NootNoot At SemEval-2024 Task 6: Hallucinations and Related Observable Overgeneration Mistakes Detection

Sankalp Bahad[1]
IIIT Hyderabad
sankalp.bahad@research.iiit.ac.in

Yash Bhaskar[1]
IIIT Hyderabad
yash.bhaskar@research.iiit.ac.in

Parameswari Krishnamurthy[2]
IIIT Hyderabad
param.krishna@iiit.ac.in

Abstract

Semantic hallucinations in neural language generation systems pose a significant challenge to the reliability and accuracy of natural language processing applications. Current neural models often produce fluent but incorrect outputs, undermining the usefulness of generated text. In this study, we address the task of detecting semantic hallucinations through the SHROOM (Semantic Hallucinations Real Or Mistakes) dataset, encompassing data from diverse NLG tasks such as definition modeling, machine translation, and paraphrase generation. We investigate three methodologies: fine-tuning on labelled training data, fine-tuning on labelled validation data, and a zero-shot approach using the Mixtral 8x7b instruct model. Our results demonstrate the effectiveness of these methodologies in identifying semantic hallucinations, with the zero-shot approach showing competitive performance without additional training. Our findings highlight the importance of robust detection mechanisms for ensuring the accuracy and reliability of neural language generation systems.

1 Introduction

The modern NLG landscape is plagued by two interlinked problems: On the one hand, our current neural models have a propensity to produce inaccurate but fluent outputs; on the other hand, our metrics are most apt at describing fluency, rather than correctness. This leads neural networks to "hallucinate", i.e., produce fluent but incorrect outputs that we currently struggle to detect automatically. For many NLG applications, the correctness of an output is however mission critical. For instance, producing a plausible-sounding translation that is inconsistent with the source text puts in jeopardy the usefulness of a machine translation pipeline. With our shared task, we hope to foster the growing interest in this topic in the community.

With SHROOM (Mickus et al., 2024) we adopt a post hoc setting, where models have already been trained and outputs already produced: participants will be asked to perform binary classification to identify cases of fluent overgeneration hallucinations in two different setups: model-aware and model-agnostic tracks. That is, participants must detect grammatically sound outputs which contain incorrect or unsupported semantic information, inconsistent with the source input, with or without having access to the model that produced the output. To that end, we will provide participants with a collection of checkpoints, inputs, references and outputs of systems covering three different NLG tasks: definition modeling (DM), machine translation (MT) and paraphrase generation (PG), trained with varying degrees of accuracy. The development set will provide binary annotations from at least five different annotators and a majority vote gold label.

2 Dataset

SHROOM (Semantic Hallucinations Real Or Mistakes) dataset, aimed at addressing the challenge of detecting semantic hallucinations in neural language generation systems. SHROOM encompasses data from three diverse NLG tasks: data modeling (DM), machine translation (MT), and paraphrase generation (PG). Each task presents unique nuances in identifying semantic inaccuracies within generated text.

2.1 Model Aware and Agnoustic Data

There were two types of data entries, one in which the model information was present, which was the model aware dataset and the other where the model information was absent, the model agnousitc dataset.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 964–968

Example 1: Definition Modeling (DM)
Source: The sides of the casket were covered with heavy black broadcloth, with velvet caps, presenting a deep contrast to the rich surmountings. What is the meaning of surmounting? **Target:** A decorative feature that sits on top of something. **Hypothesis:** A sloping top. **Model:** ltg/flan-t5-definition-en-base **Label:** Hallucination **Probability (Hallucination):** 0.6

Example 2: Definition Modeling (DM)
Source: And the mower whets his sithe, What is the meaning of whet? **Target:** To hone or rub on with some substance, as a piece of stone, for the purpose of sharpening – see whetstone. **Hypothesis:** To cause to whirl. **Model:** ltg/flan-t5-definition-en-base **Label:** Hallucination **Probability (Hallucination):** 1.0

Example 3: Definition Modeling (DM)
Source: The corporation was accused of unethical behavior for knowingly producing a product suspected of harming health. What is the meaning of unethical? **Target:** Not morally approvable; morally bad; not ethical. **Hypothesis:** Not ethical; not conforming to ethical principles. **Model:** ltg/flan-t5-definition-en-base **Label:** Not Hallucination **Probability (Hallucination):** 0.0

Example 4: Paraphrase Generation (PG)
Source: We must create a society where everyone is able to enjoy a good education. **Hypothesis:** Everyone is capable of enjoying a good education in a society. **Model:** tuner007/pegasus$_{p}araphrase$ **Label:** Hallucination **Probability (Hallucination):** 0.8

Example 5: Paraphrase Generation (PG)
Source: Schooling is a fundamental issue that we should today reaffirm. **Hypothesis:** We should reiterate the importance of schooling. **Model:** tuner007/pegasus$_{p}araphrase$ **Label:** Not Hallucination **Probability (Hallucination):** 0.2

