Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28242

Assessment of the Factual Response Ability of LLMs

Shuhan Wang

School of Mathematics, Hefei University of Technology, Hefei, China
2023215614@mail.hfut.edu.cn

Abstract. When large language models (LLMs) are applied in medical, legal and other fields, the phenomenon of "hallucinations" seriously restricts their reliability. This paper reviews the relevant studies and reviews the evaluation of LLMs' factual response ability. The evaluation method includes the internal state analysis method, which detects the factuality by analyzing the static activation characteristics of the model and the dynamic output mode. The confidence estimation evaluation method compares the training probe, sequence probability and other methods, and finds that the training probe performs best in fact-verification and Q&A tasks, but relies on model weight and training data. The dynamic consistency evaluation method evaluates the output consistency of the model under the disturbance of entities and attributes based on dynamic benchmarks such as DyKnow, and reveals that the model is highly sensitive to input changes. The robustness evaluation method uses MONITOR indicators to quantify the frame effect and contextual interference effect, which has both high efficiency and accuracy. The external knowledge-assisted evaluation method, combined with retrieval augmented generation (RAG), evaluates the model's dependence on external information through external knowledge matching. This paper provides a systematic reference for understanding the factual flaws of LLMs, optimizing the evaluation framework, and improving the reliability of models, and helps them to be used credibly in high-risk fields.

Keywords: Reinforcement learning, Large language models, Factual reply, Hallucination, Evaluation methodology

1. Introduction

With the rapid development of large language models (LLMs) in the field of natural language processing, they are becoming more and more widely used in scenarios such as intelligent customer service, medical diagnosis, and legal consultation. However, the "illusion" phenomenon of LLMs, that is, the generation of content that does not match the facts, is still a core obstacle to its reliable application. Research shows that even the most advanced LLMs still have 20% outputs containing outdated information [1], and the hallucination rate in complex question-and-answer tasks exceeds 30% [2]. In medical scenarios, misuse of drug use advice can be life-threatening; in the field of law, misinterpretation of clauses may lead to judicial injustice; in the field of education, misleading content can affect knowledge acquisition. Therefore, accurately evaluating the factual response

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD28242

capabilities of LLMs is not only a technical need to improve model performance, but also a key to ensuring AI ethics and social trust.

Existing research shows that LLMs have multiple manifestations of factual defects. First of all, its output stability is susceptible to prompt word expression, context interference and entity name variants. For example, the output consistency of some models is only 28% under entity perturbation [1]. Secondly, the model's confidence in facts is often estimated due to surface language form fluctuations, cross-language scenarios are more prominent, and knowledge is insufficient, and static training data is difficult to cope with real-time updated world knowledge, resulting in low accuracy of time-sensitive fact responses. These issues highlight the urgency of building a comprehensive and dynamic factual assessment system.

This paper reviews the LLMs factual reply ability evaluation method from five key perspectives: internal state analysis method, detecting factuality by analyzing model neuron activation and output dynamic characteristics; confidence estimation evaluation method, focusing on the confidence stability of the model on factual output; dynamic consistency evaluation method, evaluating output consistency under input perturbation; robustness evaluation method, quantifying the impact of prompting framework and contextual interference; external knowledge-assisted evaluation method, combining search enhancement technology to evaluate the model's dependence on external knowledge. By sorting out these methods, this article provides a panoramic perspective for understanding the factual defects of LLMs, lays the foundation for optimizing the evaluation framework and improving the reliability of the model, and helps LLMs to be trustworthy in high-risk fields.

2. Evaluation method

2.1. Internal state analysis method

Inspired by human lie detectors judging lies through physiological reactions, this method assumes that there is a distinguishable pattern of internal states when large language models generate factual and non-factual content. By extracting the static and dynamic features of the model, a fact detection pipeline is constructed to evaluate the factuality of the output. Among them, in terms of static characteristics, an activation map is used to focus on the activation intensity of fact-related neurons, so that such neurons are activated more intensely when the fact output is output. In terms of dynamic characteristics, the final output ranking of each layer (the fact output is more stable later ranking), the top-k output index (the fact output has higher semantic similarity), and the top-k output probability (the fact output has greater probability difference). In addition, the model design adopts the Siamese network, combined with the convolutional neural network (ResNet18) to process static and partial dynamic features, and the gated recurrent unit (GRU) to process sequential dynamic features, and uses triple marginal loss training.

