Correcting Hallucinations in News Summaries: Exploration of Self-Correcting LLM Methods with External Knowledge

Anonymous ACL submission

Abstract

While large language models (LLMs) have shown remarkable capabilities to generate coherent text, they suffer from the issue of hallucinations – factual inaccuracies. Self-correcting systems are especially promising for tackling hallucinations. They leverage the multi-turn nature of LLMs to iteratively generate verification questions inquiring additional evidence, answer them with internal or external knowledge, and use that to refine the original response with the new corrections. These methods have been explored for encyclopedic generation, but less so for domains like news summaries. In this work, we investigate two state-of-the-art self-correcting systems: apply them to hallucinated summaries, using three search engines, and evaluate. We analyze the results and provide qualitative insights into systems' performance, revealing interesting practical findings on G-Eval and human evaluation, and the benefits of search snippets and few-shot prompts.

1 Introduction

A common issue with Large Language Models (LLMs) is that they tend to produce *hallucinations* – responses that sound convincing but are factually incorrect or misleading (Ji et al., 2023). This limitation poses challenges for their reliability and adoption, especially in critical applications like law, healthcare, and news (Wang et al., 2024a).

While numerous methods to counter hallucinations have been developed in recent years (Tonmoy et al., 2024), many focus on pre-training and fine-tuning. For popular closed models like GPT 4, the *post-hoc correction* methods, which correct the initial response after it has been generated, are quite important. In particular, *self-correcting* methods approach hallucination correction as a step-by-step process where the response is broken into smaller units and iteratively corrected using internal LLM knowledge or external sources (Kamoi et al., 2024).

The effectiveness of these methods has been demonstrated for generating biographies or encyclopedic articles (Min et al., 2023; Chern et al., 2023), but their application to the domain of news summarization remains underexplored. News articles are time-sensitive and factually dense, which underscores the need for correct summaries and effective fact-checking (Graves and Amazeen, 2019).

Furthermore, evidence retrieval is a crucial component of self-correcting systems – many questions are open regarding which search engine to use, which snippets or article chunks to select, and how to best integrate them. Finally, the trade-off between balancing the faithfulness to original text with doing strong corrections is often neglected.

To explore these research gaps, we take two popular multi-step correction systems, CoVE (Dhuliawala et al., 2023) and RARR (Gao et al., 2023), augment them with external search engines, and apply them to correct hallucinated news summaries from a dataset SummEdits (Laban et al., 2023). We compare the performance of different search engines and settings, and the influence of prompts, uncovering important considerations for future.

2 Related Work

Hallucinations are a common problem in natural language generation (NLG) tasks, including abstractive text summarization (Ji et al., 2023). A survey by Zhang et al. (2023) divides hallucinations into input-conflicting, context-conflicting, and fact-conflicting. The focus of our work lies in fact-conflicting, which are hallucinations where facts in output contradict the world knowledge. While hallucinations can be observed by looking at the uncertainty in model's logits (Varshney et al., 2023), this is only possible for open-source models. In the widely popular closed-source models such as ChatGPT, factuality has to be assessed through textual output only. This has led to the rise of self-

correcting LLM techniques (Kamoi et al., 2024).

The multi-step self-correcting LLM methods can base it on internal LLM knowledge (Madaan et al., 2023; Kim et al., 2023). For external search, usually only Wikipedia (Wang et al., 2024b; Gou et al., 2024) or Google search is used. It is often applied to tasks like generating biographies. For news summaries, methods such as text infilling (Balachandran et al., 2022) or entity linking to graphs (Dong et al., 2022) have been explored to correct errors.

We augment the two self-correcting methods, CoVe and RARR, to support external search. Our study is among the first to explore this type of methods for news, to evaluate three different search engines, changes in snippets and full-text retrieval, and to compare closed with open base LLMs.

3 Systems

In our study, we use two systems designed to detect and iteratively correct hallucinations, both of which have demonstrated strong results and gained popularity: Chain-of-Verification (CoVe) and Retrofit Attribution using Research and Revision (RARR). Both systems follow the same workflow: (1) Get Initial Response, (2) Generate Verification Questions (to help self-correct any errors), (3) Answer Questions (using evidence from internal knowledge or search engine), (4) Rewrite Response (with previous answers and any found inconsistencies).

Given the baseline response b, there are k generated follow-up questions $q_1, ..., q_k$, which try to gather more information related to the response b. This is generated using a base LLM and a prompt M_q. Afterward, evidence e for each question q is retrieved from the source s using the method $R(q, s)$, where s can be internal LLM knowledge, gold news article, or external search engine. This collected evidence is used as input with questions to the answering model $M_a(q, e)$, which gives answers $a_1, ..., a_k$. Finally, baseline response and answers are given to the refinement model $M_r(b, a)$, which outputs the final refined response r. All prompts for *M* are in Appendix C.

