SmurfCat at SemEval-2024 Task 6: Leveraging Synthetic Data for Hallucination Detection

Elisei Rykov[1,2], **Yana Shishkina**[2,3], **Kseniia Petrushina**[1,4] ,
Kseniia Titova[1,5], **Sergey Petrakov**[1], **and Alexander Panchenko**[1,6]
[1]Skolkovo Institute of Science and Technology, [2]Tinkoff,
[3]HSE University, [4]Moscow Institute of Physics and Technology, [5]MTS AI, [6]AIRI
{e.rykov, y.a.shishkina}@tinkoff.ai, {kseniia.petrushina, kseniia.titova, sergey.petrakov, a.panchenko}@skol.tech

Abstract

In this paper, we present our novel systems developed for the SemEval-2024 hallucination detection task. Our investigation spans a range of strategies to compare model predictions with reference standards, encompassing diverse baselines, the refinement of pre-trained encoders through supervised learning, and an ensemble approaches utilizing several high-performing models. Through these explorations, we introduce three distinct methods that exhibit strong performance metrics. To amplify our training data, we generate additional training samples from unlabelled training subset. Furthermore, we provide a detailed comparative analysis of our approaches. Notably, our premier method achieved a commendable 9th place in the competition's model-agnostic track and 17th place in model-aware track, highlighting its effectiveness and potential.

1 Introduction

Large language models are proficient in generating human-like text across various styles. However, even the most advanced models can produce hallucinations, leading users to question their reliability. There are two primary types of hallucinations: factuality hallucinations, which involve the generation of content that deviates from actual facts, and faithfulness hallucinations, when the model fails to solve tasks correctly following specific instructions (Huang et al., 2023).

The SemEval 2024 Shared-task on Hallucinations and Related Observable Overgeneration Mistakes (Mickus et al., 2024) has integrated both types into three tasks. The Definition Modeling task (DM) focused on fact-related hallucinations by challenging models to generate contextually relevant word definitions. Both the Machine Translation (MT) and Paraphrase Generation (PG) tasks included faithfulness hallucinations, with models asked to produce translations or paraphrases for given sentences. Evaluation labelled datasets for these tasks were provided and the training dataset consisted only of source sentences and model generations, without corresponding labels.

Motivated by the lack of annotated resources and the efficacy of other language models trained on synthetic data, we developed two synthetic datasets that replicate the targeted domain. First, we collected data through a proprietary GPT-4 model (OpenAI, 2023), but our methods trained on the achieved data did not yield the desired results as prompt engineering made maintaining the domain challenging. As a second approach, we trained LLaMA2-7b (Touvron et al., 2023) adapters using a small set of annotated examples and applied them to the unlabeled training data. This method proved to be a more effective form of in-domain data augmentation.

While the competition was run on two tracks, we focus mainly on the model-agnostic track. In our methods we utilized the most effective models with varied sizes and architectures, which we had evaluated beforehand. Our experiments involved fine-tuning a pre-trained embedding model, repurposing it to function as a binary classifier across a number of open-source datasets, including our synthetic sets. We also experimented with a promising method for evaluating paraphrases by modifying its design and fine-tuning the model on different data. Finally, we tested different combinations of the highest-performing approaches in an ensemble setting. Generated synthetic data and code published on GitHub[1].

2 Related work

In the field of text representation, the E5 (Wang et al., 2022) family represents a group of cutting-edge sentence embedding models trained through

[1] https://github.com/s-nlp/shroom

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 869–880
June 20-21, 2024 ©2024 Association for Computational Linguistics

contrastive methods. The E5-Mistral[2] model, a powerful embedding model that has been fine-tuned on a selection of annotated data, is currently recognized as the leading open-source model by the Multitask Text Embedding Benchmark (Muennighoff et al., 2023).

