International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 11 | Nov 2026 www.irjet.net p-ISSN: 2395-0072

Impact of Hallucination Detection Techniques on the Factual Accuracy & Reliability of Large Language Models: A Study among Students in Belapur CBD, Navi Mumbai

Dhairya Kamble[1], Jagruti Juvatkar[2], Dhanashree Shelke[3]

[12]*Student Department of MCA, D. Y. Patil Deemed to be University, School of Humanities and Sciences, Belapur, Maharashtra, India.*

[3]*Student Department of MCA, D. Y. Patil Deemed to be University, School of Humanities and Sciences, Belapur, Maharashtra, India.*

Abstract - *Large Language Models (LLMs) such as GPT, LLaMA, and Gemini have emerged as powerful tools in education, information retrieval, coding, and content generation. Although these models can produce fluent and meaningful responses, they may sometimes generate incorrect, unsupported, or fabricated information, commonly known as hallucinations. Such hallucinations can affect the factual accuracy and reliability of information generated by AI, particularly when students use LLMs for academic purposes. This study focuses on the impact of hallucination detection techniques on the factual accuracy and reliability of Large Language Models among students in Belapur CBD, Navi Mumbai. The study examines hallucination detection approaches such as uncertainty estimation, self-consistency checking, retrieval-based verification, LLM-as-a-judge, and internal model-state analysis. It further studies students' awareness, usage, perception, and trust regarding these techniques while using LLM-generated information. A quantitative survey-based method is employed to collect primary data from students through a structured questionnaire. Likert-scale items are used to measure awareness of hallucinations, usage of detection techniques, perceived factual accuracy, reliability, and trust in LLM-generated responses. Secondary data from research papers, academic journals, and existing studies are used to provide theoretical support and contextual understanding. The study attempts to contribute to the growing research on trustworthy and reliable Artificial Intelligence by examining hallucination detection from the perspective of students. The findings are expected to provide insights into how verification and hallucination detection practices can improve confidence in AI-generated information and support more responsible use of LLMs in academic environments.*

Key Words: Large Language Models, LLM Hallucination, Hallucination Detection, Factual Accuracy, Reliability, Artificial Intelligence, Students, AI Safety.

1. INTRODUCTION

Large Language Models (LLMs) have become an important part of modern Artificial Intelligence (AI) and are widely used in education, research, coding, information retrieval, and content generation. Models such as GPT, LLaMA, and Gemini can understand user queries and generate human-like responses quickly. Due to their ability to provide explanations, summaries, and solutions, these tools have become increasingly popular among students for academic purposes.

Despite their usefulness, LLMs sometimes generate incorrect, misleading, unsupported, or fabricated information that appears convincing and accurate. This phenomenon is known as AI Hallucination. Such hallucinations can affect the factual accuracy and reliability of AI-generated information, particularly when students depend on these tools for assignments, projects, and research without verifying the responses.

To address this issue, several hallucination detection techniques have been developed, including uncertainty estimation, self-consistency checking, retrieval-based verification, LLM-as-a-judge, and internal model-state analysis. These techniques help identify potentially incorrect information and improve the reliability of AI-generated responses by checking consistency and comparing information with reliable sources.

However, the effectiveness of these techniques also depends on users' awareness and their willingness to verify AI-generated information. Students need to understand the limitations of AI tools and develop responsible verification practices before using AI-generated content for academic purposes.

The present study focuses on analysing the impact of hallucination detection techniques on the factual accuracy and reliability of Large Language Models among students in Belapur CBD, Navi Mumbai. It examines students' AI-tool usage, awareness of hallucinations, knowledge of verification techniques, and perceptions of the reliability of AI-generated information.

The study adopts a quantitative research approach using a structured questionnaire to collect primary data from students. Secondary data is collected from research papers, academic journals, and existing studies. Overall, the research

highlights the importance of hallucination detection, information verification, and responsible AI usage in improving confidence in AI-generated information.

1.1 Problem Statement

Large Language Models such as GPT, LLaMA, and Gemini are increasingly used by students for academic activities, including assignments, projects, research, and information gathering. Although these tools provide quick and meaningful responses, they sometimes generate incorrect, misleading, unsupported, or fabricated information, commonly known as hallucinations. Such responses may appear highly confident and accurate, making it difficult for students to identify factual errors.