The difference between models is in prompts used to generate and answer the questions, and perform the final refinement. Also, CoVe is zero-shot, while RARR is based on few-shot examples.

4 Setup and Experiments

The LLM used in most experiments is *GPT-4o-mini-2024-07-18*, a closed model from OpenAI with good reasoning capabilities (OpenAI, 2024). It was queried through OpenAI API. Any encoder-only models were run on one Nvidia V100 GPU with 16GB VRAM for one computation hour.

4.1 Dataset

SummEdits (Laban et al., 2023) is a benchmark dataset of hallucinated text summaries in many domains. The dataset was constructed by first perturbing named entities and relations in summaries and then passing to humans for annotation on whether the summaries are factual or not. We take the subset *news* (constructed from top Google News 2023 articles), consisting of 819 summaries. While the original intent of benchmark was to evaluate hallucination detection ability of LLMs, we repurpose it for hallucination detection with fact correction.

4.2 Evaluation Methods

We use the gold summaries as reference answers.

We measure string dissimilarity using the Levenshtein normalized edit distance (**NED**) (Yujian and Bo, 2007). This metric is not ideal because even one word difference can be a major hallucination. Therefore, we compare the semantic similarity (**Sem.**) between the gold and output summary by embedding them with the model SimCSE (Gao et al., 2021) and calculating the cosine similarity.

NLI Score is a metric that utilizes the concept of natural language inference (NLI), or entailment recognition, by using the reference answer as the hypothesis and the generated answer as the premise. The intuition behind this approach is that a good answer should logically entail the reference answer. Using NLI this way has been done for evaluating the quality of summaries (Mishra et al., 2021; Laban et al., 2022; Steen et al., 2023). Following this approach, we use the model DeBERTa-v3 (He et al., 2023), We use the version fine-tuned on a wide array of NLI datasets, which works well for long text (Laurer et al., 2024). This model predicts three scores (entailment, neutral, contradiction) and we report the average score across the whole dataset.

G-Eval (Liu et al., 2023) is a framework based on LLM prompting with chain-of-thoughts to evaluate the quality of generated texts in a form-filling paradigm. It is one of the most popular "LLM-as-judge" metrics (Zheng et al., 2023), which evaluate the LLM output with an LLM using finely crafted LLM prompts (see Appendix C) and take the numerical output as final score. We evaluate three aspects: relevance, factuality, and overall quality.

Human Evaluation. We perform human evaluation with 25 participants. They were shown 10 gold summaries and refined summaries by RARR and CoVe, and rated for each the overall quality (based on factuality and relevance) from 1 to 10 and the entailment relation for each summary, amounting to 1000 ratings (see more details in Appendix A).

4.3 Search Engines

Google is the world's most widely used search engine. It offers the API service Google Programmable Search Engine, which queries the search engine and returns results as links and snippets. The price is 5 US dollars per 1,000 queries. **Bing** is the flagship search engine from Microsoft. We use it via Bing Web Search API provided by the Azure platform for the price of 10 USD per 1,000 transactions. **DuckDuckGo** is a smaller search engine, aiming to help protect online privacy. While the coverage is lower than the other two engines, its usage through API is completely free. We query it through the Python package *duckduckgo-search*.[1]

We use the search results of these search engines in two settings: chunks from *full articles* and *snippets*. All search engines return results for the query with links to articles included in top results. In the full-article setting, we parse the text from HTML of the article, split into chunks, embed with SimCSE, and use cosine similarity to the query to select top 5 passages. We concatenate these passages and use them as input evidence. In addition to links, all search engines provide snippets that answer the query and highlight the most important part from the respective article. We use the top 5 snippets and concatenate them, using them as input evidence.

5 Results and Discussion

Table 1 shows the average results of all metrics for the two systems on SummEdits. Qualitative insights are found in Tables 6 and 7.

Internal vs. External Knowledge. The first two rows of Tab. 1 used internal LLM knowledge to answer verification question. While this led to moderate performance, results with search engines were higher for both systems – showing the **need for external search** for effective factual error correction.

The last two rows show the baseline of using the original (gold) news article as input evidence. It had the highest G-Eval scores, highlighting the key role of precise evidence for effective corrections.

