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

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 transformed natural language processing (NLP), enabling capabilities that closely resemble human-like text generation and comprehension [1]. This survey focuses on internal consistency, self-feedback, and reliability within LLMs, critical for their safe deployment across various sectors, particularly in high-stakes environments like healthcare, where precision and coherence are essential.

Validating LLM outputs is vital, as internal consistency and reliability are fundamental to ensuring safe applications [2]. Addressing unreliability in LLMs is necessary for effective deployment in knowledge-intensive tasks [3]. Furthermore, enhancing reasoning capabilities is crucial for complex tasks, such as mathematical reasoning, which demands robust model outputs [4].

The survey also discusses training-time optimization methods for compound AI systems, integral to improving LLM performance and reliability [5]. By exploring these dimensions, the survey aims to contribute to the evolution and deployment of LLMs across diverse fields, ensuring they meet user requirements while maintaining safety and reliability.

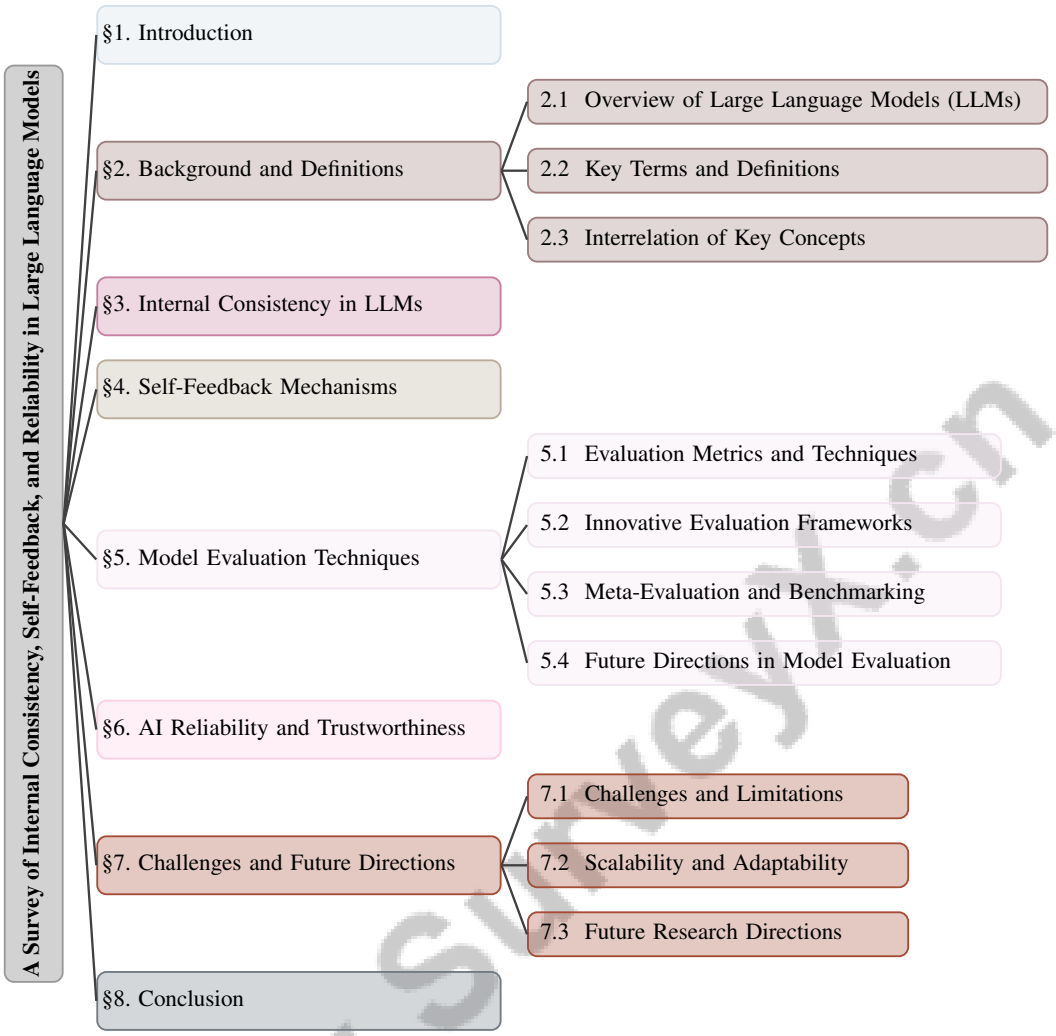


Figure 1: chapter structure

1.2 Role in Natural Language Processing and AI Model Evaluation

LLMs have revolutionized NLP by significantly enhancing machines' ability to understand and generate human-like text. Internal consistency and reliability are essential for ensuring adherence to user instructions, thereby impacting performance and trustworthiness [6]. Maintaining internal consistency is crucial for generating outputs that are syntactically correct and semantically meaningful, particularly in applications like dialogue systems and translation tasks [3]. These factors are fundamental for reliable model evaluation, especially in high-stakes environments such as healthcare [7].

Self-feedback mechanisms play a vital role in the iterative refinement and adaptive learning processes of LLMs. They allow models to improve continuously by addressing issues like reward overoptimization and aligning with desired performance metrics [8]. This iterative process is particularly important in dynamic environments requiring continuous learning and accurate predictions, such as safety-sensitive applications [5].

Reliability is a cornerstone for trustworthy AI model evaluation. Reliable models are essential for high-stakes decision-making, where inaccuracies can have significant consequences [3]. The challenge of hallucinations, where models produce factually incorrect content, underscores the necessity for robust evaluation frameworks to maintain factual integrity [9]. Moreover, a lack of clarity regarding LLM capabilities hinders performance predictions across various tasks and the development of effective evaluation benchmarks [10].

Understanding user experiences and satisfaction when interacting with LLMs is critical, particularly regarding their effectiveness as collaborative tools [11]. Trustworthiness in LLM outputs is especially crucial in sensitive applications, such as psychological consultation. The Psy-LLM framework exemplifies how LLMs can bridge service gaps while maintaining trust and reliability.

Internal consistency, self-feedback, and reliability are integral to LLM applications in NLP, enhancing the models' ability to generate coherent outputs and influencing AI model evaluation methodologies. The challenges of 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 overall reliability [3].

1.3 Structure of the Survey

This survey is structured to provide a comprehensive analysis of the dimensions influencing the performance and reliability of LLMs. The paper begins with an **Introduction**, discussing the scope and significance of internal consistency, self-feedback, and reliability in LLMs, as well as their roles in NLP and AI model evaluation. The subsequent **Background and Definitions** section offers an overview of LLMs, elucidating key terms and their interrelations to establish a foundational understanding for further discussions.