The major problem is that students may depend on AI-generated information without properly verifying its correctness. This can affect their understanding of academic concepts, the quality of their assignments, and their trust in AI-based applications. Although several hallucination detection techniques, such as uncertainty estimation, self-consistency checking, retrieval-based verification, and LLM-as-a-judge, are available, students may not have sufficient awareness or knowledge of how to use these techniques effectively.

Therefore, there is a need to examine students' awareness and use of hallucination detection and verification techniques and understand their relationship with the perceived factual accuracy and reliability of AI-generated information.

The present study addresses this problem by investigating the impact of hallucination detection and verification practices on the perceived factual accuracy, reliability, and trustworthiness of Large Language Model outputs among students in Belapur CBD, Navi Mumbai.

1.2 Research Gap

Existing research has examined the causes, categories, detection methods, and evaluation of hallucinations in language generation systems. Technical approaches such as retrieval-based verification, consistency checking, semantic uncertainty, and automated evaluation have also been studied. However, the present study addresses a complementary user-centred area: students' awareness of hallucinations and their reported verification practices in an academic setting. It focuses on the relationship students perceive between verification, reliability, and trust. The survey does not benchmark detection algorithms or independently test the factual correctness of model outputs; rather, it provides descriptive evidence about student awareness and behaviour.

1.3 Objectives

- To examine students' AI-tool usage and awareness of verification.
- To investigate the association between verification-related responses and perceived reliability.
- To explain hallucinations and their implications for LLM reliability.
- To describe major hallucination-detection and verification techniques.
- To recommend practices for responsible academic use of LLMs.

1.4 Research Questions

- How frequently do students use AI tools for academic purposes, and how aware are they of hallucinations?
- To what extent do students report verifying AI-generated information using reliable sources?
- Which hallucination-detection and verification practices are relevant to evaluating LLM-generated responses?
- How do students perceive the usefulness of verification in improving confidence in AI-generated information?
- How can students use LLMs more responsibly in academic activities?

2. METHODOLOGY

This study adopts a quantitative, survey-based research design to examine the impact of hallucination detection and verification techniques on the factual accuracy and reliability of Large Language Models (LLMs). The research is supported by a review of secondary sources, including research papers, journals, articles, and existing studies related to AI hallucinations, detection techniques, and the reliability of AI-generated information. The primary data for this study were collected through a structured online questionnaire prepared using Google Forms and distributed among students in Belapur CBD, Navi Mumbai.

The questionnaire was designed to understand students' usage of AI tools, their awareness of AI hallucinations, and their practices for verifying AI-generated information. It also focused on students' opinions regarding the usefulness of verification techniques, their perceived improvement in the reliability of AI-generated responses, and their level of trust in verified information. The collected responses were analyzed quantitatively to identify patterns, understand user perceptions, and evaluate the role of verification techniques in improving the factual accuracy and reliability of AI-generated content.

2.1 Research Design

A quantitative descriptive approach was adopted to summarize respondents' answers through frequencies and percentages. The questionnaire included a frequency-of-use question and Likert-scale statements. The analysis reports combined positive responses (Strongly Agree and Agree) where these percentages were available in the original survey summary.

2.2 Sample Size and Sampling Method

The study includes 85 respondents, as indicated in the Google Forms response summaries. Purposive sampling, a non-probability sampling method, was used to obtain responses from students with experience using AI-based applications such as ChatGPT, Gemini, and other LLM tools.

2.3 Primary Data Collection

Primary data were collected through a structured online questionnaire. Questions covered frequency of AI-tool usage, awareness that AI tools can generate incorrect or fabricated information, recognition that confident responses may contain errors, use of reliable sources for verification, awareness of detection practices, perceived usefulness of verification, perceived improvement in accuracy and reliability, and trust in verified information.

2.4 Secondary Data Collection

Secondary information was drawn from research papers, academic journals, conference publications, and existing studies concerning hallucinations in natural language generation, retrieval-augmented generation, factuality evaluation, self-consistency checking, model-state analysis, and LLM-based evaluation. The literature provides the conceptual basis for describing detection techniques and interpreting the survey findings.

