Measuring LLMs' Sensitivity to Paraphrased Opinion Prompt

Bushra Alhetelah
Department of Computer Engineering
King Fahd University of Petroleum & Minerals (KFUPM)
Dhahran 31261, Saudi Arabia
g202401980@kfupm.edu.sa

Irfan Ahmad
Department of Information and Computer Science, King Fahd University of Petroleum & Minerals (KFUPM),
SDAIA–KFUPM Joint Research Center for Artificial Intelligence
Dhahran 31261, Saudi Arabia
irfan.ahmad@kfupm.edu.sa

Abstract

Large language models (LLMs) are now widely used in applications that depend on closed-ended decisions, including automated surveys, policy screening, and decision-support tools. In such contexts, these models are typically expected to produce consistent binary or ternary responses (for example, Yes, No, or Neither) when presented with questions that are semantically equivalent. However, recent studies show that LLM outputs can be influenced by relatively minor changes in prompt wording, raising concerns about the reliability of their decisions under paraphrasing. In this paper, we conduct a systematic analysis of paraphrase robustness across five widely used LLMs. To support this evaluation, we develop a controlled dataset consisting of 200 opinion-based questions drawn from multiple domains, each accompanied by five human-validated paraphrases. All models are evaluated under deterministic inference settings and constrained to a fixed Yes/No/Neither response format. We assess model behavior using a set of complementary metrics that capture the stability of each evaluated model. DeepSeek Reasoner and Gemini 2.0 Flash show the highest stability when responding to paraphrased inputs, whereas Claude 3.7 Sonnet exhibits strong internal consistency but produces judgments that differ more frequently from those of other models. By contrast, GPT-3.5 Turbo and LLaMA 3 70B display greater sensitivity to surface-level variations in prompt phrasing. Overall, these findings suggest that robustness to paraphrasing is driven more by alignment strategies and reasoning design choices than by model size alone.

1 Introduction

LLMs have become an enabling technology across many aspects of life. They can generate natural language and assist across a wide range of domains, including healthcare, education, public policy, and e-commerce (Naveed et al., 2025; Siino et al., 2025; Raiaan et al., 2024). However, in many real-world scenarios LLMs are not mainly used for long-form text generation. Instead, they are also deployed to produce direct decisions, typically in the form of binary or ternary responses such as Yes, No, or Neither. These decisions can influence significant processes, including clinical triage, policy assessment, and public opinion analysis, making the reliability and stability of model outputs especially important (Buhnila et al., 2024; Karanjai et al., 2025). Despite their strong performance, LLMs are known to be sensitive to variations in prompt formulation (Wang et al., 2021; Errica et al., 2025). Multiple studies have shown that semantically equivalent inputs may cause models to generate different outputs which slightly differ in wording, syntax, or structure. This has been examined extensively in open-ended generation tasks, while it has received less attention in closed-ended settings. In practice, even minor rephrasing—such as substituting synonyms or reordering clauses—can lead to changes in binary judgments, potentially undermining trust in downstream systems.

This study addresses this gap by systematically evaluating the robustness of five state-of-the-art LLMs—OpenAI GPT-3.5 Turbo, Claude 3.7 Sonnet, DeepSeek Reasoner, Google Gemini 2.0 Flash, and LLaMA 3 70B—when responding to paraphrased opinion prompts. We construct a dataset of 200 opinion questions spanning multiple domains, each paired with five LLM-generated and human-validated paraphrases. Model behavior is analyzed under controlled conditions using a fixed Yes/No/Neither response format. By measuring consistency, agreement, and variability across paraphrased inputs, this work provides insight into how different model design and alignment choices influence decision stability under semantic equivalence.

Our findings provide new insights into how LLMs respond to paraphrased opinion prompts,

The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2026), pages 52–59

highlighting both recent progress in improving decision stability and the continued need for careful evaluation of robustness in closed-ended, decision-oriented applications.