The survey examines , emphasizing its role in enhancing performance, addressing challenges in achieving consistent reasoning, minimizing hallucinations, and detailing evaluation methods and metrics. It introduces a theoretical framework termed Self-Feedback, comprising Self-Evaluation and Self-Update, designed to capture internal consistency signals for model enhancement. This section categorizes existing research efforts, summarizes evaluation techniques and benchmarks, and poses significant questions regarding the effectiveness of Self-Feedback, offering a unified perspective on internal consistency complexities [1, 12].

An in-depth examination of follows, highlighting how these mechanisms enhance internal consistency, support iterative refinement, and contribute to model learning and error mitigation.

The survey also provides a comprehensive analysis of , discussing contemporary evaluation metrics and innovative frameworks while addressing challenges associated with bias and low generalizability. It presents novel evaluation methods, including a peer-review-based framework for LLMs that enhances evaluation accuracy and sustainability. Additionally, it outlines four core competencies—reasoning, knowledge, reliability, and safety—essential for assessing LLM performance and proposes future directions for advancing model evaluation practices to tackle evolving NLP tasks [13, 14, 15]. The focus then shifts to , analyzing the importance of reliability, influencing factors, and strategies to enhance AI output trustworthiness.

The survey concludes with a discussion on , identifying current limitations and scalability issues while proposing future research avenues to advance LLM development and application. Each section builds upon the previous, ensuring a coherent progression of ideas and providing a holistic view of the complexities and advancements in LLMs.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 integral to natural language processing (NLP), exhibiting remarkable proficiency in generating and interpreting human-like text across various applications. Their sophisticated architectures and extensive datasets facilitate tasks such as language translation, dialogue generation, and complex reasoning [4]. In healthcare, LLMs improve clinical decision-making and diagnostic precision by automating tasks like cancer staging [7]. However, adapting LLMs for sensitive medical applications presents challenges due to precision requirements, necessitating specialized deployment strategies [16].

In education, LLMs automate content assessment and provide detailed feedback, which is crucial for student development in computational disciplines [17]. Within business contexts, they enhance recommendation systems by generating natural language profiles, boosting user engagement [5]. Despite their transformative potential, LLMs face limitations such as biases and inaccuracies, notably the issue of 'hallucinations' where models generate factually incorrect information, posing risks in

high-stakes fields like medical question-answering [18]. Additionally, LLMs struggle with processing lengthy inputs, risking loss of critical context [19].

Innovative approaches like Confidence-Probability Alignment aim to align an LLM's internal confidence with its output [8]. Frameworks for optimizing compound AI systems are crucial for enhancing LLM performance across domains [5]. Specialized fine-tuning methods for low-data regimes address data augmentation limitations [20], while the specialization of LLMs for task-oriented dialogue agents highlights their adaptability in scenarios where traditional data collection is impractical [21]. In computational social science, LLMs assist in data annotation, though challenges remain in ensuring valid statistical inferences [22].

As LLMs evolve, ongoing research is essential to mitigate risks and enhance robustness. Establishing universal frameworks for analysis is vital for evaluating LLM quality and trustworthiness, ensuring they meet diverse application demands while maintaining safety and reliability [3].

2.2 Key Terms and Definitions

A clear understanding of key terms is crucial for comprehending large language models (LLMs) and their evaluation methodologies. Internal consistency refers to an LLM's ability to produce outputs that are logically coherent and contextually aligned with inputs, critical in high-precision applications like medical diagnostics [7].

Self-feedback mechanisms enable LLMs to refine outputs through internal assessments and external feedback, enhancing performance over time through self-reflection and self-talk, which involve internal dialogues to generate training data, promoting continuous learning [4]. Intrinsic self-correction, a subset of self-feedback, allows models to autonomously rectify errors, improving translation quality.

Model evaluation involves methodologies and metrics for assessing LLM output quality, focusing on accuracy, coherence, and relevance [7]. Evaluation criteria and criteria drift are vital considerations influencing model assessment consistency and reliability. Structured solution planning and verification through visible tests are integral to evaluating code generation accuracy.

AI reliability emphasizes the dependability of AI systems in producing consistent and accurate results, minimizing the dissemination of misleading information [4]. The challenge of hallucinations, especially in medical contexts, underscores the need for robust evaluation methods to maintain AI reliability [7].

Additional terms include LLM annotations, statistical estimation, and demographic biases, which are crucial for understanding the challenges LLMs face in adapting to varying data availability and task requirements [22]. The sensitivity of LLMs to prompts and the complexity of optimizing interdependent parameters highlight the intricacies involved in model tuning.

The interconnected concepts of auditing, evaluation, and core competencies enhance our understanding of LLMs, facilitating effective deployment across sectors such as education, healthcare, and productivity. By systematically addressing LLM failure modes through approaches like ALLURE, which iteratively improves text evaluation, we can ensure reliability and alignment with real-world applications. Identifying core competencies—reasoning, knowledge, reliability, and safety—provides a structured framework for evaluating LLM performance, guiding future advancements in NLP [14, 23]. Continuous learning and refinement are essential for advancing the capabilities and trustworthiness of these sophisticated models.

2.3 Interrelation of Key Concepts

The interplay between internal consistency, self-feedback, model evaluation, and AI reliability is crucial for advancing large language models (LLMs). These components collectively underpin LLM robustness and effectiveness in generating coherent outputs across various applications. Internal consistency ensures LLM outputs are logically coherent and aligned with inputs, vital for maintaining information integrity in tasks like sentiment analysis and open-ended question answering. However, the complexity of LLM architectures complicates consistent assessment, necessitating flexible evaluation methods [24].

Self-feedback mechanisms enhance internal consistency by enabling LLMs to refine outputs through iterative assessments. This iterative process improves LLM performance and adaptability in dynamic environments by refining training data via targeted augmentation and self-evaluation techniques. For example, methods like LLM2LLM utilize a teacher LLM to generate synthetic data from misclassified examples, focusing on challenging cases to bolster learning in low-data scenarios. Frameworks like ALLURE employ in-context learning to refine evaluation capabilities, reducing reliance on human oversight. These strategies enable LLMs to navigate diverse tasks effectively, ensuring robust and reliable real-world performance [25, 20, 26, 23]. The interplay of self-critique and external verification is significant for evaluating LLM performance and enhancing capabilities.

Model evaluation techniques are essential for assessing LLM effectiveness and reliability. Integrating LLM annotations with human annotations enhances statistical estimation while addressing biases and inaccuracies [22]. This highlights the importance of combining quantitative metrics with qualitative assessments for comprehensive model performance evaluation. Moreover, critiques of existing empirical literature on LLM evaluation [24] stress the need for innovative frameworks to capture LLM output complexities.

