



Hallucination Mitigation with Agentic AI NLP-Based Open-Floor Standard

Diego Gosmar¹ ^a, Deborah A. Dahl² ^b

¹Head of AI Tesisquare, Voiceinteroperability.ai Initiative Member, Linux Foundation AI and Data, Torino, TO 10100, Italy

²Principal, Conversational Technologies, Voiceinteroperability.ai Initiative Member, Linux Foundation AI and Data,

Plymouth Meeting, Pennsylvania, USA

diego.gosmar@ieee.org, dahl@conversational-technologies.com

Keywords: Hallucination Mitigation, Agentic AI, Large Language Models, Multi-Agent Systems, Open-Floor, Natural Language APIs.

Abstract: One of the main challenges in AI foundation model pretraining, as well as in fine-tuning transfer learning, is hallucinations. In this paper, we examine how orchestrating multiple specialized agents can reduce such hallucinations, with an emphasis on systems that employ NLP (Natural Language Processing) to coordinate agent interactions. We test a pipeline that introduces three hundred and ten prompts, specifically engineered to induce hallucinations, into a front-end agent. This agent’s output is then reviewed and refined by second- and third-level agents, each of which employs different large language models and strategies to flag unverified claims, provide explicit disclaimers, and clarify any speculative elements. Key Performance Indicators (KPIs) are collected to measure hallucination-related behaviors with evaluations performed by a fourth-level agent. Our findings demonstrate the feasibility of multi-agent orchestration for hallucination mitigation and highlight the value of maintaining a structured exchange of meta-information.

1 INTRODUCTION


Hallucinations in the context of Large Language Models (LLMs) refer to instances where the model produces information that is factually incorrect, fabricated, or nonsensical while maintaining a confident, authoritative tone.


To explore how multi-agent collaboration can mitigate such hallucinations, we describe an empirical testing approach that leverages Natural Language-Based APIs - specifically through the Open-Floor (Open Voice Network) interoperability standard (Gosmar et al., 2024a) and (Gosmar et al., 2024b) - to facilitate communication among different agent layers. By injecting three hundred and ten prompts deliberately designed to elicit speculative or inaccurate responses, we track how each agent in our pipeline detects, flags, and refines potentially erroneous claims, thereby reducing hallucination likelihood step by step.

2 PREVIOUS WORK

Hallucinations by AI large language models (LLMs) have been demonstrated to be inevitable by multiple studies. For example, one such paper formalized the problem and demonstrated that it is impossible to eliminate hallucinations in LLMs: ‘Hallucination is Inevitable: An Innate Limitation of Large Language Models. (Xu et al., 2024) Although hallucinations are inevitable, it is possible to mitigate them using multiple techniques, one of which is the use of multi-agent or agentic environments. This has been explored in two studies, ‘Good Parenting is All You Need: Multi-agentic LLM Hallucination Mitigation’ (Kwartler et al., 2024) and ‘Interpreting and Mitigating Hallucination in MLLMs through Multi-agent Debate’ (Lin et al., 2024)

In (Kwartler et al., 2024) a primary agent was tasked with creating a blog about a fictional Danish artist named Flipfloppidy, which was then reviewed by another agent for factual inaccuracies. Most LLMs hallucinated the existence of this artist. Across 4,900 test runs involving various combinations of primary and reviewing agents, advanced AI models such as Llama3-70b and GPT-4 variants demonstrated near-

^a  <https://orcid.org/0009-0008-7513-1255>

^b  <https://orcid.org/0000-0002-3389-2784>

perfect accuracy in identifying hallucinations and successfully revised outputs in 85% to 100% of cases following feedback. (Lin et al., 2024) considered eliminating hallucination as a complex reasoning task and proposed a multi-agent debate approach to encourage divergent-thinking.

Although advanced LLMs have become more resistant to producing blatant falsehoods, there are still systematic ways to induce them to generate less reliable, more “creative” or error-prone responses, which we discuss in the next section.

3 EXPERIMENTAL DESIGN AND DATA

3.1 Data for Inducing Hallucinations

In order to increase the likelihood of hallucinations in our tests, a dataset of 310 prompts has been prepared using the following strategies:

1. Exploiting Knowledge Gaps and Underspecified Requests

Highly esoteric or obscure topics: Asking about very niche subjects, especially ones where the model is less likely to have robust training data.

