Evaluating the Factuality of Large Language Models Using Multiple Plug-and-Play Fact Sources

Zhaoheng Huang, Yutao Zhu*, Ji-Rong Wen, Zhicheng Dou

Gaoling School of Artificial Intelligence, Renmin University of China
{huangzh, ytzhu, jrwen, dou}@ruc.edu.cn

Abstract

Large language models (LLMs) often produce factually inaccurate content, or *hallucinations*, which undermines their reliability. Existing factuality evaluation systems usually rely on a single predefined fact source, making them task-specific and hard to extend. We present UFO, a unified framework for factuality evaluation that supports multiple plug-and-play fact sources. UFO integrates human-written evidence, web search results, and LLM knowledge within a single evaluation pipeline, and allows users to flexibly select, reorder, and even define customized sources. The system is accessible through both a Python interface and a web-based demo, offering interactive claim-level verification and visualization. Experiments show that UFO system achieves moderate consistency with human annotations. Overall, UFO serves as a transparent and extensible platform for benchmarking fact sources, comparing LLMs, and enabling real-world fact-checking applications across diverse domains.

Introduction

Large language models (LLMs) have enabled a wide range of generative AI applications such as ChatGPT (Hurst et al. 2024) and DeepSeek (DeepSeek-AI 2024), but their outputs still often contain factual inaccuracies or *hallucinations* (Bang et al. 2023; Ji et al. 2023; Kalai et al. 2025). These errors undermine the reliability of LLMs, limiting their use in domains where factual correctness is essential. Existing factuality evaluation systems (Iqbal et al. 2024; Zhao et al. 2024) typically rely on a single or fused source (e.g., Wikipedia) and are designed for specific tasks, which makes them difficult to scale or adapt when new domains and hallucination detection tasks arise.

In practice, fact sources may vary in cost, reliability, and availability. For instance, collecting human-written references can be expensive (Min et al. 2023), while web search results and an LLM's internal knowledge may be more scalable but can be less consistent and noisier compared to curated human-written references. However, prior systems typically rely on a single source or apply fixed fusion strategies, without supporting flexible substitution or systematic combination. As a result, they cannot benchmark the effectiveness of different fact sources and their combinations, leaving open questions about how these sources complement each other across various hallucination detection tasks.

To address this challenge, we present UFO, a unified and flexible framework for factuality evaluation. UFO supports plug-and-play integration of multiple fact sources, including human-written evidence, web search results, and LLM knowledge, and allows users to flexibly select, reorder, or even define custom fact sources. UFO also allows users to directly input texts or upload files of LLM outputs for evaluation. The system applies a unified verification process to generate claim-level factuality scores, and provides both a Python interface and an interactive web demo for visualization and comparison. With UFO, users can evaluate factual accuracy, compare the contribution of different fact sources, and benchmark LLMs across tasks, offering a transparent and extensible platform for factuality assessment.

Related Work

Factuality evaluation has progressed from n-gram metrics to LLM-based approaches such as FactScore (Min et al. 2023) and FacTool (Chern et al. 2023), which decompose text into claims and verify them against fixed sources. Follow-up work has explored fact editing and correction (Wang et al. 2023), efficient verification (Tang, Laban, and Durrett 2024; Shan, Bauer, and Manning 2025), tool-augmented verification in multimodal settings (Chen et al. 2024), and verifiability enhancement (Song, Kim, and Iyyer 2024). Prior systems such as OpenFactCheck (Iqbal et al. 2024) and Loki (Li et al. 2024) explore pipelines for factuality evaluation, but remain tied to a single or fused fact source (e.g., Wikipedia or web search) and offer limited flexibility and extensibility.

Our system UFO fills this gap by enabling plug-and-play integration of multiple fact sources, including human-written evidence, web results, and LLM knowledge within a unified evaluation framework. Unlike prior work, the UFO system further allows users to define customized fact sources through a simple interface, making factuality evaluation more flexible, extensible, and adaptable across domains.

UFO System Overview

The UFO system follows a four-step pipeline, guiding users through uploading evaluated texts, configuring fact sources, running evaluation under scenarios, and exploring results.

*Corresponding author.