Vectara's *hallucination_detection_model*[3] is a fine-tuned DeBERTa focused on summarization datasets that includes annotations for factual consistency. TrueTeacher (Gekhman et al., 2023) is a family of models and an associated dataset designed for evaluating factual consistency. The dataset was created by first fine-tuning various-sized T5 models on summarization tasks. These models were then employed to generate hypotheses, which were subsequently automatically annotated using a 540B Large Language Model (LLM). This annotated dataset was then utilized to train multiple models to assess factual consistency.

The Mutual Implication Score (MIS) (Babakov et al., 2022) is a metric devised for evaluating the quality of text style transfer and paraphrasing systems, grounding its assessment on content similarity between the prediction and the reference text. It leverages a RoBERTa-NLI (Nie et al., 2020) model that has been fine-tuned and incorporates it into an architecture that processes two input texts sequentially in both forward and reverse directions. The final hidden states from these two passes are merged and forwarded to a classification head to determine the MIS score. Initially, the MIS metric was trained using the Quora Question Pairs dataset (QQP) (Sharma et al., 2019).

SimCSE (Similarity-based Contrastive Self-supervised Learning) (Gao et al., 2021) is a self-supervised learning method for text embeddings. It is used for creating embeddings of text data that are semantically meaningful and can be used in various downstream tasks. It involves training a neural network to maximize the similarity between embeddings of similar sentences and minimize the similarity between embeddings of dissimilar sentences. LaBSE (Language-agnostic BERT Sentence Embedding) (Feng et al., 2022) is a method for generating multilingual sentence embeddings using the BERT architecture.

Other metrics for evaluating content preservation, such as BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002), CHRF (Character n-gram F-score) (Popović, 2015), METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee and Lavie, 2005), and BLEURT (Bilingual Evaluation Understudy for Ranking and Tuning) (Sellam et al., 2020), also stand out. BLEU utilizes a modified unigram precision score, CHRF evaluates the quality of machine translation by comparing character n-grams in candidate translations against reference translations to compute an F-score, METEOR calculates the harmonic mean of precision and recall at the single-word level, and BLEURT employs a fine-tuned BERT model in a cross-encoder setup, using synthetic data to assess semantic similarity.

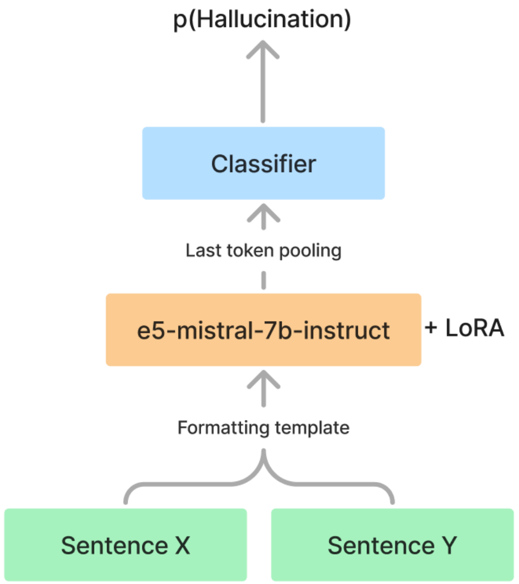

Figure 1: Classifier architecture when using synthetic data.

3 Data

3.1 Existing datasets

The QQP dataset consists of pairs of questions from the Quora forum. For each pair, it is indicated whether the questions are paraphrases, i.e. they ask about the same thing. PAWS (Zhang et al., 2019) is a paraphrase detection dataset that contains complex cases with both paraphrase and non-paraphrase samples that have high lexical overlap.

We postulated that other pre-existing datasets, such as QQP and PAWS, might exhibit particular biases due to their distinct task domains (for instance, QQP dataset includes only questions). To mitigate this potential issue, we generated synthetic

[2] https://huggingface.co/intfloat/e5-mistral-7b-instruct

[3] https://huggingface.co/vectara/hallucination_evaluation_model

	DM	MT	PG
Not Hallucination	188	211	132
Hallucination	175	179	132
Total	363	390	264

Table 1: Adapter train sample sizes.

data taking unlabeled training samples as starting points.

