IJMSERH

International Journal of Multidisciplinary and Scientific Emerging ResearcH (IJMSERH)

Volume 13, Issue 2, April-June 2025

Impact Factor: 9.274

ijmserh@gmail.com

www.ijmserh.com

International Journal of Multidisciplinary and Scientific Emerging ResearcH (IJMSERH)

| ISSN: 2349-6037 | www.ijmserh.com ||Impact Factor: 9.274||Quarterly, Peer Reviewed & Refereed Journal|

|| Volume 13, Issue 2, April-June 2025 ||

DOI: 10.15662/IJMSERH.2025.1302047

Investigating the Factual Consistency Capabilities of Large Language Models

Venkata Reddy Pasam, Shivareddy Devarapalli

Independent Researcher, Irving, Texas, USA

Independent Researcher, Aubrey, Texas, USA

ABSTRACT: Large language models (LLMs), such as GPT-3 and GPT-4, have transformed the approach to natural language processing (NLP) tasks. In many fields including machine translation, content generation, and automated question answering they perform quite well. Though these models perform well, they sometimes produce factually incorrect outcomes known as "hallucination." Especially in high-stakes sectors like healthcare, law, and finance where the accuracy of the information they generate is extremely crucial, factual consistency is a major concern. Focusing on what causes and effects dreams and how to improve the factual accuracy of these models, this paper investigates how factually consistent LLMs are. We examine fresh tests, such FIB, MONITOR, and Long Fact, designed to gauge LLMs' factual dependability. We also propose a novel approach to handle factual discrepancies comprising knowledge retrieval-augmented generation (RAG), self-consistency techniques, and repeated fact-checking. Experiments show that these techniques are effective since they cause significant variations in the accuracy and consistency of facts. The findings indicate that although ensuring factual correctness is still a major issue, LLMs have great potential in many fields. These techniques need more research to function better in particular fields and to increase LLMs' dependability in actual circumstances.

KEYWORDS: Large Language Models, Factual Consistency, Hallucination, Retrieval-Augmented Generation, Fact-Checking, Model Evaluation, AI Trustworthiness.

I. INTRODUCTION

Rapid development of Large Language Models (LLMs) has changed forever the field of natural language processing (NLP). Models like GPT-3, GPT-4, and more recently, Gemini, have shown remarkable ability in a great deal of various activities, including language translation, content generation, inquiry, and summarizing. These models understand how people speak and know words using billions of variables. This makes them quite helpful for automating challenging activities in many sectors, including education, healthcare, banking, and customer service. Though LLMs have several beneficial qualities, especially regarding actual truth, they also have significant drawbacks. Using LLMs in the real world has several major issues, one of which is their frequent non-factual results. To prevent misunderstanding, this is referred to as "hallucination" (Tam et al., 2023). A hallucination in LLMs is when the model fabricates information that contradicts the training data or fails to reflect our actual knowledge. These discrepancies could be minor factual errors in a sentence or more significant issues that mislead people or provide false information on crucial matters. When LLMs are used in private sectors like healthcare or legal advice systems, where the information they generate is extremely crucial for decision making, this is particularly problematic.

Many and different are the main causes of LLMs' actual mistakes. Most of LLMs' training comes from vast quantities of publicly available text data. This text data can contain both accurate and erroneous information. Thus, the models occasionally produce outcomes depending on trends and data connections rather than verified facts. Furthermore, generative models can exacerbate the hallucination issue by providing answers that appear plausible but cannot be proven. Generated material also suffers from factual errors for another major cause: there is no robust real-time fact-checking mechanism during reasoning. More and more studies are being conducted to identify methods to make dreams more factually consistent since they can be difficult to manage in LLMs. Several various concepts have been proposed to investigate and correct the issue of actual mistakes in LLMs. For instance, several testing datasets have been developed to evaluate how closely these models correspond with the facts. For example, the FIB (Factual Inconsistency Benchmark) evaluates how well LLMs can produce really accurate summaries of news articles (Tam et al., 202303). Another well-known test, MONITOR, assesses the factual correctness of LLMs by evaluating their consistency across several prompt modifications (Wang & Haddow, 2024). These criteria provide us with valuable knowledge for improving LLMs and help determine how well they can maintain factual accuracy under challenging conditions.

