FACTS&EVIDENCE: An Interactive Tool for Transparent Fine-Grained Factual Verification of Machine-Generated Text

Varich Boonsanong♠   **Vidhisha Balachandran**◇   **Xiaochuang Han**♠
Shangbin Feng♠   **Lucy Lu Wang**♠   **Yulia Tsvetkov**♠
♠University of Washington   ◇Microsoft Research
{varicb, xhan77, shangbin, lucylw, yuliats}@cs.washington.edu
vidhishab@microsoft.com

Abstract

With the widespread consumption of AI-generated content, there has been an increased focus on developing automated tools to verify the factual accuracy of such content. However, prior research and tools developed for fact verification treat it as a binary classification or a linear regression problem. Although this is a useful mechanism as part of automatic guardrails in systems, we argue that such tools lack transparency in the prediction reasoning and diversity in source evidence to provide a trustworthy user experience.

We develop FACTS&EVIDENCE—an interactive and transparent tool for user-driven verification of complex text. The tool facilitates the intricate decision-making involved in fact-verification, presenting its users a breakdown of complex input texts to visualize the credibility of individual claims along with explanation of model decisions and attribution to multiple, diverse evidence sources. FACTS&EVIDENCE aims to empower consumers of machine-generated text and give them agency to understand, verify, selectively trust and use such text.

1 Introduction

In our time of proliferating use and adoption of AI-generated text and AI-assisted writing, the need for robust and transparent fact verification tools to help users verify the accuracy of machine-generated content has never been more pressing. Errors and hallucinations generated by AI models are already translating into real-world problems through errors in generated code introducing unintended bugs (Shrestha, 2024), nonexistent books and references propagating misinformation (Ryan, 2022; Dolan, 2024; Bohannon, 2023), and incorrect medical advice causing health complications (Alstin, 2023), which cost stakeholders time, money, and in some cases leading to physical harm (Metz and Blumen-

thal, 2019; Johnson et al., 2022). Although fact-verification of machine-generated texts has been studied extensively in recent research (Manakul et al., 2023; Gao et al., 2023; Huang et al., 2023; Li et al., 2024), this work has not resulted in practical tools that are accessible to users, easy to use, and transparent (Do et al., 2024). As such, developing fact-verification tools *aimed at informing users of the veracity of AI-generated content* is necessary to enable users to understand, verify, and trust such content.

Recent research has developed methods for automatically detecting factual errors and hallucinations through self-checking with Large Language Models (LLMs) (Manakul et al., 2023), uncertainty calibration (Farquhar et al., 2024; Zhang et al., 2023; Feng et al., 2024) or fine-tuning specific factuality classifiers (Mishra et al., 2024), and tools for visualizing such errors for users (Fatahi Bayat et al., 2023; Krishna et al., 2024). The primary intent of this research has been the binary classification (factual vs. non-factual) of complex text to serve as an intrinsic guardrail in AI systems. As a result, their decisions are categorical, often not explainable, and can be noisy (Mishra et al., 2024).

On the other hand, real-world human fact-checking is significantly more complex, involving verification against diverse sources with varying levels of reliability and leanings (Augenstein et al., 2019; Dias and Sippitt, 2020; Glockner et al., 2022). Evidence for verifying facts is traced through various sources, and claims are verified against each source. The authenticity and reliability of each source or news outlet is considered when aggregating and deciding the veracity of each claim. Such fine-grained views into each claim per source provide a detailed understanding of the correctness of long, complex texts, allowing consumers of the text to assess which specific claims to trust and use.