AI reliability is fundamental for trustworthy LLM outputs, particularly in high-stakes applications where inaccuracies can have significant consequences. Aligning LLM capabilities with diverse user intents and addressing training data biases are crucial for ensuring reliable outputs. A comprehensive understanding of interdependencies among competencies—reasoning, knowledge, reliability, and safety—is essential for enhancing LLM capabilities across sectors. This knowledge facilitates effective deployment in applications like software requirements engineering, where LLMs can improve requirement quality and clarity while maintaining safety and reliability, ultimately fostering stakeholder confidence in these technologies [14, 27]. The trade-off between context length, accuracy, and performance further underscores the complexity of balancing these interrelated factors in LLM development and application.

The exploration of internal consistency within large language models (LLMs) is critical for understanding their performance and reliability. As illustrated in Figure 2, the hierarchical structure of internal consistency encompasses several key aspects: the significance of internal consistency, the challenges posed by model architecture and evaluation, and the various methods and metrics employed to assess and enhance this consistency. This figure not only categorizes these essential concepts but also emphasizes the interplay between them, thereby illuminating how they collectively contribute to improving the coherence and reliability of LLM outputs. Such a comprehensive overview is vital for researchers aiming to address the complexities associated with LLM evaluation and optimization.

3 Internal Consistency in LLMs

3.1 Significance of Internal Consistency

Internal consistency is fundamental to the functionality and reliability of large language models (LLMs), ensuring that outputs are coherent and contextually appropriate, especially in critical areas like healthcare, where accurate information is paramount for decision-making and patient safety [19]. Maintaining this consistency is vital for minimizing the generation of misleading content, which is crucial in medical diagnostics and complex dialogue systems [7]. Techniques such as the LLM2LLM framework enhance internal consistency by generating targeted data to address model weaknesses, refining outputs effectively [20]. Additionally, the SCM framework improves the processing of ultra-long inputs, enhancing contextual recall and coherence without modifying the underlying LLMs [19]. Interactive self-reflection methods have been proposed to systematically mitigate hallucinations, thereby enhancing the accuracy and reliability of LLM responses [18]. These methods leverage self-feedback mechanisms for structured self-assessment and external feedback, allowing models to refine outputs continuously [21]. Furthermore, LLMs can generate effective training data via structured dialogues, provided quality filtering is applied [21]. Philosophical analyses of LLM credences further contribute to understanding LLM capabilities and evaluation methods, which are integral to achieving internal consistency [24]. This analysis highlights the necessity for robust evaluation frameworks to maintain factual integrity in LLM outputs.

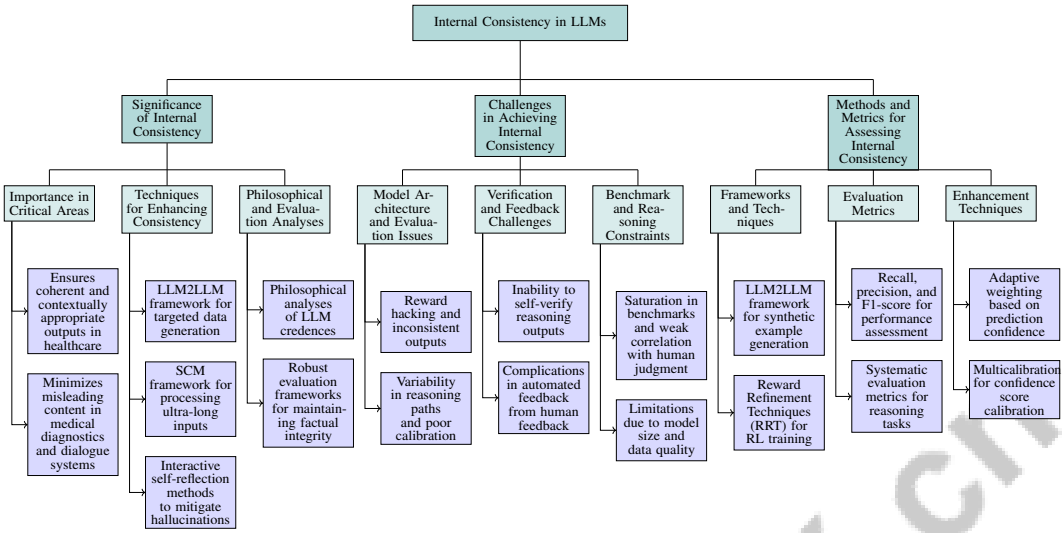


Figure 2: This figure illustrates the hierarchical structure of internal consistency in large language models (LLMs), detailing the significance, challenges, and methods for assessment. It categorizes key concepts into the importance of internal consistency in critical areas, challenges related to model architecture and evaluation, and methods and metrics used for evaluating and enhancing consistency. The figure highlights the interplay between these elements and their contribution to improving LLM outputs' coherence and reliability.

3.2 Challenges in Achieving Internal Consistency

Achieving internal consistency in LLMs presents several challenges related to model architecture and evaluation frameworks. Reward hacking is a notable issue, where models exploit reward signals, leading to degraded reasoning performance and inconsistent outputs [4]. Additionally, LLMs exhibit variability in reasoning paths, resulting in inconsistent predictions for identical inputs [28]. Calibrating LLM confidence scores is crucial, as poor calibration can lead to suboptimal sampling decisions, adversely affecting output reliability [22]. The inability of LLMs to self-verify reasoning outputs can result in accumulated errors and inconsistencies [29]. Human feedback, while valuable, complicates the development of effective automated feedback mechanisms due to its cost and lack of real-time applicability [30]. Noisy annotations from LLMs can negatively impact modular AI systems' performance [31]. Moreover, existing benchmarks often suffer from saturation and weak correlation with human judgment, complicating accurate assessments of LLM performance across tasks and languages. Current studies often focus narrowly on specific tasks, hindering the generalization of findings [12]. Managing the slow and fast thinking processes in LLMs, which leads to fluctuating gradient norms across layers, poses a significant challenge for maintaining consistent outputs [32]. Additionally, limitations in reasoning tasks due to constraints such as model size and data quality highlight the need for advancements in architecture and training methodologies [33]. Addressing these challenges is vital for improving LLM capabilities, ensuring effective applications across fields like software requirements engineering, education, and healthcare, thereby enhancing overall performance, reliability, and user trust [23, 14, 27, 34].