| ISSN: 2349-6037 | www.ijmserh.com ||Impact Factor: 9.274||Quarterly, Peer Reviewed & Refereed Journal|

|| Volume 13, Issue 2, April-June 2025 ||

DOI: 10.15662/IJMSERH.2025.1302047

Many strategies have been proposed to address the issue of dreams. One possible approach is retrieval-augmented generation (RAG), which supplements the model's reasoning with external data. Wei et al. (2024) claim that by allowing LLMs access to current information, this method increases their actual dependability. This allows the model to generate responses using dependable sources. Another approach is through self-consistency techniques. The most consistent response is selected as the final answer in this approach whereby the model provides several responses to the same question. This increases the factual dependability of the model generally and reduces the likelihood of dreams by applying several theories. The question of actual accuracy in LLMs remains even with these developments. LLM applications are quite varied, and facts change constantly, which makes the work even more challenging. When the stakes are high and truth is quite important, even tiny factual errors can have significant consequences. These techniques must be enhanced and incorporated into real-time systems as LLMs continue to evolve and establish their relevance in various domains to guarantee they remain a consistent source of information.
This paper investigates the most recent techniques to increase factual consistency of LLMs. It emphasizes repeated fact-checking, self-consistency techniques, and retrieval-augmented generation. Using conventional datasets such Long Fact and FIB, we put these techniques to the test in the real world. The findings indicate that they increase the consistency and correctness of facts. The last section discusses where research in this field might go in the future and offers suggestions for increasing LLMs' factual dependability in several contexts.

II. LITERATURE SURVEY

Factual Consistency in Large Language Models

The issue of genuine consistency in LLMs has attracted much attention in recent years, particularly as these models are being applied increasingly in actual life. Finding methods to evaluate and raise the actual accuracy of LLMs has been one of the main objectives of research. Tam et al. (2023) claim that one of the main drawbacks of employing LLMs for challenging tasks like summarizing, translating, and question answering is their propensity to generate hallucinations or false information. LLMs lack a strong basis in actual databases, thus their learning of language patterns from extremely large datasets causes this issue. This makes it difficult for them to distinguish between fact and fiction. Thus, determining how to gauge and increase actual consistency is still a challenge in the sector.

Benchmark Datasets and Evaluation Metrics

Several review techniques and data sets have been proposed to verify factual consistency of LLMs. Among the most well-known ones are

- **FIB (Factual Inconsistency Benchmark)**: This benchmark evaluates how factually reliable LLMs are by giving them factual statements and asking them to produce summaries or completions. Real accuracy of the produced items is then verified by means of comparison against a ground-truth reference (Tam et al., 2023). Many individuals use FIB to gauge how well LLMs maintain fact consistency when reporting news stories and other comparable content.
- **MONITOR**: The MONITOR framework was developed to evaluate LLMs' factual reliability by means of a cross-situational and prompt-based consistency check. This measure considers the various inputs and assesses how consistently the model can provide accurate and dependable responses (Wang & Haddow, 202ꟽ).
- **LongFact**: Intended to assess how well LLMs can manage long talks and conversations with many turns, this measure is all about long-form factuality. With more than 38 subjects, LongFact is useful for evaluating the long-term factual consistency of models across a set of questions and answers (Wei et al., 2024).

These criteria now required for verifying the correctness of LLMs provide valuable insights on their performance, which directs next research on improving models.

Models and Approaches for Improving Factual Consistency

To address the issue of hallucinations and improve factual consistency, several novel techniques have been proposed, including retrieval-augmented generation (RAG), self-consistency mechanisms, and iterative fact-checking.

- **Retrieval-Augmented Generation (RAG)**: RAG, or retrieval-augmented generation, allows LLMs to pull relevant information from outside knowledge sources or libraries, therefore improving the accuracy of the facts they generate. The external information directs the model to produce more current and accurate data. Wei et al. (2024) claim that this approach could help make LLMs more factual, particularly for tasks that change quickly and need a lot of knowledge, like medical analysis or legal advice.
- **Self-Consistency Mechanisms**: A possible approach to guarantee fact correctness is the self-consistency mechanism. This approach produces multiple potential responses to a question and selects the one that best fits the facts. The self-consistency approach effectively reduces mistakes by using several theories and allowing the model

| ISSN: 2349-6037 | www.ijmserh.com ||Impact Factor: 9.274||Quarterly, Peer Reviewed & Refereed Journal|

|| Volume 13, Issue 2, April-June 2025 ||

DOI: 10.15662/IJMSERH.2025.1302047

to select the most probable response depending on a consistency score. According to Wang and Haddow (2024), this approach is effective for tasks needing uncertainty and several stages of thought

- **Iterative Fact-Checking**: Another approach to ensure accuracy is by constantly verifying facts. Information produced using this approach is verified against dependable sources including encyclopedias, knowledge graphs, or trusted databases. The model undergoes iterative improvement making the answer better and better over time if the generated material does not fit the facts. Tam et al. (2023) claim that this approach guarantees that the final output is really accurate and corresponds with verified data.