The remainder of this paper is organized as follows. Section 2 reviews prior work on LLM robustness, paraphrasing, and opinion-focused question answering. Section 3 describes the methodology and experimental setup. Section 4 presents the results along with a comparative analysis, followed by discussion, ethical considerations and limitations.

Prior studies have shown that LLM outputs can change under paraphrased inputs. However, most existing work primarily examines response stability within specific application domains, such as psychometric questionnaires or political opinion surveys (Haller et al., 2024). For example, recent large-scale analyses generate multiple paraphrases per item and evaluate response validity and stability across models. While these studies provide important insights into paraphrase robustness, they typically focus on domain-specific settings or evaluate robustness using a single perspective, such as response variance or validity.

In contrast, this work systematically measures sensitivity to paraphrased opinion prompts using controlled prompt groups and consistency-based evaluation metrics across multiple modern LLM families. Rather than evaluating stability only at the individual prompt level, we analyze consistency patterns across structured paraphrase groups and across models. This enables a comparative and model-agnostic analysis of paraphrase sensitivity in closed-ended opinion tasks. Furthermore, the dataset and evaluation framework introduced in this work are designed to support reproducible cross-model analysis under semantically equivalent prompt variations, providing a complementary perspective to prior stability-focused studies.

Fine-tuning and instruction modification are as important as pretraining when adapting LLMs for practical deployment (Wu et al., 2025). Instruction tuning ensures that models behave in accordance with human intent by training them on carefully curated prompt–response pairs, often combined with reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). Modern LLMs such as ChatGPT and Gemini demonstrate that instruction tuning improves instruction-following, coherence, and safety alignment (Kalla et al., 2023; Deng et al., 2025; Team et al., 2024). These alignment and tuning choices strongly influence how models respond to variations in prompt formulation.

Paraphrasing preserves the semantic meaning of a question while modifying its surface form through lexical, syntactic, or grammatical changes (Bhagat and Hovy, 2013). An increasing body of research demonstrates that surface-level variations in prompt wording can significantly influence LLM behavior. These findings highlight the strong sensitivity of LLMs to prompt phrasing. Paraphrasing is commonly used to improve model performance and robustness by rephrasing prompts while preserving semantic intent. However, this also exposes an important limitation: semantically equivalent prompts do not always produce consistent model outputs, particularly in decision-making and classification settings. The CheckList behavioral testing framework (Ribeiro et al., 2020) evaluates model robustness using capability-based test suites designed to expose systematic model weaknesses.Recent studies further confirm that LLM outputs can vary substantially across paraphrased prompts, even when response formats are constrained and tasks are closed-ended (Haller et al., 2024).

In the context of opinion-focused question answering, paraphrasing introduces additional challenges beyond sentiment classification, including the identification of opinion holders. Kim and Hovy 2005 established opinion-holder recognition as a ranking problem among candidate entities, demonstrating that grammatical structure and contextual signals are critical for accurate attribution. Prior work on human respondents has also shown that paraphrased question formulations can alter how individuals interpret opinion expressions, affecting categorization and extraction performance (Uemlianin, 2000). Similarly, studies on LLMs show that semantically equivalent paraphrases can lead to different model outputs and reduced decision stability (Haller et al., 2024).

Paraphrasing also presents challenges for robustness and detection. Prior work shows that adversarial paraphrasing can significantly degrade the performance of AI-generated text detectors, while training on paraphrased examples can improve detection robustness (Kurt Pehlivanoğlu et al., 2024; Lau and Zubiaga, 2025). Beyond automated systems, prompt formulation has also been shown to shape human–AI interaction, influencing user trust, reliance, and decision diversity in mixed human–AI decision-making settings (Lai et al., 2023; Chen et al., 2023). Together, these findings demonstrate

that prompt phrasing affects both model behavior and downstream human interpretation.