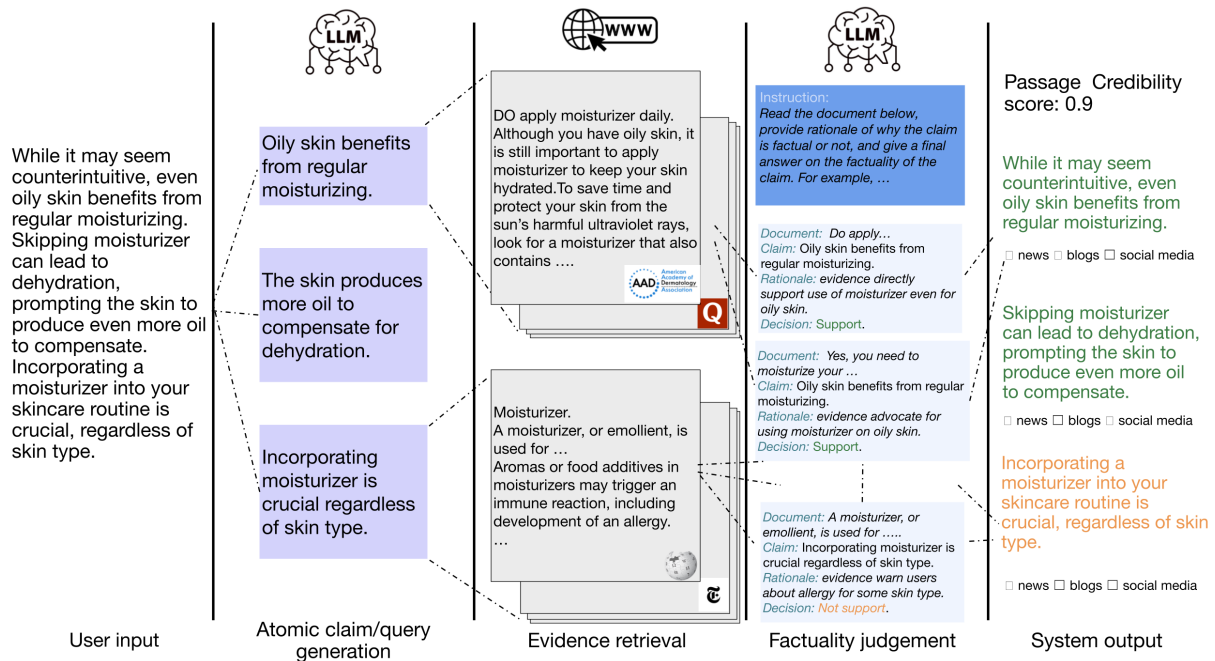In this work, we take a step towards addressing this gap. We present FACTS&EVIDENCE—a user-

437

Figure 1: The pipeline figure of FACTS&EVIDENCE. The user input for verification goes through atomic claim and query generation, evidence retrieval, and factuality judgement processes. The system output is aggregated from the judgements over atomic claims, providing users with sources of evidence and an overall credibility score.

driven, transparent and interactive fact-verification tool to empower consumers of AI-generated text to understand the factual accuracy of the text. FACTS&EVIDENCE takes any long, complex text from a user as input and presents a breakdown of the text into individual claims. Each claim is searched on the web to identify multiple, diverse evidence items and verify against each of them. Ultimately, decisions from multiple pieces of evidence and claims are aggregated to compute a credibility score for the entire text. In addition to aggregate decisions, FACTS&EVIDENCE presents the per-claim, per-evidence decision, along with a model-generated rationale explaining the decision and a source tag indicating the source category of the evidence. Further, FACTS&EVIDENCE is interactive, allowing users to include or exclude any source type from the verification process based on personal preferences and specific needs.

We evaluate the reliability of decisions produced by FACTS&EVIDENCE using the FAVA dataset (Mishra et al., 2024) and show that the predicted scores are competitive with strong baselines, outperforming previous systems by ~ 40 F1 points, and establishing a high-quality system in claim verification of open-ended generation tasks. In summary, our contribution is FACTS&EVIDENCE— a high-quality, fine-grained, and interactive tool

providing transparency and access to real-time information in the fact-verification process which can inform and empower users, giving them the ability to inspect and validate the veracity of each part of the content they use.[1]

2 FACTS&EVIDENCE – User Interface

This section presents the user interface of FACTS&EVIDENCE when it is used to inspect the veracity of machine-generated texts.