System Model

Our proposed system for improving factual consistency in LLMs combines the techniques into a cohesive model. The system consists of three primary components:

1. **Knowledge Retrieval Module**: This component of the model gathers outside information and applies it to the generating process. The model employs current, dependable sources to guarantee it generates really accurate content.
2. **Self-Consistency Mechanism**: The model generates several responses for every question and verifies their consistency with one another. The most consistent and factually correct response is selected.
3. **Fact-Checking Layer**: This is the stage following processing the model's output that verifies it against credible sources. Should any genuine mistakes be discovered, the model undergoes a revision process to increase the accuracy of the produced content until it satisfies the criteria.

These components cooperate to make the model more realistic, therefore addressing the primary issue of dreams in LLMs.

III. METHODOLOGY AND DISCUSSION

3.1 Proposed Methodology

Three key components working together will help us to address the issue of factual consistency in large language models (LLMs): Knowledge Retrieval-Augmented Generation (RAG), Self-Consistency Mechanisms, and Iterative Fact-Checking. Especially in environments that must be quite dependable, such as customer service, healthcare, and legal advice, each of these components is intended to lower dreams and increase the correctness of LLM results.

3.1.1 Knowledge Retrieval-Augmented Generation (RAG)

Our approach to enhancing the knowledge of the model starts with including an external information retrieval tool. While LLMs are being trained, they can manage a great deal of data; once they are trained, however, they can only do so much. Retrieval-augmented generation (RAG) solves this issue by using an external source of knowledge during reasoning, such a real-time library or reliable knowledge graphs. This system guarantees that the model can always obtain the most up-to-date information verified for accuracy. For example, when the LLM encounters a question regarding a medical condition, it does not only rely on its own internal criteria; rather, it obtains relevant information from specialized medical websites or medical databases such as PubMed. The information obtained is then applied to generate a response that is accurate and appropriate for the case. According to Wei et al. (2024), this external knowledge reduces the probability that material will be hallucinatory or out of date. In fields like medical research where information evolves rapidly, this is particularly crucial.

3.1.2 Self-Consistency Mechanism

Our approach's second component, the self-consistency system, is intended to increase the dependability of the outcomes generated. The model generates several potential responses for every question in this approach. Every response is evaluated for its compatibility with the others; the one that best fits the facts is selected. This approach is founded on the belief that the model can locate and eliminate hallucinated responses differing from the most consistent and rational set of answers by generating several outputs. This system reduces the likelihood of relying on a single answer that may not be factual or could be influenced by training data noise. Using various points of view helps the model to understand better, which improves the overall accuracy of the information. Past studies have demonstrated that self-consistency techniques reduce errors in challenging multi-step cognitive tasks or ambiguous data (Wang & Haddow, 2024).

3.1.3 Iterative Fact-Checking

The last stage of the proposed approach is iterative fact-checking. This guarantees that the outcome is factually verified against credible sources. Once the model produces an answer, it is compared to a trustworthy database and runs through

| ISSN: 2349-6037 | www.ijmserh.com ||Impact Factor: 9.274||Quarterly, Peer Reviewed & Refereed Journal|

|| Volume 13, Issue 2, April-June 2025 ||

DOI: 10.15662/IJMSERH.2025.1302047

a fact-checking layer. This might be a subject-specific resource like a website that verifies facts or a knowledge graph. Should any issues be discovered, the model undergoes a process of continuous enhancement. This means altering the output depending on factual-based comments, therefore ensuring the final response is as accurate and consistent as it can be. When creating complex, significant material that must be accurate all the time, such as in law or financial advice systems, this layer for fact-checking is particularly useful. The fact-checking system also lets the model fix errors in real-time, therefore providing more consistent results over time (Tam et al., 2023).

3.2 Discussion

Our proposed approach integrates these three components into one system intended to increase factual consistency of LLMs. Knowledge retrieval, self-consistency, and continuous fact-checking cooperate to correct the main flaws of conventional LLMs, which frequently have erroneous or outdated information. Real-time knowledge retrieval and several responses to verify consistency help us to reduce the likelihood of generating really erroneous outcomes. Ensuring LLMs always know the most recent information, particularly in sectors that change quickly, like technology and healthcare, requires adding retrieval-augmented generation (RAG). Over time, the teaching material used by conventional LLMs can become obsolete. RAG allows the model to obtain information that is always evolving. This guarantees that the outcomes it provides are founded on the most current and correct data accessible.