3. BACKGROUND: LARGE LANGUAGE MODELS AND HAALUCINATIONS

LLMs learn statistical patterns from large collections of text and generate responses by predicting likely token sequences in a given context. Their fluency does not guarantee that every statement is factually correct. A hallucination may involve an incorrect fact, an unsupported claim, a fabricated citation, or information that conflicts with the supplied context. Hallucinations can arise from limitations in training data, incomplete or outdated knowledge, ambiguity in prompts, weak grounding in external evidence, and the model's tendency to generate plausible continuations.

3.1 Why Hallucinations Matter in Academic Use

In academic settings, inaccurate explanations, invented references, incorrect dates, or unsupported claims may affect assignments and research. Students therefore need to distinguish between using an LLM as an aid and treating its output as an authoritative source. Important claims should be checked against original research papers, textbooks, official publications, and other reliable sources.

3.2 Types of Hallucination

Type	Description	Possible academic example
Factual error	A statement conflicts with verifiable facts.	An incorrect date or definition.
Unsupported claim	A claim is presented without adequate evidence.	A broad conclusion with no supporting source.
Fabricated reference	A citation or publication detail is invented or inaccurate.	A non-existent journal article.
Contextual inconsistency	The response contradicts information in the prompt or source.	A summary changes a reported result.
Reasoning error	The response reaches an incorrect conclusion from stated information.	An incorrect calculation or inference.

4. MAJOR HALLUCINATION DETECTION AND VERIFICATION TECHNIQUES

Detection approaches operate at different stages and require different levels of technical access. Some are model-level methods that estimate uncertainty or inspect internal representations, while others use external evidence or compare generated responses. Students can also apply simpler verification practices without access to a model's internal architecture.

Technique	Basic working principle	Strength / use	Limitation
Uncertainty estimation	Estimates how uncertain a model is about a generated answer, using signals such as probability or	Can flag responses that may require additional checking.	Uncertainty is not identical to factual incorrectness; confident errors can still occur.

	semantic uncertainty.		
Self-consistency checking	Generates or compares multiple responses to the same prompt and looks for contradictions or semantic variation.	Can identify unstable answers without requiring internal model access.	Repeated responses may share the same error.
Retrieval-based verification	Compares generated claims with information retrieved from external documents or knowledge sources.	Provides evidence-grounded checking and supports source verification.	Depends on source quality, retrieval coverage, and correct interpretation.
LLM-as-a-judge	Uses another language model to assess the quality, relevance, or factual support of a response.	Can assist with scalable evaluation.	The evaluator may also be biased or hallucinate.
Internal model-state analysis	Examines hidden representations or internal signals that may correlate with truthfulness.	May provide signals unavailable from text alone.	Often requires model access and specialized technical methods.
Manual source verification	Checks important claims against reliable publications, textbooks, official sources, and original references.	Accessible to students and directly supports academic checking.	Requires time and the ability to assess source credibility.

5. IMPACT ON FACTUAL ACCURACY AND RELIABILITY

Hallucination detection is intended to identify statements that may be incorrect, unsupported, or inconsistent with evidence. When a questionable claim is detected and checked against reliable sources, users can avoid repeating some inaccurate information. Retrieval-based verification can connect claims to external evidence, while consistency checks and uncertainty signals can indicate when an answer deserves further scrutiny.

Reliability is broader than whether a single statement is correct. It also concerns whether information is consistent, supported, appropriately qualified, and dependable across repeated use. Detection methods can contribute to reliability, but no single method guarantees that every hallucination will be identified. The quality of the reference sources, the task, the model, and the evaluation procedure all affect the result.

In this study, factual accuracy and reliability are examined through students' perceptions and reported behaviours. The questionnaire does not independently score LLM responses against a verified answer key. Accordingly, positive survey responses indicate that students believe verification is useful; they should not be interpreted as measured improvements in model accuracy.

6. SURVEY FINDINGS AND INTREPRETATION

The following results are based on 85 Google Forms responses included in the source research paper. Percentages are reproduced from the supplied survey summaries. The analysis is descriptive and reflects respondents' self-reported awareness, practices, and perceptions.

