MARiA at SemEval 2024 Task-6: Hallucination Detection Through LLMs, MNLI, and Cosine similarity

Reza Sanayei* **Abhyuday Singh*** **MohammadHossein Rezaei** **Steven Bethard**
University of Arizona
rsanayei,abhyudaysingh,mhrezaei,bethard@arizona.edu

Abstract

The advent of large language models (LLMs) has revolutionized Natural Language Generation (NLG), offering unmatched text generation capabilities. However, this progress introduces significant challenges, notably hallucinations—semantically incorrect yet fluent outputs. This phenomenon undermines content reliability, as traditional detection systems focus more on fluency than accuracy, posing a risk of misinformation spread.

Our study addresses these issues by proposing a unified strategy for detecting hallucinations in neural model-generated text, focusing on the SHROOM task in SemEval 2024. We employ diverse methodologies to identify output divergence from the source content. We utilized Sentence Transformers to measure cosine similarity between source-hypothesis and source-target embeddings, experimented with omitting source content in the cosine similarity computations, and Leveragied LLMs' In-Context Learning with detailed task prompts as our methodologies. The varying performance of our different approaches across the subtasks underscores the complexity of Natural Language Understanding tasks, highlighting the importance of addressing the nuances of semantic correctness in the era of advanced language models.

1 Introduction

The SHROOM task (Mickus et al., 2024) aims to address the challenge of detecting grammatically sound outputs containing incorrect semantic information in NLG systems. This task is crucial due to the prevalent issue of neural models producing fluent but inaccurate outputs, referred to as "hallucinations" (Maynez et al., 2020). Given the critical importance of correctness in NLG applications, SHROOM aims to foster interest in automating the detection of these hallucinations. Participants were tasked with the binary identification of such hallucinations across different NLG tasks, including Definition Modeling (DM), Machine Translation (MT), and Paraphrase Generation (PG).

Our system leveraged three distinct approaches to tackle the SHROOM task: A baseline cosine similarity, MultiNLI classification (Williams et al., 2018), and Large Language Models (LLMs), specifically Mixtral-8x7B-Instruct (Jiang et al., 2024). Each approach was tailored to identify hallucinations in NLG outputs by comparing them with the source input and detecting inconsistencies in semantic information. Through various combinations of these approaches, we aimed to accurately identify grammatically sound but incorrect outputs generated by neural models.

2 Background and Related Work

Previous research efforts have attempted to detect and control (Filippova, 2020) hallucinations. Dziri et al. (2022) worked on the origins of hallucinations, concluding that $> 60\%$ of the standard benchmarks consist hallucinated responses. Xiao and Wang (2021) proposed a simple extension to beam search to reduce hallucination. Obaid ul Islam et al. (2023) proposed a natural language inference (NLI) based method to preprocess the training data to reduce hallucinations.

The most similar to our work, Guerreiro et al. (2023) studied hallucinations in Neural Machine Translation. They analyzed multiple methods to detect hallucinations and developed `DeHallucinator` which overwrites the translation detected as a hallucination with a better one.

3 Dataset

In this section, we provide a detailed overview of the SHROOM dataset, highlighting its composition, structure, and associated challenges.

* Equal Contribution.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1584–1588

3.1 Composition

The SHROOM dataset consists of two main tracks: model-aware and model-agnostic, encompassing three subtasks: paraphrase generation, machine translation, and definition modeling. The test dataset comprises 3000 objects, with 1500 belonging to each of the model-aware and model-agnostic tracks. Additionally, the development data consists of 500 objects for each track, while the trial data comprises 80 objects. The unlabeled training data comprises 5000 objects for both the model-aware and model-agnostic tracks.

3.2 JSON Object Structure

Each dataset object contains the following components:

Task Description: Indicates the subtask to which the object belongs. Source (src): Input passed to be processed by the NLP model. Target (tgt): Intended correct processed "gold" text. Hypothesis (hyp): Actual output produced by the NLP model. Reference (ref): Specifies whether the reference includes the source, target, both, or neither. Labels: Each object is labeled by five human annotators as hallucination or not hallucination. Probability of Hallucination (p(hallucination)): Represents the probability of the hypothesis being a hallucination, ranging from 0.0 to 1.0. This probability is determined based on the consensus of the annotators. Label: Indicates the majority vote among the annotators, labeling the object as hallucination or not hallucination.