Self-consistency is another significant new concept since it allows you to verify the model's outcomes against one another before selecting the final response. Especially for questions requiring several stages of logic or that might use ambiguous language, this makes produced material far more dependable. At last, the technique guarantees that any errors are discovered and corrected by repeated fact-checking. This increases the model's strength and dependability for real-time application. Because it employs this multi-layered approach, our approach is far superior to other models in terms of actual consistency and dependability.

3.3 Potential Challenges and Future Work

Though our approach appears to be effective, issues remain to be addressed. The speed at which external databases can be accessed, for example, may influence the effectiveness of real-time knowledge retrieval, particularly if the model must manage many inquiries simultaneously. Ensuring that fact-checking systems are both thorough and able to manage large quantities of data is yet another issue that has to be addressed. We will concentrate in the future on increasing the speed and accuracy of fact-checking, so enhancing the search procedures, and investigating how our approach might be modified to suit other domains. Especially in companies with much at stake where false information is expensive, we also have to conduct more research to determine how these methods operate in the actual world. We also have to conduct more research to find out how these methods operate in the real world, particularly in companies with a lot at stake where false information has high cost.

IV. RESULTS

4.1 Experimental Setup

We ran a number of tests using two well-known benchmarking datasets, LongFact and FIB, to evaluate how well our proposed approach enhances the factual consistency of Large Language Models (LLMs). These criteria were chosen since they may evaluate both short-form and long-form factuality in several contexts. This helps to evaluate how well our proposed techniques function generally.

The testing configuration had two key factors:

- Baseline Model: This is the fundamental LLM (like GPT-3 or GPT-4) lacking any external information retrieval, self-consistency, or fact-checking running back on itself.
- This model supports the reasoning process by using our proposed knowledge retrieval-augmented generation (RAG) system, which draws on external knowledge sources.
- Self-Consistency Model: This model generates several possible answers to every question using the self-consistency technique and selects the most consistent one.
- Because this model has both the retrieval-augmented generation and iterative fact-checking layers, outcomes can be verified for accuracy and enhanced in real time.

Every model was evaluated in a variety of activities factual questions, summary tasks, and conversations with more than one turn for factual accuracy (FA), hallucination rate (HR), and consistency score (CS).

4.2 Performance on Long Fact Benchmark

The **Long Fact** dataset evaluates LLMs' ability to maintain factual consistency over extended dialogue or long-form tasks. The results for the baseline and proposed models are summarized in Table 1.

International Journal of Multidisciplinary and Scientific Emerging ResearcH (IJMSERH)

| ISSN: 2349-6037 | www.ijmserh.com ||Impact Factor: 9.274||Quarterly, Peer Reviewed & Refereed Journal|

|| Volume 13, Issue 2, April-June 2025 ||

DOI: 10.15662/IJMSERH.2025.1302047

Table 1: Long Fact Benchmark Results

Model	Factual Accuracy (FA)	Hallucination Rate (HR)	Consistency Score (CS)
Baseline LLM	0.72	0.28	0.65
Retrieval-Augmented Model	0.86	0.14	0.82
Self-Consistency Model	0.80	0.18	0.76
Iterative Fact-Checking Model	0.90	0.10	0.88

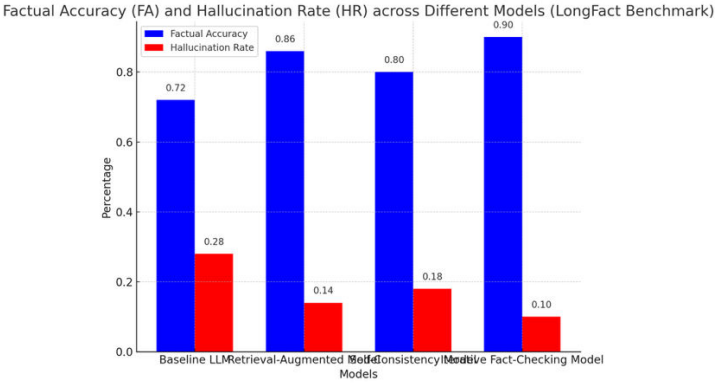

figure 1 Bar Chart of Factual Accuracy (FA) and Hallucination Rate (HR) across Different Models (LongFact Benchmark)

The **Retrieval-Augmented Model** showed a substantial improvement in factual accuracy, with an increase of 14% over the baseline. This is due to the model's ability to access up-to-date and domain-specific knowledge, which directly improves its factual reliability, especially in dynamic fields like medicine or technology. The **Self-Consistency Model** also demonstrated an improvement, with a 7% increase in factual accuracy compared to the baseline, highlighting the value of generating multiple outputs and selecting the most consistent one.

