Howard University - AI4PC at SemEval-2025 Task 3: Logit-based Supervised Token Classification for Multilingual Hallucination Span Identification Using XGBOD

Saurav K. Aryal and **Mildness Akomoize**
EECS, Howard University
Washington, DC, USA
saurav.aryal@howard.edu

Abstract

This paper describes our system for SemEval-2025 Task 3, Mu-SHROOM, which focuses on detecting hallucination spans in multilingual LLM outputs. We reframe hallucination detection as a point-wise anomaly detection problem by treating logits as time-series data. Our approach extracts features from token-level logits, addresses class imbalance with SMOTE, and trains an XGBOD model for probabilistic character-level predictions. Our system, which relies solely on information derived from the logits and token offsets (using pretrained tokenizers), achieves competitive intersection-over-union (IoU) and correlation scores on the validation and test set.

1 Introduction

SemEval-2025 Task 3, Multilingual Shared-task on Hallucinations and Related Observable Overgeneration Mistakes (Vázquez et al., 2025) addresses the critical challenge of detecting hallucinations in instruction-tuned Large Language Model (LLM) outputs. The challenge of hallucination extends beyond text-based LLMs to multimodal large language models (MLLMs) as well, posing significant obstacles to their real-world applications (BAI et al., 2025). This task is crucial for ensuring the reliability and trustworthiness of LLMs in real-world applications, especially in multilingual contexts. Mu-SHROOM encompasses 14 languages: Arabic, Basque, Catalan, Chinese, Czech, English, Farsi, Finnish, French, German, Hindi, Italian, Spanish, and Swedish, reflecting the growing need for robust multilingual LLM evaluation (Ji et al., 2022).

Our system tackles this span detection task by framing it as a point-wise anomaly detection problem. We hypothesize that hallucinated text spans exhibit anomalous patterns in the LLM's output logits compared to factual or consistent text. Inspired by recent work demonstrating the potential of Large Language Models for time series anomaly detection (Liu et al., 2024), we leverage the XGBOD (eXtreme Gradient Boosting for Outlier Detection) algorithm (Zhao, 2019), trained on features extracted directly from the LLM's logit sequences, to identify these anomalous points indicative of hallucinations. By participating in Mu-SHROOM, we discovered that a relatively simple, data-driven anomaly detection approach can achieve effective hallucination span detection across diverse languages, relying solely on model logits without prompts, text, tokens, or Retrieval-Augmented Generation (RAG).

2 Background

Recent work has also explored the use of LLMs directly for time series anomaly detection. (Liu et al., 2024) proposed LLMAD, a framework that uses LLMs for few-shot anomaly detection in time series, achieving both high accuracy and interpretability. Their work, while focused on general time series data, further motivates our exploration of anomaly detection techniques for hallucination detection in LLM text output, particularly by leveraging the LLM's own logit representations. The survey by (Luo et al., 2024) provides comprehensive overviews of various hallucination detection techniques. The field of time series anomaly detection itself is a well-established area, with extensive research into various methodologies, as highlighted in comprehensive surveys by (BLÁZQUEZ-GARCÍA et al., 2020; DARBAN et al., 2024). These surveys cover a wide range of techniques, including deep learning approaches, and discuss applications across diverse domains. Anomaly detection techniques have been successfully applied to various time-series data domains, such as system log analysis (Du et al., 2017).

The task input consists of:

1. `model_input` (instruction prompt)

2. `model_output_text` (LLM generated text)

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1790–1794

3. `model_output_logits` (logit values for each token in the output)

4. `model_output_tokens`

5. Human annotations in the form of `soft_labels` (probabilistic hallucination spans) and `hard_labels` (definite hallucination spans)

The expected output from participating systems is, for each character in the `model_output_text`, a probability indicating whether it is part of a hallucination span.

The dataset is split into Sample, Validation, Unlabeled Train, Unlabeled Test, and Labeled Test sets. We utilized the Validation set for training our anomaly detection model and the Unlabeled Test set for evaluation. The dataset covers 14 languages and utilizes outputs from various public-weight LLMs. Our submission focused on span detection across all 14 languages using a single model.

3 System Overview

Our system employs a three-stage process: feature extraction from logits, anomaly detection using XGBOD and prediction of anomaly scores for each token.

3.1 Feature Extraction

Our system extracts six features for each token in the LLM's output, including the raw logit value, its normalized position in the sequence, and the logit difference from the previous token, all derived from the `model_output_logits` sequence. We hypothesize that tokens within hallucinated spans will display distinctly anomalous logit patterns compared to tokens in non-hallucinated spans. These features are designed to capture both the individual token's logit behavior and its context within the overall logit sequence.

3.2 Anomaly Detection with XGBOD

We employ XGBOD, an efficient and effective outlier detection algorithm based on eXtreme Gradient Boosting (XGBoost), for anomaly detection. We chose XGBOD for its ability to handle complex feature interactions and robustness against data imbalance (Zhao, 2019).