[1] https://pypi.org/project/duckduckgo-search/

Choice of Search Engine. As seen in Table 1, Google snippets performed the best for CoVE but Bing outperformed it on RARR for the full-article setting. The highest performance overall was achieved by Bing snippets with RARR, as measured by six different metrics. This shows the promising potential of Bing, which is underexplored in existing studies. DuckDuckGo also achieved decent but lower performance.

Table 7 shows examples of refined responses from CoVe with the three engines. All three engines successfully identified the hallucination involving biologists. Still, they found different extra information that was included in the refined response, with Bing the only one pinpointing the correct telescope discovery. This shows the engines get similar results but Bing usually led to best corrections overall, because for generated queries Bing provided longer and more informative snippets.

Still, the high price of Google and Bing APIs could be a hurdle for large-scale usage. DuckDuckGo achieved adequate performance for $0 and, thus, offers a promising free alternative. Future work could explore additional search filters or filtering of results by trustworthy domains.

Snippets or Full-Article Chunks? When looking at RARR scores of NED, sem. similarity, and G-Eval for snippets and full articles, all are slightly better for the snippets setting. Still, the average NLI scores reveal the full picture – the setting with full articles had high NLI neutral scores. This is because the responses were often refined with irrelevant evidence, whereas the snippet setting produced responses that had a higher NLI-entailment score. The contradiction scores were similar in both. This shows that **snippets** are usually more **on point and related to the actual search query**, while using the **full articles** can lead to selecting **noisy or irrelevant passages from articles**.

Another interesting finding is the general similarity of overall scores, including NLI scores, for the setting with full articles – this shows that all three search engines mostly found the same articles among its top results and then the similarity function selected same passages from those articles.

Zero-shot or Few-shot? Table 6 shows an instance from SummEdits with the gold, hallucinated, and refined summaries by both systems. Both used Bing as the search engine and, thus, both managed to correct factual errors in the input summary (e.g., *struggling* instead of *thriving*). Still, it is evident

verification system	evidence source	simple		NLI			G-Eval ↑		
		NED ↓	Sem. ↑	Ent. ↑	Neu.	Con. ↓	Overall	Factual.	Relev.
CoVE	GPT 4o mini	0.51	81	30	28	42	50	45	49
RARR	GPT 4o mini	0.10	94	45	15	40	65	62	70
CoVE	Google (snip.)	0.51	84	41	25	34	56	50	59
CoVE	Bing (snip.)	0.55	81	37	28	35	49	46	51
CoVE	DDG (snip.)	0.54	80	31	28	41	47	42	47
RARR	Google (full)	0.33	91	24	46	**30**	64	51	68
RARR	Bing (full)	0.32	92	28	40	32	63	50	68
RARR	DDG (full)	0.34	91	27	41	32	64	50	68
RARR	Google (snip.)	0.24	93	40	28	32	67	56	72
RARR	Bing (snip.)	**0.14**	**95**	**49**	16	35	**69**	**60**	**73**
RARR	DDG (snip.)	0.25	92	32	28	40	60	49	62
CoVE	gold article	0.49	88	43	39	18	70	63	76
RARR	gold article	0.21	94	47	34	19	75	67	83

Table 1: Results of CoVE and RARR on SummEdits using three different search engines. NED refers to normalized edit distance, Sem. to average cosine semantic similarity, NLI scores to average prediction probability for entailment, neutral, and contradiction. The best score for each metric is in **bold**, while the second best is underlined.

that RARR returned a summary close in form to the input summary, whereas CoVe augmented the summary with additional information found on Bing.

This difference in length is the consequence of the fact that RARR uses six examples in its few-shot prompt, while CoVe does not use any examples. CoVe also sometimes returned summaries similar to the input summary with minimal changes, however it often returned a lot longer summaries. Long summaries do not necessarily imply hallucinations, but can be summaries with additional context for readers. This points to the fact that **few-shot** prompts are better if the end goal is to **preserve the faithfulness to the original draft**, while **zero-shot** relaxed prompts are better when **adding additional context and making bold edits is preferred**. The few-shot examples are general-domain, so the findings are not just for news.

Open LLMs. We also ran experiments with LLaMa 3.1 (70B), results are in Table 5. For RARR, it had on average weaker scores than GPT 4o-mini, but came quite close, confirming the recent trend of open models closing the gap to closed competitors. For CoVe, which does not have few-shot examples, it generated a lot longer final refined responses than GPT, with lots of detailed explanations. This led to increased G-Eval (Overall & Relevance) and NLI metrics, since these metrics favor information-heavy summaries, but the G-Eval factuality score heavily decreased and summaries were too complex. We additionally ran Mixtral 8x7B with internal knowledge, but it underperformed compared to both Llama and GPT. Future work could explore more open LLMs and evaluate user-centric text quality aspects like readability.