This method enables the evaluation accuracy of more than 96% on various architectures such as GPT2-XL-1.5B and Llama2-7B [3], verifying the effectiveness of using the internal state of the model for fact detection.

2.2. Confidence estimation assessment method

This method analyzes the reliability of the confidence estimation of its own output factuality by analyzing the large language model and judging its factual response ability. This method mainly focuses on two scenarios: fact verification (P(T), judging the authenticity of the statement) and

DOI: 10.54254/2755-2721/2025.LD28242

question-and-answer (P(IK), judging whether the correct answer can be given), and compares the performance of different confidence estimation methods. This method trains the classifier based on the model's hidden state, predicts the factuality of the output, and also requires model weights and labeled data. The sequence probability of this method uses the average logarithmic probability of the output sequence as a confidence indicator, and allows the model to directly prompt the model to generate confidence scores (such as "1-10 points") and replace the token probability to measure confidence by the probability of a specific token (such as "yes/no". In addition, the method also evaluates confidence based on the consistency of multiple sampling outputs.

The training probes of this method performed optimally in P(T) and P(IK) tasks, with AUPRC significantly higher on the Lama T-REx dataset than other methods. The verbalization method of its instruction fine-tuning model has good performance, but overall it is lower than that of the training probe. The model confidence is poor in semantic equivalent inputs (such as paraphrase, translation), with a standard deviation of up to 0.5 [4].

2.3. Dynamic consistency assessment method

This method evaluates the consistency of the output of large language models when input changes in inputs to measure the stability of their factual replies by introducing entity perturbations (such as aliases, different expressions) and attribute perturbations (such as synonyms descriptions). In terms of dynamic datasets, the method extracts 130 time-sensitive facts (covering categories of countries, athletes, organizations, etc.) from Wikidata [1], containing synonymous expressions of different lexical forms and attributes of entities. Its evaluation index is propt consistency, that is, the proportion of outputting the same facts, and is divided into consistency under entity perturbation (different alias) and attribute perturbation (different relationship description).

Even if advanced models (such as GPT-4) still have 20% outputs containing outdated information [1], some models (such as OpenELM 270M) have only 26% consistency under entity perturbation [1]. However, this method makes the consistency of instruction fine-tuning models (such as ChatGPT and Llama-2 Chat) significantly higher than that of pre-trained models, with an average increase of more than 30% [1].

2.4. Robustness assessment method

This method quantifies the sensitivity of the large language model to the cue framework effect (different cue forms) and context interference (irrelevant information), measures the robustness of factual replies by the distance of the output probability distribution, and proposes a comprehensive evaluation of the MONITOR index. The core indicator of this method, MONITOR, consists of PFD (prompt framework degree, measuring the difference in output distribution of different cue forms) and IRD (interference correlation, measuring the impact of contextual interference), and is calculated by probability distribution distance. The evaluation process mainly uses "anchor point" (including the correct output probability of the forward context) as the benchmark, calculates the distance from the output probability under other prompts or contexts, and combines the MONITOR value.

This method makes MONITOR highly correlated with accuracy (Pearson coefficient -0.846) [5], which can effectively distinguish model robustness. The MONITOR value of LLaMA-30b-ins. Has the lowest (0.479) [5], and the calculation cost is only 1/3 of the traditional accuracy evaluation [5], and is suitable for large-scale model evaluation.

DOI: 10.54254/2755-2721/2025.LD28242

2.5. External knowledge-assisted assessment method

This method combines search-enhanced generation (RAG) technology to evaluate the dependence of large language models on external information through the matching degree of external knowledge (such as a domain knowledge base), and then measures their factual response capabilities, which are especially suitable for domain-specific and time-sensitive queries. This method constructs 34,000 QA pairs on the dataset [2], covering college courses, research projects, etc., and is generated through web crawlers and automated annotation. Its RAG process includes, embedding a model (mixedbread-ai/mxbai-embed-large-v1) to search related external knowledge, reordering the model to optimize the search results, and the core model (LLaMA-2) to generate replies based on the search content. The evaluation indicators of this method mainly include recall, F1 score (the balance between fact matching accuracy and recall) and BLEU score (generating fluency).

This method makes the RAG pipeline increase the F1 score from 0.186 to 0.289 [2] compared with the baseline model without RAG, and the factual error rate of domain-specific queries is reduced by 40% [2]. In addition, fine-tuning the embedding model can further improve the search accuracy and increase the recall by 7% [2].