2.1 Landing Page:

Upload Panel: The first screen (ref: Figure 2) presents a panel for the user to upload the text they want to verify. Additionally, the user can configure different parameters associated with their search: (i) LLM: choice of 2 backend LLMs (gpt-3.5-turbo-0301 and Llama3-8B) used for verifying claims; (ii) Retrieval Mode: choice of dense (embedding-based) or sparse (keyword-based) retrieval for extracting relevant evidence from web search text; (iii) Evidence Configuration: choice of top-k web results to parse, top-k relevant passages within each search result, and context size

---

[1]Demo URL:
https://factsandevidence.cs.washington.edu/
Video URL:
https://www.youtube.com/watch?v=_MaAI2H7c3w

Figure 2: FACTS&EVIDENCE Upload Panel

of each passage. For each of these parameters, default values are set based on internal evaluations for best claim verification performance and user experience as outlined in §4. Once the text is provided and the user is satisfied with the parameters, they can submit their request to verify the text.

Readme: The home screen also includes a README that explains the objectives of the tool and a brief explanation of how the system conducts fact verification behind the scenes to provide full transparency to users. It describes how a user can use the tool and what level of interaction they can expect from the tool. In addition, it includes an explanation for each configurable parameter and possible choices available to the user.

2.2 Fact Visualization Page:

Once the user submits their text, we run our fact verification engine on our back-end and present a detailed view of the final results to the user.

Credibility Panel: First the credibility panel (ref: Figure 3) presents a view of the input text with the credibility of individual sentences highlighted. A *credibility score* is computed for each sentence in the text. Each sentence is visualized with a color coded with [red] for low credibility [0-0.3], [orange] for medium credibility [0.3-0.6] and [green] for high credibility [0.6-1]. Finally, a credibility score for the entire text is also presented based on aggregated sentence level scores.

Evidence Panel: For each sentence, individual atomic claims are displayed in an evidence panel (ref: Figure 4) with an expanded view showing multiple evidences from the Web used for prediction, a claim-evidence level factuality prediction (supported/not supported/irrelevant), and a rationale which

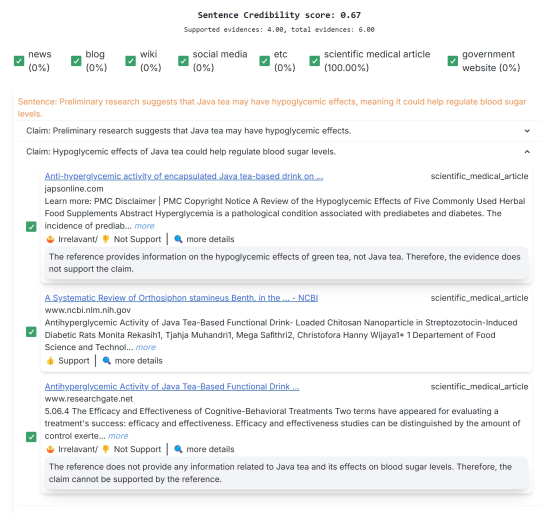Figure 3: Credibility Panel of FACTS&EVIDENCE

Figure 4: Evidence Panel of FACTS&EVIDENCE

explains the prediction in free-form text. Each evidence has a source type tag associated with it. Source types indicate the general type of web page the evidence was extracted from: news, scientific article, blogs, etc.

The Evidence Panel gives users the option to include specific evidence documents or sources of evidence based on their preferences. Checkboxes can be selected to include or exclude each evidence document and recompute the credibility score, allowing users to inspect how specific evidence impacts the summary score. Users are also provided source-type check boxes at a global level to include or exclude certain sources they do not wish to use. For example, when verifying healthcare-related text users may choose to exclude blogs and social media posts as evidence and only focus on scientific articles or government websites for information. Credibility scores are automatically updated based on the source and evidence selections made by the users. This fine-grained view and user control into the evidences and source types allows user agency in customizing the fact verification process for different domains and inputs.