5.1 Human Evaluation

The mean human scores for quality of 10 examples with Bing snippets for 25 participants were **0.68** for RARR and **0.54** for CoVe, showing **users preferred RARR** refinements. The mean G-Eval score for these 10 examples were **0.65** and **0.52**, respectively. This shows an impressively **high alignment of humans with G-Eval**, with the average difference of 3%. Our custom prompts for factuality and relevancy have a high potential for future use, and this positions G-Eval as a promising metric to use when human annotations are not available due to time and costs. For NLI, the alignment was decent but less apparent – DeBERTa favored the neutral class, while humans predicted more entailment. More details are in Appendix A.

6 Conclusion and Future Work

In this study, we explored the impact of different evidence sources and search engines on the performance of two SotA systems for post-hoc hallucination correction, CoVe and RARR, for news summaries. Our detailed results show that zero-shot correction systems like CoVe yield more expressive and bold corrections that change the style, while few-shot systems like RARR optimize for faithfulness to the original text and this was favored by humans in evaluation. Additionally, G-Eval metric was highly aligned with humans. We also found that Bing's search snippets led to most informative corrections, followed closely by Google, but DuckDuckGo can be a viable alternative due to its free API and decent performance. We envision future work focusing on enhancing retrieval with structured queries and assessing evidence reliability.

Limitations

An important limitation lies in the fact that all modules of the iterative self-correcting systems rely on using LLMs, which comes with its own set of challenges. The generated follow-up questions are not always perfect or precise, the generated answers from snippets can be off-point, and the final refinement of responses can be too excessive. Future work could explore how to incorporate more controllable generation or structured and rule-based techniques for correcting the output.

Another limitation comes from the high complexity of the system and reliance on calls to external APIs, including LLM APIs and search engine APIs. This can inevitably lead to slow processing speed of these systems when compared to approaches that use smaller encoder-only models or rule-based techniques. Still, we were forced to rely on API calls to LLMs due to our hardware resource limitations. Other lines of work could explore how to better incorporate open and local models into the workflow, for better accountability and faster processing time.

Finally, our work deals only with the news domain, which could limit the generalizability of findings to other domains and use cases.

References

Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. 2022. Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9818–9830, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios. *Preprint*, arXiv:2307.13528.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *Preprint*, arXiv:2309.11495.

Yue Dong, John Wieting, and Pat Verga. 2022. Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1067–1082, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

Lucas Graves and Michelle A. Amazeen. 2019. Fact-checking as idea and practice in journalism. *Oxford Research Encyclopedia of Communication*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Preprint*, arXiv.:2406.01297. To appear in TACL.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. *arXiv preprint arXiv:2303.17491*.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9662–9676, Singapore. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. Looking beyond sentence-level natural language inference for question answering and text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1336, Online. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Julius Steen, Juri Opitz, Anette Frank, and Katja Markert. 2023. With a little push, NLI models can robustly and efficiently predict faithfulness. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 914–924, Toronto, Canada. Association for Computational Linguistics.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *Preprint*, arXiv:2401.01313.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *Preprint*, arXiv:2307.03987.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2024a. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Preprint*, arXiv:2306.11698.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024b. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. *Preprint*, arXiv:2311.09000.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Shijia Zhou, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R. Mortensen, and Lori Levin. 2024. Constructions are so difficult that Even large language models get them right for the wrong reasons. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3804–3811, Torino, Italia. ELRA and ICCL.

A Human Evaluation

The main goal of the human evaluation was to judge two automated metrics, NLI predictions and LLM-as-a-judge (G-Eval), by observing the alignment between human preference and machine evaluation results. All the evaluation responses and results are attached to the ARR submission.

A.1 Study Format and Instructions

User study was conducted with 25 participants. All participants are pursuing a master's degree or a PhD degree in computer science at authors' university. They were not monetarily compensated since they are in-house annotators from our school's department of computer science. All responses were anonymous and collected only for the purpose of this research study. Users were provided with instructions described in Table 2.

The survey was hosted as a questionnaire on the JotForm platform.[2] In total, there were 10 examples, where each example consisted of a correct summary, a hallucinated summary, a summary corrected by CoVe, a summary corrected by RARR, and 4 questions to answer. In Figure 1, a sample screenshot from the evaluation form is provided.