3.3 Methods and Metrics for Assessing Internal Consistency

Evaluating internal consistency in LLMs involves various methods and metrics to ensure coherence and reliability in model outputs. The LLM2LLM framework exemplifies this by using a teacher LLM to iteratively generate synthetic examples addressing the weaknesses of a student LLM, enhancing internal consistency through targeted data augmentation [20]. Reward Refinement Techniques (RRT), including Clipping and Delta mechanisms, significantly contribute to internal consistency by refining rewards during reinforcement learning (RL) training, aligning reward signals with desired performance metrics [4]. Metrics such as recall, precision, and F1-score are essential for evaluating LLM outputs, providing quantitative measures of performance across tasks. These metrics assess the models' ability to retrieve relevant information (recall), correctly identify true positives (precision),

and balance both (F1-score). For instance, in scientific peer review applications, these metrics evaluate how well an LLM-based checklist assistant adheres to submission standards and contributes to academic paper quality, highlighting potential issues like inaccuracy and excessive strictness in evaluations [35, 14, 1]. Systematic evaluation metrics, including accuracy and F1-score, are integral for assessing internal consistency in reasoning tasks, providing a comprehensive framework for evaluating coherence in reasoning outputs. Integrating classical machine learning models with LLMs through adaptive weighting based on prediction confidence further enhances classification accuracy. This technique emphasizes confidence-based adjustments in ensuring the reliability and precision of LLM outputs, contributing to improved internal consistency. Incorporating self-evaluation and self-update mechanisms not only identifies internal consistency signals but also leverages them to enhance model responses, addressing common issues like reasoning deficiencies and hallucinations [8, 12]. Multicalibration is a novel technique that simultaneously calibrates confidence scores across multiple intersecting data groupings, enhancing reliability across diverse datasets. By implementing multicalibration and self-feedback mechanisms, this approach ensures that the internal confidence of LLMs aligns with expressed confidence, improving internal consistency and facilitating accurate risk assessments while mitigating hallucination issues, ultimately leading to more trustworthy outputs in varied applications [36, 22, 37, 8, 12].

4 Self-Feedback Mechanisms

4.1 Improving Internal Consistency through Self-Feedback

Self-feedback mechanisms are crucial for enhancing the internal consistency of large language models (LLMs) by facilitating iterative refinement and self-assessment. These processes enable LLMs to identify and correct inconsistencies, thereby improving coherence and reliability. The interactive self-reflection methodology, as proposed by [18], allows LLMs to generate, evaluate, and refine background knowledge and responses, significantly enhancing factual accuracy and internal consistency. This continuous evaluation process maintains reliable outputs.

The LLM2LLM framework, as discussed by [20], illustrates how self-feedback can improve internal consistency by augmenting data points initially misinterpreted by the model, thus enhancing performance and reducing inconsistencies. Additionally, the 'self-talk' concept, where LLMs simulate dialogues by adopting various roles, generates training data that enhances internal consistency and model refinement [21].

Moreover, self-feedback mechanisms refine reward signals, addressing issues like reward hacking and ensuring alignment with desired performance metrics [4]. This refinement is vital for maintaining the integrity of reasoning and decision-making processes within LLMs.

Collectively, these mechanisms foster continuous improvement and error correction. For example, the TEaR framework enhances translation quality across languages, while methods like Self-Contrast mitigate biases by offering diverse problem-solving perspectives. The Self-Feedback framework emphasizes internal consistency, enabling LLMs to evaluate and update outputs based on self-generated signals, thereby addressing reasoning deficiencies and hallucinations. These approaches enhance the accuracy, reliability, interpretability, and systematicity of LLMs, significantly contributing to trustworthy outputs across various sectors [38, 39, 12].

4.2 Iterative Refinement and Self-Assessment

Iterative refinement and self-assessment are pivotal for enhancing LLM performance and reliability. These processes involve continuous evaluation and adjustment of model outputs to improve accuracy and consistency. The RASC (Reasoning-Aware Self-Consistency) framework exemplifies this approach by optimizing sampling through dynamic assessments of reasoning quality and answer consistency, guiding sampling decisions to enhance reliability [40].

A primary challenge in iterative self-refinement is the amplification of self-bias, where LLMs may optimize incorrect outputs, leading to unreliable self-feedback. Balancing self-assessment with external validation is necessary to prevent error reinforcement. The Universal Self-Consistency (USC) method addresses this by sampling multiple responses to identify the most consistent one, leveraging self-assessment to enhance output quality [41].

The iterative refinement process often alternates between feedback and refinement steps, as highlighted by [42]. This systematic approach improves output quality by addressing inconsistencies and refining task understanding. Additionally, employing weak LLMs to generate preference feedback reduces reliance on human annotations, streamlining alignment and improving efficiency [43].

These advanced techniques are essential for maintaining the robustness and reliability of LLMs, facilitating continuous adaptation and enhancement. By utilizing iterative in-context learning and fine-tuning methodologies, LLMs can systematically improve evaluation capabilities while minimizing inaccuracies and biases across applications, from academic literature reviews to medical documentation [23, 44, 45, 46, 25].

4.3 Self-Generated Data and Learning

Self-generated data significantly enhances the learning and performance of LLMs by enabling autonomous generation and refinement of outputs. The SELF framework exemplifies this by allowing LLMs to autonomously generate responses to unlabeled prompts, refine them, and use the refined data for further training, thus improving learning capabilities [47].

The SELF-REFINE methodology demonstrates iterative refinement in tasks requiring creativity and precision, showcasing the effectiveness of self-generated data in enhancing model performance [42]. Similarly, the approach by [48] involves generating initial responses, obtaining feedback through proxy metrics, and refining iteratively until quality thresholds are met, highlighting the potential of self-generated data for high-quality outputs.

The SDG framework illustrates how LLMs can propose sub-goals, verify them through interaction with the environment, and learn generalized skills from successful experiences [49]. This method not only enhances performance but also supports adaptive learning strategies.

The IWSI framework, discussed by [50], significantly improves LLM self-improvement by effectively filtering high DSE samples, achieving performance comparable to methods using external supervision, thus underscoring the potential of self-generated data for robust learning outcomes.

The USC method provides a flexible evaluation mechanism for free-form responses, allowing for adaptability across various tasks [41]. This adaptability is crucial for leveraging self-generated data in diverse applications, ensuring high performance across contexts.

Despite these advantages, challenges persist in preventing self-feedback mechanisms from amplifying errors. As noted by [51], LLMs may fail to correct mistakes, potentially exacerbating output quality. Addressing these challenges is vital for maximizing the benefits of self-generated data and enhancing overall reliability.