Our experimentation with synthetic data creation was divided into two main approaches: the first involved training LoRA (Hu et al., 2022) adapters for the LLaMA2-7b model using the annotated data derived from the validation set. The second approach involved the generation of both correct and incorrect hypotheses by employing GPT-4 and specific prompts.

All tasks were distilled down to the paraphrase evaluation task. Consequently, we only used targets (sources for paraphrase generation) and hypotheses as inputs for the models.

3.2 LLaMA2-7b adapter

We trained 6 LoRA adapters, pairing them to specialize in either generating hallucinations or producing correct responses for each task. Due to the limited amount of labeled data, we made use of model's ability of in-context learning by prepending samples with instructions: *Paraphrase* for non-hallucinations and *Provide an incorrect paraphrase* for hallucinations. The number of samples for each adapter is shown in Table 1.

Training and generation hyperparameters are displayed in Table 2. For each task and label we manually selected the best epoch by analyzing a small set of generated samples. These checkpoints were further employed to synthesize hypotheses for their task's training set. A small sample of the generated data using LLaMA2-7b adapter is provided in the Appendix C.

3.3 GPT-4 prompting

In addition, we created two distinct prompts for the PG task. In these prompts, we directed GPT-4 to generate a paraphrase of a source sentence extracted from an unlabeled training sample. The nature of the paraphrase, whether it should contain hallucinations and overgeneration errors or not, was determined by the specific prompt we used.

We enriched the prompt structure for few-shot learning purposes, incorporating several illustrative examples drawn from both the validation and trial data splits. Alongside each *incorrect* example, we included an explanation to clarify why the provided hypothesis did not meet the criteria.

Stage	Hyperparameter	Value
Training	lr	4e-4
	warmup_steps	1
	optimizer	AdamW
	scheduler	linear
	LoRA alpha	16
	LoRA dropout	0.05
	LoRA r	16
	batch size	32
Inference	num_beams	3
	do_sample	true
	repetition_penalty	1.2
	top_k	50
	max_new_tokens	512

Table 2: Training and inference hyperparameters for LoRA adapters.

Moreover, we tasked GPT-4 to execute its reasoning step-by-step: to iterate through several examples with accompanying explanations, and, by leveraging those explanations, to discern and select the most suitable paraphrase.

We utilized the *gpt-4-1106-preview* model, adhering to the default generation parameters stipulated by the OpenAI API service.

3.4 Data filtration

In the process of evaluating the synthetic data we generated, we encountered multiple issues that necessitated an extra layer of filtering:

- A number of the samples produced by the LLaMA2-7B model were excessively lengthy, containing up to 1024 tokens.
- The labeling of samples by the LLaMA2-7B as *Hallucination* was frequently incorrect. Samples designated as hallucinations were often devoid of any such content, and conversely, non-hallucination samples sometimes contained hallucinations.
- A peculiar pattern was observed in the DM task generations from LLaMA2-7B, where more than 9,000 samples started with the word *any* or *anything* denoting a biased starting point which may impact the diversity and neutrality required for effective training.
- In the subsets of synthetic data generated by GPT-4 and labeled as *Not Hallucination* the resulting examples were deemed too straightforward, potentially leading to a training dataset

that cannot robustly challenge and thereby improve the model's discriminatory capabilities.

To tackle the identified issues with the synthetic data, we adopted a systematic filtering methodology. We began by eliminating any hypothesis that exceeded a length of 200 tokens, ensuring the data remained succinct. For the samples that started with *any* or *anything*, we decided to limit the number to 500 to minimize bias.

With the aim of refining the data quality, we then annotated all the synthetic samples using MIS. We set specific thresholds for these MIS scores to filter the data further. In the subset containing hallucinations, we removed samples that had a score lower than 0.1 or higher than 0.5. For non-hallucinated samples, we only retained those with a score between 0.7 and 0.9. These score ranges were established empirically to ensure a balance between discernibility and ambiguity in both the hallucinated and non-hallucinated examples.