Survey aspect	**Positive response / result**
AI-tool usage for academic purposes	77.6% very frequently; 14.1% frequently (91.7% combined)
Awareness of AI hallucinations	77.6%
Recognition that confident AI responses may contain errors	73.0%
Verification using reliable sources	68.2%
Awareness of hallucination-detection techniques	68.2%
Perceived usefulness of verification techniques	76.4%
Belief that verification improves accuracy and reliability	75.3%
Greater trust in verified AI-generated information	78.8%

6.1 Frequency of AI-Tool Usage

The results indicate that 77.6% of respondents use AI tools very frequently for academic purposes, while 14.1% use them frequently. Thus, 91.7% report frequent or very frequent use. This indicates that AI applications are regularly

used by a substantial proportion of the surveyed students, making awareness of hallucinations and verification particularly relevant.

6.2 Awareness of AI Hallucinations

A combined 77.6% of respondents agree or strongly agree that AI tools can generate incorrect, unsupported, or fabricated information. This suggests that most respondents recognise hallucination as a limitation of AI-generated content. However, some respondents remain neutral or disagree, indicating scope for further AI-literacy guidance.

6.3 Recognition of Factual Errors in Confident Responses

A total of 73.0% agree or strongly agree that AI-generated information may contain factual errors even when it appears confident and accurate. This finding highlights the importance of evaluating the evidence behind a response rather than relying only on its presentation or fluency.

6.4 Verification Using Reliable Sources

The survey shows that 68.2% of respondents report verifying AI-generated information using reliable sources before using it in assignments, projects, or research. The remaining responses indicate that verification is not a consistent practice for every respondent. Students may benefit from clear guidance on checking references and confirming important claims.

6.5 Awareness of Detection Techniques

A total of 68.2% report awareness of practices such as cross-checking information, checking references, and comparing multiple AI responses. This suggests that many students know basic verification methods, although further practical instruction may help students apply these methods more consistently.

6.6 Perceived Effectiveness of Verification

A combined 76.4% believe that verification techniques help identify incorrect or misleading AI-generated information. This result measures perceived usefulness rather than independently tested detection performance.

6.7 Perceived Improvement in Accuracy and Reliability

A total of 75.3% believe that checking AI-generated information improves its factual accuracy and reliability for academic use. This indicates a generally positive attitude towards verification, while remaining a perception-based result.

6.8 Trust in Verified AI-Generated Information

A combined 78.8% report greater trust in AI-generated information when it has been verified using reliable sources or appropriate techniques. This is the highest positive response among the seven Likert-scale statements and suggests that evidence-based checking is important to students' confidence.

7. Detection Gaps and Limitations

The survey and reviewed literature point to several practical challenges. Awareness does not always translate into consistent verification. Students may not know which sources are authoritative, may accept references without opening them, or may assume that multiple similar AI responses prove correctness. Automated evaluators can also make mistakes, and retrieval-based approaches depend on the relevance and quality of retrieved sources.

- Confident wording may make inaccurate information appear trustworthy.
- Students may have limited time or skills to verify every statement in a long response.
- Repeated AI responses may reproduce the same underlying error.
- Retrieved sources may be outdated, incomplete, irrelevant, or unreliable.
- An LLM used as an evaluator may share limitations or biases with the model being assessed.
- Advanced techniques may require computational resources or access to model internals.

8. EVALUATION OF DETECTION AND VERIFICATION APPROACHES

The approaches below are compared conceptually using the descriptions in the reviewed literature. This is not an experimental ranking: the present student survey did not test the algorithms or measure their computational cost.

