduce evaluator model biases [20, 81].

Safety concerns, particularly in critical applications, highlight the need for reliable LLM outputs. Explainability in LLMs is crucial for user trust and deployment in high-stakes environments [80]. The inconclusiveness of evidence regarding LLM credences raises doubts about current assessment techniques, emphasizing the need for robust evaluation frameworks [30].

6.3 Strategies to Enhance AI Reliability

Enhancing LLM reliability is essential for effective deployment in high-stakes applications. Improved calibration techniques, like multicalibration, ensure reliable confidence scores across diverse datasets [46]. Systematic frameworks for evaluating LLM confidence, such as those proposed by [8], improve reliability by aligning internal confidence with expressed outputs. Rejection mechanisms, like the Reinforcement Learning from Knowledge Feedback (RLKF) framework, train LLMs to refuse

questions beyond their knowledge, reducing erroneous outputs [82, 8, 83, 46]. These strategies enhance AI reliability, aligning confidence with performance and improving output interpretability for better decision-making.

Integrating explainability features into LLMs is vital for reliability, especially in sectors like healthcare and finance where transparency is crucial for user trust. LLMs' potential to generate natural language explanations can redefine interpretability, particularly in auditing and risk assessment contexts. Addressing challenges like hallucinated explanations and computational costs is necessary to realize this potential. Incorporating human domain knowledge into LLMs enhances predictive analytics and decision-making, bridging the gap between machine learning techniques and expert insights [84, 1, 85].

6.4 Reliability and Trustworthiness Concerns

The reliability and trustworthiness of AI outputs, particularly from LLMs, remain under scrutiny. Hallucinations, where LLMs produce incorrect content, are a significant issue exacerbated by biases in training datasets [3]. This challenges the need for robust evaluation frameworks ensuring factual integrity, especially in high-stakes applications like medical diagnostics [7]. Reward hacking during training, where models exploit signals to maximize perceived performance, can result in unreliable outputs [4]. Human evaluations introduce concerns of subjectivity and inconsistency, affecting reliability assessments [72]. Existing benchmarks, with high costs and low adaptability, may not adequately capture LLM task dynamics [81]. The lack of transparency and explainability in LLM outputs poses significant trustworthiness concerns, particularly where decision rationale is critical [80].

7 Challenges and Future Directions

In the rapidly evolving landscape of large language models (LLMs), addressing multifaceted challenges is imperative to enhance their effectiveness and reliability. A thorough understanding of these challenges provides a foundational context for exploring the limitations inherent in LLMs, particularly regarding their performance metrics and operational efficacy. The following subsection delves into specific challenges and limitations faced by LLMs, highlighting critical areas requiring attention for future advancements in the field.

7.1 Challenges and Limitations

LLMs face numerous challenges and limitations that hinder their ability to consistently achieve desired performance metrics across diverse applications. A significant challenge is the inherent self-bias within LLMs, which undermines their self-refinement capabilities and leads to suboptimal performance [25]. This self-bias complicates the models' ability to manage multiple sub-tasks effectively, a critical requirement for comprehensive performance in complex recommendation systems [86].

LLMs' reliance on simple projects for assessments may not generalize well to complex requirements, necessitating manual verification of outputs [40]. The challenge is compounded by difficulties in integrating qualitative expert knowledge into predictive analytics, particularly in standardizing subjective insights [85]. The focus on grammatical and lexical complexity in research overlooks other factors influencing argument comprehensibility, complicating performance metrics achievement [87].

The inadequate exploration of LLM memory capabilities and the absence of a theoretical framework to explain these mechanisms limit the models' potential in tasks requiring advanced reasoning, as existing benchmarks fail to capture long-term trends and complexities [24, 88]. Moreover, the lack of error correction mechanisms poses significant obstacles to achieving effective reasoning capabilities [35].

Achieving truly representative diversity among evaluators is challenging, affecting LLM reliability, as evaluator diversity is crucial for unbiased assessments [89]. The small sample size of LLMs in some studies may affect the generalizability of findings, complicating the establishment of robust evaluation frameworks [9].