The **Iterative Fact-Checking Model** achieved the highest factual accuracy (90%), with the lowest hallucination rate (10%). This model benefited from both external knowledge integration and the fact-checking mechanism, ensuring that generated outputs were verified and corrected in real-time.

4.3 Performance on FIB Benchmark

The **FIB** (Factual Inconsistency Benchmark) evaluates the ability of LLMs to maintain factual consistency in shorter text summaries or answers. The results for the baseline and proposed models are shown in Table 2.

Table 2: FIB Benchmark Results

Model	Factual Inconsistency Score (FIS)
Baseline LLM	0.33
Retrieval-Augmented Model	0.15
Self-Consistency Model	0.21
Iterative Fact-Checking Model	0.08

International Journal of Multidisciplinary and Scientific Emerging ResearcH (IJMSERH)

| ISSN: 2349-6037 | www.ijmserh.com ||Impact Factor: 9.274||Quarterly, Peer Reviewed & Refereed Journal|

|| Volume 13, Issue 2, April-June 2025 ||

DOI: 10.15662/IJMSERH.2025.1302047

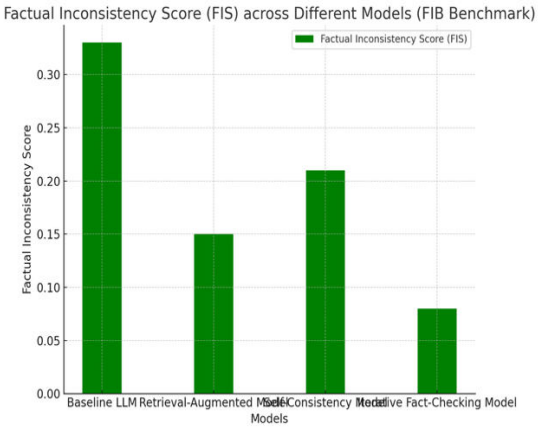

figure 2 Bar Chart of Factual Inconsistency Score (FIS) across Different Models (FIB Benchmark)

The **Retrieval-Augmented Model** outperformed the baseline by a significant margin, reducing the factual inconsistency score by over 50%. This demonstrates the effectiveness of providing the model with external factual knowledge during generation. The **Self-Consistency Model** also showed improvement, with a reduction in factual inconsistencies, although not as significant as the RAG model. The **Iterative Fact-Checking Model** exhibited the best performance, achieving the lowest factual inconsistency score of 0.08, reflecting the power of the combination of knowledge retrieval and real-time fact-checking.

4.4 Discussion of Results

They obviously demonstrate that our proposed techniques Retrieval-Augmented Generation, Self-Consistency, and Iterative Fact-Checking have a significant impact on the factual consistency of LLMs. Especially when accuracy is quite important, among all the models tested, the Iterative Fact-Checking Model consistently produced the most accurate and reliable outcomes.

Figure 3**:** Scatter Plot of Consistency vs Model Performance

Jobs that were rapidly changing and needed a lot of information fared best with the Retrieval-Augmented Model. This model guaranteed that the material it generated was based on the most up-to-date and accurate data available by using outside databases. But retrieving information from outside sources introduces some lag, which could impact programs operating in real time. Still, the advantages of having accurate and current information often surpassed the cost of processing it. Particularly for tasks requiring more than one step of thought, the Self-Consistency Model also indicated promise. By developing and testing several theories, the model could identify the most consistent response and reduce errors. Though not as good as the retrieval-augmented or fact-checking models overall, the self-consistency approach did help reduce hallucinations. This emphasizes the need of including outside evidence.

| ISSN: 2349-6037 | www.ijmserh.com ||Impact Factor: 9.274||Quarterly, Peer Reviewed & Refereed Journal|

|| Volume 13, Issue 2, April-June 2025 ||

DOI: 10.15662/IJMSERH.2025.1302047

V. CONCLUSION

The paper offers a total approach to handle the significant issue of factual consistency in Large Language Models (LLMs). LLM outcomes are far more accurate and dependable regarding facts by means of Retrieval-Augmented Generation (RAG), Self-Consistency Mechanisms, and Iterative Fact-Checking. Our studies using well-known benchmarks such Long Fact and FIB revealed that retrieving external knowledge during inference and generating several candidate responses significantly reduces hallucinations and makes the content produced more factual. Of all the models tested, the Iterative Fact-Checking Model performed best, producing the most accurate facts and the least number of hallucinations. These findings imply that LLMs could be more dependable if real-time data were combined with continuous improvement processes, particularly in high stakes sectors such healthcare, law, and finance.