439

3 FACTS&EVIDENCE – Behind the Scenes

This section details the back-end framework for verifying facts in model-generated texts. For providing a detailed view of facts, evidence items, and their factual accuracy, the user text is broken down to individual claims, and for each claim, multiple, diverse supporting evidence items are retrieved from the web. A factual accuracy prediction and explanation for each claim is produced using an LLM and the resulting details are displayed to the user as outlined in §2. Figure 1 presents an outline of our process which we describe in detail below.

Atomic Claim Generation   To break down the user text into individual atomic claims, we prompt an LLM to break a paragraph into its constituent sentences and break each sentence into a list of atomic claims. To accommodate context-length limits of different LLMs, we syntactically break long paragraphs into smaller ones and assume that each paragraph's context is self-contained. We prompt an LLM in a few-shot setting to decontextualize and generate a claim that can be used for web search. Atomic claims often assume context of the larger text, including pronouns and other references that can adversely impact our search for relevant evidence. For example, the claim 'Headaches are a common side effect of this treatment' does not include relevant information about which treatment the claim refers to. Incorporating decontextualization instructions addresses this by specifically instructing the LLM to include any additional context required to make the claim sufficiently detailed for querying the Web.

Evidence Retrieval   Given a list of atomic claims and their queries, we conduct a Google web search using the `Serper`[2] service and scrape the top-N websites as evidence using the `readability`[3] library hosted on `Cloudflare Worker`[4]. We identify the most relevant text within each source evidence using a text retrieval model (`jina-embeddings-v3`). The evidence context is constructed by taking M sentences before and after each relevant evidence sentence. In the case where the user prefers multiple passages within a document, this process can be repeated for top-K similar sentences instead of just the best match

[2]https://serper.dev/
[3]https://github.com/mozilla/readability
[4]https://workers.cloudflare.com/

sentence. In the end, the evidence contexts for an atomic claim can be multiple paragraphs of highly relevant text chunks from diverse sources of evidence.

Source categorization   To provide users with the option of including or excluding an entire source or category of evidence, we classify each evidence source into a set of predefined categories. We extract the hostname of the website and ask an LLM in a zero-shot fashion to select one of the closest related categories among {news, blog, wiki, social media, scientific/medical article, government website, other}. These categories were chosen to cover a wide range of evidence sources and can be expanded further in the future.

Factuality Judgement   We use chain-of-thought style prompting (Wei et al., 2022) by giving the retrieved evidence and atomic claim to LLM to determine whether the evidence can support the atomic claim or not. We extract the model's decision and rationale to present in the UI so that users can validate the decision and decide whether or not it should be included in the aggregate factuality assessment. We weigh each piece of evidence equally and calculate a sentence-level and passage-level credibility score as the *percentage of evidence that support the claims* in a sentence/passage.

$$\text{credibility score} = \frac{\text{\# support evidence}}{\text{\# total evidence}}$$

To promote reusability, we create our back-end API to comply with open API standards. Any developer can call any intermediate steps of our tools through this API interface. Additionally, all prompts used to query LLMs in the various steps are included in Appendix §A.