The computational cost associated with generating multiple samples and the requirement for numerous LLM calls, as noted in the Atomic Self-Consistency (ASC) approach, represent a significant limitation, especially in resource-constrained scenarios [90]. Achieving reliable confidence estimates is challenging due to poor calibration performance and increased inference latency [91].

In medical applications, the reliance on the quality and comprehensiveness of training datasets impacts model performance [21]. The reliance on initial LLM predictions, which may contain errors, limits the ability to achieve desired performance metrics [34].

Misalignment between the model's expressed self-reasoning and its true internal confidence presents a primary challenge in achieving desired performance metrics [8]. Current self-reflection methods in LLMs face limitations in achieving desired performance metrics [15]. Moreover, the lack of mechanisms to verify the quality of LLM-generated evaluations is a significant limitation in current evaluation methods [2].

The reliance on LLM reasoning capacity can lead to inaccuracies in generated code despite structured approaches, as seen in the LPW method [22]. Current studies fall short in addressing multimodal approaches and do not capture the full scope of optimization methods applicable to compound AI systems [5]. The main challenge faced by LLMs is the lack of a single model that can generalize across different task formats and datasets, preventing effective hallucination detection [3].

Addressing these challenges is essential for advancing LLM capabilities and reliability. This includes developing more effective instruction-following mechanisms, enhancing evaluation frameworks to accommodate diverse languages and writing styles, and improving model editing techniques. Enhancing transparency and understanding user experiences are crucial for developing LLMs that are reliable, trustworthy, and capable of addressing ethical challenges like bias, accountability, and privacy. By clarifying LLM decision-making processes and engaging stakeholders in evaluating user interactions, we can foster a more responsible integration of LLMs into various applications, ultimately improving software quality and ethical AI practices [92, 40].

7.2 Scalability and Adaptability

Scalability and adaptability are critical factors in deploying and applying LLMs across diverse sectors. LLM scalability is often constrained by computational resources, impacting their deployment in real-time applications and environments with limited infrastructure [90]. The computational demands associated with generating multiple samples and making numerous LLM calls, as highlighted in the ASC approach, underscore the challenges in achieving scalable solutions [90].

Adaptability involves LLMs' capacity to adjust to varying contexts and tasks without extensive retraining. This adaptability is essential for LLMs to function effectively in dynamic environments where task requirements frequently change. The reliance on specific training datasets and initial conditions, as noted in medical applications, limits LLM adaptability, particularly in novel or unforeseen scenarios [21].

Integrating qualitative expert knowledge into predictive analytics poses additional challenges for scalability and adaptability, particularly in standardizing subjective insights across diverse applications [85]. Effective integration of such knowledge is crucial for developing LLMs that can generalize across different domains and applications.

The lack of mechanisms to verify the quality of LLM-generated evaluations presents a significant barrier to scalability, complicating the ability to adapt models to new tasks and environments [2]. Achieving reliable confidence estimates due to poor calibration performance further impacts LLM scalability, limiting their ability to provide consistent and trustworthy outputs across varying conditions [91].

Addressing these challenges requires developing more efficient computational techniques, creating adaptive learning frameworks, and establishing robust evaluation methods that accommodate diverse application needs. Enhancing LLM scalability and adaptability allows these models to evolve into versatile tools that meet diverse user needs and application scenarios while significantly improving software requirements quality assurance. Recent advancements in LLM serving systems focus on optimizing performance and efficiency while maintaining core decoding mechanisms, enabling these models to support stakeholders in requirements engineering more effectively. Research demonstrates that LLMs can evaluate software requirements quality according to established standards, propose im-

provements, and explain their decision-making processes, reducing development costs and enhancing software quality [40, 78].

7.3 Future Research Directions

Future research in LLMs should focus on several critical areas to address existing challenges and leverage technological advancements. One promising direction involves enhancing LLM feedback mechanisms to improve robustness against erroneous feedback and extending these methods to multilingual models or different domains [13]. This extension is crucial for developing adaptable and robust models capable of functioning effectively across diverse linguistic and contextual settings.