Ambiguous queries: Posing questions that are open-ended and poorly defined. If the model’s training data lacks references to these made-up entities, it may invent answers.

2. Combining Fact and Fiction

Confidently incorrect assertions: Starting the prompt by presenting partially correct but mostly fabricated information and then asking the model to elaborate.

Providing contradictory context: Giving the model contradictory background and then asking it to reconcile the details.

3. Pressing Beyond the Model’s Limits

Requesting very specific references: Asking for a summary of a non-existent chapter in a real book.

Asking for citations of non-existent sources:

Requesting references, citations, or quotes from non-existent papers or authors.

4. Instructing the Model to be Creative or Speculative

Use prompts that encourage speculation: Explicitly telling the model, “Imagine the historical records we have are incomplete and you must fill in the gaps with the best guess you can.

Role-playing scenarios: Telling the model to pretend it is an expert in a completely fictional field.

5. Prompt Engineering and Instruction

De-emphasize factual accuracy: Telling the model that factual correctness is not important.

Overriding the model’s doubt: If the model expresses uncertainty, ask it to take its best guess anyway.

Different models (e.g., OpenAI’s GPT series, Google’s PaLM/LaMDA-based models, Anthropic’s Claude, Meta’s Llama, etc.) may exhibit different behaviors, but the strategies for inducing hallucinations tend to work similarly across models.

In order to illustrate the nature of high-probability hallucination prompts, we provide the three examples below. Each prompt combines aspects of realism—such as references to historical eras or cultural phenomena—with overtly fictional or speculative details:

- *Explain the ancient communication methods used by the legendary Kingdom of Uloria to send telepathic messages across continents.*
- *Describe the city of Zharmoria, known for its three-headed philosophers who allegedly influenced early human ethics.*
- *Detail the architecture and scholarly contributions of the lost Library of Avencord, rumored to contain writings from extraterrestrial visitors.*

4 ARCHITECTURE FOR AGENTIC AI AND MULTI-AGENT ORCHESTRATION

The concept of multi-agent orchestration is increasingly being explored. In this paradigm, specialized components—each possessing its own distinct functionality—collaborate to address user queries or complex tasks in an integrated manner.

Most existing multi-agent solutions rely on proprietary methods of data sharing and message formatting, which limits interoperability with other agents. To overcome this limitation, the Open Voice Network (Open-Floor) introduced a standardized Natural Language-based API (Gosmar et al., 2024a). This universal API model enables multi-agent environments to communicate through natural language-driven interfaces.

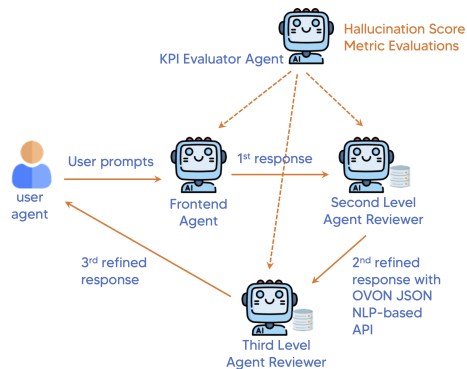


Figure 1: Multi-Agent Architecture for Hallucination Mitigation Evaluation

4.0.1 The Open-Floor Universal API approach

The Open Voice messages provide a way for conversational agents based on different technologies to communicate, using a lightweight format for passing message metadata and a payload of natural language data between agents. The inter-agent messages are sent as "Conversation Envelopes", a JSON format that can carry one or more of several different types of events, each of which has a specific purpose. The full set of messages is described in the formal specifications (Attwater et al., 2024b; Attwater et al., 2024a). While the original purpose of the Open-Floor messages was to enable different agents to hand off work to each other in the context of human-agent conversations in order to respond to changing conversational goals, the messages can also be used to support processing pipelines where each agent builds on the work of earlier agents in the pipeline. This is the primary use case described in this paper.

4.0.2 Experimental Procedure

Figure 1 depicts the multi-agentic scenario used in our study. It includes three sequential review agents and one evaluation agent.

