LCTeam at SemEval-2025 Task 3: Multilingual Detection of Hallucinations and Overgeneration Mistakes Using XLM-RoBERTa

Araya Kiros Hailemariam[1] **and Jose Maldonado Rodriguez**[1] **and Ezgi Başar**[1] **and Roman Kovalev**[1] **and Hanna Shcharbakova**[2]
[1]University of Groningen
[2]Saarland University
{a.k.hailemariam,j.e.maldonado.rodriguez,e.basar,r.kovalev}@student.rug.nl
hash00004@stud.uni-saarland.de

Abstract

In recent years, the tendency of large language models to produce hallucinations has become an object of academic interest. Hallucinated or overgenerated outputs created by LLMs contain factual inaccuracies which can potentially invalidate textual coherence. The Mu-SHROOM shared task sets the goal of developing strategies for detecting hallucinated parts of LLM outputs in a multilingual context. We present an approach applicable across multiple languages, which incorporates the alignment of tokens and hard labels, as well as training a multi-lingual XLM-RoBERTa (Conneau, 2019) model. With this approach we managed to achieve 2nd in Chinese and top-10 positions in 7 other language tracks of the competition.

1 Introduction

In recent years, due to the development of transformer-based architectures (Vaswani, 2017), Natural Language Generation models saw immense advancements. However, the field is currently struggling with the tendency of neural systems to produce fluent, yet factually inaccurate outputs, aggravated by lack of adequate accuracy metrics. All of the above causes the models to "hallucinate".

The overgeneration of inaccurate facts puts in jeopardy the practical applications based on NLG (Mickus et al., 2024), which in turn prompts more interest in tackling the problem of detecting hallucinations in the outputs of NLG models.

The problems mentioned above were the motivation for the Mu-SHROOM shared task, aimed at identifying hallucinations and related overgeneration mistakes (Vázquez et al., 2025). The task builds upon ints previous iteration, but with focus on more languages and LLM outputs.

In this paper we present our approach to detecting hallucinated output using the multilingual XLM-RoBERTa model (Conneau et al., 2019). We use the model inputs and outputs from the provided dataset, which are concatenated and aligned with hard labels and then passed through the model. Apart from that, we also provide the description of the shared task and the datasets, as well as the discussion of results.

2 Related Work

In this section, we offer a short overview of methods used in previous work on detecting hallucinations in LLMs' outputs. We will examine model-agnostic and model-aware approaches, as well as black-box detection methods and prompting-based techniques.

Model-agnostic methods, such as prompt engineering and few-shot learning, focus on improving hallucination detection without relying on model internals. These techniques leverage strategies like meta-regression frameworks and automatic label generation, which allow them to be applied across different LLMs and tasks, offering a flexible solution to hallucination detection (Mehta et al., 2024; Chen et al., 2024; Allen et al., 2024; Arzt et al., 2024; Rykov et al., 2024).

On the other hand, model-aware approaches take advantage of internal signals from LLMs, such as layer activations and attention values, to detect hallucinations more precisely. By directly analyzing the model's internal workings, these methods can provide deeper insights into how LLMs generate outputs. Techniques like Retrieval-Augmented Generation (RAG) and chain-of-verification strategies have been explored to validate generated content against external knowledge sources, ensuring higher accuracy in detecting hallucinations (Liu et al., 2024; Varshney et al., 2023). However, these methods are generally less effective than model-agnostic approaches including (Mehta et al., 2024) and (Obiso et al., 2024) at SHROOM shared task. They are limited by their dependence on open ac-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1325–1331

cess to the model's internal architecture, which may not be possible for closed-source models like ChatGPT (Azaria and Mitchell, 2023).

Prompting-based metrics are also used in hallucination detection, leveraging the instruction following capabilities of LLMs. These methods involve providing LLMs with evaluation guidelines and both the generated and source content (Luo et al., 2023). Various strategies have been explored, including direct and chain-of-thought prompting, and in-context learning (Jain et al., 2023).

3 Task Description and Datasets

3.1 Task Description

The task presented by the organizers concerns the detection of hallucinated spans within a text. In contrast to the binary nature of SemEval-2024 Task 6, where whole texts were labeled as either containing or not containing hallucinations, this year's task concerns the dimension of detecting the position of hallucinations within the text.