Exploring hyperparameter optimization within the LLM2LLM framework and integrating it with other LLM techniques such as prompt tuning and few-shot learning presents significant opportunities for improving model performance [25]. These advancements could lead to more efficient and effective LLMs capable of handling various tasks with minimal additional training.

In dialogue systems, future research should focus on maintaining general conversational skills while utilizing self-talk data and investigate using negative signals for further training improvements [26]. This approach could enhance the naturalness and effectiveness of LLMs in generating dialogue, making them more reliable and versatile conversational agents.

Enhancing commit message quality and exploring additional methods to generate informative reasoning paths are critical for improving program repair outcomes [93]. Future research could focus on refining these aspects to ensure more accurate and contextually relevant code suggestions, thereby enhancing LLM utility in software development.

Exploring enhancements in LLM calibration and extending the method's applicability to other NLP tasks and languages is an important area for future research [27]. Improved calibration techniques could lead to more reliable model outputs, ensuring LLMs produce consistent and trustworthy content across various applications.

Applying Reward Refinement Techniques (RRT) in larger models and various inference-time search strategies to improve reasoning capabilities is another promising research avenue [4]. These strategies could enhance LLM reasoning accuracy, making them more effective in complex decision-making tasks.

By pursuing these research directions, LLMs can evolve towards developing models that are more robust, adaptable, and ethically sound. This evolution is critical for addressing unique challenges such as hallucination, accountability, and bias, particularly pertinent to LLMs. Interdisciplinary collaboration and establishing tailored ethical frameworks will ensure these models can be deployed effectively across diverse sectors, enhancing transparency and reliability in applications ranging from academic research to information dissemination. Integrating fine-tuned LLMs in systematic literature reviews exemplifies the potential for these advancements to streamline research methodologies, setting new standards for accuracy and efficiency in the face of growing academic demands [51, 10, 92].

8 Conclusion

This survey has elucidated the critical roles of internal consistency, self-feedback, and reliability in the development and application of large language models (LLMs), emphasizing their significance in enhancing AI system performance across various domains. Internal consistency is essential for ensuring coherent and contextually appropriate outputs, particularly in high-stakes environments such as healthcare and education, where precision and coherence are critical [23]. Self-feedback mechanisms empower LLMs to iteratively refine their outputs, thereby improving their ability to generate accurate and reliable responses. The interactive self-reflection method significantly reduces hallucination in LLM-generated answers, showcasing its effectiveness and scalability [23].

Reliability is a cornerstone for the trustworthiness of LLMs, with frameworks like VERITAS demonstrating the potential for a unified approach to judging the reliability of LLMs, outperforming existing models while maintaining competitive performance [3]. The SaySelf method successfully reduces calibration errors and improves LLMs' confidence estimates, highlighting its potential impact on future AI systems [91].

The integration of diverse feedback in reinforcement learning with human feedback (RLHF) is crucial for developing AI systems that are more aligned with human values and capable of operating in complex moral and social landscapes [89]. The combination of human intuition with advanced machine learning techniques presents a promising direction for future research, advocating for empirical validation of the proposed frameworks [85].

The potential impact of these elements on the future development of AI systems is profound. Larger models, while showing better performance, face saturation without advancements in pretrained models, reiterating the importance of continuous innovation in LLM capabilities [88]. The self-verification method significantly enhances the reasoning capabilities of LLMs, providing a promising direction for future research in improving AI systems [35].

References

[1] Alexander Goldberg, Ihsan Ullah, Thanh Gia Hieu Khuong, Benedictus Kent Rachmat, Zhen Xu, Isabelle Guyon, and Nihar B. Shah. Usefulness of llms as an author checklist assistant for scientific papers: Neurips'24 experiment, 2024.

[2] Shreya Shankar, J. D. Zamfirescu-Pereira, Björn Hartmann, Aditya G. Parameswaran, and Ian Arawjo. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences, 2024.