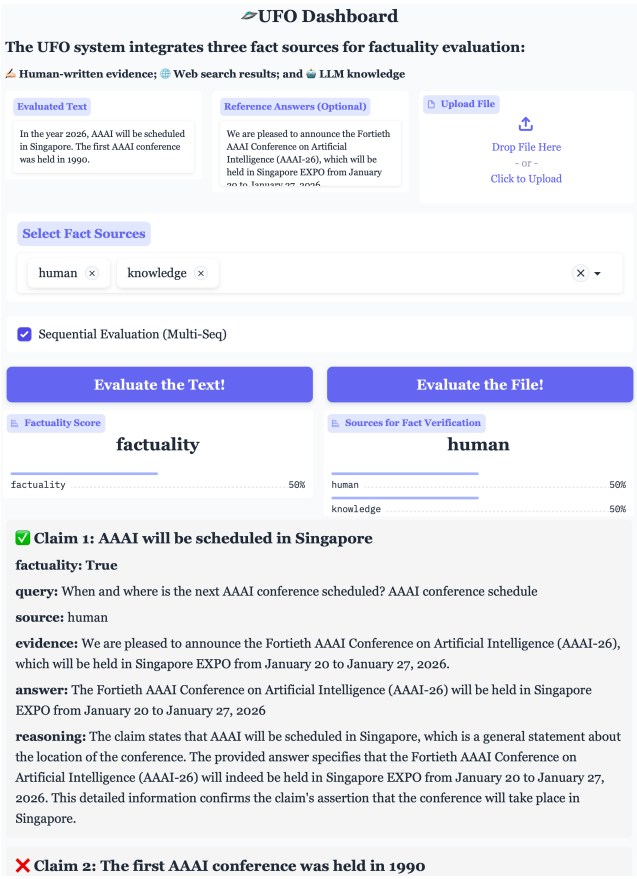

Figure 1: The UFO dashboard and workflow.

Text Upload and Fact Source Setup. The UFO system allows users to either upload texts directly or provide a JSONL file with optional curated human-written evidence. Each line in the input JSONL file contains two keys: the "response" for the evaluated LLM text, and the optional "reference answers" for the human-written evidence.

The UFO system provides three pre-defined fact sources: (1) **Human-written evidence**, which can be uploaded by the user or retrieved from a general corpus such as a Wikipedia dump collected prior to the release of ChatGPT; (2) **Web search results**, obtained via the Google Search API provided by Serper; and (3) **LLM knowledge**, constructed by sampling $n = 5$ passages per query from an LLM with a temperature $t = 1$. This design leverages the observation that hallucinated outputs often show low consistency across multiple generations and higher perplexity (Miao, Teh, and Rainforth 2024). The web interface allows users to reorder the selected sources interactively. For advanced use, the Python interface supports further customization: users can freely adjust parameters or add new sources by implementing lightweight adapters under the project directory.

Factuality Evaluation. With fact sources configured, the UFO system evaluates factuality through three LLM-based modules: (1) **Fact Unit Extraction** decomposes the evaluated text into several atomic fact units, each consisting of a claim c and a corresponding query q; (2) **Fact Source Verification** processes each query q against passages P from a fact source S, extracting one candidate answer a from each passage (or "NOANS" if no valid answer can be found); and (3) **Fact Consistency Discrimination** determines whether a claim c and its candidate answer a are factually consistent, assigning a binary label of 1 (factual) or 0 (hallucinated).

By combining these modules, UFO evaluates the contribution of different fact sources at the claim level. The overall factuality score of a text is then calculated as the average of the binary labels across all fact units.

Evaluation Scenarios. The UFO system supports two evaluation scenarios for assessing the factuality of model outputs. **Single-Source Evaluation.** This setting is equivalent to prior work, where each claim in the text is verified against a fixed fact source. The outcome for each claim is determined by majority voting over the binary labels of all valid answers (excluding "NOANS") within the chosen source. **Multi-Source Evaluation.** To go beyond fixed-source evaluation, UFO enables users to combine multiple fact sources in two ways. In the *Multi-MV* mode, all valid answers from different sources are aggregated and a decision is made by majority voting over their binary labels. In the *Multi-Seq* mode, fact sources are checked in a user-specified order; if no valid answer can be found in the current fact source, the UFO system proceeds to the next. This sequential strategy mirrors human fact-checking behavior and provides a systematic way to benchmark how different sources complement one another.

Results Exploration The workflow is illustrated in Figure 1. The UFO system processes the factuality evaluation pipeline and produces three types of outputs: (1) the overall factuality score, (2) the proportion of each fact source used during evaluation, and (3) detailed verification results for each decomposed claim. We collect LLM outputs from diverse domains with human annotations of factual accuracy. Experimental results show that the *Multi-Seq* evaluation scenario achieves the best consistency with human annotations. In addition, UFO logs the entire evaluation process to a result file, allowing users to view fact-source contribution pie charts via the Python interface and benchmark sources and LLMs in a plug-and-play manner. Detailed implementation and results are provided in the project documentation.

Conclusion and Future Work

In this paper, we introduced UFO, a unified framework for evaluating the factuality of LLM outputs using multiple plug-and-play fact sources. The UFO system supports both a Python interface and a web-based demo, allowing users to flexibly select, reorder, or define their own fact sources through a simple configuration interface. This makes factuality evaluation more transparent, customizable, and accessible for researchers and practitioners. Future work includes exploring real-world applications in domains such as education and news verification, and extending the plug-in mechanism to support additional modalities and domain-specific sources such as scientific databases.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62402497).