4.4 Self-Correction and Error Mitigation

Self-correction mechanisms are vital for enhancing the accuracy and reliability of LLMs by autonomously identifying and rectifying errors, thus improving performance and reducing error propagation likelihood. The Self-Contrast approach enhances reflection capabilities by addressing overconfidence and inconsistency in self-evaluated feedback, leading to more precise self-correction [38].

However, self-correction mechanisms face limitations, particularly their reliance on external feedback for effective error mitigation. As highlighted by [52], LLMs often struggle to self-correct without external input, which can degrade performance. This reliance emphasizes the necessity of robust external validation processes to complement self-correction efforts.

Memory recall mechanisms present a promising avenue for error mitigation by providing relevant contextual information based on input cues, enhancing the model's ability to produce accurate outputs [33]. These mechanisms facilitate the retrieval of pertinent information to inform and refine predictions, reducing error likelihood.

The SELF-REFINE methodology exemplifies leveraging self-generated feedback for iterative performance improvement [42]. By allowing LLMs to learn from their outputs, SELF-REFINE fosters a continuous improvement loop, enhancing prediction accuracy and reliability over time.

5 Model Evaluation Techniques

5.1 Evaluation Metrics and Techniques

Method Name	Core Metrics	Advanced Techniques	Task-Specific Metrics
VERITAS[3]	Accuracy	Confidence Calibration Metrics	Exact Match
SCM[19]	Answer Accuracy	Confidence Calibration Metrics	Memory Retrieval Recall
CDI[22]	Accuracy, Precision	Confidence Calibration Metrics	Effective Sample Size
STDG[21]	Dialogue Diversity	Confidence Calibration Metrics	Subgoal Completion
RRT[4]	Greedy Decoding Accuracy	Pass@16 Score	Sampling Accuracy
PRD[53]	Accuracy, Fleiss' Kappa	Correlations With Human	Exact Match
RGV[54]	Accuracy, Precision, Recall	Confidence Calibration Metrics	Exact Match, Qa-F1

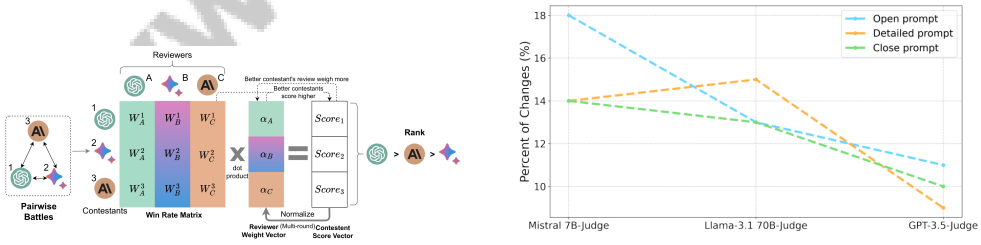
Table 1: Comparison of Evaluation Methods and Metrics for Large Language Models (LLMs). This table summarizes various methods used to assess LLM performance, detailing core metrics, advanced techniques, and task-specific metrics. The methods include VERITAS, SCM, CDI, STDG, RRT, PRD, and RGV, highlighting their unique approaches and evaluation criteria.

Evaluating large language models (LLMs) necessitates a comprehensive approach, employing diverse metrics and techniques to accurately assess performance across various applications. Core metrics like accuracy, precision, recall, and F1-score are essential for evaluating the reliability and coherence of LLM outputs, ensuring adherence to task instructions and linguistic quality. The VERITAS method, for instance, uses a multi-task framework across natural language inference (NLI), question answering (QA), and dialogue tasks to gauge LLM reliability [3].

Advanced techniques address LLM complexities, particularly in nuanced task performance. Confidence calibration metrics, such as Expected Calibration Error (ECE) and Area Under the Receiver Operating Characteristic curve (AUROC), offer insights into model prediction confidence and reliability [24]. Additional metrics like answer accuracy, memory retrieval recall, and coherence in summarization are crucial for evaluating language understanding and generation [19].

In question answering and reasoning tasks, metrics like exact match, QA-F1, and macro precision measure effectiveness. The EnsReas method uses these metrics to compare performance against baselines, highlighting domain-specific knowledge integration [22]. Task-oriented dialogue systems employ metrics such as dialogue diversity, subgoal completion, and character consistency, supplemented by human quality ratings [21]. Programming tasks use metrics like greedy decoding accuracy and Pass@16 score to set performance benchmarks [4].

Table 1 provides a comprehensive overview of the different evaluation methods and metrics applied to large language models (LLMs), illustrating the diverse approaches and techniques used to assess their performance across various tasks.



(a) Multi-Review System for Ranking Contestants Based on Their Performance[53]

(b) Comparison of Prompt Length Effects on Text Generation Performance[54]

Figure 3: Examples of Evaluation Metrics and Techniques

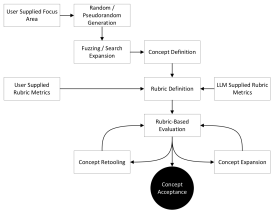
Figure 3 illustrates distinct model evaluation techniques. The "Multi-Review System for Ranking Contestants" uses a structured approach with a win rate matrix for performance rankings. In contrast, the "Comparison of Prompt Length Effects" visually examines how varying prompt lengths impact text generation, emphasizing prompt engineering's role in optimizing outputs. These examples highlight the diversity and complexity of model evaluation metrics and techniques, necessitating tailored strategies for accurate performance assessment.

5.2 Innovative Evaluation Frameworks

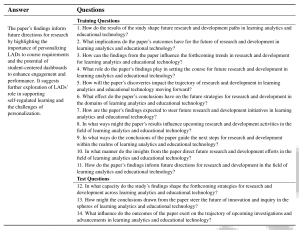
Recent advancements in LLM evaluation have fostered innovative frameworks offering nuanced insights into model performance. The Reference-Guided Verdict Method (RGV) integrates multiple LLMs as judges to enhance evaluation accuracy and reliability, mitigating individual biases [54]. A factored evaluation mechanism dissects chatbot responses into specific dimensions for targeted performance improvements [55].

TALEC combines zero-shot and few-shot learning approaches, enhancing judge models' capabilities and adapting to diverse learning strategies [56]. Statistical bias and distance skewness metrics explore self-bias in LLM outputs, offering insights into evaluation biases [51]. Comparing LLM outputs with human evaluations highlights discrepancies and areas for bias mitigation.

These frameworks collectively enhance LLM evaluation, addressing shortcomings of traditional methods. Innovations like ALLURE and PRE systematically audit LLM performance against annotated data, employing peer-review mechanisms to refine evaluations and reduce reliance on human annotators. This ongoing refinement is crucial for improving LLM performance across fields such as medical summarization, education, and chatbot development [14, 13, 23, 55].