The number of samples generated using both synthetic methods, before and after the filtering stage, is given in Table 3. After generating the synthetic data, we performed several experiments with different combinations of synthetic data.

4 Methods

4.1 Black-box baselines

First, we started with an assessment of various baseline models that are detailed in Section 2, including a new addition, GPT-4. These baseline models were utilized as-is, in a *black-box* fashion, without any further fine-tuning specifically for our tasks.

For all models other than GPT-4, we employed the inference code available on the official HuggingFace Hub pages. For GPT-4, we created specific prompts for each task. Within these prompts, we instructed GPT-4 to methodically process the information and ascertain the presence of hallucinations within the sample. We provided all pertinent data (source, hypothesis, and, when available, target) within the prompt. It is important to note that the collection and evaluation of predictions were conducted strictly within the model-aware track. The prompt is available in Appendix A.

4.2 SFT E5-Mistral

The obtained synthetic data was used to fine-tune the E5-Mistral model on our domain. In our experiments, we adjusted the data inputs by adding or omitting certain subsets of synthetic data to create the final blend used for training. The choice of the E5-Mistral model as the foundation for our work was based on its superior performance compared to other models.

The design of our classifier is depicted in Figure 1. In simple terms, we prepare two sample sentences with a specific format and input them into an model with LoRA. Afterwards, we obtain the embedding of the last token and pass it to the classification head.

4.3 Mutual Implication Score

In this setup we experimented with some improvements to the original Mutual Implication Score model architecture. Even though MIS was already trained on a large amount of paraphrase detection data, QQP dataset biased to the questions. Therefore, we thought that we can fine-tune it to decrease this bias.

In Table 4 we present default training hyperparameters used for experiments with MIS. Unless stated otherwise, we chose to train with the RoBERTa encoder, classifier and QQP dataset from original MIS study.

We tried various experiment configurations, ranging from the use of new datasets to alterations in architecture and training methods. We will describe all the modifications presented:

1. **MIS**: Vanilla MIS from HuggingFace Hub without any fine-tuning.
2. **MIS trained with LoRA**: Add LoRA adapters instead of partially unfreezing layers.
3. **MIS with Vectara**: Replace the original RoBERTa encoder with Vectara's model.
4. **MIS with one encoder**: Change MIS two-folded architecture with a single one.
5. **MIS trained on the PAWS**: Add 108,463 human-labeled paraphrase adversaries from PAWS.
6. **MIS trained on our synthetic data**: Add our synthetic data obtained previously.

4.4 Content Preservation Measures

We conducted a separate analysis on several NLP techniques as examined in the original MIS study. This exploration aimed to assess their suitability for the task of hallucination detection, considering the inherent connection between style transformation, paraphrase generation, and hallucination detection. A well-executed paraphrase should retain the essence of the original text without introducing

Source method	Task	Label	# before filtering	# after filtering
LLaMA2-7B	MT	Hallucination	18 093	7 758
		Not Hallucination	17 056	3 572
	PG	Hallucination	13 961	2 839
		Not Hallucination	14 928	3 952
	DM	Hallucination	19 224	5 939
		Not Hallucination	20 000	12 032
GPT-4	PG	Hallucination	7 439	-
		Not Hallucination	6 279	-

Table 3: The number of samples in the synthetic datasets. No filtering was performed for GPT-4.

Hyperparameter	Value
lr	1e-4
lr scheduler	constant
optimizer	AdamW
batch size	32

Table 4: Training hyperparameters for MIS experiments.

extraneous elements, which is particularly crucial given that one of the competition's subtasks involved paraphrasing. Specifically, our investigation involved LaBSE, SimCSE, and the metrics for evaluating content preservation described in Section 2.