The model is trained on the labeled validation set. SMOTE is used to oversample the minority class and the labels provided in the validation set are used as the ground truth for anomaly/non-anomaly classification at the token level. Specifically, if a token's character span overlaps with any hallucination span, it's labeled as anomalous (1), otherwise non-anomalous (0).

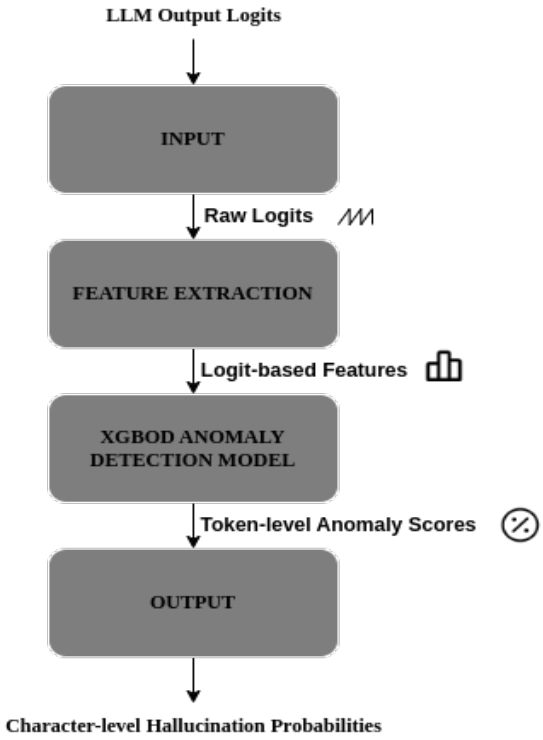

Figure 1: Overview of the system pipeline for hallucination span identification, illustrating data flow from LLM output logits to character-level probability predictions.

3.3 Inference and Prediction

For the unlabeled test set, we apply the trained XGBOD model to predict anomaly scores for each token. The process mirrors the feature extraction step used in training. For each instance in the test set:

1. Extract logits and model output text.

2. Tokenize the text to obtain token offsets.

3. Extract the same six logit-based features for each token as during training.

4. Use the trained XGBOD model to predict an anomaly score (probability) for each token.

5. Map token-level anomaly scores back to character-level probabilities.

The design choices for our system is motivated by several key considerations. First, we leverage

LLM logits as a rich internal representation, reflecting the model's predictive probabilities, confidence, and uncertainty. While hallucinations can sometimes present as linguistically plausible text, we hypothesize that these instances may still correspond to deviations from the logit patterns typically observed during factual or consistent generation. We are not necessarily looking for individual "surprising" or out-of-distribution logit values, but rather for subtle shifts in the distribution, sequence, or relationships among logits, which we aim to capture through our extracted features. This lightweight design translates to significantly reduced compute requirements for both model training and inference compared to large transformer models or methods involving extensive external knowledge bases. Training our XGBOD model is substantially faster and less resource-intensive, enabling efficient processing and potential suitability for real-time hallucination detection. This data-driven methodology, training an anomaly detection model on the validation set, allows us to learn hallucination patterns directly from the data, rather than relying on heuristics. Finally, the language-agnostic nature of logit-based features enables multilingual applicability. Drawing inspiration from the successful use of anomaly detection in time-series data (Du et al., 2017) and the recent application of LLMs to time-series anomaly detection (Liu et al., 2024), our system provides a targeted and efficient solution for multilingual hallucination span identification while utilizing a distinct gradient boosting model and focusing specifically on text hallucination.

4 Experimental Setup

4.1 Data Splits and Preprocessing

We used the validation set provided by the Mu-SHROOM task organizers to train our XGBOD model. The unlabeled test set was used to generate our predictions for the competition. We did not use any additional data or external resources beyond the provided datasets and pre-trained tokenizers.

The preprocessing steps included the following:

- Tokenization: For each language and `model_id`, we used the corresponding Hugging Face Transformers tokenizer (AutoTokenizer) (Wolf et al., 2019) to obtain token offsets for feature extraction and label alignment.

- Feature extraction: As described in Section 3.1, we extracted six logit-based features for each token.

- Label Generation: The validation set la WObels were used to generate token-level anomaly labels (0 or 1) as described in Section 3.2.

- Addressing Class Imbalance: Due to the imbalanced nature of the data, we employed SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al., 2002) to oversample the minority class in the training set, improving model robustness.

4.2 Hyperparameter Tuning

We performed hyperparameter tuning for the XGBOD model using GridSearchCV with 3-fold cross-validation on the resampled training data (after SMOTE).

Parameter	Search Space
n_estimators	{100, 200, 300}
max_depth	{3, 5, 7}
learning_rate	{0.01, 0.05, 0.1}

Table 1: Hyperparameter search space for XGBOD using GridSearchCV.

4.3 Evaluation Metrics

The Mu-SHROOM task evaluates system performance using two character-level metrics:

1. Intersection-over-Union (IoU): Measures the overlap between predicted and gold hallucination spans.

2. Correlation (Cor): Measures the correlation between the system's predicted hallucination probabilities and the empirical probabilities derived from human annotations.