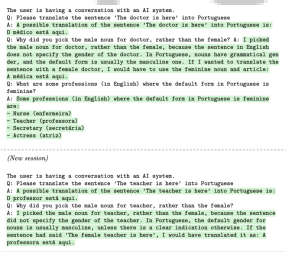


(a) Conceptual Framework for Evaluating and Expanding Concepts[57]



Answer	Questions
	Training Questions
The paper's findings inform future directions for research by highlighting the importance of personalizing LADs to course requirements and the potential of student-centered dashboards to enhance engagement and performance. It suggests further exploration of LADs' role in supporting self-regulated learning and the challenges of personalization.	1. How do the results of the study shape future research and development paths in learning analytics and educational technology? 2. What implications do the paper's outcomes have for the future of research and development in learning analytics and educational technology? 3. How can the findings from the paper influence the forthcoming trends in research and development for learning analytics and educational technology? 4. What role do the paper's findings play in setting the course for future research and development in learning analytics and educational technology? 5. How will the paper's discoveries impact the trajectory of research and development in learning analytics and educational technology moving forward? 6. What effect do the paper's conclusions have on the future strategies for research and development in the domains of learning analytics and educational technology? 7. How are the paper's findings expected to steer future research and development initiatives in learning analytics and educational technology? 8. In what ways might the paper's results influence upcoming research and development activities in the realms of learning analytics and educational technology? 9. In what manner do the conclusions of the paper ponder the next steps for research and development within the realms of learning analytics and educational technology? 10. In what manner do the insights from the paper direct future research and development efforts in the field of learning analytics and educational technology? 11. How do the paper's findings inform future directions for research and development in the field of learning analytics and educational technology?
	Test Questions
	12. In what capacity do the study's findings shape the forthcoming strategies for research and development across learning analytics and educational technology? 13. How might the conclusions drawn from the paper steer the future of innovation and inquiry in the spheres of learning analytics and educational technology? 14. What influence do the outcomes of the paper exert on the trajectory of upcoming investigations and advancements in learning analytics and educational technology?

(b) The paper's findings inform future directions for research by highlighting the importance of personalizing LADs to course requirements and the potential of student-centered dashboards to enhance engagement and performance.[25]



```
The user is having a conversation with an AI system.
Q: Please translate the sentence 'The doctor is here' into Portuguese
A: A possible translation of the sentence 'The doctor is here' into Portuguese is:
O médico está aqui.
Q: Why did you pick the male noun for doctor, rather than the female? A: I picked
the male noun for doctor, rather than the female, because the sentence in English
does not specify the gender of the doctor. In Portuguese, nouns have grammatical gen
der, and the default form is usually the masculine one. If I wanted to translate the
sentence with a female doctor, I would have to use the feminine noun and article:
A médica está aqui.
Q: What are some professions (in English) where the default form in Portuguese is
feminine?
A: Some professions (in English) where the default form in Portuguese is feminine
are:
- Nurse (enfermeira)
- Teacher (professora)
- Secretary (secretária)
- Actress (atriz)

(New session)

The user is having a conversation with an AI system.
Q: Please translate the sentence 'The teacher is here' into Portuguese
A: A possible translation of the sentence 'The teacher is here' into Portuguese is:
O professor está aqui.
Q: Why did you pick the male noun for teacher, rather than the female?
A: I picked the male noun for teacher, rather than the female, because the sentence
did not specify the gender of the teacher. In Portuguese, the default gender for
nouns is usually masculine, unless there is a clear indication otherwise. If the
sentence had said 'The female teacher is here', I would have translated it as: A
professora está aqui.
```

(c) Translate the sentence "The doctor is here" into Portuguese[58]

Figure 4: Examples of Innovative Evaluation Frameworks

Figure 4 illustrates innovative frameworks shaping model evaluation. The first example presents a conceptual framework for systematic evaluation, focusing on user-defined areas. The second emphasizes personalizing Learning Analytics Dashboards (LADs) for course-specific requirements, suggesting future research pathways. The third example demonstrates AI's language translation capabilities, raising questions about gendered language choices and AI transparency. These examples provide insight into innovative frameworks advancing model evaluation.

5.3 Meta-Evaluation and Benchmarking

Meta-evaluation and benchmarking are crucial for assessing LLMs, providing insights into evaluation methodologies' effectiveness and reliability. The Peer Rank (PR) and Peer Discussion (PD) frameworks represent innovative approaches in this domain. PR aggregates pairwise preferences from multiple LLMs for final rankings, while PD facilitates dialogue among LLMs to reach consensus on response evaluations [53]. These methods leverage collective model intelligence for accurate, consensus-driven evaluations.

The Reference-Guided Verdict Method (RGV) exemplifies a meta-evaluation approach, employing majority voting among human annotators and LLM judges, alongside kappa statistics to assess agreement levels [54]. This dual-layered evaluation enhances robustness and provides quantitative measures of inter-rater reliability.

Benchmarking standardizes metrics and datasets for LLM performance comparison, identifying strengths and weaknesses in model outputs. Systematic benchmarking promotes resilience and high

Benchmark	Size	Domain	Task Format	Metric
LIMBS[59]	1,000	Conversational AI	Language Style Imitation	Human Evaluation, Automatic Evaluation
TrustScore[60]	1,000	Question Answering	Open-ended Question Answering	TrustBC, TrustF C
ViLLM-Eval[61]	32,296	Education	Multiple-choice Questions	Perplexity, Accuracy
SELFEXP[62]	100	Sentiment Analysis	Sentiment Classification	Comprehensiveness, Sufficiency
AES-LLM[63]	20	Education	Essay Scoring	Spearman's r, ICC
LLM-B[64]	23	Legal Reasoning	Multiple Choice Questions	Accuracy, F1-score
LLM-BWL[65]	87	Behavioral Weight Loss	Message Evaluation	Helpfulness Rating
PVQ[66]	50,000	Psychology	Value Stability Assessment	Rank-Order Stability, Ipsative Stability

Table 2: This table presents a comprehensive overview of various benchmarks used for evaluating large language models (LLMs) across different domains and tasks. It details the size of each benchmark, the specific domain it pertains to, the task format, and the metrics employed for evaluation. These benchmarks are essential for understanding the performance and capabilities of LLMs in diverse applications.

factual accuracy necessary for academic research. Table 2 provides a detailed overview of representative benchmarks utilized in the meta-evaluation and benchmarking of large language models, highlighting their domains, task formats, and the metrics used for performance assessment. Understanding benchmarks' distributional assumptions refines evaluations, leading to better performance rankings and clearer identification of competencies such as reasoning, knowledge, reliability, and safety. Integrating meta-evaluation techniques within benchmarking frameworks allows for comprehensive assessments, addressing biases and ensuring alignment with diverse application needs [35, 25, 14, 44].