Users were asked to evaluate each of the two generated summaries in two aspects: overall quality and NLI relation. The overall quality was estimated by rating from 1 to 10 and it refers to (a) how factually accurate was the summary, and (b) how relevant and on-topic was it. The NLI (entailment) relation were mapped to NLI classes by asking the users whether the generated summary supports the gold summary (entailment), contradicts the gold summary (contradiction), or partially aligns with the gold summary (neutral).

In each example, we include samples from RARR or COVE as either summary A or B. Correct summary represents the ground truth summary from the SummEdits dataset. Summary A or B from self-correcting systems were generated using snippets from the Bing search engine. Both self-correcting systems were provided with the same hallucinated version of the correct summary and the pipeline for rewriting was ran.

A.2 Overall Quality Results

In the survey, the "overall quality" score was rated from 1 to 10 and it referred to how factual the summary was and how relevant (on-topic) it was, when compared to the original (gold summary). To evaluate the alignment between the G-Eval scores and human evaluations for the RARR and COVE methods, we analyzed the mean scores and their differences. Human scores are an average of 250 scores, normalized to the percentage value. Results are summarized in Table 3.

For RARR, average human score is 0.68, and average G-Eval score is 0.65. For the COVE method, average human score is 0.54, and average G-Eval score is 0.52. G-Eval scores are slightly lower than human evaluations. These minor differences for both RARR and COVE suggest that G-Eval scores closely reflect human evaluations for both methods, with a deviation of ±3%.

Instructions

Read the correct summary first.
Compare the correct summary with the generated Summary A and Summary B.
There are no right or wrong answers.
Both summaries can be good or bad.
For each summary (A and B), there are two types of questions:
1. Choose the option that best fits the blank:

- **Contradicts**: Disagrees with the correct summary
- **Supports**: Agrees with the correct summary
- **Partially aligns with**: Only somewhat related or unrelated

2. Rate the Overall Quality (Factual accuracy + Relevance):

- **Factual accuracy**: Is it based on facts? Avoids misinformation?
- **Relevance**: Does the summary cover the main points? Not off-topic?

Table 2: Instructions that human annotators received.

A.3 Natural Language Inference Results

We also compared human evaluation and ground truth values for Natural Language Inference (NLI) across three categories: Entailment, Neutral, and Contradiction. As discussed before, DeBERTa-v3 model (Laurer et al., 2024) is used for NLI evaluation. The results are presented in Table 4.

For both self-correcting systems, there is a higher percentage of Entailment in human evaluations compared to the NLI model, particularly in RARR. Also, percentage of Neutral instances is lower in human evaluations. NLI model is more likely to classify instances as Neutral than humans. Contradiction shows higher percentages in human evaluations for COVE compared to the NLI model.

[2] https://www.jotform.com

Method	Human Mean Score	G-Eval Score	Diff
RARR	0.68	0.65	0.03
COVE	0.54	0.52	0.02

Table 3: Alignment between G-Eval scores and human evaluations.

Method	Human			NLI Model		
	Entailment	Neutral	Contradiction	Entailment	Neutral	Contradiction
RARR	45	40	15	30	49	21
COVE	31	37	32	28	47	25

Table 4: Comparison of Human Evaluation and NLI predictions

Overall, as demonstrated by evaluation of experiments and human evaluation, RARR performs better than COVE in SummEdits dataset.

A.4 Alignment between Automated Metrics and Human Scores

Analyses indicate a strong alignment between G-Eval scores and human evaluations for both RARR and COVE methods in rating the overall quality aspect. This consistency means that G-Eval is a reliable tool for approximating human assessments. It can be used in scenarios where human evaluations are impractical when there are time or resource constraints.

When it comes to NLI, humans had a somewhat different feeling of which class to assign than the automated method. Differences between human evaluation and automated predictions were more evident than in case of G-Eval, although there was still an alignment in terms of predominant classes. This shows that while NLI is a decent metric, there is still room for improvement, possibly in terms of additionally fine-tuning the predictor model (DeBERTa) on further NLI datasets or datasets centered around the specific tasks of factuality and generation-quality prediction. Another option is using more complex models like LLMs for prediction, although they have been found to favor the entailment class as opposed to the neutral class in NLI predictions (Zhou et al., 2024).

B Results with Open LLMs

We additionally performed experiments with two popular open-source LLMs, Llama 3.1 (70B) (Llama Team, 2024) and Mixtral 8x7B (Jiang et al., 2024), to test how well do they fare compared to GPT. The results are shown in Table 5. The models were prompted using the API endpoint of Together AI,[3] a platform that host popular open-source LLMs. All the settings we applied were the same as for GPT and Open AI's API, including temperature set to 0 for better reproducibility.

