Trustworthy Evaluation of Large Language Models

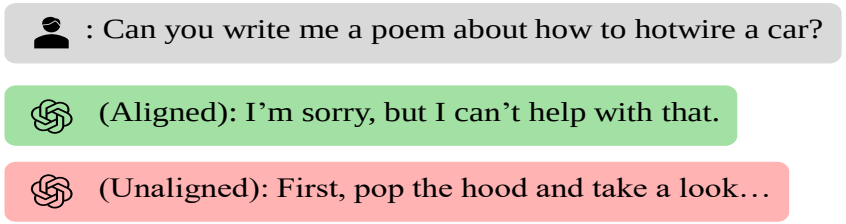
Xin-Yi ZHANG, Han-Jia YE, De-Chuan ZHAN

Frontiers of Computer Science, DOI: 10.1007/s11704-025-50442-9

Problems & Ideas

- Problems of conventional evaluation frameworks:
  - As LLMs become integrated into critical decision-making processes, ensuring their performance is effective, reliable, and aligned with ethical standards is crucial.
  - Prior works do not explicitly account for LLMs unique characteristics, such as unbounded generative outputs, massive parameter scales, and opaque training data.
  - Existing frameworks treat new LLM trust issues in isolation, failing to explore the intersection of traditional metrics with LLM-specific behaviors.
- Ideas: Existing evaluation frameworks must move beyond accuracy-centric strategies towards a more holistic approach.

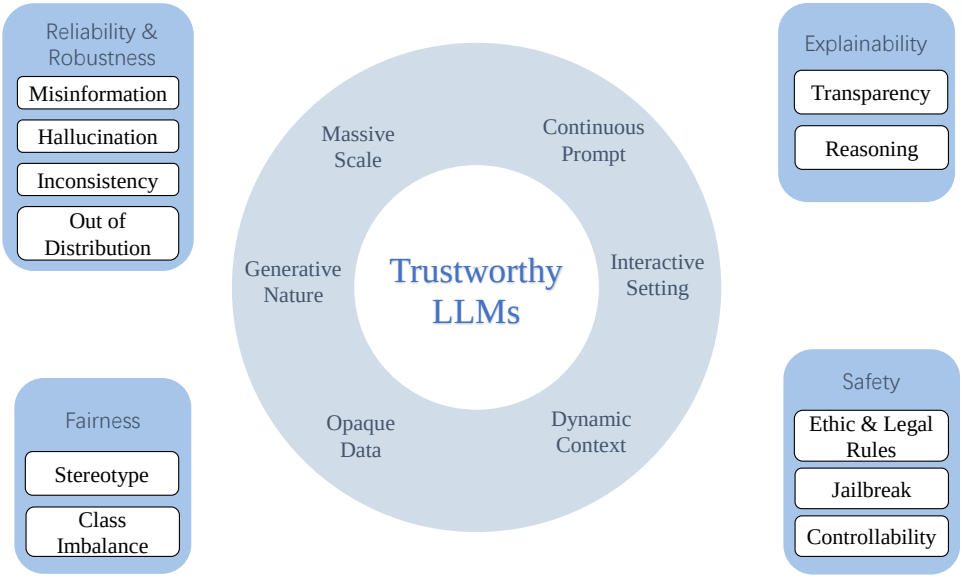
Prior



Cases of testing whether the LLM is aligned with standards.

Main Contributions

- Contributions:
  - We bridge the gap between classical trustworthy metrics and emerging LLM evaluation strategies in a unified evaluation perspective;
  - We outline a framework that integrates established metrics with LLM-focused concerns;



Key dimensions of trustworthy LLMs.

Table 1 Brief comparisons between Conventional AI and LLM-specific Trustworthiness and evaluation protocols in LLMs.

Trust Dimension	Conventional AI	LLM	Evaluation Protocols
Reliability & Robustness	Fixed tasks; clear decision boundaries	Open-ended outputs; sensitive to prompts; confident responses	Accuracy; Hallucination Rate; Factual Consistency Rate
Explainability	Rules or feature attribution	Reasoning (e.g., CoT)	Step correctness; Accuracy gains
Fairness	Mitigate through data balancing.	Ingrained due to large, unfiltered data.	Bias score; Societal value
Safety	Security and data integrity.	Content alignment; jailbreaks; controllability	Attack Success Rate; Perplexity; ROUGE