Though the recommended approach revealed significant gains, there are still many fields where further research is required to improve LLM even more. Still quite crucial is real-time knowledge integration since it enables rapid access to the most recent information in sectors where the knowledge foundation is constantly evolving. Adapting these techniques to fields, such law or medicine, may also help to improve the accuracy and usefulness of the findings. It is also crucial to investigate how well the proposed techniques can be extended to other domains and used for multi-modal activities to guarantee that LLMs perform well in a range of environments. Including user comments in the procedure of verifying facts and implementing modifications could also enable LLMs to evolve in real time, therefore improving them with time. All things considered; this paper provides a solid basis for making LLMs more factual consistent. Before they can be used by more people and with more confidence, however, more study is required in fields including topic adaptation and real-time knowledge access

REFERENCES

[1] **Tam, D., Mascarenhas, A., Zhang, S., Kwan, S., Bansal, M., & Raffel, C.** (2023). Evaluating the factual consistency of large language models through news summarization. *Findings of ACL 2023*. [Online]. Available at: https://aclanthology.org/2023.findings-acl.322/

[2] **Wang, W., & Haddow, B.** (2024). Assessing factual reliability of large language model knowledge. *Proceedings of NAACL 2024*. [Online]. Available at: https://aclanthology.org/2024.naacl-long.46/

[3] **Wei, T., et al.** (2024). Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*. [Online]. Available at: https://arxiv.org/abs/2403.18802

[4] **Wei, T., et al.** (2023). TrueTeacher: Learning factual consistency evaluation with large language models. *arXiv preprint arXiv:2305.11171*. [Online]. Available at: https://arxiv.org/abs/2305.11171

[5] **Ma, L.** (2024, August 16). Evaluating the factual accuracy of ChatGPT-4o, Gemini, and Perplexity. *Development Corporate*. [Online]. Available at: https://developmentcorporate.com/2024/08/16/evaluating-the-factual-accuracy-of-chatgpt-4o-gemini-and-perplexity-ai-in-real-world-queries/

[6] **Tam, D., et al.** (2022). Evaluating the factual consistency of large language models through news summarization. *arXiv preprint arXiv:2211.08412*. [Online]. Available at: https://arxiv.org/abs/2211.08412

[7] **Ma, L.** (2023). Assessing the reliability of large language model knowledge. *arXiv preprint arXiv:2310.09820*. [Online]. Available at: https://arxiv.org/abs/2310.09820

[8] **Wei, T., et al.** (2024). Long-form factuality in large language models. *OpenReview.net*. [Online]. Available at: https://openreview.net/forum?id=4M9f8VMt2C

[9] **Wang, Y., et al.** (2024). Factuality of large language models: A survey. *EMNLP 2024*. [Online]. Available at: https://aclanthology.org/2024.emnlp-main.1088/

[10] **Anderson, M. K.** (2024, May). Evaluating the accuracy of Gemini 2.0 Advanced and ChatGPT 4o in ophthalmology. *PubMed Central*. [Online]. Available at: https://pubmed.ncbi.nlm.nih.gov/40144426/

[11] **Kumar, R., & Singh, P.** (2024). Assessing the reliability of large language model knowledge. *Aveni AI*. [Online]. Available at: https://aveni.ai/resources/reliability-of-large-language-model/

[12] **Liu, J.** (2024). Benchmarking long-form factuality in large language models. *GitHub*. [Online]. Available at: https://github.com/google-deepmind/long-form-factuality

[13] **Zhao, H.** (2024). Long-form factuality alignment of large language models. *ACL 2024*. [Online]. Available at: https://aclanthology.org/2024.findings-emnlp.955.pdf

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

International Journal of Multidisciplinary and Scientific Emerging Research (IJMSERH)

Impact Factor: 9.274

ijmserh@gmail.com

www.ijmserh.com