4.5 Ensembling

To enhance the performance of different pre-trained models, we combined them into an ensemble. The final decision on the presence of hallucinations is based on the predictions of multiple independent models.

The predictions of separate models were normalized so that the decision boundary was the same for all models. Thus, differences in the scale of the threshold value did not introduce bias into the final decision.

We have chosen the best set of models for the ensemble from the possible options: E5-Mistral, fine-tuned E5-Mistral, Vectara, TrueTeacher, *all-mpnet-base-v2*[§] and also Mutual Implication Score. We calculated cosine between the encoded representations of the model's hypothesis and the target sentence. To obtain a prediction, this score was compared with a descision boundary. For each model we select the optimal classification threshold on validation subset for each track and task. For Vectara we used a threshold of 0.5.

We employed different strategies on aggregating individual hallucination scores: Normalized averaging and Voting.

4.5.1 Normalized averaging

The predictions of separate models were normalized so that the decision boundary was the same for all models. Thus, differences in the scale of the threshold value did not introduce bias into the final decision.

Individual model scores are normalized as follows:

$$\hat{p} = \begin{cases} kp + b, & p \geq \text{thr} \\ \frac{p}{2\text{thr}}, & p < \text{thr} \end{cases}$$

where $k = \frac{1}{2(1-\text{thr})}$, $b = 1 - k$ and thr is the optimal decision boundary on validation.

This transformation allows to keep the score within $[0, 1]$, at the same time, the decision boundary for all models becomes 0.5.

4.5.2 Voting

Another strategy is to aggregate the binary predictions of the models in an ensemble. The presence of hallucinations was determined by voting models, depending on the number of votes in favor. At the verification stage, we determine the minimum number of model votes required to acknowledge the pair of sentences, model hypothesis and ground truth, as a paraphrase, for example, at least one, two or three models voted in favor. That is, we predicted a hallucination if an insufficient number of models compared to the optimal validation threshold classified the sample as a paraphrase.

5 Results

The comparative analysis of the performance across all baselines, our proposed methods, and the leading approaches derived from the official rankings is collated in Table 5.

[§] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Method	val		test	
	agnostic	aware	agnostic	aware
ahoblitz*	-	-	**0.85**	**0.81**
zackchen*	-	-	0.84	**0.81**
liuwei*	-	-	0.83	0.80
Voting	0.85	0.82	0.82	0.78
Normalized averaging	0.81	0.81	0.81	0.79
MIS + PAWS	0.82	0.82	0.81	0.78
SFT E5 Mistral	0.83	0.77	0.80	0.77
MIS	0.81	0.78	0.80	0.77
E5 Mistral	0.81	0.80	0.76	0.78
Vectara	0.76	0.76	0.75	0.77
TrueTeacher	0.79	0.79	0.76	0.80
GPT-4	-	0.74	-	-
SimCSE	0.80	0.80	0.76	0.76
BLEURT	0.77	0.77	0.74	0.74
LaBSE	0.72	0.75	0.69	0.73
METEOR	0.68	0.71	0.67	0.69
chrF	0.63	0.72	0.65	0.67
BLEU	0.67	0.70	0.64	0.65
Official baseline	-	-	0.70	0.74

Table 5: Performance of described approaches. Accuracy is observed as evaluation score. *Top approaches from the official rankings.

Method	Models	val		test	
		agnostic	aware	agnostic	aware
Voting	MIS + E5-Mistral + SFT E5-Mistral + all-mpnet + Vectara	**0.85**	**0.82**	**0.82**	0.78
	MIS + E5-Mistral + SFT E5-Mistral + all-mpnet	**0.85**	0.80	**0.82**	0.77
Normalized averaging	MIS + E5-Mistral + SFT E5-Mistral	**0.85**	0.79	0.81	0.78
	MIS + all-mpnet + Vectara + TrueTeacher	0.81	0.81	0.81	**0.79**

Table 6: Ensembling results. Accuracy is observed as evaluation score.