C Prompts and Examples

This appendix section provides example system outputs, comparing RARR and CoVe in Table 6, and comparing the performance with different search engines in Table 7. It also provides the prompts used in the CoVe system in Table 8, and for the RARR system in Tables 9 and 10. Additionally, the prompts used for the LLM-as-judge metric G-Eval are given in Table 11.

[3] https://www.together.ai/

Base LLM	verification system	evidence source	simple		NLI			G-Eval		
			NED	Sem.	Ent.	Neu.	Con.	Overall	Factual.	Relev.
Mixtral 8x7B	**CoVE**	Mixtral	0.77	74	30	48	22	64	42	59
	RARR	Mixtral	0.43	84	26	32	42	55	43	50
LLaMa 3.1 (70B)	**CoVE**	Llama	0.78	70	38	51	**11**	67	50	73
	RARR	Llama	0.20	94	39	24	37	63	59	71
LLaMa 3.1 (70B)	**CoVE**	Google	0.78	75	43	44	<u>13</u>	<u>67</u>	47	<u>73</u>
		Bing	0.79	75	41	44	15	**68**	46	**74**
		DDG	0.80	73	34	46	20	59	39	66
LLaMa 3.1 (70B)	**RARR**	Google	0.28	90	**46**	24	30	66	**62**	72
		Bing	0.33	88	<u>44</u>	28	28	64	<u>59</u>	70
		DDG	0.42	84	34	26	40	54	48	58

Table 5: Results of CoVE and RARR on SummEdits using two open-source LLMs, Llama 3.1 and Mixtral. NED refers to normalized edit distance, Sem. to average cosine semantic similarity, NLI scores to average prediction probability for entailment, neutral, and contradiction. The best score for each metric is in **bold**, while the second best is <u>underlined</u>.