Although previous studies have investigated the effects of paraphrasing in open-ended generation, retrieval, detection, and relevance judgment tasks, comprehensive evaluation of paraphrase robustness in closed-ended, decision-oriented opinion prompts remains limited. In particular, the consistency of fixed Yes/No/Neither responses under controlled paraphrase variation has not been systematically evaluated across multiple modern LLM families. This gap directly motivates the empirical evaluation presented in this work, which measures response consistency under controlled paraphrase variation across diverse LLM architectures.

2 Methodology

This section describes the experimental design used to evaluate the robustness of large language models (LLMs) to paraphrased opinion prompts. We outline the construction of the paraphrase dataset, the selection of evaluated models, the prompt execution protocol, and the quantitative metrics used to measure consistency and agreement.

The experimental setup used in this work was designed specifically to address the research objective of measuring paraphrase sensitivity in closed-ended opinion prompts, rather than directly adopting a single existing evaluation methodology. However, the design is conceptually aligned with prior work on robustness and behavioral evaluation of NLP systems, which emphasizes evaluating model behavior under semantically equivalent input transformations and paraphrased prompts (Ribeiro et al., 2020; Haller et al., 2024).

Building on these evaluation principles, this work introduces a task-specific framework based on controlled paraphrase groups and consistency-based metrics across multiple modern LLM families. This enables systematic cross-model comparison while maintaining consistency with established robustness evaluation practices.

All experiments were conducted under controlled and deterministic conditions to ensure that observed variations in model outputs were attributable to paraphrase differences rather than sampling randomness or decoding variability.

2.1 Data Collection

A set of 200 distinct opinion prompts spanning technology, public policy, work culture, and education was generated using OpenAI's ChatGPT-4o-mini model. Prompt generation was guided by domain-balanced topic selection to ensure coverage across major societal and decision-oriented themes. For each original prompt, five paraphrased variants were generated using the same model, yielding a total of 1,200 items (200 original prompts and 1,000 paraphrases). All automatically generated paraphrases were manually reviewed and verified by the authors to ensure semantic equivalence with the original question. To illustrate the nature of the paraphrasing process, Table 1 presents a representative example consisting of one original opinion question and its corresponding paraphrased variants.

The paraphrase validation process was conducted through manual review by one of the authors to ensure semantic equivalence between each paraphrased prompt and its corresponding original question. During this process, paraphrases were evaluated for preservation of intent, opinion framing, and decision context. Paraphrases that introduced semantic drift, ambiguity, or unintended bias were either corrected or regenerated. Overall, approximately 7% of the automatically generated paraphrases required minor manual adjustments to ensure semantic consistency with the source prompt.

Because validation was performed by a single annotator, no inter-annotator agreement metric was computed. However, a consistent validation protocol was applied across all prompts, focusing on semantic equivalence, grammatical correctness, and preservation of response intent.

To support reproducibility and facilitate future research, the dataset is publicly released via Hugging Face under the name Paraphrased Opinion Prompt Sensitivity (POPS) Dataset[1]

2.2 Model Selection and Specifications

We deliberately selected a diverse set of language models to capture variation across model size, reasoning capability, and deployment properties, as summarized in Table 2. The selected models represent diversity across deployment type, reasoning specialization, and alignment design rather than strict parameter scale comparisons. The evaluation includes both open-weight and closed, provider-managed systems, as well as models optimized

[1] https://huggingface.co/datasets/mazew2000/Paraphrased_Opinion_Prompt_Sensitivity_Dataset_POPS

1	Original Opinion Question	Do you believe higher education should be provided at no cost to students?
2	Paraphrased Questions	Should governments eliminate tuition fees for university-level education?
3		In your opinion, should college education be free for everyone?
4		Would it be fair to make university education accessible without financial burden?
5		Do you support the idea of publicly funded college education for all citizens?
6		Should students have the right to attend university without paying tuition fees?

Table 1: Sample Original Opinion Prompt and Corresponding Paraphrased Questions

for general-purpose reasoning and models specifically designed for structured reasoning tasks. This diversity enables comparative analysis across performance, operational cost, latency, and real-world deployment constraints.

