Maha Bhaashya at SemEval-2024 Task 6: Zero-Shot Multi-task Hallucination Detection

Patanjali Bhamidipati[†]
IIIT Hyderabad
patanjali.b@research.iiit.ac.in

Advaith Malladi[†]
IIIT Hyderabad
advaith.malladi@research.iiit.ac.in

Manish Shrivastava
IIIT Hyderabad
m.shrivastava@iiit.ac.in

Radhika Mamidi
IIIT Hyderabad
radhika.mamidi@iiit.ac.in

Abstract

In recent studies, the extensive utilization of large language models has underscored the importance of robust evaluation methodologies for assessing text generation quality and relevance to specific tasks. This has revealed a prevalent issue known as hallucination, an emergent condition in the model where generated text lacks faithfulness to the source and deviates from the evaluation criteria. In this study, we formally define hallucination and propose a framework for its quantitative detection in a zero-shot setting, leveraging our definition and the assumption that model outputs entail task and sample specific inputs. In detecting hallucinations, our solution achieves an accuracy of 0.78 in a model-aware setting and 0.61 in a model-agnostic setting. Notably, our solution maintains computational efficiency, requiring far less computational resources than other SOTA approaches, aligning with the trend towards lightweight and compressed models.

1 Introduction

The contemporary landscape of Natural Language Generation (NLG) is marked by a confluence of complexities, wherein two primary challenges emerge as focal points of concern. Firstly, the prevalent neural models within NLG frameworks consistently produce outputs that exhibit linguistic fluency yet suffer from inaccuracies (Huang et al., 2023). Secondly, the current evaluation metrics, vital for evaluating the effectiveness of NLG systems, demonstrate a significant inclination towards fluency measures while neglecting to prioritize accuracy. So, this highlights the need to consider the "truthfulness" of the model's output, i.e its alignment with the source to ensure a comprehensive assessment.(Dale et al., 2022)

In the realm of NLG applications, the criticality of output accuracy cannot be overstated. A divergence between the fluency and factual correctness of generated content not only undermines the utility of NLG systems but also engenders substantial risks across various domains. Consider, for instance, the domain of machine translation,the production of seemingly plausible yet inaccurate translations not only compromises the integrity of the translated content but also defeats the purpose of facilitating correct translations.

Likewise, in tasks like definition modeling and paraphrase generation, where accurately conveying semantic meaning is crucial, the presence of incorrect outputs presents notable challenges in upholding the integrity and dependability of the generated content.

2 SHROOM Dataset

SHROOM (a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes) dataset is a task-based hallucination detection dataset which is divided into two major categories:

2.1 Model Aware (MAw)

Model Aware (MAw) refers to situations where the model under study is known.

2.2 Model Agnostic (MAg)

Model Agnostic (MAg) refers to situations where the model under study is not known.

The dataset encompasses three major Natural Language Generation tasks, namely:

1. Definition Modeling (DM): In this task, models are trained to generate a definition for a given example in context.

[†] The authors contributed equally to this work.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1685–1689
June 20-21, 2024 ©2024 Association for Computational Linguistics

2. Machine Translation (MT): In this task, models aim to generate translations of the given samples.

3. Paraphrase Generation (PG): In this task, models aim to produce paraphrases of the given text samples.

Further, each sample in the train set is populated with information such as task ($task$): indicating what objective the model was trained for, source (src): the input passed to the models for the generation, target (tgt): the intended reference "gold" standard text that the model ought to generate, hypothesis (hyp): the actual model production,also the model-aware dataset is populated with model name ($model$) used for the task, with the val set additionally being populated with majority-based gold-label ($label$), based on the annotator labels along with the probability values of the sample being hallucination ($p(Hallucination)$) based on the proportion of annotators who claim that the sample is an hallucination.

3 Definitions

As described earlier, the SHROOM shared task encompasses of three different taks, Definition Modelling (DM), Paraphrase Generation (PG), and Machine Translation. We define the **Hallucination** in the context of the specific task at hand. Defining hallucinations individually in the context of a specific task enables detecting hallucinations quantitatively and qualitatively. We offer distinct definitions of hallucinations and methodologies for detecting hallucinations within the context of each of the aforementioned task.