Example 6: Machine Translation (MT)
Source: Malo osungilako asilikali ankhondo amaluso osiyanasiyana ku depatimenti ya zachitetezo yaku U.S. ikutsata ziduswa. **Target:** The United States Strategic Command of the U.S. Department of Defense office is tracking the debris. **Hypothesis:** The U.S. Department of Defense's military intelligence facility is tracking the targets. **Model:** facebook/nllb-200-distilled-600M **Label:** Hallucination **Probability (Hallucination):** 1.0

Table 1: Examples from SHROOM Val dataset

2.2 Data Analysis

The dataset compilation involved sourcing data from a variety of sources to ensure its robustness and generalizability. For DM, definitions were gathered from various domains, covering a wide range of topics. MT data consisted of parallel corpora from multiple language pairs to capture translation nuances effectively. Finally, for PG, a collection of sentences and corresponding paraphrases from various genres was curated to represent natural language variation comprehensively.

2.3 Annotation

Annotating the dataset for semantic hallucinations followed a binary scheme, where each instance was labeled by 5 annotators as either containing semantic hallucinations or being free of such errors. To ensure the reliability of annotations, each instance underwent assessment by at least five annotators, with a majority vote determining the gold label.

2.4 Dataset Statistics

The SHROOM dataset comprises of multiple instances across all tasks. The distribution of instances for each NLG task is summarized below:

NLG Task	Train Set	Test Set
Definition Modeling (DM)	10000	563
Machine Translation (MT)	10000	562
Paraphrase Generation (PG)	10000	375

Table 2: Distribution of Instances by Task

For the Model Agnostic Dataset and Model Aware Dataset, each has:

Validation Set (Labelled):

NLG Task	Instances
Data Modeling	187
Paraphrase	125
Machine Translation	187

Train Set (Unlabelled):

NLG Task	Instances
Data Modeling	10000
Paraphrase	10000
Machine Translation	10000

Test Set:

NLG Task	Instances
Data Modeling	563
Paraphrase	375
Machine Translation	562

2.5 Example Instances

Table 2 provides examples from the SHROOM dataset, showcasing instances with and without semantic hallucinations for each NLG task.

Our participation in this shared task involves leveraging the SHROOM dataset to develop and evaluate models for detecting semantic hallucinations in NLG systems. This dataset serves as a valuable resource for benchmarking and advancing research in this area.

3 Methodology

Our Methodology involved first Labelling the Training Data, fine tune a model on the test data then evaluating the model on test data. We chose Roberta-base as our base model for fine tuning as XLM-RoBERTa is a multilingual language model optimized for classification tasks. It is pretrained on massive multilingual data, and has a robust architecture and performance enable efficient fine-tuning across diverse text classification problems with state-of-the-art accuracy. For Labelling the Training Data, we used Mixtral 8x7B (Jiang et al., 2024), specifically mixtral-8x7b-instruct-v0.1.Q5_K_M.gguf

3.1 Labelling Training Dataset

The prompt (Zamfirescu-Pereira et al., 2023) used for Labelling Training Dataset using Mixtral-8x7b-instruct is:

```
if task == "PG":
    context = f"Context: {src}"
else: # i.e. task == "MT" or task == "DM":
    context = f"Context: {tgt}"

sentence = f"Sentence: {hyp}"
message = f"{context}\n{sentence}\nIs
the Sentence supported by the Context
above?
Answer using ONLY yes or no:"
```

```
prompt = f"[INST] {message} [/INST]"
```

3.2 Finetune Roberta-base on the Mixtral labelled train dataset

We chose to fine-tune Roberta-base on the Mixtral labeled train dataset to adapt the model specifically for the task of detecting semantic hallucinations. The Mixtral labeled training dataset provided binary labels indicating whether a given sentence exhibited semantic hallucinations or not. The probability label for hallucination ranged from 0 to 1, derived from the log probability of the Mixtral model output. Therefore, we formulated the task as a binary classification problem: distinguishing between sentences containing semantic hallucinations and those that do not.

During fine-tuning, we modified the last layer of Roberta-base to accommodate the binary classification task. We used techniques such as cross-entropy loss and gradient descent to update the model's parameters based on the labeled training data. By fine-tuning on the Mixtral labeled dataset, we aimed to enhance Roberta-base's ability to identify semantic hallucinations in natural language generation outputs.

3.3 Finetune Roberta-base on the Pre-Annotated Data

In addition to fine-tuning on the Mixtral labeled train dataset, we performed fine-tuning on preannotated data, specifically the development dataset. This dataset had been annotated by five annotators, and each instance was assigned a probability label for hallucination ranging from 0 to 1 in increments of 0.2. The probability labels were based on the consensus among the annotators.