4 Experiments

Evaluation: We evaluate our predictions using the FavaBench dataset (Mishra et al., 2024). FavaBench includes human factuality annotations at the span level for approximately 1,000 model-generated responses from two popular LLMs (ChatGPT and Llama2-Chat) for 200 diverse information-seeking and knowledge-intensive queries. For ChatGPT and Llama2-Chat, 59.8% and 65.5% of the respective responses include at least one hallucination, respectively. We obtain sentence-level labels by associating any sentence with a factually incorrect span with an incorrect sentence label. We run the FavaBench inputs
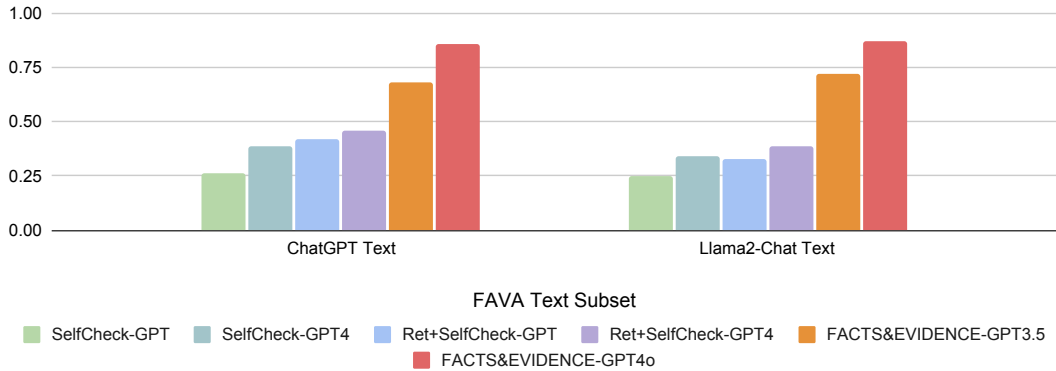
Figure 5: Binary F1 Results on factual error detection. **FACTS&EVIDENCE improves error prediction accuracy by ~44 points on average across the two subsets.**

Setting	ChatGPT Text	Llama2-Chat Text
No:Ev 1, Ctxt W 15	0.23	0.25
No:Ev 1, Ctxt W 30	0.23	0.22
No:Ev 3, Ctxt W 15	0.68	0.72
No:Ev 3, Ctxt W 30	0.64	0.67
No Atomic Claim Gen	0.40	0.43

Table 1: Results of Ablation Study

Retriever	ChatGPT Text	Llama2-Chat Text
BM25	0.46	0.47
Distilbert	0.59	0.72
Snowflake	0.84	0.86
Jina	0.84	0.87

Table 2: Analysis of varying evidence retrievers

through our system using `gpt-3.5-turbo-0301` for all LLM calls, and for each sentence we predict Not-Factual if *any* of the evidences does not support the claims (credibility score $< t$)[5]. Following Mishra et al. (2024), we report sentence-level F1 against human-annotated judgments on the ChatGPT and Llama2-Chat subsets of the data separately.

Baselines: We compare predictions from our tool against multiple strong baselines: (i) SelfCheck-ChatGPT (Manakul et al., 2023) prompts ChatGPT (`gpt-3.5-turbo-0301`) to reflect and identify factual errors in the generated text; (ii) SelfCheck-ChatGPT4 (`gpt-4`) follows the same setup with GPT4 instead; (iii) Ret+SelfCheck-GPT follows the setup from Mishra et al. (2024) and Min et al. (2023) to prompt ChatGPT to identify factual errors in the text given the retrieved document from a static retrieval system (Contriever-MSMarco with Wikipedia documents) to augment the original prompt; (iv) Ret+SelfCheck-ChatGPT4 follows the same setup with GPT4 instead.

Overall Results: In Figure 5, we present the sentence-level binary F1 results on FavaBench. Our results show that FACTS&EVIDENCE performs better than prior work, showing that the tool predictions are high quality and reliable. In par-

ticular, FACTS&EVIDENCE outperforms all baselines that directly use LLMs with or without evidence to identify factual errors in complex text (SelfCheck-GPT, Ret+SelfCheck-GPT, SelfCheck-GPT4, Ret+SelfCheck-GPT4). Specifically, compared to previous works, which rely on static knowledge sources and single evidence to verify facts, our tool breaks complex text to individual claims, dynamically retrieves diverse evidence from the web, and uses multiple evidence for verifying each fact, which improves the quality of our predictions.

Ablation Study: In Table 1, we present an ablation study that varies the number of evidences and the size of the context window of each evidence to study the effect of evidences on performance. Our results indicate that our performance increases with an increased number of evidence, validating our hypothesis that complex fact checking requires multiple diverse evidences for high-quality predictions. However, the increase in the context window does not significantly increase the F1 score. This enables us to build a more efficient demo by using a smaller number of tokens as evidence, reducing the load time, and improving the user experience. Additionally, to validate the importance and quality of the atomic claim generator, we ablate the pipeline by removing the claim generation. Our results show significant drop in performance confirming that breaking complex claims to atomic claims is essential for fact verification.