This detection relies on two different types of labels, namely soft and hard labels, which are derived from a manual annotation process. Soft labels include all hypothesized spans along with their predicted probability of being an hallucination, whereas hard labels include only the spans that are decisively categorized as hallucination, as visualized in Table 1. In this example, only the last span in the soft labels is included in the hard labels due to achieving a probability higher than 0.5.

Participating teams are ranked based on their intersection-over-union (IoU) scores for hard labels while ties are broken using the Spearman correlation between predicted and true soft labels.

Model input	When was the Swedish Navy founded?
Model output	The Swedish navy was founded in 1625.
Soft labels	[{"start":1,"prob":0.0909090909,"end":18}, {"start":18,"prob":0.1818181818,"end":33}, {"start":33,"prob":1.0,"end":37}]
Hard labels	[[33,37]]

Table 1: An example of soft and hard labels.

3.2 Datasets

Task organizers supply participants with training, validation, and test datasets in multiple languages. Table 2 illustrates the number of samples for each language across the different dataset splits. Notably, the training set contains samples in English, French, Spanish, and Chinese while the validation set covers 6 additional languages. Basque, Catalan, Czech, and Farsi are test-only languages.

Language	Train	Validation	Test
Arabic	-	50	150
Basque	-	-	99
Catalan	-	-	100
Chinese	200	50	150
Czech	-	-	100
English	809	50	154
Farsi	-	-	100
Finnish	-	50	150
French	1850	50	150
German	-	50	150
Hindi	-	50	150
Italian	-	50	150
Spanish	492	50	152
Swedish	-	49	147

Table 2: Distribution of the training, validation, and test set samples for each language.

Without labeled data in the training set, teams are prompted to devise systems that do not rely on the availability of labeled hallucination data. Each data point in the training set includes the input prompt, the corresponding output text, the HuggingFace identifier of the model which has generated the output, the tokenized version of the output, and the logit values for the tokens.

The validation and test sets follow the same structure as the training data with the addition of providing hard and soft labels for the generated output. Hallucination labels are given at the character level, meaning that each output text may contain one or more hallucination spans. The spans were determined by human annotators with at least 3 people annotating each data point. Soft labels are based on probabilistic scores reflecting the level of consensus among annotators regarding hallucinated spans. These scores are calculated using the proportion of annotators who marked a given span as hallucinated and the resulting labels contain the start and end characters of sequences sharing the same score. Hard labels indicate spans which the majority of the annotators identified as hallucinated.

4 Methodology

Our system utilizes the validation dataset provided by the task organizers, which comprises ten languages. To ensure balanced representation across languages, we first split each language-specific dataset into training and validation subsets using a 90/10 ratio and then merged them to form combined training and validation sets. To address data scarcity, we implemented a cross-lingual label transfer mechanism based on neural machine translation and semantic span alignment. We employed Helsinki-NLP's Opus-MT bilingual models (Tiedemann and Thottingal, 2020) to translate labeled text from a source language into target languages.

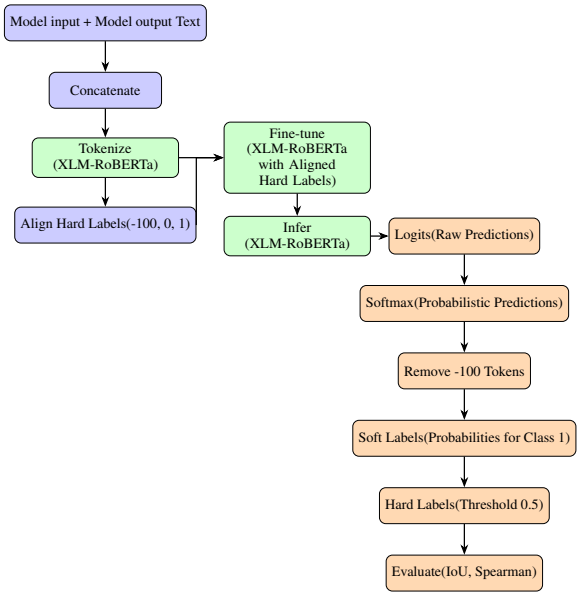

Figure 1: Hallucination detection flowchart.