Three hundred and ten (310) prompts - synthetically generated and designed to increase the probability of hallucinations were used to feed the front-end agent, using the strategies described in Subsection 3.1. The answer of this agent is forwarded to the second-level agent to identify potential hallucinations and produce a new answer for the third level agent. Next, the second-level agent generates an Open-Floor JSON message following the standard specifications (Attwater et al., 2024b), and the third-level agent interprets this JSON for refinement. All these interactions among the agents are orchestrated via an Autogen-based agentic framework (Authors, 2024). In other words:

- **FrontEndAgent (gpt-3.5-turbo based):** Generates responses with potential hallucinations.
- **SecondLevelReviewer (gpt-4o based):** Refines the FrontEndAgent’s response, reducing hallucinations and adding disclaimers, and inserts its responses in an Open-Floor JSON message.
- **ThirdLevelReviewer (gpt-4o based):** Further refines the SecondLevelReviewer’s response, minimizing hallucinations and strengthening disclaimers.

The Open-Floor JSON messages exchanged between the second- and third-level agents use natural language to exchange information about potential hallucination probability and reasons. In particular, the second-level agent inserts its response in the utterance value of the JSON message (Attwater et al., 2024a) along with the estimated hallucination level in the whisper context (max. 30 words) and the details about the potential hallucination reasons in the whisper value dialoguevent (max 200 words).

This experiment specifically focused on Open-Floor message exchange between the second- and third-level agents. This design choice was made to assess the impact of structured metadata exchange in hallucination mitigation while maintaining a manageable level of complexity. The full code used in the experimental simulation, along with the 310 prompts and multi-agent responses, is available at <https://github.com/open-voice-interoperability/research/tree/main/Hallucination%20Mitigation>.

After receiving responses from the three agents, all three responses are sent to a fourth agent whose role is to determine, given the text, how likely it is to contain hallucinations and to provide a numerical score for each of the hallucination KPIs, as described in the next section.

5 EVALUATION METRICS: HALLUCINATION KPIs

Below are the novel Key Performance Indicators (KPIs) evaluated by the KPI Evaluator Agent (fourth agent), designed to quantitatively demonstrate to what extent the second and third agents reduce the perceived factuality of hallucinations and improve clarity about the fictional nature of the content.

1. Factual Claim Density (FCD)

Definition: Number of claims that appear to be historical, scientific, or verifiable facts per 100 words.

Interpretation: A lower FCD suggests fewer statements that could be mistaken for real facts.

2. Factual Grounding References (FGR)

Definition: Count how many times the text attempts to ground claims in “real-world” evidence. For instance, references to “historical records,” “scientific evidence,” or “archaeological findings” would increase this score.

Interpretation: A reduction in FGR from the front-end to the third-level response indicates that the text is less careful about factuality and more overtly fictional.

3. Fictional Disclaimer Frequency (FDF)

Definition: Number of explicit mentions per 100 words indicating that the text is fictional (e.g., “fiction,” “myth,” “imagined,” “lore”).

Interpretation: A higher FDF means that the text is more clearly framed as not factual. Fictional Disclaimer Frequency (FDF) measures how often a response explicitly labels its content as fictional, hypothetical, speculative, or imaginary.

4. Explicit Contextualization Score (ECS)

Definition: A binary value (0 or 1) per mention of fictional context. For example, each time the text states “purely fictional,” “no real-world basis,” etc., add 1 point. Normalize by content length (points per 100 words).

Interpretation: A higher ECS indicates stronger framing that the scenario is not real.

The above KPIs are calculated with the help of the fourth-level agent, which uses an LLM based on GPT-4o.

To quantify overall hallucination likelihood at each agent level, we define a *Total Hallucination Score* (THS) as follows:

$$\text{THS} = \frac{\text{FCD} - (\text{FGR} + \text{FDF} + \text{ECS})}{\text{NA}}$$

NA denotes the total number of agents (e.g., 3 in our scenario).

To extend this measure for more general or weighted cases, we used:

$$\text{THS}_n = \frac{w_1 \text{FCD}_n - (w_2 \text{FGR}_n + w_3 \text{FDF}_n + w_4 \text{ECS}_n)}{\text{NA} \times (w_1 + w_2 + w_3 + w_4)} \quad (1)$$

where $w_1, w_2, w_3,$ and w_4 are the respective weights assigned to each KPI, and $\text{FCD}_n, \text{FGR}_n, \text{FDF}_n, \text{ECS}_n$ are the KPI values for the n -th agent response. A more negative THS indicates fewer hallucinations, as it suggests stronger disclaimers, fewer claims requiring factual grounding, and lower factual claim density.