[5]$t = 0.3$ considering 3 evidences

Sentence: Java tea is commonly used as a diuretic, meaning it may increase urine production. The property has led to its traditional use in managing conditions such as edema (swelling) and UTIs.		
Error Type	**Example**	**% Occurrence**
Incorrect Atomic Claim	It is traditionally used to manage conditions such as edema.	26.7%
Incorrect Evidence Retrieval	There's a lot of different types of unflavored black tea, and the type and where it's grown makes a big difference.	53.3%
Incorrect Factuality Judgment	Not Factual prediction for 'Java tea is commonly used as a diuretic'	20.0%

Table 3: Error types and their occurrences. Predominant failures in FACTS&EVIDENCE arises in evidence retrieval.

In addition to evidence representation, we also conducted a study on the retriever model in Table 2. The Jina's embedding model (`jinaai/jina-embeddings-v3`) has the highest F1 score, followed by Snowflake embedding model (`snowflake-arctic-embed-l`), and other industry standard retrieval systems such as BM25 and Distillbert (`msmarco-distilbert-dot-v5`).

Qualitative Analysis We also performed a qualitative error analysis on 30 randomly sampled incorrect predictions from FACTS&EVIDENCE to identify parts of the pipeline that led to incorrect predictions Table 3. Our analysis indicates that most of the errors in the pipeline are due to evidence retrieval failures. Such failures often stem from search engine failing to find relevant documents or website scraper failures due to security blocks or parsing errors. These failures are external to the system and could potentially be mitigated by increasing the number of evidence at a higher cost. Other errors in the pipeline were due to incorrect atomic claims (improper decontextualization, complex claims) or incorrect factuality judgements. Future work could explore custom finetuned models for the individual steps to mitigate these errors.

5 Related Work

Factuality and Hallucination in LLMs: Recently, a focus has been on addressing the problem of model hallucinations (Kumar et al., 2023; Mallen et al., 2023; Min et al., 2023). Previous research has studied the variety of errors and hallucinations produced by different models and characterized them using taxonomies for summarization (Pagnoni et al., 2021), text simplification (Devaraj et al., 2022), dialogue generation (Dziri et al., 2022; Gupta et al., 2021), and open-ended generation (Mishra et al., 2024). Automated methods were developed to identify and flag model errors and hallucinations (Min et al., 2023; Mishra et al.,

2024; Gao et al., 2023; Yuksekgonul et al., 2023). Although hallucination detection methods are primarily used as safety guardrails in systems, early tools have been developed to detect and present users with factual errors in model-generated text (Krishna et al., 2024; Fatahi Bayat et al., 2023). While these tools primarily present single evidence and final model decisions, FACTS&EVIDENCE provides a transparent and fine-grained view of the factual accuracy of model-generated text with an interactive interface for consumers with diverse evidences, detailed credibility scores, and explanations.

Fact-Verification: In parallel, there has been active research on automated fact verification of human written text and claims as well (Walter et al., 2019; Guo et al., 2021; Bekoulis et al., 2021). Various methods, tools, and systems have been built for automated fact checking of human-written claims (Thorne and Vlachos, 2018; Nakov et al., 2021). However, these techniques and systems have not been tested for LM-generated text. Although the aim of FACTS&EVIDENCE is to support the verification and understanding of facts in the text generated by a model and has been explicitly evaluated for this setting, it can be directly adapted and used to verify written human claims.

6 Conclusion

We presented FACTS&EVIDENCE—an interactive tool for user-driven fact verification of AI-generated text. The tool enables consumers of outputs from various LLMs to visualize and understand individual claims and ground the claims in diverse evidence from the Web. We evaluate the decisions made by the FACTS&EVIDENCE system and show that our system outperforms strong baselines in the claim verification task. FACTS&EVIDENCE promotes transparency and user agency in the verification process of facts and aims to empower con-

sumers to understand fine-grained claims and appropriately use text generated by AI systems.