5.1 Ensembling

According to the results, the Voting approach we developed surpasses all baselines as well as other methods we devised. Nevertheless, the performance narrowly trails the foremost methods from the model-agnostic track in the official rankings by a minimal margin of 0.01. In regards to the application of Ensembling methods, a detailed evaluation delineating the constituent models employed is documented in Table 6. It was discerned that the incorporation of our SFT E5-Mistral model enhances overall performance metrics.

5.2 MIS

Succeeding in performance ranking is the MIS model, refined through training on the PAWS dataset. As previously elucidated, an assortment of configurations was examined, the details of which are exhaustively represented in Table 7. It is observed that the original MIS model's performance was not substantially uplifted; modifications yielded no marked increment in accuracy. Nonetheless, it is notable that the integration of the PAWS dataset into the training process marginally amplified accuracy for both tracks. Simultaneously, a minor enhancement on the aware track was observed upon the deployment of the Vectara encoder in place of the RoBERTa model.

5.3 SFT E5-Mistral

The next approach by performance is our SFT E5-Mistral. The accuracy for different configurations in our synthetic data experiments can be found in Table 8. The combination of PG and DM synthetic data achieves the best results. Unexpectedly, the use of synthetic data from GPT-4 does not yield as good outcomes. This suggests that GPT-4's synthetic data may contain some inherent biases.

Method	val		test	
	agnostic	aware	agnostic	aware
MIS (original)	0.80	0.78	0.77	**0.80**
+ LoRA	0.79	0.79	0.78	**0.80**
+ Vectara	0.79	0.81	**0.81**	0.77
+ Single fold	0.78	0.77	0.75	0.78
+ PAWS	**0.82**	**0.82**	**0.81**	0.78
+ Synthetic data	0.79	0.77	0.77	0.74

Table 7: MIS ablation study results. Accuracy is observed as evaluation score.

Source	Subset	agnostic	aware
GPT	PG	0.76	0.72
LLaMA	PG	0.81	0.75
	DM	0.63	0.51
	MT	0.79	0.71
	PG + DM	**0.83**	**0.77**
	PG + MT	0.81	0.76
	MT + DM	0.75	0.71
	All	0.77	0.71
GPT + LLaMA	All	0.77	0.73

Table 8: Synthetic data ablation study on E5-Mistral. Accuracy is observed as evaluation score.

We carried out a detailed evaluation of a particular subset and identified probable causes for bias:

- For texts generated without hallucinations, they tend to be overly formal and intricate.
- In cases with hallucinations, numerous instances are exceedingly convoluted, sometimes to the extent that the sentences convey the opposite meaning. Our investigation revealed that such hallucinations might not be readily detectable.

It is also clear that relying solely on DM synthetic data does not sufficiently address other tasks. By contrast, a model checkpoint trained with PG synthetic data shows promising performance. Just like the MIS approach, it appears that having PG data is sufficient to address hallucinations in other tasks, provided that the target is accessible.

5.4 Black-box baselines

All our advanced methods outperform black-box baselines on model-agnostic track. Even though, we observe that the E5-Mistral and MIS methods sets a solid baseline on model-agnostic track, maintaining a high level of performance even without any fine-tuning. Considering model-aware track, all baseline models except of GPT-4 show similar performance. The GPT-4 model does not do as well as the others in terms of the average score with our specific prompts. Finally, there is the official baseline that our approaches outperform.

5.5 Content Preservation Measures

Across preservation measures, SimCSE demonstrates the most notable results. In the model-agnostic track, it performs at the same level as more sophisticated approaches such as TrueTeacher, Vectara, or E5 Mistral, without any fine-tuning. However, other preservation measures do not perform as well. Most of them, with the exception of BLEURT, perform even worse than the official baseline in the model-agnostic track.