[3] Rajkumar Ramamurthy, Meghana Arakkal Rajeev, Oliver Molenschot, James Zou, and Nazneen Rajani. Veritas: A unified approach to reliability evaluation, 2024.

[4] Jiaxuan Gao, Shusheng Xu, Wenjie Ye, Weilin Liu, Chuyi He, Wei Fu, Zhiyu Mei, Guangju Wang, and Yi Wu. On designing effective rl reward at training time for llm reasoning, 2024.

[5] Matthieu Lin, Jenny Sheng, Andrew Zhao, Shenzhi Wang, Yang Yue, Yiran Wu, Huan Liu, Jun Liu, Gao Huang, and Yong-Jin Liu. Llm-based optimization of compound ai systems: A survey, 2024.

[6] Juyeon Heo, Christina Heinze-Deml, Oussama Elachqar, Shirley Ren, Udhay Nallasamy, Andy Miller, Kwan Ho Ryan Chan, and Jaya Narain. Do llms "know" internally when they follow instructions?, 2024.

[7] Andrew M. Bean, Karolina Korgul, Felix Krones, Robert McCraith, and Adam Mahdi. Do large language models have shared weaknesses in medical question answering?, 2024.

[8] Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. *arXiv preprint arXiv:2405.16282*, 2024.

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

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

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

[12] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

[13] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

[14] Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Rong-Hua Li, Yi Wang, Zhonghao Wang, Feiyu Xiong, and Zhiyu Li. Internal consistency and self-feedback in large language models: A survey, 2024.

[15] Zhiheng Xi, Dingwen Yang, Jixuan Huang, Jiafu Tang, Guanyu Li, Yiwen Ding, Wei He, Boyang Hong, Shihan Do, Wenyu Zhan, Xiao Wang, Rui Zheng, Tao Ji, Xiaowei Shi, Yitao Zhai, Rongxiang Weng, Jingang Wang, Xunliang Cai, Tao Gui, Zuxuan Wu, Qi Zhang, Xipeng Qiu, Xuanjing Huang, and Yu-Gang Jiang. Enhancing llm reasoning via critique models with test-time and training-time supervision, 2024.

[16] Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. Self-contrast: Better reflection through inconsistent solving perspectives. *arXiv preprint arXiv:2401.02009*, 2024.

[17] Marco AF Pimentel, Clément Christophe, Tathagata Raha, Prateek Munjal, Praveen K Kanithi, and Shadab Khan. Beyond metrics: A critical analysis of the variability in large language model evaluation frameworks, 2024.

[18] 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. *arXiv preprint arXiv:2308.07902*, 2023.

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

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

[21] Qiang Li, Xiaoyan Yang, Haowen Wang, Qin Wang, Lei Liu, Junjie Wang, Yang Zhang, Mingyuan Chu, Sen Hu, Yicheng Chen, Yue Shen, Cong Fan, Wangshu Zhang, Teng Xu, Jinjie Gu, Jing Zheng, and Guannan Zhang Ant Group. From beginner to expert: Modeling medical knowledge into general llms, 2024.

[22] Chao Lei, Yanchuan Chang, Nir Lipovetzky, and Krista A. Ehinger. Planning-driven programming: A large language model programming workflow, 2025.

[23] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, 2023.

[24] Bing Wang, Xinnian Liang, Jian Yang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. Enhancing large language model with self-controlled memory framework, 2024.

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

[26] Dennis Ulmer, Elman Mansimov, Kaixiang Lin, Justin Sun, Xibin Gao, and Yi Zhang. Bootstrapping llm-based task-oriented dialogue agents via self-talk, 2024.

[27] Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candès, and Dan Jurafsky. Can unconfident llm annotations be used for confident conclusions?, 2024.

[28] Yi Luo, Zhenghao Lin, Yuhao Zhang, Jiashuo Sun, Chen Lin, Chengjin Xu, Xiangdong Su, Yelong Shen, Jian Guo, and Yeyun Gong. Ensuring safe and high-quality outputs: A guideline library approach for language models, 2024.