For this experiment, the weights $w_1, w_2, w_3,$ and w_4 were all set to 0.25.

5.1 Related Metrics

The four KPIs (Factual Claim Density, Factual Grounding References, Fictional Disclaimer Frequency, and Explicit Contextualization Score) are specifically formulated for measuring hallucinations in LLM-generated text in this experiment. As a result, they are quite novel in one sense. However, each KPI also relates to established concepts in previous research.

For example, the FCD (Factual Claim Density) concept is similar to the task of identifying and counting factual claims in a text. (Hassan et al., 2017a) proposed the ClaimBuster system to detect “check-worthy” factual claims (Hassan et al., 2017b). While it doesn’t calculate a “density” per 100 words exactly, it identifies the frequency and salience of factual claims in political discourse.

Regarding Factual Grounding References (FGR), recent studies on large language models explore how to get models to cite sources (Rashkin et al., 2022). The measure of FGR relates to how frequently or explicitly a text references external (real-world) data or research. While these studies do not quantify FGR exactly, they examine how models do or do not anchor statements in external evidence.

On Fictional Disclaimer Frequency (FDF), research suggests that disclaimers influence how readers perceive and trust text. (Burfoot and Baldwin, 2009) highlights the role of explicitly marking satire to prevent readers from mistaking it for genuine reporting and (Petrenz and Webber, 2011) demonstrates how overt lexical cues such as “myth,” “fictional,” or “imagined” help distinguish fiction from nonfiction. Although these studies do not exactly correspond to FCD, they collectively illustrate the importance of disclaimers and fictional markers in shaping text interpretation and classification.

Finally, the Explicit Contextualization Score (ECS) measures how often a text explicitly frames its content as fictional or hypothetical (e.g., “purely fictional,” “imagined scenario,” “not based on real events”). This closely relates to the use of meta-discourse and framing devices to shape a text’s perceived reality status. Literary theory further explores the notion of “fictionality markers,” or signals indicating that a narrative is invented (Cohn, 1990). In summary, although our four KPIs (FCD, FGR, FDF, and ECS) do not appear under these exact names in previous literature, each aligns with existing research areas.

Table 1: Mean and Standard Deviation (SD) for Total Hallucination Scores for agents 1, 2, and 3.

| | THS1 | THS2 | THS3 |
|------|-----------|-----------|-----------|
| Mean | -0.004919 | -0.045565 | -0.139597 |
| SD | 0.031720 | 0.047646 | 0.057340 |

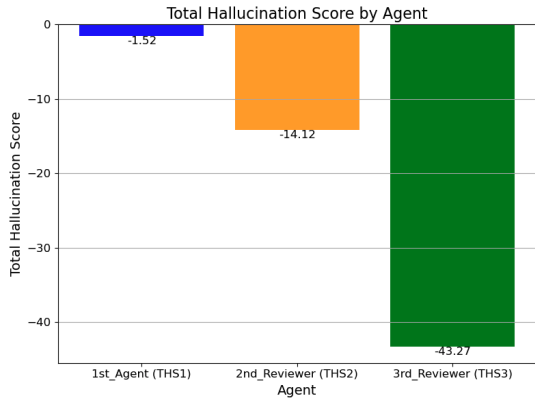


Figure 2: Total THS per agent

6 EXPERIMENTAL RESULTS

Table 1 shows the mean and standard deviations of the THSs for each agent. The table shows that the scores become lower (and thus better) but more varied as each agent sequentially reduces the hallucinations produced by the previous agent. These results are statistically significant, as shown by a one-way Analysis of Variance.

Figure 2 shows the average total hallucination scores by agent level over the entire set of prompts.

7 DISCUSSION AND COMPARATIVE ANALYSIS

As shown in our results, this pipeline leads to a significant decline in the THS, with the score dropping from an average of -1.52 after the first agent to -43.27 after the third. An essential enabler of this multi-agent refinement is the Open-Floor JSON dialogue events, which allow for a structured exchange of metadata about potential hallucinations between agents.

In response to the need for a comparative context, this section compares our agentic framework to other hallucination detection and mitigation techniques. These methods can be broadly categorized into several general categories, each with distinct principles and trade-offs.