6 Conclusion

We conducted a comparative analysis involving six baseline models (MIS, E5-Mistral, Vectara, TrueTeacher, GPT-4, and the official baseline from the participant kit) alongside four sophisticated approaches (Voting and Normalized Averaging in Ensembling, as well as the refined MIS and SFT E5-Mistral). Of all methods evaluated, Ensembling demonstrated the highest performance. Nonetheless, the refined MIS and the SFT E5-Mistral exhibited only a minor shortfall in performance when compared to these leading methodologies.

Indeed, there appear to be several avenues for enhancing our synthetic data to potentially exceed the performance of other methods:

- Instead of training separate adapters for each task, centralized training with one adapter across multiple tasks could enrich the learning context and expand the size of the training dataset.
- Exploring a range of other models, such as Mistral-7b (Jiang et al., 2023), Mixtral-8x7b[¶], or LLaMA models of various larger sizes (LLaMA-13b, LLaMA-30b), could identify more efficient architectures or models that are better suited to handle the synthetic data effectively.
- For improving the quality of GPT-generated synthetic data, incorporating a more extensive range of examples within few-shot prompts and providing detailed explanations for the *correct* samples could help in mitigating bias and increasing the fidelity of the generated

[¶] https://huggingface.co/mistralai/Mixtral-8x7B-v0.1

data.

The potential use of our adapters to generate both positive and negative samples aimed at a specific target is indeed promising. By assembling datasets that offer these contrasting examples, we could refine the training process through contrastive fine-tuning. Such a method is hypothesized to yield superior performance by facilitating the model's ability to discern and learn from the nuanced differences between correct and incorrect instances.

References

Nikolay Babakov, David Dale, Varvara Logacheva, and Alexander Panchenko. 2022. A large-scale computational study of content preservation measures for text style transfer and paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 300–321, Dublin, Ireland. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. 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.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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. *CoRR*, abs/2311.05232.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

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.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2006–2029. Association for Computational Linguistics.

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, ACL 2020, Online, July 5-10, 2020*, pages 4885–4901. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. 2019. Natural language understanding with the quora question pairs dataset. *ArXiv*, abs/1907.01041.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A GPT-4 prompt for PG task evaluation

```
Read the source sentence and the paraphrased hypothesis and answer whether there are any hallucinations
or related observable overgeneration errors for the paraphrasing task.
Before answering, think step by step and write why you chose the answer you did.
Answer the last string with 'The hypothesis is correct' if there are no hallucinations or misgenerations.
Otherwise, answer with 'The hypothesis is false'.

Example 1:
Source sentence: "The European Parliament does not approve the budget."
Paraphrased hypothesis: "The budget cannot be adopted against the will of the European Parliament."
The hypothesis is false

Example 2:
Source sentence: "Everyone is capable of enjoying a good education in a society."
Paraphrased hypothesis: "We must create a society where everyone is able to enjoy a good education."
The hypothesis is correct
```

Figure 2: Prompt for GPT-4 evaluation on PG task.