5 Results

As shown in Table 1, our system shows a significant improvement in intersection-over-union (IoU) scores in most languages compared to the baseline. In particular, for Arabic (AR), Spanish (ES), Finnish (FI), French (FR), and Italian (IT), our system achieves IoU scores that are substantially higher than the baseline. This indicates that our anomaly detection approach is considerably more effective in identifying and accurately delineating hallucination spans in these languages.

Moving to Correlation (Cor) scores, the comparison is more nuanced. our system generally achieves competitive or superior correlation scores, indicating a better alignment between our predicted hallucination probabilities and the human-annotated

Lang	Id	Metrics	
		IoU	Cor
AR	XGBOD	0.2138	0.3844
	Baseline	0.0001	0.2235
DE	XGBOD	0.2522	0.2763
	Baseline	0.2716	0.1288
EN	XGBOD	0.1325	0.2751
	Baseline	0.0802	0.3061
ES	XGBOD	0.1341	0.3642
	Baseline	0.0715	0.0774
FI	XGBOD	0.3996	0.3432
	Baseline	0.0843	0.2625
FR	XGBOD	0.4164	0.3990
	Baseline	0.1130	0.0911
HI	XGBOD	0.2586	0.3216
	Baseline	0.2421	0.1452
IT	XGBOD	0.2675	0.4020
	Baseline	0.0010	0.2004
SV	XGBOD	0.1109	0.0668
	Baseline	0.1893	0.1696
ZH	XGBOD	0.2152	0.1119
	Baseline	0.0776	0.1502

Table 2: Comparison of our system and baseline results on the test set (IoU and Cor)

probabilities. For languages like Arabic, Spanish, French, and Italian, our system exhibits higher correlation values. However, for English and Swedish, the baseline shows slightly higher correlation scores, suggesting it might be somewhat better at ranking the likelihood of hallucination at the character level in these languages, even if its span identification (IoU) is weaker.

Considering both IoU and Correlation metrics, our system presents a significant imporvement over the provided baseline, particularly in its ability to better identify hallucination spans (as reflected by the IoU metric) across a wide range of languages. While the baseline shows some comparable results in specific languages in terms of correlation, our anomaly detection approach provides a more robust and generally superior language-agnostic solution for the Mu-SHROOM task, especially for span detection.

6 Conclusion

In this paper, we presented our system for SemEval-2025 Task 3: Mu-SHROOM. Our approach reframes multilingual hallucination span detection as a point-wise anomaly detection problem on LLM output logits, utilizing the XGBOD algorithm (Zhao, 2019). Experimental results demonstrate the effectiveness of this simple yet powerful approach, achieving competitive performance across 14 diverse languages. While our system shows promising results, we acknowledge several limitations and directions for future work. A key limitation is that our point-wise anomaly detection with XGBOD, while effective, does not explicitly model the temporal dependencies within the logit sequence. Furthermore, accurately calculating token offsets proved challenging across diverse models due to varying tokenizer support, sometimes necessitating reliance on less precise string parsing approaches. Future research will explore directly learning from these temporal dependencies by treating logit sequences as time series, potentially using sequence models within frameworks like Ludwig. We are also actively investigating integrating Retrieval-Augmented Generation (RAG) and textual information to provide richer context for hallucination detection. This includes exploring a multimodal approach to leverage diverse data sources. Addressing the inherent data scarcity and the challenges of obtaining consistent human labels across multiple languages and models remains a crucial long-term goal for advancing research in this domain. We believe that further exploration of these directions will lead to more robust and accurate multilingual hallucination detection systems.

Acknowledgement

This research project was supported in part by the Office of Naval Research grant N00014-22-1-2714. The work is solely the responsibility of the authors and does not necessarily represent the official view of the Office of Naval Research.

References

ZECHEN BAI, PICHAO WANG, TIANJUN XIAO, TONG HE, ZONGBO HAN, ZHENG ZHANG, and MIKE ZHENG SHOU. 2025. Hallucination of multimodal large language models: A survey. *arXiv*.

ANE BLÁZQUEZ-GARCÍA, ANGEL CONDE, USUE MORI, and JOSE A. LOZANO. 2020. A review on outlier/anomaly detection in time series data. *arXiv*.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16.

ZAHRA ZAMANZADEH DARBAN, GEOFFREY I. WEBB, SHIRUI PAN, CHARU C. AGGARWAL,

and MAHSA SALEHI. 2024. Deep learning for time series anomaly detection: A survey. *arXiv*.

Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. 2017. Deeplog: Anomaly detection and diagnosis from system logs through deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1285–1298, Dallas, TX, USA. ACM.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *arXiv*.

Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2024. Large language models can deliver accurate and interpretable time series anomaly detection. *arXiv*.

Junliang Luo, Tianyu Li, Di Wu, Michael Jenkin, Steve Liu, and Gregory Dudek. 2024. Hallucination detection and hallucination mitigation: An investigation. *arXiv*.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yue Zhao. 2019. Xgbod: Improving supervised outlier detection with unsupervised representation learning. *arXiv preprint arXiv:1912.00290*.