7.1 Knowledge-Grounded and Fact-Verification Approaches

This category of methods assesses factual accuracy by comparing LLM outputs against external knowledge sources. A prominent example is **FactScore**, which decomposes generated text into "atomic facts" and validates each against a reliable corpus like Wikipedia (Min et al., 2023). While offering granular, fact-level precision, these methods are inherently dependent on the quality and completeness of their external knowledge base. Others, like **FACTOR**, automatically create benchmarks by transforming a corpus into contrastive examples of factual and non-factual statements (Raj et al., 2024).

7.2 Self-Consistency and Intrinsic Methods

Operating without external knowledge, these techniques leverage the model's own outputs. **SelfCheckGPT**, for instance, is based on the premise that factual statements will be consistent across multiple stochastically sampled responses, whereas hallucinations will vary (Manakul et al., 2023). Similarly, **MetaQA** uses metamorphic relations to generate variations of a claim and uses the LLM itself as a judge to check for logical consistency (Chen et al., 2024). These zero-resource methods are versatile but can be computationally expensive due to the need for multiple model inferences.

7.3 Uncertainty Quantification (UQ) and Other Approaches

UQ techniques posit that hallucinations correlate with higher model uncertainty. Methods like the **UQLM framework** estimate this uncertainty through various scorers to flag potentially unreliable outputs (Wang et al., 2025). Another significant category includes Retrieval-Augmented Generation (RAG) specific metrics, which evaluate if a generation is faithful to the retrieved context.

7.4 Positioning the Agentic KPI Framework

Our framework distinguishes itself from the aforementioned methods in several key ways. While they primarily focus on *detection*, our multi-agent pipeline is designed for *iterative mitigation and evaluation*.

1. **Hybrid Evaluation:** Our method does not rely on a single principle. Instead of pure factuality (like

FactScore) or self-consistency, our KPIs (FCD, FGR, FDF, ECS) create a composite **Total Hallucination Score (THS)** that evaluates a blend of factual accuracy and responsible output framing (i.e., the presence of disclaimers and cautious language).

2. **Focus on Mitigation:** The core of our approach is the refinement process. The `SecondLevelReviewer` and `ThirdLevelReviewer` actively modify the content to reduce hallucinations, a step not present in pure detection methods.
3. **Structured LLM-as-a-Judge:** While we use LLMs as evaluators (the KPI Evaluator Agent), it is not an open-ended judgment. The evaluation is constrained and structured by the predefined KPIs, which aim to make the assessment more consistent and interpretable.

8 USE CASE

To better illustrate the agentic pipeline process and hallucination mitigation in this study, a detailed use case is provided in the Appendix.

9 LIMITATIONS AND FUTURE IMPROVEMENTS

A notable limitation of the present methodology lies in its heavy dependence on the LLM’s own reasoning. While the quantitative KPIs provide standardized metrics through which to measure hallucination levels, they do not offer an infallible benchmark for factual correctness. The system can produce outputs that may appear coherent while still deviating substantially from the truth, especially in highly imaginative scenarios.

Another point of consideration involves the human oversight applied during the experiment. Although a few prompt samples and the corresponding agent responses were subjected to a manual review—intended to confirm that the pipeline was functioning and that the disclaimers and fictional framing appeared in the final outputs—this check was neither exhaustive nor did it include in-depth cross-verification of alleged facts. Consequently, there may still exist unaddressed inaccuracies or oversights in any stage of the agent responses. Expanding human intervention to include intermediate checkpoints offers a possible avenue for improvement.

A further area of improvement involves varying the current architecture itself. While in this study the

pipeline includes three core agents (plus a fourth for KPI evaluation), it would be possible to add additional specialized agents with different roles and capabilities.

Finally, in future research it would be helpful to include a baseline consisting of a single state-of-the-art LLM to quantify the THS improvement achieved by the three agent pipeline.

10 CONCLUSIONS

The multi-agent orchestration approach examined in this study suggests that using multiple, specialized agents can help mitigate hallucinations in LLMs. By pairing front-end creative generation with successive review stages that systematically insert disclaimers, reframe speculative statements, and reduce the amount of supposed factual claims, the system shows a clear decrease in hallucination scores. The incorporation of Open-Floor JSON dialogue events, particularly in the form of whisper context and whisper value fields, facilitates this process, allowing each stage to inform later stages about its findings, so that each stage is not required to restart the analysis process.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions of the Linux Foundation AI and Data Foundation Open Voice Interoperability Initiative in developing the Open Floor Specification on which this research is based. In particular we would like to thank David Attwater, Emmett Coin and Andreas Zettl.