B GPT-4 prompt for synthetic paraphrased data generation with hallucinations

```
Your aim is to produce an incorrectly paraphrased sentence that contains a hallucination for the given
source sentence. Hallucinations in a paraphrase can add new information that wasn't present in the
source sentence, or exclude some important information, or reverse the meaning of the source sentence.
Remember that reversing source sentence has the lowest level of priority, so use it only if there is no
other way to make a hallucination. Usually it's much better to misrepresent some information, add new
or exclude something important. If there is some quantitative information in the source, feel free to
change them slightly. Complete the task using the examples below. The examples also show the correct
paraphrase for the source sentences. Note that there are no hallucinations in the correct paraphrase,
whereas your aim is to corrupt the source and produce a false paraphrase.

Examples:
Source: "I have a permit."
The correct paraphrase: "Uh, I’m validated."
The incorrect paraphrase: "I have a permit to carry it."
Explanation: The incorrect paraphrase adds information that is not present in the source sentence ("to
carry it")

Source: "Easy, easy."
The correct paraphrase: "Watch it now."
The incorrect paraphrase: "The process is easy."
Explanation: The incorrect paraphrase introduces additional information ("The process is")

Source: "A five, six, seven, eight."
The correct paraphrase: "And 5, 6, 7, 8."
The incorrect paraphrase: "A number between five and eight."
Explanation: While the source sentence is a rhythmic count or sequence of specific numbers, the incorrect
paraphrase generalizes it to "a number between five and eight".

Source: "A lot safer that way."
The correct paraphrase: "Because it’s safer."
The incorrect paraphrase: "That is a safer way to travel."
Explanation: The major hallucination lies in the addition of "That is," which wasn't present in the
original source sentence. This introduces a new element and changes the focus from the general concept
of safety to a specific way of travel

Source: "You’re a scam artist."
The correct paraphrase: "You are an imposter."
The incorrect paraphrase: "You’re not a good scam artist."
Explanation: While the source sentence simply states "You’re a scam artist," the incorrect paraphrase
implies a judgment on the person's skill as a scam artist

Don't answer now, read the source and think step by step how to make a false paraphrase for the source
sentence. Before answering, provide several examples with explanations and choose the best one. Answer
starting with 'The incorrect paraphrase:
```

Figure 3: Prompt for PG data with hallucinations generation using GPT-4.

C Synthetic data examples

Task	Target	Hypothesis	Label
MT	I know you kissed Tom.	I know you're going to beat someone.	Hallucination
	I'll talk to Tom today.	I'll talk to John tomorrow.	Hallucination
	I should've waited.	I'll be better tomorrow.	Hallucination
	He was mistaken for his younger brother	He was taken for his brother.	Not Hallucination
	She passed the examination.	She successfully completed the certification.	Not Hallucination
	A plane popped up out of nowhere.	A plane just came out of the blue.	Not Hallucination
	No one has seen Tom since.	No one has seen Tom for a long time.	Not Hallucination
PG	Wanna talk?	Would you want to talk with me?	Not Hallucination
	We must not allow this to happen.	We must prevent this from happening.	Not Hallucination
	Have you found her?	Have you seen her?	Hallucination
	A word of advice.	Give me some advice.	Not Hallucination
	Hold your course.	You're going the wrong way, man.	Hallucination
	Can I take a message?	Can I take a message for you,	Not Hallucination
	My job?	My job is to carry out the trash.	Hallucination
DM	Delicious .	(scrambley) A scrambley dish.	Hallucination
	To increase the level or amount of .	To increase in volume.	Not Hallucination
	Causing the air to be hot .	Hot. Something that is hot.	Not Hallucination
	(slang, derogatory) schizoid, schizophrenic; crazy	(transitive) Crazy	Not Hallucination
	Covered with petals or petal-like objects.	planted.	Hallucination
	Alternative form of midstream	Middle stream	Not Hallucination
	To require	take time to finish something.	Hallucination

Table 9: Sample of synthetic data generated using LLaMA2-7B

Target	Hypothesis	Label
That cannot be in our interest!	It's not beneficial for us!	Not hallucination
The written language should be made more user-friendly.	The spoken language should be made more user-friendly.	Hallucination
I do not think that is quite what the agreement is.	I do not think that's the contract we signed.	Hallucination
The vote will take place tomorrow at 11.30 a.m.	Tomorrow, the voting process is scheduled for 11.30 in the morning.	Not hallucination
Mrs Green, you have the floor.	Mrs. Green, you own the flooring.	Hallucination
I was also in a northern industrial suburb in Milan.	I too have been to one of Milan's northern industrial neighborhoods.	Not hallucination
Mr President, I should like to make a further remark.	Mr. President, I would like to add another comment.	Not hallucination
Mrs Bonino tells me that no response is necessary.	Mrs. Bonino informed me a response isn't required.	Not hallucination

Table 10: Sample of synthetic data generated using GPT-4