[29] Shichao Sun, Junlong Li, Weizhe Yuan, Ruifeng Yuan, Wenjie Li, and Pengfei Liu. The critique of critique, 2024.

[30] Geoff Keeling and Winnie Street. On the attribution of confidence to large language models, 2024.

[31] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

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

[33] Danna Zheng, Danyang Liu, Mirella Lapata, and Jeff Z. Pan. Trustscore: Reference-free evaluation of llm response trustworthiness, 2024.

[34] Chia-Hsuan Chang, Mary M. Lucas, Yeawon Lee, Christopher C. Yang, and Grace Lu-Yao. Beyond self-consistency: Ensemble reasoning boosts consistency and accuracy of llms in cancer staging, 2024.

[35] Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. *arXiv preprint arXiv:2212.09561*, 2022.

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

[37] Karan Taneja and Ashok Goel. Can active label correction improve llm-based modular ai systems?, 2024.

[38] Ming Li, Yanhong Li, and Tianyi Zhou. What happened in llms layers when trained for fast vs. slow thinking: A gradient perspective, 2024.

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

[40] Sebastian Lubos, Alexander Felfernig, Thi Ngoc Trang Tran, Damian Garber, Merfat El Mansi, Seda Polat Erdeniz, and Viet-Man Le. Leveraging llms for the quality assurance of software requirements, 2024.

[41] Yash Saxena, Sarthak Chopra, and Arunendra Mani Tripathi. Evaluating consistency and reasoning capabilities of large language models. *arXiv preprint arXiv:2404.16478*, 2024.

[42] Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring, 2024.

[43] Zhihui Xie, Jizhou Guo, Tong Yu, and Shuai Li. Calibrating reasoning in language models with internal consistency, 2024.

[44] Yuhang Wu, Yingfei Wang, Chu Wang, and Zeyu Zheng. Large language model enhanced machine learning estimators for classification, 2024.

[45] Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. Data-efficient fine-tuning for llm-based recommendation, 2024.

[46] Gianluca Detommaso, Martin Bertran, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in llms, 2024.

[47] Zhaopeng Feng, Yan Zhang, Hao Li, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. Improving llm-based machine translation with systematic self-correction. *arXiv preprint arXiv:2402.16379*, 2024.

[48] Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. Reasoning aware self-consistency: Leveraging reasoning paths for efficient llm sampling, 2025.

[49] Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation, 2023.

[50] Leitian Tao and Yixuan Li. Your weak llm is secretly a strong teacher for alignment, 2024.

[51] 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.

[52] Jianqiao Lu, Wanjun Zhong, Wenyong Huang, Yufei Wang, Qi Zhu, Fei Mi, Baojun Wang, Weichao Wang, Xingshan Zeng, Lifeng Shang, Xin Jiang, and Qun Liu. Self: Self-evolution with language feedback, 2024.

[53] Keshav Ramji, Young-Suk Lee, Ramón Fernandez Astudillo, Md Arafat Sultan, Tahira Naseem, Asim Munawar, Radu Florian, and Salim Roukos. Self-refinement of language models from external proxy metrics feedback, 2024.

[54] Shaohui Peng, Xing Hu, Qi Yi, Rui Zhang, Jiaming Guo, Di Huang, Zikang Tian, Ruizhi Chen, Zidong Du, Qi Guo, Yunji Chen, and Ling Li. Self-driven grounding: Large language model agents with automatical language-aligned skill learning, 2023.

[55] Chunyang Jiang, Chi-min Chan, Wei Xue, Qifeng Liu, and Yike Guo. Importance weighting can help large language models self-improve. *arXiv preprint arXiv:2408.09849*, 2024.

[56] Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. Pride and prejudice: Llm amplifies self-bias in self-refinement, 2024.

[57] Jeremy Straub and Zach Johnson. Initial development and evaluation of the creative artificial intelligence through recurring developments and determinations (cairdd) system, 2024.

[58] Sher Badshah and Hassan Sajjad. Reference-guided verdict: Llms-as-judges in automatic evaluation of free-form text, 2024.