Limitations

Though our goal is to provide a transparent user facing fact-verification tool for consumers of model generated text, our system internally relies on models for multiple steps in the pipeline. Inherently, there can arise concerns of accuracy and reliability of the pipeline. Through evaluations we showed that our system produces high-quality predictions for most of the cases. But as with any automated system, errors can arise. We hope that by designing and interactive user interface that exposes all the intermediate steps of the pipeline, our users will be able to view and judge our system decisions. Our goal is to maintain the final agency with the user directly. In future, we would like to include measures of confidence in model generations and predictions and present that information to the user as well to support them in their decision making.

Another limitation of our current system is efficiency. To improve the reliability, accuracy and transparency of our system, we introduced multiple sequential steps in the pipeline with multiple LLM calls to external APIs. While this enabled us to inspect and improve each step, we had to trade off some speed. Our demo currently takes 30s-1min for passages of 5-7 sentences. We plan to focus on improving the efficiency of the system in the future by introducing local small LMs that are finetuned for individual steps.

Acknowledgments

References

Chad Van Alstin. 2023. ChatGPT's 'hallucinations' make it difficult to trust for advice on cancer care. https://healthexec.com/topics/artificial-intelligence/chatgpt-hallucinations-cancer-care-chatbot-ai. [Online; accessed 24-August-2023].

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. *arXiv preprint arXiv:1909.03242*.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *ACM Comput. Surv.*, 55(1).

Molly Bohannon. 2023. Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions. https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/. [Online; accessed 5-August-2024].

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

Nicholas Dias and Amy Sippitt. 2020. Researching fact checking: Present limitations and future opportunities. *The Political Quarterly*, 91(3):605–613.

Hyo Jin Do, Rachel Ostrand, Justin D. Weisz, Casey Dugan, Prasanna Sattigeri, Dennis Wei, Keerthiram Murugesan, and Werner Geyer. 2024. Facilitating human-llm collaboration through factuality scores and source attributions.

Eric W Dolan. 2024. ChatGPT hallucinates fake but plausible scientific citations at a staggering rate, study finds. https://www.psypost.org/chatgpt-hallucinates-fake-but-plausible-scientific-citations-at-a-staggering-rate-study-finds/. [Online; accessed 14-April-2024].

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. FaithDial: A Faithful Benchmark for Information-Seeking Dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyy, Samira Khorshidi, Fei Wu, Ihab Ilyas, and Yunyao Li. 2023. FLEEK: Factual error detection and correction with evidence retrieved from

external knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 124–130, Singapore. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *ACL*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhijiang Guo, M. Schlichtkrull, and Andreas Vlachos. 2021. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Dialfact: A benchmark for fact-checking in dialogue. *arXiv preprint arXiv:2110.08222*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv*, abs/2311.05232.

Thaddeus L. Johnson, Natasha N. Johnson, Denise McCurdy, and Michael S. Olajide. 2022. Facial recognition systems in policing and racial disparities in arrests. *Government Information Quarterly*, 39(4):101753.

Kundan Krishna, Sanjana Ramprasad, Prakhar Gupta, Byron C. Wallace, Zachary Chase Lipton, and Jeffrey P. Bigham. 2024. Genaudit: Fixing factual errors in language model outputs with evidence. *ArXiv*, abs/2402.12566.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Language generation models can cause harm: So what can we do about it? an actionable survey. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3299–3321, Dubrovnik, Croatia. Association for Computational Linguistics.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. *ArXiv*, abs/2401.03205.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark John Francis Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *EMNLP*, abs/2303.08896.

Cade Metz and Scott Blumenthal. 2019. How A.I. Could Be Weaponized to Spread Disinformation. https://www.nytimes.com/interactive/2019/06/07/technology/ai-text-disinformation.html. [Online; accessed 05-August-2024].