To leverage the fine-grained annotations provided by multiple annotators, we formulated the task as a multi-class classification problem. We fine-tuned Roberta-base (Conneau et al., 2020) to classify instances into one of six categories corresponding to the six probability levels (0, 0.2, 0.4, 0.6, 0.8, or 1). This approach allowed the model to learn from the nuanced annotations provided by the annotators and make more nuanced predictions about the presence of semantic hallucinations.

By fine-tuning Roberta-base on both the Mixtral labeled train dataset and the preannotated development dataset, we aimed to create a robust model capable of accurately detecting semantic hallucinations across a range of natural language generation tasks and datasets.

4 Results

We present the results of our experiments using three different methodologies for detecting semantic hallucinations in neural language generation systems.

4.1 Methodology 1: Fine-tune on the labelled Training Data (2 Class)

We fine-tuned our model on the labelled training data, treating the task as a binary classification problem. The results over multiple epochs are summarized in Table 3. We observed an improvement in both agnostic and aware accuracy over epochs, with agnostic accuracy reaching 76.47% and aware accuracy reaching 61.27% by the third epoch. However, the Matthews correlation coefficient (rho) showed less consistent improvement, with agnostic rho peaking at 0.58 and aware rho at 0.38 in the second epoch.

Epoch	**Agnostic Acc.**	**Aware Acc.**
1	0.753	0.609
2	0.759	0.601
3	0.765	0.613
Epoch	**Agnostic ρ**	**Aware ρ**
1	0.568	0.346
2	0.580	0.381
3	0.584	0.355

Table 3: Results for Methodology 1: Fine-tune on labelled Training Data (2-Class)

4.2 Methodology 2: Fine-tune on the labelled Validation Data (6 Class)

In this methodology, we fine-tuned the model on the labelled validation data, treating the task as a six-class classification problem. Results are presented in Table 4. Agnostic accuracy fluctuated around 45-51% over different epochs, while aware accuracy showed similar fluctuations around 47-58%. Matthews correlation coefficient (rho) varied between 0.43 and 0.52 for agnostic classification and between 0.48 and 0.52 for aware classification.

4.3 Methodology 3: Zero-shot Mixtral 8x7b

For the zero-shot approach (Yue et al., 2023), where we directly applied the Mixtral 8x7b model without fine-tuning, results are shown in Table 5. Agnostic accuracy achieved 78.73%, while aware accuracy

Epoch	Agnostic Acc.	Aware Acc.
3	0.515	0.578
5	0.473	0.487
10	0.449	0.483
15	0.463	0.473
Epoch	**Agnostic ρ**	**Aware ρ**
3	0.477	0.490
5	0.477	0.490
10	0.502	0.524
15	0.434	0.512

Table 4: Results for Methodology 2: Fine-tune on labelled Validation Data (6-Class)

reached 77.73%. The Matthews correlation coefficient (rho) for agnostic classification was 0.50, and for aware classification, it was 0.48.

Overall, the zero-shot approach demonstrated competitive performance compared to fine-tuning on labelled data, indicating the effectiveness of the Mixtral 8x7b model in detecting semantic hallucinations without additional training.

Approach	Agnostic Acc.	Aware Acc.
Zero-shot	0.787	0.777
Approach	**Agnostic ρ**	**Aware ρ**
Zero-shot	0.499	0.485

Table 5: Results for Methodology 3: Zero-shot Mixtral 8x7b

5 Conclusion

In this study, we investigated three different methodologies for detecting semantic hallucinations in neural language generation systems. We fine-tuned a model using labelled training data, labelled validation data, and also explored a zero-shot approach using the Mixtral 8x7b instruct model.

Our results indicate that fine-tuning on labelled data, whether it is the training data or the validation data, led to improvements in both agnostic and aware accuracy over multiple epochs. However, the effectiveness of fine-tuning on validation data seemed to diminish as the number of epochs increased, suggesting potential overfitting.

Interestingly, the zero-shot approach using the Mixtral 8x7b instruct model achieved competitive performance compared to fine-tuning on labelled data. This indicates the robustness of the Mixtral model in detecting semantic hallucinations without additional training.

Overall, our findings suggest that while fine-tuning on labelled data can lead to improvements in detection accuracy, the zero-shot approach with pre-trained models like Mixtral 8x7b instruct provides a viable alternative, especially when labeled data is limited or unavailable. Future research could explore further optimization of fine-tuning strategies and investigate the generalizability of pre-trained models across different domains and tasks.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. 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, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Timothee Mickus, Elaine Zosa, Raúl Vázquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. Semeval-2024 shared 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)*, pages 1980–1994, Mexico City, Mexico. Association for Computational Linguistics.

Zhenrui Yue, Huimin Zeng, Mengfei Lan, Heng Ji, and Dong Wang. 2023. Zero-and few-shot event detection via prompt-based meta learning. *arXiv preprint arXiv:2305.17373*.

JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.