[59] Ruosen Li, Teerth Patel, and Xinya Du. Prd: Peer rank and discussion improve large language model based evaluations, 2024.

[60] Kaiqi Zhang, Shuai Yuan, and Honghan Zhao. Talec: Teach your llm to evaluate in specific domain with in-house criteria by criteria division and zero-shot plus few-shot, 2024.

[61] Bhashithe Abeysinghe and Ruhan Circi. The challenges of evaluating llm applications: An analysis of automated, human, and llm-based approaches, 2024.

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

[63] Trong-Hieu Nguyen, Anh-Cuong Le, and Viet-Cuong Nguyen. Villm-eval: A comprehensive evaluation suite for vietnamese large language models, 2024.

[64] Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207*, 2023.

[65] 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.

[66] Zhuoran Huang, Michael P. Berry, Christina Chwyl, Gary Hsieh, Jing Wei, and Evan M. Forman. Comparing large language model ai and human-generated coaching messages for behavioral weight loss, 2023.

[67] Grgur Kovač, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, and Pierre-Yves Oudeyer. Stick to your role! stability of personal values expressed in large language models. *Plos one*, 19(8):e0309114, 2024.

[68] Qusai Khraisha, Sophie Put, Johanna Kappenberg, Azza Warraitch, and Kristin Hadfield. Can large language models replace humans in the systematic review process? evaluating gpt-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages, 2023.

[69] Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim, Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. Mm-eval: A multilingual meta-evaluation benchmark for llm-as-a-judge and reward models, 2024.

[70] Bram M. A. van Dijk, Tom Kouwenhoven, Marco R. Spruit, and Max J. van Duijn. Large language models: The need for nuance in current debates and a pragmatic perspective on understanding, 2023.

[71] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*, 2023.

[72] Sonia Meyer, Shreya Singh, Bertha Tam, Christopher Ton, and Angel Ren. A comparison of llm finetuning methods evaluation metrics with travel chatbot use case, 2024.

[73] Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following, 2024.

[74] Baizhou Huang, Shuai Lu, Weizhu Chen, Xiaojun Wan, and Nan Duan. Enhancing large language models in coding through multi-perspective self-consistency. *arXiv preprint arXiv:2309.17272*, 2023.

[75] Lincoln Murr, Morgan Grainger, and David Gao. Testing llms on code generation with varying levels of prompt specificity, 2023.

[76] Maia Kotelanski, Robert Gallo, Ashwin Nayak, and Thomas Savage. Methods to estimate large language model confidence. *arXiv preprint arXiv:2312.03733*, 2023.

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

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

[79] Hannah Rose Kirk, Andrew M Bean, Bertie Vidgen, Paul Röttger, and Scott A Hale. The past, present and better future of feedback learning in large language models for subjective human preferences and values. *arXiv preprint arXiv:2310.07629*, 2023.

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

[81] 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.

[82] Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback. *arXiv preprint arXiv:2403.18349*, 2024.

[83] Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. Calibrating long-form generations from large language models. *arXiv preprint arXiv:2402.06544*, 2024.

[84] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models, 2024.

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

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

[87] Carlos Carrasco-Farre. Large language models are as persuasive as humans, but how? about the cognitive effort and moral-emotional language of llm arguments, 2024.

[88] Chanjun Park and Hyeonwoo Kim. Understanding llm development through longitudinal study: Insights from the open ko-llm leaderboard, 2024.

[89] Kristian González Barman, Simon Lohse, and Henk de Regt. Reinforcement learning from human feedback: Whose culture, whose values, whose perspectives?, 2025.

[90] Raghuveer Thirukovalluru, Yukun Huang, and Bhuwan Dhingra. Atomic self-consistency for better long form generations, 2024.

[91] Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. Sayself: Teaching llms to express confidence with self-reflective rationales, 2024.

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

[93] Toufique Ahmed and Premkumar Devanbu. Better patching using llm prompting, via self-consistency. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 1742–1746. IEEE, 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.