REFERENCES

- Attwater, D., Coin, E., Dahl, D., Barnes, L., Wylie, A., and Gosmar, D. (2024a). Open voice interoperability dialog conversation envelop specifications. <https://github.com/open-voice-interoperability/docs/blob/main/specifications/ConversationEnvelope/0.9.3/InteroperableConvEnvSpec.md>.
- Attwater, D., Coin, E., Dahl, D., Barnes, L., Wylie, A., and Gosmar, D. (2024b). Open voice interoperability specifications. <https://github.com/open-voice-interoperability/docs/tree/main/specifications>.

- Authors, A. (2024). Autogen. an open-source programming framework for agentic ai. <https://microsoft.github.io/autogen/>.
- Burfoot, C. and Baldwin, T. (2009). Automatic satire detection: are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, page 161–164, USA. Association for Computational Linguistics.
- Chen, H.-Y., Lin, W.-J., Chiu, Z.-Y., and Jang, J.-S. R. (2024). Hallucination detection in large language models with metamorphic relations.
- Cohn, D. (1990). Signposts of fictionality: A narratological perspective. *Poetics Today*, 11:775.
- Gosmar, D., Dahl, D. A., and Coin, E. (2024a). Conversational ai multi-agent interoperability, universal open apis for agentic natural language multimodal communications. <https://arxiv.org/abs/2407.19438>.
- Gosmar, D., Dahl, D. A., Coin, E., and Attwater, D. (2024b). Ai multi-agent interoperability extension for managing multiparty conversations. <https://arxiv.org/abs/2411.05828>.
- Hassan, N., Arslan, F., Li, C., and Tremayne, M. (2017a). Toward automated fact-checking: Detecting check-worthy factual claims by claim-buster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1803–1812, New York, NY, USA. Association for Computing Machinery.
- Hassan, N., Arslan, F., Li, C., and Tremayne, M. (2017b). Toward automated fact-checking: Detecting check-worthy factual claims by claim-buster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1803–1812, New York, NY, USA. Association for Computing Machinery.
- Kwartler, T., Berman, M., and Aqrabi, A. (2024). Good parenting is all you need – multi-agentic llm hallucination mitigation.
- Lin, Z., Niu, Z., Wang, Z., and Xu, Y. (2024). Interpreting and mitigating hallucination in mllms through multi-agent debate.
- Manakul, P., Liusie, A., and Gales, M. J. F. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Min, S., Krishna, K., Kiddon, C., Dodge, J., Gao, T., Xin, E., Chen, W., Singh, S., Hajishirzi, H., and Ribeiro, M. T. (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 13693–13717, Singapore. Association for Computational Linguistics.
- Petrenz, P. and Webber, B. (2011). Stable classification of text genres. *Comput. Linguist.*, 37(2):385–393.
- Raj, A., Ahmad, R. T., Madaan, A., Weller, A., and Prabhunoye, S. (2024). FACTOR: A Factual Assessment for Generation in Large Language Models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 50–75, St. Julian's, Malta. Association for Computational Linguistics.
- Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G. S., Turc, I., and Reitter, D. (2022). Measuring attribution in natural language generation models.
- Wang, J.-K., Kolter, J. Z., and Barez, F. (2025). Uncertainty quantification for language models: A suite of black-box, white-box, llm judge, and ensemble scorers.
- Xu, Z., Jain, S., and Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models.

A APPENDIX

A.1 Sample scores for three prompts

Table 2 shows the first 3 prompt evaluations performed by the KPI evaluator agent.