Approach	Evidence used	Technical requirement	Practical relevance to students
Uncertainty estimation	Model confidence or uncertainty signals	May require model outputs or internal access, depending on method	Useful as a warning signal when exposed by a system
Self-consistency	Multiple generated responses	Repeated generation and comparison	Students can compare answers, but must still check facts
Retrieval	External	Search/retrie	Highly relevant

-based verification	documents and sources	val system or manual source lookup	to checking claims against reliable references
LLM-as-a-judge	Evaluation generated by another LLM	Access to an evaluator model and suitable criteria	Can support review but should not replace source checking
Internal-state analysis	Hidden representations or model signals	Specialized model access and analysis	Primarily a research/developer technique
Manual verification	Original sources and trusted publications	Information literacy and time	Directly applicable to student academic work

9. PROPOSED STUDENT-CENTRED VERIFICATION MODEL

Based on the literature and survey findings, this paper proposes a Student-Centred LLM Verification Model (SCLVM). The model is a practical synthesis for academic use, not a technically validated detection algorithm. It translates the central idea of hallucination detection into a sequence of checks that students can apply before using AI-generated information.

9.1 Model Stages

1. Stage 1 - Identify important claims: Separate factual statements, dates, statistics, quotations, definitions, and references from general explanatory text.
2. Stage 2 - Assess uncertainty: Treat unclear, unusually specific, or unsupported claims as requiring additional checking. If an AI system provides uncertainty indicators, use them as prompts for review rather than proof of error.
3. Stage 3 - Cross-check consistency: Compare the response with another answer or ask for clarification where useful, while recognising that repeated outputs can share the same mistake.
4. Stage 4 - Retrieve independent evidence: Check claims against textbooks, original research papers, official websites, and other credible sources.
5. Stage 5 - Verify references: Open cited publications and confirm that the title, author, date, and cited content are genuine and relevant.
6. Stage 6 - Make a responsible-use decision: Use only claims that are adequately supported, correct or qualify uncertain statements, and acknowledge sources according to academic requirements.

10. OPERATIONAL RECOMMENDATIONS

10.1 Recommendations for Students

- Do not treat an LLM response as an automatically authoritative source.
- Verify factual claims, statistics, quotations, and references using reliable original sources.
- Open cited links or publications to confirm that they exist and support the stated claim.
- Compare multiple responses only as an additional check, not as a substitute for evidence.
- Clearly distinguish verified facts from explanations or suggestions generated by AI.
- Follow institutional rules for disclosure and responsible use of AI tools.

10.2 Recommendations for Educational institutions

- Include AI literacy and hallucination awareness in student orientation and digital-literacy programmes.
- Teach students how to evaluate source credibility and verify references.
- Provide examples of fabricated citations, unsupported claims, and confident but incorrect responses.
- Encourage responsible AI use that supports learning without replacing independent thinking.

10.3 Recommendations for AI Developers and System Providers

- Provide clear source links and distinguish retrieved evidence from generated explanation.
- Communicate uncertainty and system limitations in a way users can understand.
- Evaluate detection methods across different tasks and domains, including their false-positive and false-negative rates.
- Monitor whether automated evaluation systems introduce their own factual errors or biases.

11. DISCUSSION

The survey results indicate that AI tools are used frequently by the respondents for academic purposes. This makes the reliability of generated information a practical concern rather than only a technical research issue. The high level of awareness of hallucinations suggests that many students understand that fluent AI responses can contain errors.
At the same time, the reported use of reliable sources and awareness of detection techniques are each 68.2%, lower than the 77.6% awareness of hallucinations. This difference suggests a possible gap between recognising the problem and consistently applying verification practices. The survey also shows that 76.4% perceive verification techniques as

useful and 78.8% report greater trust in verified information.

These findings are consistent with the general purpose of detection and verification research: to identify unsupported or inaccurate content and improve the evidence available to users. However, the study does not test detection algorithms or establish that verification directly causes higher factual accuracy. The results should therefore be interpreted as evidence about student perceptions and reported behaviour in the surveyed group.

11.1 Awareness and Practice

Awareness of hallucinations is relatively high in the sample, but reported source verification is lower. This indicates an opportunity for practical training that moves beyond explaining what hallucinations are and demonstrates how to check claims and references.

11.2 Reliability and Trust

The positive response to trusting verified information suggests that students value evidence-supported AI outputs. Trust should nevertheless remain conditional: verification needs to be performed against credible and relevant sources, and important claims should not be accepted solely because another AI system approves them.

11.3 Implications for Academic Work

Students can use LLMs to support brainstorming, explanations, and drafting, but should retain responsibility for checking facts and citations. Institutions can support this approach through clear guidelines, source-evaluation instruction, and assignments that reward critical evaluation.