For accurate projection of labeled spans (hard labels) onto the translated text, we adopted a semantic similarity-based alignment approach. We used the `paraphrase-multilingual-MiniLM-L12-v2` model from Sentence Transformers (Reimers and Gurevych, 2020) to generate embeddings for each labeled span in the source text, as well as for candidate spans in the translated output. To accommodate natural variations in translation length, we introduced a variational span matching strategy: for a source span of length l, we evaluated candidate spans in the translated text with lengths in the range $[l-\delta, , l+\delta]$, where δ is a tunable parameter (set to 5 in our experiments). Cosine similarity between embeddings was used to identify the most semantically aligned span, which was then adjusted to remove any leading or trailing whitespace, ensuring precise label transfer. The resulting multilingual dataset includes model input text, model output text, hard labels, and relevant metadata. We focus on three key components: the model input text, which provides contextual grounding; the model output text, which is analyzed for hallucinations and overgenerations; and the hard labels, which annotate hallucinated or overgenerated spans for use in evaluation and supervision. The overall architecture of our system is depicted in Figure 1.

The flowchart outlines the entire process, starting from concatenating model input and output text, followed by tokenization using XLM-RoBERTa (Conneau, 2019). The next steps involve hard label token alignment, model training, and inference. Post-inference, the logits are processed to calculate soft labels, followed by thresholding to obtain hard labels. Finally, we evaluate performance using the Intersection over Union (IoU) and Spearman correlation metrics as provided by the task organizers.

4.1 Preprocessing Pipeline

The first step of preprocessing is concatenating the model input text with the model output text to form a unified sequence. This ensures that the model retains the necessary context from the input while focusing on evaluating the generated output. The concatenated sequence is then tokenized into tokens or subword units using the model's tokenizer. After tokenization, token-hard-label alignment is performed to map the hard labels onto the tokenized sequence.

Hard labels are provided as nested lists, with each list specifying the start and end indices of hallucinated or overgenerated spans within the model output text. To align these spans with the tokenized sequence, offset mappings are generated for each token, indicating their start and end positions within the concatenated text. Initially, all tokens are assigned a label of -100, marking those that do not belong to the model output text, such as tokens from the model input context and special tokens (e.g., [CLS], [SEP], and [PAD]). The model input text is tokenized separately to determine where the tokens from the model output begin in the concatenated sequence. The start of the model output tokens is identified as the position immediately following the last token of the model input (excluding the special separator token).

To align the hard labels with the tokenized sequence, the model input text length is added to both the start and end indices of the hard-labeled spans. Tokens overlapping these adjusted spans are

labeled as 1 (hallucination), while those outside are labeled 0. Tokens labeled -100 are excluded from the loss calculation, ensuring the model focuses on the relevant output tokens during training. However, tokens labeled -100 still contribute to the context and are used by the model to make predictions for other tokens. This label indicates that these tokens do not affect the loss calculation, without diminishing their role in the model's contextual understanding. For testing, where hard labels are unavailable, all model output tokens are initialized only as 0 or -100. This setup helps distinguish model output tokens from others during prediction.

4.2 Model Training and Inference

We fine-tuned the XLM-RoBERTa model using the aligned hard labels in a supervised learning framework. During training, the model learns to identify hallucinated or overgenerated spans based on the provided hard labels for the model output text, while also using the contextual information from the model input text. This approach helps the model distinguish between hallucinated and non-hallucinated spans at the token level.

During inference, the model's logits are passed through a softmax activation function to generate probabilistic predictions, indicating the likelihood of each token being classified as Class 1 (hallucinated or overgenerated) or Class 0 (neither). Tokens labeled -100 are excluded to focus on model output tokens. Class 1 probabilities are aggregated for soft label evaluation, and a 0.5 threshold is applied to derive binary hard labels. Finally, predictions are evaluated using the mean Intersection over Union (IoU) and Spearman correlation values for each language's test set.

5 Experiments and Results

5.1 Experiments

To provide a more comprehensive evaluation of our model's performance, we present post-submission test results alongside the best IoU scores from the official leaderboard for comparison. While these results were not obtained during the submission window, they reflect improvements that were achieved during our subsequent experiments. The post-submission setting only uses the original validation data, as opposed to transferring labels from the data available in other languages.