In the context of definition modelling, the model is expected to generate the definition of a word which has been used in the provided context by making use of distributional semantics. Definition modelling models such as flan-t5-definition-en-base (Giulianelli et al., 2023) are not fully capable of making use of distributional semantics to define a word as used in a context. In a sample where the word W has been used in a setting contrasting to the definition the model has learnt during its training process, the models fails to provide a contextual definition of the word W. Examples for the same have been demonstrated in Table 1 and Table 2. We observe the model outputs a definition of word W which is very similar what it has learnt during its training process. Based on this observation, we assume that the targets provided in the SHROOM dataset have been extracted from the training data of the definition modelling dataset. With this assumption, we define **"Hallucination to be an instance where the output generated by the definition modelling model *does not* entail the target output."** Thereby reducing the hallucination detection task to a Natural Language Inference task in the context of definition modelling.

In the context of paraphrase generation and machine translation, the model's inputs and outputs are anticipated to exhibit semantic equivalence. If the generated paraphrase or translation diverges from semantic equivalence with the source text, they are deemed imperfect paraphrases or translations. Therefore, in the context of paraphrase generation and machine translation, we define **"Hallucination to be an instance where the paraphrase or translation generated by the model is *not semantically equivalent* to the source."** This reduces the hallucination detection in the context of given tasks to a semantic equivalence detection task, which could also be framed as bidirectional entailment detection, a variation of the Natural Language Inference task.

The aforementioned definitions of hallucination allow us to simplify the hallucination detection task to a Natural Language Inference task, thereby enabling us to qualitatively and quantitatively detect hallucinations.

In a more generic setting, we provide a definition of hallucinations that can be adapted to any task to effectively detect them. We define **"hallucinations as instances where the output generated by the model is *not faithful* to the input or the training data of the model. If the model generates information that is contradictory to the model's training data or the input to the model, it can be termed as a hallucination."**

4 Methodology

Grounding to the above definitions, the experimental setup we designed goes on to quantify the alignment of the model's output (hyp) with either the source (src) or the target (tgt) based on the task ($task$) the data sample corresponds to.

We propose that examining the **entailment** relationship between the model's output (hyp) and either the source (src) or the target (tgt) (which is

Example 1: Definition Modeling (DM)
Model Input: I went into the *water bottle* to withdraw cash. What is the definition of *water bottle*? **Model Output:** A container for holding liquids. **Expected Output:** A financial institution such as a bank or ATM to withdraw cash **Model:** flan-t5-definition-en-base

Table 1: A Table showing distinction between the model output and the expected output where the model fails to understand the contextual definition of the term *water bottle*

Example 2: Definition Modeling (DM)
I jumped into the flaxcron to do some swimming. What is the meaning of *flaxcron*? **Model Output:** A slender, slender **Expected Output:** A pool of water. **Model:** flan-t5-definition-en-base

Table 2: A Table showing distinction between the model output and the expected output where the model fails to understand the contextual definition of the term *flaxcron*

also inherently linked to the source (src)), depending on the task, sheds light on data samples that are **not** "detached" from the source. Consistent with our initial hypothesis that hallucinations occur when samples are "detached" from the source, this approach based on Natural Language Inference (NLI) can effectively aid in hallucination classification.

- In the context of definition modelling, **if the** hyp **does not entail the** tgt, the sample has been classified as **Hallucination**.

- In the context of machine translation and paraphrase generation, we check equivalence through bidirectional entailment. **If the** hyp **does not bidirectionally entail the** src, the sample has been classified as **Hallucination**

- In the context of machine translation and paraphrase generation, we verify our hypothesis of semantic equivalence between the src and hyp by comparing the performance metrics in the case of both unidirectional and bidirectional entailment.

Recent research heavily relies on large language models (LLMs) to benchmark various natural language understanding and generation (NLG) tasks. However, this practice extends to hallucination detection as well, which we find ironic and counterproductive, considering LLMs' inherent tendency to hallucinate. Using LLMs for hallucination detection presents two major drawbacks. Firstly, their computational demands are significant, making them an expensive solution (Bai et al., 2024). Secondly, the lack of complete interpretability in LLMs renders them unreliable for this task (Singh et al., 2024).