3. Commonly used data sets and evaluation criteria

3.1. Mainstream datasets

Table 1. Characteristics of datasets and applicable scenario tables for large language model factual evaluation

Dataset	Release time	Scale	Features	Applicable scenarios
FKTC	2023	210,000 prompt words [5]	Covering 20 fact data sets, focusing on common facts	Robustness assessment (such as MONITOR)
DyKnow	2024	130 Time-sensitive facts[1]	Dynamically update Wikidata data, with entity perturbation samples	Dynamic consistency assessment
T-REx	2018	34K triple (facts)[1]	Includes entity-relationship-attribute structure, supports propt generation	External knowledge-assisted assessment
PopQA	2023	14K Question-The answer is correct [4]	Contains synonymous answer variants, suitable for confidence estimation assessment	Confidence stability analysis
CMU domain dataset	2024	34,000 QA [2]	Focus on university fields (such as course time, research projects)	Domain-specific factual assessment

Table 1 shows the characteristics of datasets and applicable scenarios for large language model factual evaluation. Through exploration, it can be seen that there are currently five mainstream data sets, namely FKTC, DyKnow, T-REx, PopQA, and CMU domain data sets.

The FKTC dataset is constructed based on 16,167 triplets of T-REx, containing 210,171 Q&A tips, covering 20 fact categories [5]. Seven synonymous prompts were generated through GPT-4, and diversity was ensured through BLEU threshold filtering and manual verification. Positive and negative interference contexts were also added to evaluate the robustness of the model under the prompt framework and context interference, and supporting MONITOR index calculation. The DyKnow dataset is a dynamic benchmark dataset. 130 time-sensitive facts are extracted from

DOI: 10.54254/2755-2721/2025.LD28242

Wikidata [1], covering categories such as country, athlete, organization, etc., including perturbation samples such as entity alias and attribute synonyms. The data is updated in real time to reflect fact changes, and is used to evaluate the accuracy of the model to time-sensitive facts and consistency under input perturbations. The T-Rex dataset contains 34K entity-relationship-attribute triplets [1], which provides structured fact data and supports the generation of diversified hints. It is the basis for building FKTC and also provides the original factual basis for external knowledge-assisted evaluations (such as RAG). The PopQA dataset contains 14K questions-Answer correct [4]. Each answer has multiple synonyms, covering different popular entities. It is used to evaluate the reliability of the model's confidence estimation in question-and-answer tasks. It is an important test dataset for the confidence estimation evaluation method. The CMU field data set contains 34,000 QA pairs [2], focusing on the university field, and generated through network crawlers and automated annotation, is used to evaluate the factuality of field-specific queries and verify the effectiveness of RAG and other technologies in vertical fields.

3.2. Evaluation criteria

The evaluation indicators of the factual response ability of large language models cover multiple dimensions, and these indicators reflect the performance of the model in factuality from different angles. Accuracy is one of the basic indicators, which refers to the proportion of correct fact output in the total output. For example, LLM Factoscope has an accuracy of more than 96% on a custom data set [3], which can intuitively reflect the degree of matching between model output and fact. The F1 score is used to measure the balance between fact matching accuracy and recall. For example, after RAG optimization, the F1 score increased from 0.186 to 0.289 on the CMU field dataset [2], taking into account the model's ability to capture and cover facts. The consistency indicator focuses on the stability of the model's output under input perturbation. In the DyKnow dynamic benchmark, the consistency of the optimization model under entity and attribute perturbation can exceed 50% [1], reflecting the model's stable response ability to similar inputs. The MONITOR index combines PFD (prompt frame degree) and IRD (interference correlation) to quantify the impact of cue framework effects and context interference. The MONITOR value of LLaMA-30b-ins is as low as 0.479 [5], reflecting its low sensitivity to cue perturbation, and the calculation cost is only 1/3 of the traditional accuracy evaluation [5], and it suitable for large-scale model evaluation. AUPRC (area under the Precision-Recall curve) mainly evaluates the reliability of confidence estimation. The AUPRC of the training probe method reaches 0.8 or above on Lama T-REx [4], which can effectively reflect the model's confidence estimation quality of its own output. These indicators complement each other and together form a comprehensive evaluation system for the factual response capabilities of large language models, providing a multi-dimensional reference basis for understanding model performance and optimizing model performance.