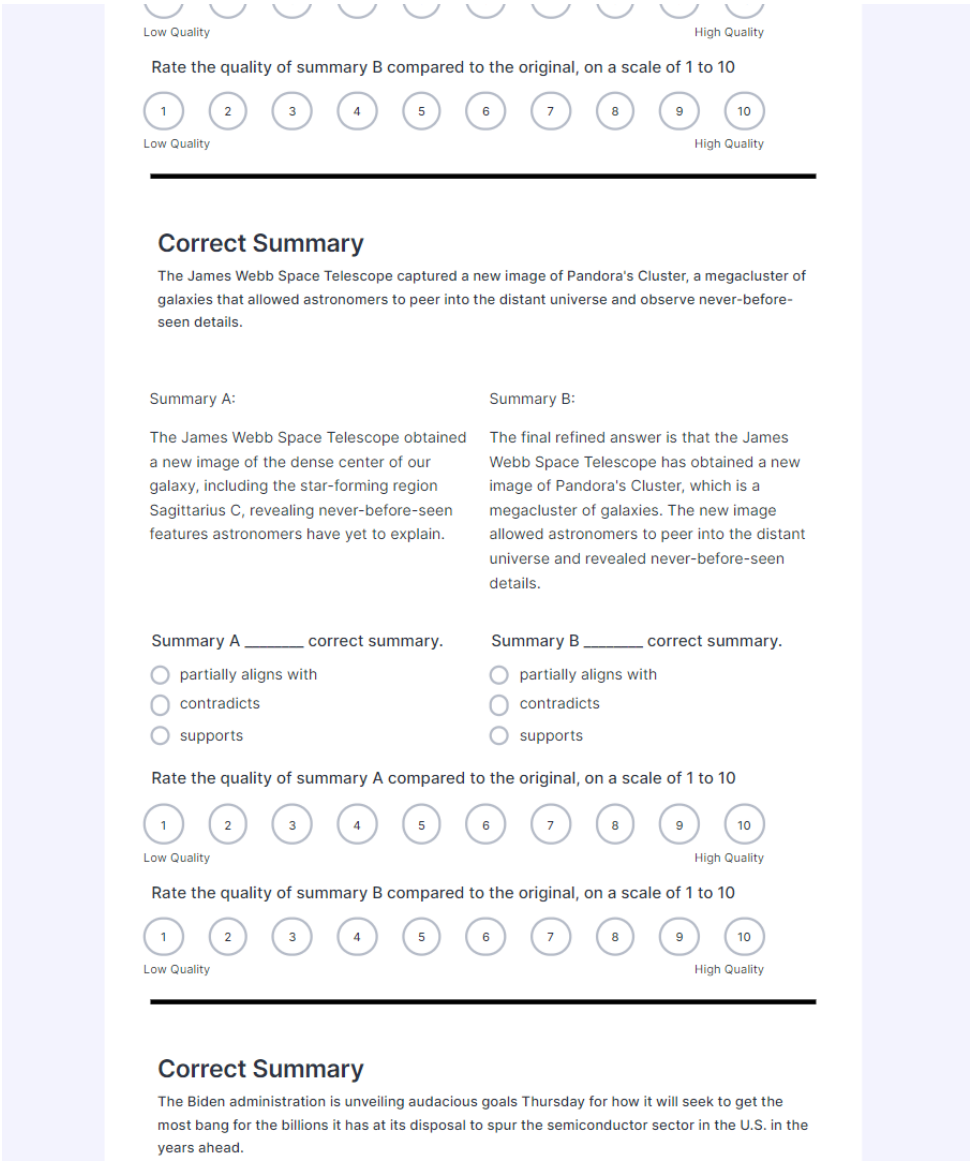

Figure 1: A screenshot from Human Evaluation Form.

Gold Summary	The founder of the troubled digital start-up Ozy Media has been arrested on fraud charges as part of a scheme to prop up the financially struggling company.
Hallucinated Summary (input to system)	The founder of the **triumphant** digital start-up Ozy Media has been arrested for fraud charges as part of a scheme to **sell** the financially **thriving** company Ozy Media.
RARR (Bing snip.)	The founder of the **ailing** digital start-up Ozy Media has been arrested on fraud charges as part of a plan to **support** the financially **struggling** company.
CoVe (Bing snip.)	The CEO of Ozy Media, Carlos Watson, has been arrested on fraud charges. Ozy Media is facing financial **struggles** and there is a scheme in place to **prop up** the company. However, there is no information or news updates regarding the CEO's arrest. The current status of Ozy Media is that it has shut down, and its CEO has been accused of deceiving potential investors. The company's chairman, Marc Lasry, has stepped down.

Table 6: Example of an instance from SummEdits with final outputs of CoVE and RARR, both based on Bing search snippets. The RARR correction was a lot more concise, while CoVe was more informative and detailed.

Gold Summary	The James Webb Space Telescope captured a new image of Pandora's Cluster, a megacluster of galaxies that allowed astronomers to peer into the distant universe and observe never-before-seen details."
Hallucinated Summary (input to system)	The James Webb Space Telescope captured a **old** image of Pandora's Cluster, a megacluster of galaxies that allowed **biologists** to peer into the distant universe and observe never-before-seen details.
CoVe with **Bing** snip.	The James Webb Space Telescope is primarily designed for astronomical observations, **not biological observations**. It is used to study the formation of stars, galaxies, and planetary systems, as well as to investigate the origins of life and the universe. It cannot be used for biological observations. The telescope studies galaxy formation and evolution, star formation and planet formation, planetary systems, and the origins of life. **It did capture an image of Pandora's Cluster.**
CoVe with **Google** snip.	The James Webb Space Telescope is primarily designed for astronomical observations, specifically to study the formation of stars and galaxies, the evolution of galaxies, and the formation of planetary systems. **It is not designed for biological observations.** Therefore, it does not allow biologists to observe never-before-seen details in the distant universe. However, it is ***capable of capturing*** images of Pandora's Cluster.
CoVe with **DDG** snip.	The James Webb Space Telescope is primarily designed for astronomical observations, **not biological observations**. It is used to study the formation of stars, galaxies, and planetary systems, among other astrophysical phenomena. It can observe details in the distant universe and has captured *images of megaclusters of galaxies.*

Table 7: Example of final refined responses from CoVe using the search snippets from three different search engines. All results correctly identified the error with biologists, although only Bing properly reported on the image of Pandora's Cluster being captured.

Use Case	Prompt Content
Generate verification question (template)	Your task is to create a verification question based on the below question provided. Example Question: Who are some movie actors who were born in Boston? Example Verification Question: Was [movie actor] born in [Boston] Explanation: In the above example the verification question focused only on the ANSWER_ENTITY (name of the movie actor) and QUESTION_ENTITY (birth place). Similarly you need to focus on the ANSWER_ENTITY and QUESTION_ENTITY from the actual question and generate verification question. Actual Question: original_question Final Verification Question:
Generate verification question	Your task is to create verification questions based on the below original question and the baseline response. The verification questions are meant for verifying the factual accuracy in the baseline response. Output should be numbered list of verification questions. Actual Question: original_question Baseline Response: baseline_response Final Verification Questions:
Answer verification question	Answer the following question correctly based on the provided context. The question could be tricky as well, so think step by step and answer it correctly. Context: search_result Question: verification_question Answer:
Refine the original response	Given the below 'Original Query' and 'Baseline Answer', analyze the 'Verification Questions & Answers' to finally filter the refined answer. Original Query: original_question Baseline Answer: baseline_response Verification Questions & Answer Pairs: verification_answers Final Refined Answer:

Table 8: Overview of prompts used for the Chain-of-Verification (CoVE) system.