Model configuration details for both the submission and post-submission phases are provided in Appendix A. Our official submission used the base variant of XLM-RoBERTa, while post-submission experiments used the larger version for further testing and validation. Training and validation batch sizes were adjusted accordingly to optimize the training process. Notably, the training batch size was increased from 18 to 26 and the validation batch size was increased from 8 to 16 during the post-submission phase.

5.2 Results

Our results (Table 3 and Table 4) show the effectiveness of label alignment in hallucinated span detection across various languages. Leveraging the context provided by the prompt in the fine-tuning stage seems to effectively guide our inference model in the detection of hallucinated spans. With comparable results throughout both of the measured metrics, we do not identify a preference in our system's capability of predicting soft or hard labels.

Language	IoU	ρ	Ranking
Arabic	0.5335	0.5537	11/29
Basque	0.4804	0.5499	13/23
Catalan	0.4924	0.4917	12/21
Chinese	0.5232	0.5171	2/26
Czech	0.4051	0.4357	9/23
English	0.4725	0.5538	16/41
Farsi	0.6018	0.4559	8/23
Finnish	0.4221	0.5300	21/27
French	0.5634	0.4883	10/30
German	0.5634	0.5031	8/28
Hindi	0.6601	0.5122	6/24
Italian	0.7013	0.5487	9/28
Spanish	0.4434	0.4335	7/32
Swedish	0.4183	0.3700	17/27

Table 3: Our best submission results for each language using IoU and Spearman correlation metrics along with our team ranking on the official leaderboard.

For the official submission, we applied label transfer to the target language and utilized validation data from other languages in addition to the original validation set. Post-submission results were obtained using the original validation data alone. As a result, the IoU scores increased for most languages compared to the official submission. The largest improvements were observed in Finnish (0.4221 - 0.6127), Swedish (0.4183 - 0.5728), and Catalan (0.4924 - 0.5532). Notably, Spanish and French exhibited significant drops in

Language	IoU	ρ	Reference IoU
Arabic	0.5660	0.5595	0.6700
Basque	0.4801	0.5285	0.6129
Catalan	0.5532	0.4934	0.7211
Chinese	0.5412	0.5488	0.5540
Czech	0.4189	0.4252	0.5429
English	0.4707	0.5472	0.6509
Farsi	0.6501	0.4713	0.7110
Finnish	0.6127	0.5936	0.6483
French	0.5048	0.4924	0.6469
German	0.5694	0.5694	0.6236
Hindi	0.6771	0.5004	0.7466
Italian	0.7201	0.5478	0.7872
Spanish	0.3832	0.4417	0.5311
Swedish	0.5728	0.4848	0.6423

Table 4: Results obtained after the end of the submission period alongside the highest IoU scores from the official leaderboard for reference.

performance. These findings suggest that training without label transfer generally improves performance across languages and that label transfer does not provide a clear advantage in most cases.

5.3 Discussion

Our observations indicate that our system has a tendency to predict a large number of short spans while the reference annotations contain a relatively small number of longer spans. This pattern is consistent across multiple languages. For example, Chinese test data contains an average of 10.58 spans with a mean span length of 34.21, while our Chinese model predicts 137.69 spans with an average length of 1.57.

Table 5 compares model predictions with gold annotations, revealing several behavioral patterns. In the first example, our model produces three different hallucination spans for the same entity, while missing some characters in the middle and at the end. Different tokenizers could lead to different outcomes and we think that experimenting with tokenization configurations could be beneficial. In the second example, our model labels an unrelated subword as hallucination. The following example in Italian demonstrates our model assigning two labels to the same span and failing to identify a full word as hallucination, only labeling the subword. The following German example illustrates how the model falsely produces multiple labels and partially misses a city name. In the last English example, our model labels each year individually rather than coming up with a single label for the range. It also fails to identify a whole another hallucinated span. These findings reveal that our model is still prone to both under- and over-prediction.