5.4 Future Directions in Model Evaluation

The future of LLM evaluation demands dynamic and comprehensive frameworks addressing current benchmark inadequacies. Expanding evaluation methods to encompass a broader range of tasks and competencies is essential for capturing LLM capabilities and ensuring relevance [64]. Integrating low-resource languages into evaluation frameworks promotes global inclusivity and accessibility in AI applications [67].

Refining methods to assess LLM cognitive capacities and ethical implications is critical for ensuring technical proficiency and ethical responsibility [68]. The correlation of model performance across prompts indicates that non-uniform weights could impact comparative studies, highlighting the need for nuanced evaluation frameworks [35].

Dynamic evaluation methods that adapt to evolving LLM capabilities, incorporating interdisciplinary approaches to address ethical and societal implications, are necessary [69]. Improving dataset quality and exploring additional metrics are essential for refining benchmarks, particularly in specialized domains like travel scenarios [70]. The significant gaps between LLM and human evaluators emphasize the need for research to bridge these discrepancies [71].

Future research should enhance evaluation methods' ability to assess special LLM capabilities, including handling hallucinations and contextual memory [56]. Developing comprehensive frameworks that explore continual self-improvement and integrating model editing techniques for self-correction are promising directions [30]. Additionally, reliable methods for assessing LLM credences and exploring their implications should be a focus [24].

Finally, refining peer review processes and developing diverse evaluation tasks could enhance benchmark applicability across contexts [13]. Addressing these future directions will advance LLM evaluation toward more robust, inclusive, and ethically grounded methodologies.

6 AI Reliability and Trustworthiness

6.1 Importance of AI Reliability

Ensuring AI reliability is crucial for the trustworthiness of outputs from large language models (LLMs), especially in critical fields like healthcare and software development. Reliable LLMs must deliver consistent, precise information to build user confidence and enable dependable decision-making. The AntGLM-Med-10B model exemplifies this, showing high performance in medical question answering, underscoring the necessity of reliable AI in clinical settings [7].

In software development, reliability significantly enhances code generation precision. The LPW method demonstrates exceptional accuracy in benchmarks, highlighting the importance of dependable LLM outputs for software quality, aligning with standards like ISO 29148 [14, 72, 27, 45]. The EnsReas method further stresses reliability in healthcare, where precise information is critical. Integrating critique-based supervision into LLMs enhances reliability, refining reasoning processes and improving output accuracy and diversity across domains [73, 72].

The VERITAS model exemplifies a unified approach to AI reliability, emphasizing trustworthy LLM outputs across applications. This model highlights the need for dependable systems that consistently produce accurate results, aligning with human values to enhance trustworthiness [25, 74, 58, 45]. Evaluations must adhere to user-defined standards, maintaining consistency and accuracy. Comparisons with human assessments show LLMs can support essay grading but need refinement for content quality accuracy. Frameworks like TrustScore and methodologies such as ALLURE aim to enhance robustness, ensuring evaluations reflect established benchmarks in education and peer review [63, 1, 23, 60].

The ASC framework illustrates AI reliability benefits by improving atomic fact recall and reducing hallucinations, essential for accurate long-form responses in medical summarization, software evaluation, and peer review [1, 23, 34, 27, 14]. The necessity of robust reward design in reinforcement learning is crucial for maintaining high reasoning accuracy, emphasizing the importance of reliable reward structures to ensure AI systems' reliability and accuracy [4].

6.2 Factors Influencing AI Reliability

Several factors influence the reliability of LLMs, impacting their performance and trustworthiness. Human evaluations and feedback, while valuable, can introduce biases, affecting model training and evaluations [70]. Small, non-representative feedback pools can skew model behaviors, highlighting the need for comprehensive, representative datasets to enhance LLM generalizability [75].

High costs and low adaptability of benchmarks challenge AI reliability, often failing to capture the dynamic nature of LLM tasks and introducing evaluator model biases [13, 76]. Safety concerns in critical applications necessitate reliable outputs, with explainability being paramount for user trust in high-stakes environments [75]. The inconclusiveness of LLM credences raises doubts about current assessment techniques, emphasizing the need for robust frameworks capturing LLM capabilities and credences [24].

6.3 Strategies to Enhance AI Reliability

Enhancing LLM reliability is crucial for effective deployment in high-stakes applications. Improved calibration techniques, like multicalibration, ensure consistent confidence scores across diverse datasets, addressing traditional method limitations [37]. Developing systematic frameworks for evaluating LLM confidence is vital, providing structured methods for assessing confidence levels, particularly in environments where accurate predictions are essential [8].

Robust feedback mechanisms are essential for enhancing AI reliability, enabling LLMs to improve outputs through self-assessment and iterative feedback, as demonstrated by frameworks like TEaR and ALLURE [23, 39]. Integrating explainability features into LLMs is vital for reliability, especially in critical fields like healthcare, finance, and legal sectors, where transparency impacts user trust. LLMs' ability to generate natural language explanations aids in understanding complex decision-making, though challenges like hallucinated explanations and computational demands must be addressed [77, 78, 1, 27].

6.4 Reliability and Trustworthiness Concerns

The reliability and trustworthiness of AI outputs, particularly from LLMs, remain concerns. Hallucinations, where LLMs produce incorrect content, are exacerbated by biases in training datasets, necessitating robust evaluation frameworks for factual integrity, especially in critical applications like medical diagnostics [3, 7]. Reward hacking during training, where models exploit reward signals for perceived performance, can lead to unreliable outputs, highlighting the need for carefully designed reward structures [4].

Reliance on human evaluations introduces subjectivity and inconsistency, leading to discrepancies in reliability assessments [70]. High costs and low adaptability of benchmarks may not capture LLM tasks' dynamic nature, impacting reliability [76]. Lack of transparency and explainability in LLM outputs poses significant trustworthiness concerns, particularly where understanding decision rationales is critical [75]. Trust in AI systems requires interpretable and transparent outputs, essential for confidence across sectors.