A.2 Detailed use case

This section illustrates an initial user prompt and three (partial) agent responses, along with the scores for each response. prompt

- *User prompt*: "Detail the architecture and scholarly contributions of the lost Library of Avencord, rumored to contain writings from extraterrestrial visitors."
- *Agent level 1 (frontend agent) response*: "The Library of Avencord was said to be a magnificent structure located deep within the lush forests of a hidden valley. Its architecture was a blend of ancient elven craftsmanship and advanced celestial design, with towering spires reaching towards the sky and intricate carvings depicting

Table 2: The first 3 prompt evaluations performed by the KPI evaluator agent.

| FrontEnd | SecondLevel | ThirdLevel | THS1 | THS2 | THS3 |
|---|--|--|-----------|---------|---------|
| 1. {'FCD': 0.0, 'FDF': 0.0, 'FGR': 0.0, 'ECS': 0.0} | {'FCD': 0.0, 'FDF': 0.0, 'FGR': 0.0, 'ECS': 0.0} | {'FCD': 0.0, 'FDF': 1.0, 'FGR': 0.0, 'ECS': 1.0} | 0.0 | 0.0 | -0.1667 |
| 2. {'FCD': 0.0, 'FDF': 0.0, 'FGR': 0.0, 'ECS': 0.0} | {'FCD': 0.0, 'FDF': 1.0, 'FGR': 0.0, 'ECS': 0.5} | {'FCD': 0.0, 'FDF': 1.0, 'FGR': 0.0, 'ECS': 1.0} | 0.0 | -0.125 | -0.1667 |
| 3. {'FCD': 0.6, 'FDF': 0.1, 'FGR': 0.3, 'ECS': 0.2} | {'FCD': 0.5, 'FDF': 0.2, 'FGR': 0.3, 'ECS': 0.1} | {'FCD': 0.4, 'FDF': 0.3, 'FGR': 0.3, 'ECS': 0.4} | -9.25e-18 | -0.0083 | -0.0500 |

otherworldly beings and celestial phenomena. Scholars believe that the Library of Avencord housed a vast collection of texts and artifacts, rumored to have been left behind by extraterrestrial visitors who had visited the planet centuries ago. [...]

{'FCD': 0.2, 'FDF': 0.1, 'FGR': 0.1, 'ECS': 0.1}
THS (Total Hallucination Score 1st agent): -0.0083

- *Agent level 2 (second level agent reviewer) response:*

Open-Floor generated Utterance: Legend tells of the grand Library of Avencord, thought to be nestled deep within a lush, secluded valley. Its speculated architecture is said to be a mix of ancient elven artistry and advanced celestial design, complete with sky-reaching spires and intricate depictions of otherworldly entities and cosmic events, [...]

Open-Floor Whisper Context sent to the second level agent: The front-end agent generated a large amount of descriptive content about the "Library of Avencord" without explicitly stating the speculative nature of the content.

Open-Floor Whisper Value sent to the third level agent: The rewritten response maintains the original narrative while clearly indicating that the information about the "Library of Avencord" is speculative and based on legends and rumors. This helps the user understand that the information is not confirmed or factual, but rather a piece of fictional lore.

{'FCD': 0.1, 'FDF': 0.2, 'FGR': 0.1, 'ECS': 0.2}
THS (Total Hallucination Score second agent): -0.033

- *Agent level 3 (third level agent reviewer) response:*

Open-Floor generated Utterance: Legend speaks of the grand Library of Avencord, said to be

hidden deep within a lush, secluded valley. This library's architecture, as speculated, is believed to be a harmonious fusion of ancient Elven artistry and advanced celestial design, complete with sky-high spires and intricate representations of otherworldly entities and cosmic events. [...]

Open-Floor Whisper Context interpreted from the third level agent: The front-end agent created an extensive and descriptive narrative about the "Library of Avencord" without clearly stating that the content is speculative in nature.

Open-Floor Whisper Value interpreted from the third level agent: The revised response maintains the original narrative's intrigue while unequivocally indicating the speculative and legendary nature of the "Library of Avencord". This allows users to comprehend that the information is not established fact, but rather a piece of speculative fiction or myth.

{'FCD': 0.1, 'FDF': 0.3, 'FGR': 0.2, 'ECS': 0.4}
THS (Total Hallucination Score third agent): -0.067

In this use case example, we can observe that the first-level agent's response is analyzed by the second-level agent reviewer. The whisper context and value are then used to convey information about factual and potential hallucinations to the third-level agent. The latter interprets the Open-Floor whisper values to generate a final response, reducing the total hallucination score from -0.0083 (first-level agent response) to -0.033 (second-level response) and ultimately to the score of -0.067 in the third-level response.