6 Conclusion

In this paper, we present a system for the detection of hallucinated spans in text. Our approach successfully applies the small amount of labeled data provided by the task organizers to fine-tune a multilingual XLM-RoBERTa model to this end. By aligning the tokens in a concatenated sequence including both the prompt and its resulting model output to the provided hard labels representing the start and end of a hallucinated span, we are able to leverage the context that the prompt provides at the same time that we make use of the available labeled data.

This system however has some limitations, such as its reliance on larger models for improved performance and its tendency to predict a larger number of spans which are shorter in length than desirable.

Annotated sample	Model prediction
Mouthier is located in the department of Haute-Loire .	Mouthier is located in the department of Haut e - Lo ire.
The Emdin light cruisers were built in the shipyards of the German Navy in Kiel , Germany.	The Emdin light cruisers were built in the shipyards of the German Nav y in Kiel , Germany.
Lo smorzatore presente nella torre 111 West 57th Street pesa circa 2.000 libbre .	Lo smorzatore presente nella torre 111 West 57th Street pesa circa 2.000 lib bre.
John Christopher Willies wurde am 10. April 1887 in der Stadt New York City geboren.	John Christopher Willies wurde am 10. April 1887 in der Stadt New York City geboren.
The Empressa Ferrocarril do Alem Pará was in service from 1956 to 1974 .	The Empressa Ferrocarril do Alem Pará was in service from 1956 to 1974 .

Table 5: Labeled sentences from the English, Italian, and German test sets alongside model predictions. Hallucination spans are highlighted in different colors.

Acknowledgements

We would like to thank our professor Malvina Nissim for making our participation in this shared task possible, as well as for her guidance and support throughout the entire process.

As recipients of an Erasmus Mundus scholarship, we would also like to acknowledge that this work was co-funded by the Erasmus Mundus Masters Program in Language and Communication Technologies (LCT), EU grant no. 2019-1508.

References

Bradley P Allen, Fina Polat, and Paul Groth. 2024. Shroom-indelab at semeval-2024 task 6: Zero-and few-shot llm-based classification for hallucination detection. *arXiv preprint arXiv:2404.03732*.

Varvara Arzt, Mohammad Mahdi Azarbeik, Ilya Lasy, Tilman Kerl, and Gábor Recski. 2024. Tu wien at semeval-2024 task 6: Unifying model-agnostic and model-aware techniques for hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1183–1196.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Ze Chen, Chengcheng Wei, Songtan Fang, Jiarong He, and Max Gao. 2024. Opdai at semeval-2024 task 6: Small llms can accelerate hallucination detection with weakly supervised data. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 721–729.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8487–8495, Toronto, Canada. Association for Computational Linguistics.

Wei Liu, Wanyao Shi, Zijian Zhang, and Hui Huang. 2024. Hit-mi&t lab at semeval-2024 task 6: Deberta-based entailment model is a reliable hallucination detector. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1788–1797.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models.

Rahul Mehta, Andrew Hoblitzell, Jack O'Keefe, Hyeju Jang, and Vasudeva Varma. 2024. Metacheckgpt–a multi-task hallucination detection using llm uncertainty and meta-models. *arXiv preprint arXiv:2404.06948*.

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. *arXiv preprint arXiv:2403.07726*.

Timothy Obiso, Jingxuan Tu, and James Pustejovsky. 2024. Harmonee at semeval-2024 task 6: Tuning-based approaches to hallucination recognition. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1322–1331.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Elisei Rykov, Yana Shishkina, Kseniia Petrushina, Kseniia Titova, Sergey Petrakov, and Alexander Panchenko. 2024. Smurfcat at semeval-2024 task 6: Leveraging synthetic data for hallucination detection. *arXiv preprint arXiv:2404.06137*.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Annual Conference of the European Association for Machine Translation*, pages 479–480. European Association for Machine Translation.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona de Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 Task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes.

A Model Configurations

	Model configuration	Value
Official submission	Model name	xlm-roberta-base
	Training batch size	18
	Validation batch size	8
	Learning rate	5e-5
	Number of epochs	10
	Metric for best model	IoU mean
	Max sequence length	512
Post-submission	Model name	xlm-roberta-large
	Training batch size	26
	Validation batch size	16
	Learning rate	5e-5
	Number of epochs	10
	Metric for best model	IoU mean
	Max sequence length	512

Table 6: Hyperparameters used for the official submission and post-submission experiments.