3.3 Issues with the Dataset

Several challenges were encountered while working with the SHROOM dataset:

Unlabeled Training Data: The unlabeled nature of the training data posed challenges, limiting the applicability of certain approaches and requiring alternative strategies for model training. **Format Discrepancy in Definition Modeling Task:** The test data for the definition modeling task deviated from the format of the development data, missing the <define> tag and presenting the definition as a question at the end. This inconsistency caused issues in several approaches and led to hallucinations in the Large Language Model (LLM) approach. **Imbalance in Language Representation:** The machine translation subtask lacked a balanced representation of multiple languages, potentially skewing the evaluation results and posing challenges for system development.

4 System Overview

In this section, we provide an overview of the three approaches employed in our system to address the SHROOM task.

4.1 Baseline Approach - Cosine Similarity

Our baseline approach utilizes cosine similarity to compare embeddings derived from source-hypothesis and source-target pairs. We employ Sentence Transformers (Reimers and Gurevych, 2019) to compute the cosine similarity, facilitating the comparison between the generated hypothesis and the target output. This approach serves as the foundation upon which subsequent refinements are built.

4.2 Approach 2 - MNLI Classification

In this approach, we leveraged the MultiNLI (MNLI) dataset for classification and similarity comparison between hypothesis and target outputs. We utilized the bart-large-mnli model, which is pre-trained on MNLI, to predict the entailment relationship between the hypothesis and target, subsequently examining similarity, and predict hallucination.

4.3 Approach 3 - Large Language Models

Leveraging LLMs, we prompt-engineered instructions for each subtask, utilizing the In-Context Learning power of these models, specifically Mixtral-8x7B-Instruct model, to detect hallucinations. We use the model for inference and experiment with temperature adjustments to optimize performance. This approach capitalizes on the contextual understanding and generative capabilities of LLMs to accurately identify hallucinations in NLG outputs.

5 Experimental Setup

In this section, we outline the experimental setup used for evaluating our system's performance on the SHROOM task.

5.1 Data Splits

We utilized the provided development set extensively for experimentation, as the training set was unlabeled. This allowed us to iteratively refine our approaches before selecting the final submissions for the test set evaluation.

Task	Hypothesis	Reference	Source	Target	Model	Labels	Label	P(H)
DM	(linguistics) The study of the relationships between words and their meanings.	Target	The <define> metaontology </define> debate has now migrated from discussions of composition.	The ontology of ontology.	-	H/N/N/N	N	0.4
PG	When did you see him?	Either	When did you last see him?	When was the last time you saw him?	tuner007/ pegasus_paraphrase	N/N/N	N	0.0
MT	It uses a giant rocket over 100 feet high to launch a satellite or telescope into space.	-	Ngini mamakai roket raksasa malabihi 100 kaki tingginya gasan maandak satelit atawa teleskop ka luar angkasa.	It takes a giant rocket over a 100 feet high to put a satellite or telescope in space.	-	N/N/N/H/N	N	0.4

Table 1: Examples from the dataset. The dataset includes three subtasks: Definition Modeling (DM), Paraphrase Generation (PG), and Machine Translation (MT). The dataset is labeled by crowdworkers as Hallucination (H) and Not Hallucination (N). P(Hallucination) indicates the probability of the hallucination based on the labels.

5.2 Preprocessing and Model Selection

Minimal preprocessing was applied to the data. Furthermore, we did not create separate distinctions for the model-aware and model-agnostic subtasks. Our decision was driven by the belief that a unified, model-agnostic solution would be the most optimal approach for addressing the task. For cosine similarity, we employed Sentence Transformers. MNLI classification utilized the Facebook bart-large-mnli model. In the case of LLMs, we initially employed Mixtral-8x7B-Instruct model and conducted an additional run post-evaluation with some changes to model settings like output token size and temperature, processing queries in batches.