4. Technical evaluation and optimization

4.1. Technical optimization

According to existing research, the performance of large language models (LLMs) in the field of factual knowledge processing shows that they have significant advantages and key shortcomings. The improvement of their factual capabilities needs to be coordinated from multiple dimensions, such as mechanism analysis, evaluation optimization, and technical adaptation. From the internal mechanism of factual knowledge of LLMs, the two-dimensional framework of token backtracking

DOI: 10.54254/2755-2721/2025.LD28242

and single token decoding can be revealed that the factual recall process of LLMs follows the three-stage model of "subject enhancement and dissemination-object retrieval and reordering-conclusion fusion and generation" [6]. The model does not lack relevant knowledge, but it is difficult to filter accurate information based on the context in the search and reordering stages. This discovery provides a mechanism-level basis for subsequent targeted optimization. In addition, by further proposing the concept of "hidden knowledge", it is pointed out that the factual knowledge encoded by LLMs (measured by hidden state detection classifiers) is significantly more than the knowledge reflected in their external output, with an average relative gap of 40% [7], and some facts are even completely hidden (the model can be perfectly identified but never generated in 1,000 samples), highlighting the disconnection between LLMs' knowledge expression and internal storage.

In terms of the construction of a factual evaluation system, the Pinocchio benchmark that can be constructed covers 20,713 multi-select problems and multi-dimensional, multi-language, and multi-domain scenarios [8]. The experiment found that even advanced models such as GPT-3.5-Turbo, their performance in real-world, domain-specific and multi-language tasks is still 6.4% lower than other tasks [8], and the Chain Thinking (CoT) prompt is only effective for instruction fine-tuning models. In addition, the SUMMEDITS benchmark uses the low-cost protocol of "seed verification-edit generation-marker filtering" to achieve high repeatability in 10 fields (notifier consistency 0.9) [9]. The results show that GPT-4's performance on this benchmark is still 8% lower than that of humans [9], exposing the shortcomings of existing LLMs in complex factual reasoning. People can also use the built FACT-BENCH to cover 20 fields, 134 attribute types and 3 answer formats [10], and find that instruction fine-tuning will significantly impair knowledge recall ability (the pre-trained model continues to be better than the instruction fine-tuning model), and the model's processing ability of long-tail entities and date-type attributes is particularly weak.

performance and optimizing model performance.

4.2. Reinforcement learning optimization effect

In the evaluation of the factual reply ability of large language models, reinforcement learning optimization has shown certain effects. In the legal clause Q&A task, after reinforcing learning optimization, the model's confidence fluctuation standard deviation dropped from 0.42 to 0.28[1], which shows that reinforcement learning helps stabilize the model's confidence in fact output and reduce factual errors caused by unstable confidence. In terms of external knowledge utilization, combining the strategies of retrieval enhancement generation (RAG) and reinforcement learning, the model's dependence on external knowledge is increased by 40% [2], reducing its dependence on internal error knowledge. Therefore, on the CMU field dataset, the F1 score of the RAG+RL strategy is increased by 12% [2] than that of pure RAG, and the processing capacity of low-frequency entities is also increased by 30% [2], effectively improving the model's factual reply performance in domain-specific queries and reducing the illusion caused by excessive dependence on internal knowledge. These optimization effects show that reinforcement learning has a positive role in improving the stability of the model's factual response and the effective use of external knowledge, which helps to enhance the reliability of large language models in high-risk fields.

DOI: 10.54254/2755-2721/2025.LD28242

5. Limitations and solutions

Table 2. Problems and solutions for mainstream methods of factual evaluation of large language models

Method	Question	Solution
Internal state analysis method	Depend on model architecture (such as only support Transformer)	Developing a cross-architecture universal activation feature extractor (such as general metrics based on attention weights)
Confidence estimation assessment method	The distribution of confidence across languages is significant (such as French-Polish)	Introduce cross-language alignment loss function to unify confidence value distribution [4]
Dynamic consistency assessment method	Missing low-frequency entity assessment	Expand DyKnow's low-frequency entity samples, and the proportion has increased to 30%[1]
External knowledge-assisted assessment method	RAG searches for noise interference (such as incorrect introduction of irrelevant documents)	In combination with reinforcement learning design "retrieval quality rewards" to filter low-relevance documents[2]

Table 2 shows the problems and solutions for mainstream methods of factual evaluation of large language models.