12. PROPOSED REFERENCE WORKFLOW FOR ACADEMIC USE

Step	Student action	Expected purpose
1. Generate	Use an LLM to obtain an explanation or initial draft.	Support learning and information exploration.
2. Extract	Identify factual claims, statistics, dates, and references.	Focus checking on claims that can be verified.
3. Check	Compare claims with reliable independent sources.	Identify inaccurate or unsupported content.
4. Confirm	Open references and confirm that they support the claim.	Avoid fabricated or irrelevant citations.
5. Revise	Correct, qualify, or remove unsupported statements.	Improve the quality of the final academic work.
6. Disclose responsibly	Follow institutional guidance on AI assistance and citations.	Promote academic integrity.

13. LIMITATIONS

- The study is based on 85 respondents and purposive sampling; the findings should not be generalized to all students.
- The questionnaire records self-reported awareness, behaviour, and perceptions, which may differ from actual verification behaviour.
- The study does not conduct controlled experiments comparing LLM outputs before and after applying detection techniques.
- The available survey summaries provide percentages but do not include the respondent-level dataset or inferential statistical test outputs.
- The demographic details such as age, gender, and educational qualification are not available in the supplied survey text and have therefore not been fabricated.
- The findings describe perceived accuracy and reliability; they do not provide an objective benchmark of model factual accuracy.

14. FUTURE RESEARCH

Future studies can extend this work by including a larger and more diverse sample of students, using probability-based sampling where feasible, and collecting respondent-level data for statistical testing. Experimental studies could compare AI-generated answers before and after applying retrieval-based verification, self-consistency checking, or other detection methods. Such studies could evaluate factual correctness using a verified answer key and report precision, recall, false-positive rates, false-negative rates, and response latency.

14.1 Suggested Evaluation Metrics

Measure	Purpose
Precision	Proportion of flagged responses that are actually hallucinated.
Recall	Proportion of hallucinated responses successfully identified.
F1-score	Harmonic balance of precision and recall.

False-positive rate	Frequency with which correct responses are incorrectly flagged.
False-negative rate	Frequency with which hallucinations are missed.
Latency	Time required to detect or verify a response.
Computational cost	Resources required to operate the detection method.
Human agreement	Level of agreement between automated judgments and expert or human assessments.

Further research may also examine real-time detection, multimodal hallucinations in text-image-audio systems, tool-augmented LLMs, and agent-based systems that perform multi-step tasks.

15. CONCLUSION

This study examined students' use of AI tools, awareness of hallucinations, reported verification practices, and perceptions of factual accuracy and reliability among students in Belapur CBD, Navi Mumbai. It also reviewed major approaches to hallucination detection, including uncertainty estimation, self-consistency checking, retrieval-based verification, LLM-as-a-judge, and internal model-state analysis.

The survey of 85 respondents indicates that AI tools are frequently used for academic purposes. A majority of respondents recognise that AI tools can generate hallucinations and that confident responses may still contain factual errors. The findings also show that 68.2% report verifying AI-generated information using reliable sources, 76.4% perceive verification techniques as useful, 75.3% believe checking improves accuracy and reliability, and 78.8% report greater trust in verified AI-generated information.

The results highlight the importance of moving from general awareness to consistent verification practices. Students should check factual claims and references against reliable sources and should not treat generated text or automated evaluation as proof of correctness. Educational institutions can strengthen responsible AI use through AI-literacy programmes and practical source-verification guidance.

The study's conclusions are limited to the perceptions and self-reported practices of the surveyed respondents. It does not demonstrate a causal improvement in factual accuracy or provide an experimental comparison of detection algorithms. Nevertheless, it offers a student-centred view of why hallucination awareness and evidence-based verification matter for the responsible academic use of Large Language Models.

16. REFERENCES

[1] Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when it's lying. Findings of the Association for Computational Linguistics: EMNLP 2023.

[2] Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630, 625–630.

[3] Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630, 625–630.

[4] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12).

[5] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

[6] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 3214–3252.

[7] Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 9004–9017.

[8] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, 36.