As shown in Table 2, Claude 3.7 Sonnet and GPT-3.5 Turbo represent closed, provider-managed models designed for strong safety alignment, reliability, and production integration. LLaMA 3 70B represents a large open-weight model supporting full customization, local deployment, and research flexibility (Ersoy and Erşahin, 2024; Gue et al., 2024). DeepSeek Reasoner was included to capture reasoning-specialized behavior under structured decision tasks, while Gemini 2.0 Flash represents a low-latency model optimized for real-time interaction scenarios (Comanici et al., 2025). Together, these models enable comparative evaluation across differences in deployment architecture, reasoning behavior, alignment strategy, and system-level trade-offs.

2.3 Prompt Design and Experimental Execution

To ensure consistent and deterministic model behavior, all prompts were issued under a fixed system instruction: "You are a strict yes-or-no answerer. For each question, answer with exactly 'Yes', 'No', or 'Neither'. No numbering, no extra text." This instruction was applied verbatim to every request. In addition, the temperature parameter was set to 0.0 for all model calls, minimizing stochastic variation and ensuring that observed response differences arose from paraphrase variation rather than sampling randomness. To comply with rate-limit constraints and maintain execution stability, prompts were dispatched sequentially in fixed-size batches of ten, with a 200 ms delay enforced between API calls. No parallelization was employed, ensuring strict request ordering.

2.4 Evaluation Metrics

Four primary metrics were computed for each model to quantify robustness and comparison under paraphrase variation.

Consistency Rate (CR). Consistency Rate measures the proportion of paraphrased responses that match the majority answer within each question group. For a question group q with P paraphrases, it is defined as:

$$\mathrm{CR}_q = \frac{1}{P} \sum_{i=1}^{P} I\Big(a_{q,i} = a_q^{\mathrm{maj}}\Big), \tag{1}$$

where $I(\cdot)$ is the indicator function. The overall model consistency is obtained by averaging CR_q across all Q questions.

Pairwise Agreement (PA). Pairwise Agreement measures the proportion of identical responses produced by two models M_1 and M_2 on the same prompts:

$$\mathrm{PA}(M_1, M_2) = \frac{1}{N} \sum_{i=1}^{N} I\Big(a_i^{(M_1)} = a_i^{(M_2)}\Big), \tag{2}$$

where N denotes the total number of prompts.

Cohen's Kappa (κ). Cohen's Kappa evaluates agreement between a model's responses to original prompts and their paraphrases while correcting for chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{3}$$

where p_o is the observed agreement and p_e is the expected agreement under random labeling.

Aggregate Statistics. For each model, the mean consistency rate and 95% confidence intervals across question groups were computed to summarize overall robustness and uncertainty in model responses under paraphrased prompts.

Attribute	Claude 3.7 Sonnet	LLaMA 3 70B	DeepSeek Reasoner	Gemini 2.0 Flash	GPT-3.5 Turbo
Parameters	Not publicly disclosed	70B (open weights)	Not publicly disclosed	Not publicly disclosed	Not publicly disclosed
Reasoning Orientation	General-use	General-use	Reasoning-oriented	Paraphrase-robust	General-use
Context Window	Large (varies plan/API)	8k tokens (standard LLaMA 3)	Varies by offering/API	Varies by offering/API	Up to 16k (variant-dependent)
Inference Cost	Moderate–High (relative)	Moderate (deployment-dependent)	Moderate (relative)	Low (relative)	Low (relative)
Latency	Low–Moderate (deployment-dependent)	Moderate (hardware-dependent)	Moderate (relative; deployment-dependent)	Low (deployment-dependent)	Low (deployment-dependent)
Fine-tuning / Plugins	No public fine-tuning	Fully customizable (open-weight)	No public fine-tuning	Limited (prompt-based)	Supports fine-tuning and function calling
Safety & Robustness	Provider-aligned safety tuning	User/deployment controlled	Provider-dependent safety	Provider-aligned safety tuning	Provider-aligned safety tuning
Primary Strengths	Strong reasoning and writing, long-context use cases	Open-weight flexibility; on-prem deployment	Reasoning-centric behavior	Fast interactive responses	Mature tooling ecosystem
Ideal Use Cases	Document-heavy workflows, assistants	Research and custom deployments	Reasoning-centric evaluation pipelines	Real-time chat systems	Prototyping and assistants

Table 2: Model Selection and Comparative Specifications

3 Results and Comparison

This section presents a quantitative analysis of how consistently each evaluated language model responds to paraphrased versions of the same opinion questions. Using multiple complementary metrics, we examine both within-model stability and cross-model alignment under controlled rewording conditions.