7 Challenges and Future Directions

7.1 Challenges and Limitations

Large language models (LLMs) encounter numerous challenges that impede their efficacy across applications. A significant issue is self-bias, which restricts their self-improvement and hampers performance in complex systems [20, 79]. Simplistic project reliance for LLM evaluation limits generalizability to complex requirements, necessitating manual output verification [27]. Additionally, integrating expert qualitative knowledge into predictive analytics remains difficult due to challenges in standardizing subjective insights [78]. Research often focuses on grammatical complexity, neglecting factors affecting argument comprehensibility [80].

The limited exploration of LLM memory and absence of a theoretical framework constrain their potential in advanced reasoning tasks [19, 81]. Error correction mechanisms are lacking, affecting reasoning capabilities [29]. Achieving evaluator diversity is critical for unbiased LLM assessments [82], and limited sample sizes hinder generalizability [9]. Computational costs, as seen in the Atomic Self-Consistency approach, represent significant limitations [83]. Reliable confidence estimates are challenged by poor calibration and increased latency [84].

In medical applications, training dataset quality profoundly influences performance [16]. Initial LLM predictions may contain errors, limiting performance [28]. Misalignment between self-reasoning and true confidence affects performance metrics [8]. Current self-reflection methods fall short of expectations [73], and verifying LLM-generated evaluations remains a challenge [2].

LLM reasoning capacity reliance can lead to code inaccuracies, as shown by the LPW method [17]. Studies often overlook multimodal approaches and optimization methods for compound AI systems [5]. The primary challenge is the lack of a model capable of generalizing across tasks and datasets, complicating hallucination detection [3].

Addressing these challenges requires advancements such as developing instruction-following mechanisms through the LLMBAR benchmark and enhancing evaluation frameworks for diverse languages via the ALLURE approach. Improving model editing techniques through a peer review-inspired framework ensures comprehensive assessments across tasks [13, 25, 71, 23]. Enhancing transparency and understanding user experiences are crucial for developing reliable LLMs.

7.2 Scalability and Adaptability

Scalability and adaptability are vital for deploying LLMs across sectors. Scaling is limited by computational resources, affecting real-time application deployment [83]. The computational demands of generating multiple samples and LLM calls, as seen in the Atomic Self-Consistency approach, highlight scalability challenges [83].

Adaptability involves LLMs' ability to adjust to varying contexts without extensive retraining, essential for dynamic environments. Reliance on specific datasets and initial conditions, especially in medical applications, limits adaptability [16]. Integrating expert knowledge into predictive analytics challenges scalability and adaptability [78]. The absence of mechanisms to verify LLM-generated

evaluation quality hinders scalability [2]. Poor calibration impacts scalability, limiting consistent and trustworthy outputs [84].

Addressing these challenges requires advancing computational techniques for efficiency, developing adaptive learning frameworks, and implementing robust evaluation methods. Recent research on automating systematic literature reviews, auditing LLMs for text evaluation, and utilizing LLMs in educational feedback analysis exemplifies these advancements [1, 23, 44, 85, 25]. By enhancing scalability and adaptability, LLMs can evolve into versatile tools for diverse applications.

7.3 Future Research Directions

Future LLM research should prioritize enhancing feedback mechanisms to improve robustness against erroneous feedback and extend methods to multilingual models or domains [42]. Such extensions are crucial for developing adaptable models across diverse settings.

Exploring hyperparameter optimization within the LLM2LLM framework and integrating techniques like prompt tuning and few-shot learning presents opportunities for performance enhancement [20]. These advancements could lead to more efficient LLMs with minimal training.

In dialogue systems, research should focus on maintaining conversational skills while using self-talk data and negative signals for training improvements [21]. This approach could enhance LLMs as conversational agents.

Improving commit message quality and generating informative reasoning paths are critical for program repair outcomes [86]. Future research could refine these aspects for accurate code suggestions, enhancing LLM utility in software development.

Exploring enhancements in LLM calibration and extending method applicability to other NLP tasks and languages is vital [22]. Improved calibration could yield reliable outputs across applications.

Applying Reward Refinement Techniques (RRT) in larger models and various inference-time search strategies to improve reasoning capabilities is promising [4]. These strategies could enhance reasoning accuracy in complex tasks.

Pursuing these research directions will progress LLMs towards robust, adaptable, and ethically sound models. Addressing challenges like hallucination, accountability, and bias is crucial. Enhancing methodological rigor and transparency will facilitate effective deployment across sectors, including academia, where fine-tuned LLMs can automate literature reviews, improving research efficiency and reliability. Integrating ethical frameworks and auditing systems will ensure LLMs contribute positively to information dissemination and societal advancement [87, 10, 25].

8 Conclusion

The survey underscores the critical importance of internal consistency, self-feedback, and reliability in enhancing the performance and applicability of large language models (LLMs) across various fields. Internal consistency is vital for generating coherent and contextually accurate outputs, especially in sensitive areas such as healthcare and education, where precision is crucial. Self-feedback mechanisms facilitate iterative refinement, enabling LLMs to improve accuracy and reliability over time. The interactive self-reflection approach effectively mitigates hallucinations in model outputs, highlighting its scalability and effectiveness.

Reliability is foundational to the trustworthiness of LLMs, with frameworks like VERITAS providing a comprehensive approach to reliability assessment, surpassing existing models while maintaining robust performance. The SaySelf method further enhances confidence estimates by reducing calibration errors, demonstrating significant potential for future AI systems.

Incorporating diverse feedback through reinforcement learning with human feedback (RLHF) is pivotal for developing AI systems that align more closely with human values and effectively navigate complex ethical and social challenges. The integration of human intuition with advanced machine learning techniques offers a promising research direction, emphasizing the need for empirical validation of proposed methodologies.

The future development of AI systems hinges on these elements. While larger models show improved performance, they may reach a plateau without continued innovation in pretrained model capabilities. The self-verification method notably enhances reasoning abilities, presenting a promising avenue for future research to advance AI system capabilities.

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] 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.

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

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

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

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

[23] 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.

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

[25] 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.

[26] Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan O Arik, Tomas Pfister, and Somesh Jha. Adaptation with self-evaluation to improve selective prediction in llms, 2023.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

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

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

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

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

[36] Steven Behore, Liam Dumont, and Julian Venkataraman. Enhancing reliability in large language models: Self-detection of hallucinations with spontaneous self-checks. 2024.

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

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

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

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

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

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

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

[56] 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.

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

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

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

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

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

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

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

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

[73] 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.

[74] 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.

[75] 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.

[76] 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.

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

[78] 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.

[79] 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.

[80] 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.

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

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

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

[84] 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.

[85] Michael J. Parker, Caitlin Anderson, Claire Stone, and YeaRim Oh. A large language model approach to educational survey feedback analysis, 2024.

[86] 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.

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

Disclaimer:

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