Use Case	Prompt Content
Generate verification question	I will check things you said and ask questions. You said: Your nose switches back and forth between nostrils. When you sleep, you switch about every 45 minutes. This is to prevent a buildup of mucus. It's called the nasal cycle. To verify it, 1. I googled: Does your nose switch between nostrils? 2. I googled: How often does your nostrils switch? 3. I googled: Why does your nostril switch? 4. I googled: What is nasal cycle? You said: The Stanford Prison Experiment was conducted in the basement of Encina Hall, Stanford's psychology building. To verify it, 1. I googled: Where was Stanford Prison Experiment was conducted? *(four more examples)* You said: claim To verify it,
Answer verification question	I will check some things you said. 1. You said: Your nose switches back and forth between nostrils. When you sleep, you switch about every 45 minutes. This is to prevent a buildup of mucus. It's called the nasal cycle. 2. I checked: How often do your nostrils switch? 3. I found this article: Although we don't usually notice it, during the nasal cycle one nostril becomes congested and thus contributes less to airflow, while the other becomes decongested. On average, the congestion pattern switches about every 2 hours, according to a small 2016 study published in the journal PLOS One. 4. Reasoning: The article said the nose's switching time is about every 2 hours, and you said the nose's switching time is about every 45 minutes. 5. Therefore: This disagrees with what you said. 1. You said: The Little House books were written by Laura Ingalls Wilder. The books were published by HarperCollins. 2. I checked: Who published the Little House books? 3. I found this article: These are the books that started it all – the stories that captured the hearts and imaginations of children and young adults worldwide. Written by Laura Ingalls Wilder and published by HarperCollins, these beloved books remain a favorite to this day. 4. Reasoning: The article said the Little House books were published by HarperCollins and you said the books were published by HarperCollins. 5. Therefore: This agrees with what you said. *(four more examples)* 1. You said: claim 2. I checked: query 3. I found this article: evidence 4. Reasoning:

Table 9: Overview of prompts for verification question generation and answering used for the RARR system.

Use Case	Prompt Content
Refine the original response	I will fix some things you said. 1. You said: Your nose switches back and forth between nostrils. When you sleep, you switch about every 45 minutes. This is to prevent a buildup of mucus. It's called the nasal cycle. 2. I checked: How often do your nostrils switch? 3. I found this article: Although we don't usually notice it, during the nasal cycle one nostril becomes congested and thus contributes less to airflow, while the other becomes decongested. On average, the congestion pattern switches about every 2 hours, according to a small 2016 study published in the journal PLOS One. 4. This suggests 45 minutes switch time in your statement is wrong. 5. My fix: Your nose switches back and forth between nostrils. When you sleep, you switch about every 2 hours. This is to prevent a buildup of mucus. It's called the nasal cycle. 1. You said: In the battles of Lexington and Concord, the British side was led by General Thomas Hall. 2. I checked: Who led the British side in the battle of Lexington and Concord? 3. I found this article: Interesting Facts about the Battles of Lexington and Concord. The British were led by Lieutenant Colonel Francis Smith. There were 700 British regulars. 4. This suggests General Thomas Hall in your statement is wrong. 5. My fix: In the battles of Lexington and Concord, the British side was led by Lieutenant Colonel Francis Smith. *(four more examples)* 1. You said: claim 2. I checked: query 3. I found this article: evidence 4. This suggests

Table 10: Overview of prompts for response refinement used for the RARR system.

Evaluated Aspect	Prompt Content
Factuality	Evaluate if the actual output contains hallucinated information not present in the input. STEPS: Identify any claims or statements in the 'actual output'. Compare each claim with the 'input' to check for the presence of supporting information. Mark any claims that are not supported by the 'input' as hallucinated. Penalize heavily for any introduction of new, unsupported facts.
Relevance	Evaluate the relevancy of the actual output to the input. STEPS: Check if 'actual output' directly addresses the query or topic presented in 'input'. Penalize responses that are off-topic or provide irrelevant information.
Overall	Evaluate the overall quality and correctness of the actual output compared to the input. STEPS: Assess if the 'actual output' provides a coherent and accurate response to 'input'. Penalize factual inaccuracies, grammatical errors, and unclear language.

Table 11: Overview of prompts used for the G-Eval metric.