5 Results

After several experiments with the above methodology, leveraging the accuracy and the Spearman correlation (ρ) metrics, we have bench-marked the hallucination detection task on the SHROOM validation and test sets to achieve an accuracy of 0.78 in model-aware and 0.61 on model-agnostic test sets respectively. For our analysis let us take only the accuracy metric into account.
The bench-marking saw a utilisation of open-source pre-trained Natural Language Inference (NLI) models available on Hugging Face. Several experiments brought out interesting observations which are worthy discussing.

We evaluated the following models:

1. MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli **(DeBERTa-1)** (He et al., 2020)

2. MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 **(DeBERTa-2)** (He et al., 2020)

Model	Unidirectional	Bidirectional
DeBERTa - 1	0.783567	0.755511
DeBERTa - 2	0.765531	0.717435
BART - 1	0.769539	0.733467
RoBERTa - 1	0.757515	0.727455

Table 3: Model-agnostic evaluation on (Uni vs Bi) directional entailment.

Model	Unidirectional	Bidirectional
DeBERTa - 1	0.596806	0.570859
DeBERTa - 2	0.576846	0.586826
BART - 1	0.610778	0.568862
RoBERTa - 1	0.612774	0.584830

Table 4: Model-aware evaluation on (Uni vs Bi) directional entailment.

3. ynie/bart-large-snli_mnli_fever_anli_R1_R2_R3-nli **(BART-1)** (Lewis et al., 2019)

4. ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli **(RoBERTa-1)** (Nie et al., 2020)

5.1 Definition Modelling

For the task of definition modelling, our approach achieves a peak accuracy of 0.748663 using the DeBERTa-2 model in a model-agnostic setting and a peak accuracy of 0.755319 using the RoBERTa-1 model in a model-aware setting. These results are in accordance with our hypothesis that when the model hallucinates, it does not entail the target.

5.2 Paraphrase Generation and Machine Translation

For the paraphrase generation and machine translation tasks, the observed results confirm our hypothesis that if the source (src) and hypothesis (hyp) are not semantically equivalent, the hypothesis is a hallucination. In hallucination detection for the paraphrase generation task, we observe that bidirectional entailment (semantic equivalence) outperforms the unidirectional entailment approach for all models. Similar results can also be observed for the machine translation task. This provides evidence that in machine translation and paraphrase generation, hallucinations can be detected by checking for semantic equivalency between the source and hypothesis.

Model	Unidirectional	Bidirectional
DeBERTa - 1	0.728	0.752
DeBERTa - 2	0.624	0.68
BART - 1	0.696	0.72
RoBERTa - 1	0.712	0.752

Table 5: Accuracy validation on **PG** task

Model	Unidirectional	Bidirectional
DeBERTa - 1	-	-
DeBERTa - 2	0.722	0.754
BART - 1	-	-
RoBERTa - 1	-	-

Table 6: Accuracy validation on **MT** task.

Model	DM	MT	PG
DeBERTa - 1	0.721925	0.855615	0.768
DeBERTa - 2	0.748663	0.823529	0.704
BART - 1	0.748663	0.844920	0.688
RoBERTa - 1	0.711230	0.834224	0.712

Table 7: Model-agnostic evaluation on individual tasks.

5.3 Overall Analysis

In the model-agnostic setting, we achieve a peak accuracy of 0.783567 using the DeBERTa-1 model and a peak accuracy of 0.612774 in a model-aware setting using the RoBERTa-1 model. These scores exhibit satisfactory performance of models pretrained on the Natural Language Inference task for Hallucination Detection.

6 Conclusion

Our work makes two significant contributions to the study of hallucinations in language models. First, we provide a concrete definition of the term "hallucination," enabling both qualitative and quantitative study and detection of such phenomena. Second, we offer a computationally efficient approach to detect hallucinations in tasks such as definition modeling, machine translation, and paraphrase generation. We frame the hallucination detection task as a function of the input to the generation model and the data used to train it. Our definitions and approaches also provide a framework that can be utilized for hallucination detection in various Natural Language Generation tasks across the spectrum.

References

Guangji Bai, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, Ziyang Yu, Mengdan Zhu, Yifei Zhang, Carl Yang, Yue Cheng, and Liang Zhao. 2024. Beyond efficiency: A systematic survey of resource-efficient large language models.

David Dale, Elena Voita, Loïc Barrault, and Marta R. Costa-jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models.