Figure 1 presents the mean consistency rates across question groups for the five evaluated language models together with 95% confidence intervals. Overall, all models exhibit high consistency under paraphrased versions of the same opinion questions, indicating substantial robustness to superficial prompt rewording. However, measurable differences in consistency are observed across models. DeepSeek Reasoner achieves the highest mean consistency rate (approximately 91%), followed closely by Gemini 2.0 Flash (approximately 90–91%). Both models also show relatively narrow confidence intervals, indicating stable behavior across paraphrase groups. Claude 3.7 Sonnet demonstrates strong performance with a mean consistency rate around 90%, though with slightly higher variability. GPT-3.5 Turbo and LLaMA 3 70B show lower mean consistency and wider uncertainty ranges, suggesting greater sensitivity to paraphrase variation. In practical terms, while all models demonstrate generally robust behavior,

Figure 1: Mean consistency rate across question groups for each evaluated model with 95% confidence intervals. Error bars indicate uncertainty in consistency estimates across paraphrase groups.

DeepSeek Reasoner and Gemini 2.0 Flash exhibit the most stable response patterns under paraphrasing, making them particularly suitable for applications requiring consistent closed-ended decision outputs.

Figure 2 illustrates the pairwise agreement matrix across all models, revealing how frequently two models produce identical responses to the same prompts. As expected, perfect self-agreement appears along the diagonal. Among off-diagonal entries, the highest agreement is observed between GPT-3.5 Turbo and LLaMA 3 70B at 92.9%, in-

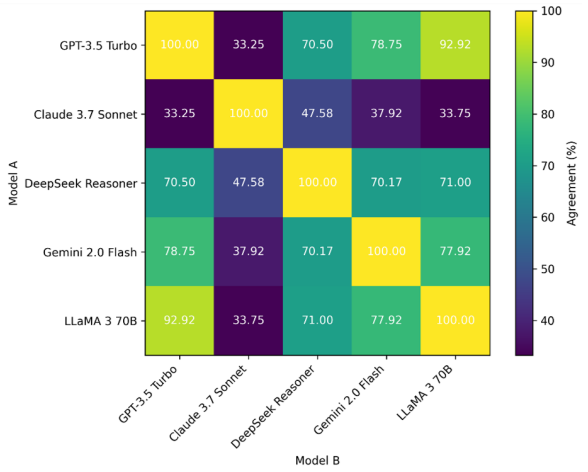

Figure 2: Pairwise Agreement Matrix Between Evaluated Language Models

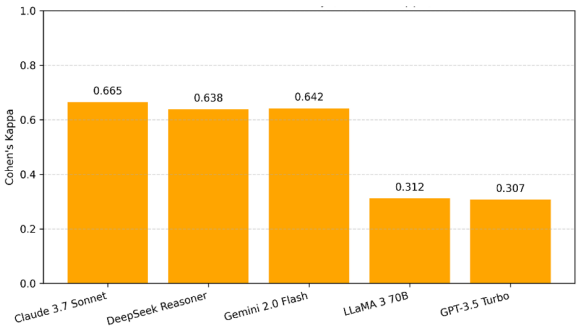

Figure 3: Self-Consistency of Language Model Outputs (Cohen's Kappa)