References

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; Do, Q. V.; Xu, Y.; and Fung, P. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. arXiv:2302.04023.

Chen, X.; Wang, C.; Xue, Y.; Zhang, N.; Yang, X.; Li, Q.; Shen, Y.; Liang, L.; Gu, J.; and Chen, H. 2024. Unified Hallucination Detection for Multimodal Large Language Models. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 3235–3252. Association for Computational Linguistics.

Chern, I.-C.; Chern, S.; Chen, S.; Yuan, W.; Feng, K.; Zhou, C.; He, J.; Neubig, G.; Liu, P.; et al. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv preprint arXiv:2307.13528*.

DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; Madry, A.; Baker-Whitcomb, A.; Beutel, A.; Borzunov, A.; Carney, A.; Chow, A.; Kirillov, A.; Nichol, A.; Paino, A.; Renzin, A.; Passos, A. T.; Kirillov, A.; Christakis, A.; Conneau, A.; Kamali, A.; Jabri, A.; Moyer, A.; Tam, A.; Crookes, A.; Tootoonchian, A.; Kumar, A.; Vallone, A.; Karpathy, A.; Braunstein, A.; Cann, A.; Codispoti, A.; Galu, A.; Kondrich, A.; Tulloch, A.; Mishchenko, A.; Baek, A.; Jiang, A.; Pelisse, A.; Woodford, A.; Gosalia, A.; Dhar, A.; Pantuliano, A.; Nayak, A.; Oliver, A.; Zoph, B.; Ghorbani, B.; Leimberger, B.; Rossen, B.; Sokolowsky, B.; Wang, B.; Zweig, B.; Hoover, B.; Samic, B.; McGrew, B.; Spero, B.; Giertler, B.; Cheng, B.; Lightcap, B.; Walkin, B.; Quinn, B.; Guarraci, B.; Hsu, B.; Kellogg, B.; Eastman, B.; Lugaresi, C.; Wainwright, C. L.; Bassin, C.; Hudson, C.; Chu, C.; Nelson, C.; Li, C.; Shern, C. J.; Conger, C.; Barette, C.; Voss, C.; Ding, C.; Lu, C.; Zhang, C.; Beaumont, C.; Hallacy, C.; Koch, C.; Gibson, C.; Kim, C.; Choi, C.; McLeavey, C.; Hesse, C.; Fischer, C.; Winter, C.; Czarnecki, C.; Jarvis, C.; Wei, C.; Koumouzelis, C.; and Sherburn, D. 2024. GPT-4o System Card. *CoRR*, abs/2410.21276.

Iqbal, H.; Wang, Y.; Wang, M.; Georgiev, G. N.; Geng, J.; Gurevych, I.; and Nakov, P. 2024. OpenFactCheck: A Unified Framework for Factuality Evaluation of LLMs. In Hernandez Farias, D. I.; Hope, T.; and Li, M., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 219–229. Miami, Florida, USA: Association for Computational Linguistics.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12): 248:1–248:38.

Kalai, A. T.; Nachum, O.; Vempala, S. S.; and Zhang, E. 2025. Why Language Models Hallucinate. arXiv:2509.04664.

Li, H.; Han, X.; Wang, H.; Wang, Y.; Wang, M.; Xing, R.; Geng, Y.; Zhai, Z.; Nakov, P.; and Baldwin, T. 2024. Loki: An Open-Source Tool for Fact Verification. *CoRR*, abs/2410.01794.

Miao, N.; Teh, Y. W.; and Rainforth, T. 2024. SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-by-Step Reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 12076–12100. Association for Computational Linguistics.

Shan, A.; Bauer, J.; and Manning, C. D. 2025. Osiris: A Lightweight Open-Source Hallucination Detection System. *CoRR*, abs/2505.04844.

Song, Y.; Kim, Y.; and Iyyer, M. 2024. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, 9447–9474. Association for Computational Linguistics.

Tang, L.; Laban, P.; and Durrett, G. 2024. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, 8818–8847. Association for Computational Linguistics.

Wang, Y.; Reddy, R. G.; Mujahid, Z. M.; Arora, A.; Rubashevskii, A.; Geng, J.; Afzal, O. M.; Pan, L.; Borenstein, N.; Pillai, A.; Augenstein, I.; Gurevych, I.; and Nakov, P. 2023. Factcheck-GPT: End-to-End Fine-Grained Document-Level Fact-Checking and Correction of LLM Output. *CoRR*, abs/2311.09000.

Zhao, X.; Yu, J.; Liu, Z.; Wang, J.; Li, D.; Chen, Y.; Hu, B.; and Zhang, M. 2024. Medico: Towards Hallucination Detection and Correction with Multi-source Evidence Fusion. In Hernandez Farias, D. I.; Hope, T.; and Li, M., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 34–45. Miami, Florida, USA: Association for Computational Linguistics.