In Addition, the dataset has some problems in the assessment of factual response capabilities of large language models, such as limited coverage, and the lack of vertical field data such as rare diseases and niche laws [1], which makes it insufficient in evaluating the model's ability to respond to facts in these specific fields. In addition, the sample balance problem is prominent, with English samples accounting for more than 80% of the cross-language samples, resulting in the evaluation results that may be biased in different language scenarios. At the same time, small samples have obvious defects. For example, the "outdated information" samples in DyKnow only account for 20% [1], making it difficult to fully evaluate the model's ability to handle time-sensitive facts. In response to these problems, people can adopt corresponding solutions to build a "vertical field fact library", integrate PubMed's rare disease cases in the medical field, and supplement local regulations in the legal field to broaden the coverage of the data set and meet the evaluation needs of different vertical fields. In addition, people can also use the "translation enhancement + human verification" method to expand small language samples, so that the proportion of samples in each language is less than 10% [4], improve the balance of samples, and generate historical fact samples through timeline slice method, increase the proportion of "outdated information" samples in DyKnow to 50% [1], solve the defects of small samples, and more comprehensively evaluate the model's ability to handle time-sensitive facts. These solutions are designed to improve the quality and applicability of data sets, thereby more accurately evaluating the factual response capabilities of large language models and providing more reliable data support for related research and applications.

6. Conclusion

This paper reviews the evaluation method from five perspectives, namely, the internal state analysis method detects the facts by analyzing static activation and dynamic output characteristics, the confidence estimation evaluation method focuses on the confidence stability of the model on its own output, the dynamic consistency evaluation method evaluates the output consistency under input perturbation, the robustness evaluation method quantifies the impact of the prompting framework

DOI: 10.54254/2755-2721/2025.LD28242

and contextual interference, and the external knowledge-assisted evaluation method combines search enhancement technology to evaluate the dependence on external knowledge.

This paper systematically sorts out these methods to provide a reference for understanding the factual defects of large language models, optimizing the evaluation framework and improving reliability, helping them to use credible applications in high-risk fields, and promoting technological iteration to achieve the upgrade from "laboratory toys" to "reliable tools".

Future research can make breakthroughs from three aspects: First, build a multimodal knowledge reward mechanism, incorporate structured knowledge such as images and tables into the RL framework, and enhance the understanding of complex facts; second, expand vertical field data sets, such as the rare disease case library in the medical field and the local regulations library in the legal field; third, establish a "dynamic evaluation-real-time optimization" closed loop, and through monthly updated benchmark tests and adaptive reward functions, the model can continuously adapt to the time evolution of knowledge.

References

[1] Seyed Mahed Mousavi, Simone Alghisi and Giuseppe Riccardi. (2025). LLMs as Repositories of Factual Knowledge: Limitations and Solutions. IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. VOL. XX, NO. X.

[2] Jiarui Li, Ye Yuan and Zehua Zhang. (2024). Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases. arXiv preprint arXiv: 2403.10446v1.

[3] Jinwen He, Yujia Gong, Zijin Lin, Cheng'an Wei, Yue Zhao and Kai Chen. (2024). LLM Factoscope: Uncovering LLMs' Factual Discernment through Measuring Inner States. Findings of the Association for Computational Linguistics: ACL 2024.

[4] Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller and Lluís Màrquez.(2024). Factual Confidence of LLMs: on Reliability and Robustness of Current Estimators. arXiv preprint arXiv: 2406.13415v1.

[5] Weixuan Wang, Barry Haddow, Alexandra Birch and Wei Peng. (2024). Assessing Factual Reliability of Large Language Model Knowledge. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).

[6] Yiqun Wang, Chaoqun Wan, Sile Hu, Yonggang Zhang, Xiang Tian, Yaowu Chen, Xu Shen and Jieping Ye. (2025). Tracing and Dissecting How LLMs Recall Factual Knowledge for Real World Questions. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

[7] Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpector, Jonathan Herzig and Roi Reichart. (2025). Inside-Out: Hidden Factual Knowledge in LLMs. Published as a conference paper at COLM 2025.

[8] Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu and Zhijiang Guo. (2024). Towards Understanding Factual Knowledge of Large Language Models. Published as a conference paper at ICLR 2024.

[9] Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander R. Fabbri, Caiming Xiong, Shafiq Joty and Chien-Sheng Wu. (2023). LLMs as Factual Reasoners: Insights from Existing Benchmarks and Beyond. arXiv preprint arXiv: 2305.14540v1.

[10] Jiaqing Yuan, Lin Pan, Chung-Wei Hang, Jiang Guo, Jiarong Jiang, Bonan Min, Patrick Ng and Zhiguo Wang. (2024). Towards a Holistic Evaluation of LLMs on Factual Knowledge Recall. arXiv preprint arXiv: 2404.16164v1.