dicating that these two models frequently reach the same binary conclusions despite originating from different providers. High agreement is also observed between Gemini 2.0 Flash and LLaMA 3 70B, as well as between Gemini 2.0 Flash and GPT-3.5 Turbo, with values in the 77–79% range. These results suggest that Gemini's decision behavior aligns closely with that of GPT-3.5 Turbo and LLaMA 3 70B. DeepSeek Reasoner demonstrates moderate agreement (approximately 70%) with most other models, indicating broadly similar but less tightly coupled decision boundaries. In contrast, Claude 3.7 Sonnet consistently exhibits low agreement with all other models, ranging from 33% to 48%. This divergence suggests that Claude applies a substantially different internal decision strategy for binary opinion questions. In ensemble settings, Claude's outputs may therefore require careful weighting or fallback mechanisms to avoid instability arising from inter-model disagreement. Figure 3 reports each model's self-consistency using Cohen's Kappa (κ), which measures agreement between responses to original prompts and their paraphrased variants while correcting for chance agreement. Claude 3.7 Sonnet achieves the highest Kappa score ($\kappa = 0.665$), indicating substantial agreement beyond chance and strong internal stability under paraphrasing. DeepSeek Reasoner ($\kappa = 0.638$) and Gemini 2.0 Flash ($\kappa = 0.642$) follow closely, confirming that these models generalize their binary decisions reliably across lexical and syntactic variations. By comparison, LLaMA 3 70B ($\kappa = 0.312$) and GPT–3.5 Turbo ($\kappa = 0.307$) fall within the fair agreement range. While these models still achieve high raw consistency rates, their lower Kappa values indicate that a larger portion of the observed agreement can be attributed to chance. This suggests greater sensitivity to paraphrase-induced variation when evaluated under a chance-corrected reliability framework, rather than implying direct flip probabilities.

Collectively, these results reveal three distinct robustness profiles:

- **High and stable:** DeepSeek Reasoner and Gemini 2.0 Flash.
- **High but moderately variable:** Claude 3.7 Sonnet and GPT–3.5 Turbo.
- **Lower and highly variable:** LLaMA 3 70B.

For applications requiring predictable binary outputs under paraphrase variation, DeepSeek Reasoner and Gemini 2.0 Flash emerge as the strongest candidates. Claude 3.7 Sonnet and GPT–3.5 Turbo remain viable but may benefit from ensemble voting or post-processing safeguards, while LLaMA 3 70B should be deployed with additional validation mechanisms when phrasing stability is critical.

4 Discussion and Conclusion

This paper presented a systematic evaluation of large language model sensitivity to paraphrased opinion prompts using a controlled dataset and consistency-based evaluation framework. The results demonstrate that while modern LLMs exhibit high overall consistency across paraphrased inputs, measurable variability remains even under semantically equivalent prompt transformations. These findings reinforce the importance of considering prompt formulation as a key factor influencing model behavior in structured decision tasks.

Beyond confirming previously observed prompt sensitivity behavior, this work contributes three primary advances. First, it introduces a new evaluation setting focused specifically on closed-ended, decision-oriented opinion prompts, which remain underexplored compared to open-ended generation and domain-specific robustness studies. Second, it proposes a reproducible evaluation methodology based on controlled paraphrase groups and consistency metrics, enabling systematic cross-model comparison under deterministic inference conditions. Third, the empirical results provide new insight into robustness mechanisms, suggesting that alignment strategies, reasoning specialization, and deployment design contribute more strongly to paraphrase robustness than parameter scale alone.

The publicly released Paraphrased Opinion Prompt Sensitivity (POPS) dataset further supports reproducibility and future research by enabling consistent benchmarking of paraphrase robustness across models and architectures. This resource allows researchers to extend evaluation beyond single-prompt testing and toward structured robustness assessment under semantically controlled variation.

Future work may expand this framework to multilingual settings, open-ended generation tasks, and interactive multi-turn decision scenarios. Additionally, integrating human evaluation and uncertainty-aware benchmarking methods could provide deeper insight into how paraphrase sensitivity influences real-world human–AI decision workflows.