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *arXiv preprint arXiv:2401.06855*.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proc. NAACL*.

Jackson Ryan. 2022. Meta Trained an AI on 48M Science Papers. It Was Shut Down After 2 Days. https://www.cnet.com/science/meta-trained-an-ai-on-48-million-science-papers-it-was-shut-down-after-two-days/. [Online; accessed 20-Nov-2022].

Asim Shrestha. 2024. How a single ChatGPT mistake cost us $10,000+. https://web.archive.org/web/20240609213809/https:

//asim.bearblog.dev/how-a-single-chatgpt-mistake-cost-us-10000/. [Online; accessed 09-June-2024].

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nathan Walter, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. 2019. Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37:350 – 375.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. 2023. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. *ICLR*, abs/2309.15098.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. *arXiv preprint arXiv:2311.13230*.

A LLM Prompts

Here we share the LLM prompts and instructions used at different stages of the pipeline:

Given a sentence, your task is to separate a given sentence into series of short claims which are standalone fully self-contained sentences. You will be given a paragraph with sentences marked and a particular target sentence within it. You should produce atomic claims for that sentences in a way that is consistent with the paragraph. Each claim should be theirs own bullet points. Each claim should be decontextualized such that it should be understandable without the previous context. Use the context of the paragraph or sentence to replace any references or claims. You don't need to repeat the questions, just directly give me claims.
Here are some examples of sentences and their atomic claims:
Paragraph: S1: <example1 sentence1> S2: <example1 sentence2> S3: <example1 sentence3> S4: <example1 sentence4>
Break the following sentence into atomic claims: S2: <example sentence2>
Claim_1: <example claim1>
Claim_2: <example claim2>
Claim_3: <example claim3>
Claim_4: <example claim4>

Paragraph: S1: <example2 sentence1> S2: <example2 sentence2> S3: <example2 sentence3> S4: <example2 sentence4> S5: <example2 sentence5>
Break the following sentence into atomic claims: S1: <example sentence1>
Claim_1: <example claim1>
Claim_2: <example claim2>
Claim_3: <example claim3>
Claim_4: <example claim4>
Claim_5: <example claim5>
Claim_6: <example claim6>

Paragraph: <input paragraph>
Break the following sentence into atomic claims: S8: <input sentence>

Table 4: **Prompt for Atomic Claim Generation.**

Given both the hostname of the website and the list of categories. Choose the best category that will best describe this hostname. Only give the final valid category from the list without any explanation.
Categories: ["news", "blog", "wiki", "social_media", "etc", "scientific_medical_article", "government_website",] Hostname: <input hostname>

Table 5: **Prompt for Classifying the source type.**

Given the claim, its context, and its evidence. You must find whether the evidence can contradict the claim in this context. The context is the original text that the claim is derived from. Critically evaluate whether the evidence shows any sign of contradiction or doesn't directly support the claim. Here are some tips on what may cause evidence contradiction:
1. if the claim is subjective, the evidence may not be supporting the claim and that is not contradiction.
2. if the evidence is inconsistent or contradictory, it is not supporting the claim and that is not contradiction.
3. if the evidence is talking about a different topic or not related to the claim, that is not contradiction.
4. if the context have contradictory information with the evidence, but not with the specific claim that are focusing on, it is not contradiction.
Here are examples of how you should respond.
Passage: <example context>
Claim: <example claim>
Evidence: <example evidence>
Rationale: <output rationale>
Final Verdict: <output prediction>
Passage: <example context>
Claim: <example claim>
Evidence: <example evidence>
Rationale: <output rationale>
Final Verdict: yes.
Let's think through this step by step.
Passage: <context>
Claim: <claim>
Evidence: <evidence>
Give your answer in the following format.
Rationale: <rationale>
Final Verdict: <yes for evidence agrees with the claim, no otherwise>
give a summarized answer as the last line with either say yes for support or no for not supported

Table 6: **Prompt for judging factuality.**