5.3 Evaluation Measures

The evaluation measures used in the task primarily revolved around accuracy percentages. We assessed the accuracy of our models in correctly identifying grammatically sound outputs containing incorrect or unsupported semantic information, inconsistent with the source input. This metric served as the primary indicator of our system's performance on the SHROOM task.

This Experimental Setup section provides essential details about our methodology and the specifics of our experimental setup, enabling reproducibility and facilitating a clear understanding of our system's performance on the SHROOM task.

6 Results

In this section, we present the quantitative analysis of our system's performance on the SHROOM task. We evaluated the performance of our system approaches on the test data for each of the three subtasks: Paraphrase Generation (PG), Machine Translation (MT), and Definition Modeling (DM). Our system comprises three distinct approaches: cosine similarity, MNLI classification, and Large Language Models (LLMs), specifically Mixtral. **We note that Mixtral is the only system submitted to the task, and other results are post-evaluation experiments.**

6.1 Model-Agnostic Setting

Table 2 provides the accuracy of model-agnostic setting. We observe that our cosine similarity approach achieved the highest accuracy, with 70.3% overall accuracy. Specifically, it performed well in PG (77.9%) and MT (75.8%) subtasks. However, the Mixtral approach yielded lower accuracy at 50.5%, with varying performance across subtasks: DM (48.4%), PG (50.1%), and MT (52.9%). After changing the settings (temperature) the accuracy improved to 60.2% (Mixtral*).

6.2 Model-Aware Setting

Table 3 provides the accuracy of model-aware setting. In the model-aware setting, the MNLI classification approach achieved the highest accuracy at

	Cosine	MNLI	Mixtral	Mixtral*
Accuracy	62.1	65.13	49.8	56.1
DM	60.4		48.3	53.5
PG	57.6		48.2	56.1
MT	66.7		52.3	58.7

Table 2: Results on model-agnostic setting. We report accuracy for all the instances and accuracy on each subtask.

	Cosine	Mixtral	Mixtral*
Accuracy	70.3	50.5	60.2
DM	59.8	48.4	56.8
PG	77.9	50.1	61.9
MT	75.8	52.9	62.2

Table 3: Results on model-aware setting. We report accuracy for all the instances and accuracy on each subtask.

65.13%, followed by the cosine similarity approach at 62.1%. The MNLI approach showed consistent performance across subtasks, while the cosine similarity approach performed particularly well in MT (66.7%). The Mixtral approach had the lowest accuracy at 49.8%, with varying performance across subtasks: DM (53.5%), PG (56.1%), and MT (58.7%). After changing the settings— (temperature) the accuracy improved to 56.1% (Mixtral*).

7 Conclusion

In conclusion, our system for addressing the SHROOM task employed three distinct approaches: baseline cosine similarity, MNLI classification, and Mixtral. Each approach was carefully designed to tackle the challenge of identifying hallucinations in natural language generation outputs.

Our experimental results demonstrated varying degrees of success across the different subtasks. While cosine similarity and MNLI classification showed promising performance, leveraging LLMs proved to be particularly effective in accurately identifying hallucinations.

Looking forward, our system's performance suggests several avenues for future work. Firstly, further exploration and refinement of each approach tailored to the specific subtleties of each subtask could potentially yield improved performance. Additionally, investigating ensemble methods or hybrid approaches that combine the strengths of different techniques may enhance overall system robustness.

Despite the challenges encountered, our system's competitive performance in the SHROOM task underscores the importance of automated, multi-expert, mechanisms for detecting and mitigating hallucinations in NLG systems. As the field continues to evolve, addressing these challenges will be crucial for advancing the reliability and accuracy of NLG applications.

In summary, our system represents a significant step towards addressing the complexities of hallucination detection in NLG outputs, and we are optimistic about the potential for future advancements in this area.

References

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, Online. Association for Computational Linguistics.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Timothee Mickus, Elaine Zosa, Raúl Vázquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024.

SemEval-2024 Task 6: SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Saad Obaid ul Islam, Iza Škrjanec, Ondrej Dusek, and Vera Demberg. 2023. Tackling hallucinations in neural chart summarization. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 414–423, Prague, Czechia. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.