Overall, this work highlights the importance of evaluating LLM robustness not only across tasks and domains, but also across semantically equivalent prompt formulations. Understanding and improving decision stability under paraphrase variation will be critical for reliable deployment of LLMs in real-world decision-support applications.

Limitations

Our evaluation was conducted under a single deterministic inference setting with the temperature fixed to zero, and therefore does not capture how paraphrase robustness may vary under stochastic decoding conditions. In addition, paraphrased prompts were generated by a language model and manually validated, which may not fully reflect the linguistic diversity of paraphrases produced by human experts. Exploring robustness across different inference settings and human-authored paraphrases is an important direction for future work.

References

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational linguistics*, 39(3):463–472.

Ioana Buhnila, Aman Sinha, and Matthieu Constant. 2024. Retrieve, generate, evaluate: A case study for medical paraphrases generation with small language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 189–203.

Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction*, 7(CSCW2):1–32.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Zehang Deng, Wanlun Ma, Qing-Long Han, Wei Zhou, Xiaogang Zhu, Sheng Wen, and Yang Xiang. 2025. Exploring deepseek: A survey on advances, applications, challenges and future directions. *IEEE/CAA Journal of Automatica Sinica*, 12(5):872–893.

Federico Errica, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. 2025. What did i do wrong? quantifying llms' sensitivity and consistency to prompt engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1543–1558.

Pınar Ersoy and Mustafa Erşahin. 2024. Benchmarking llama 3 70b for code generation: A comprehensive evaluation. *Orclever Proceedings of Research and Development*, 4(1):52–58.

Celeste Ci Ying Gue, Noorul Dharajath Abdul Rahim, William Rojas-Carabali, Rupesh Agrawal, Palvannan Rk, John Abisheganaden, and Wan Fen Yip. 2024. Evaluating the openai's gpt-3.5 turbo's performance in extracting information from scientific articles on diabetic retinopathy. *Systematic reviews*, 13(1):135.

Patrick Haller, Jannis Vamvas, and Lena Ann Jäger. 2024. Yes, no, maybe? revisiting language models' response stability under paraphrasing for the assessment of political leaning. In *First Conference on Language Modeling*.

Dinesh Kalla, Nathan Smith, Fnu Samaah, and Sivaraju Kuraku. 2023. Study and analysis of chat gpt and its impact on different fields of study. *International journal of innovative science and research technology*, 8(3).

Rabimba Karanjai, Boris Shor, Amanda Austin, Ryan Kennedy, Yang Lu, Lei Xu, and Weidong Shi. 2025. Synthesizing public opinions with llms: Role creation, impacts, and the future to edemorcacy. *arXiv preprint arXiv:2504.00241*.

Soo-Min Kim and Eduard Hovy. 2005. Identifying opinion holders for question answering in opinion texts. In *Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains*, pages 1367–1373.

Meltem Kurt Pehlivanoğlu, Robera Tadesse Gobosho, Muhammad Abdan Syakura, Vimal Shanmuganathan, and Luis de-la Fuente-Valentín. 2024. Comparative analysis of paraphrasing performance of chatgpt, gpt-3, and t5 language models using a new chatgpt generated dataset: Paragpt. *Expert Systems*, 41(11):e13699.

Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. 2023. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1369–1385.

Hiu Ting Lau and Arkaitz Zubiaga. 2025. Understanding the effects of human-written paraphrases in llm-generated text detection. *Natural Language Processing Journal*, page 100151.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, 12:26839–26874.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Ivan A Uemlianin. 2000. Engaging text: Assessing paraphrase and understanding. *Studies in Higher Education*, 25(3):347–358.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, and 1 others. 2021. Textflint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355.

Xiao-Kun Wu, Min Chen, Wanyi Li, Rui Wang, Limeng Lu, Jia Liu, Kai Hwang, Yixue Hao, Yanru Pan, Qingguo Meng, and 1 others. 2025. Llm fine-tuning: Concepts, opportunities, and challenges. *Big Data and Cognitive Computing*, 9(4):87.