BEARCAT at #SMM4H-HeaRD 2025 Task 1: No Training, No Problem - Multilingual ADE Detection

Ziqi Guo[1], **Robert Palermo**[2], **Luis M. Rocha**[*1,3]
rocha@binghamton.edu
[1]School of Systems Science and Industrial Engineering, Binghamton University, Binghamton, NY, USA.
[2]Department of Mathematics and Statistics, Binghamton University, Binghamton, NY, USA.
[3]Universidade Católica Portuguesa, Católica Biomedical Research Centre, Lisbon, Portugal.

Abstract

Adverse Drug Event (ADE) detection from user-generated content on social media offers critical real-world pharmacovigilance insights, especially when performed across multiple languages. This paper presents a zero-shot, prompt-based approach using GPT-4o for the SMM4H-HeaRD 2025 Shared Task 1: binary classification of ADE presence in posts written in English, German, French, and Russian. While our English-language approach is detailed in a separate submission titled *Uncovering Underreported Adverse Drug Reactions in Epilepsy Communities via Interpretable Social Media Mining*, this work focuses on the remaining three languages. Without using any model training, we collaboratively refined a multilingual prompt to classify batches of posts. We observed that GPT-4o achieves high recall across languages but tends to over-predict ADEs, resulting in lower precision. A rerun of the model on only predicted positives in smaller batches significantly improved F1 scores. This study highlights GPT-4o's utility for zero-shot multilingual ADE detection and suggests prompt complexity and batch size as critical parameters in zero-shot inference performance.

Introduction

The SMM4H-2025 Task 1, as described in the overview (Klein et al. 2025), aims to develop a system that predicts whether the post contains a mention of an Adverse Drug Event (ADE). The task is framed as a binary classification problem, where the system outputs 1 if an ADE is mentioned and 0 otherwise.

Existing approaches used rule-based techniques and feature-engineered classifiers such as support vector machines and LSTMs (Nikfarjam 2015), which were later surpassed by pretrained language models. Transformer-based approaches, especially BERT (Devlin et al. 2019), BioBERT (Lee 2020), and BERTweet (Nguyen 2020), became standard due to their superior performance on noisy, informal health text.

To address multilingual settings, models such as multilingual BERT (mBERT) and XLM-RoBERTa (Conneau 2020) have been employed, enabling transfer learning across languages. Researchers have also explored machine translation to augment training data (Klein 2020), domain-adaptive pretraining (Gururangan 2020), and multimodal fusion using drug embeddings (Sakhovskiy and Tutubalina 2022).

Recently, large language models (LLMs) such as GPT-4 have shown promising results due to its multilingual understanding, domain generalization, and strong reasoning abilities, even in zero-shot and few-shot learning scenarios. It can follow the instruction through prompt-based interactions and perform tasks without additional training.

In this study, we explore the ChatGPT (GPT-4o) (OpenAI 2024) for ADE classification on different languages. We develop and iteratively refine a prompt that classifies a list of posts into ADE-present or ADE-absent. Without training on any labeled dataset, we perform two-stage inference, firstly processing large batches, and then rerunning only the positive predictions in smaller batches to reduce false positives—making the process significantly more cost-effective than applying the prompt to each post individually. Our results demonstrate that the second stage reruning provide better prediction result, and ChatGPT can achieve competitive performance in a zero-shot setting and highlights its viability for multilingual health informatics task.

Methods

In this work, we explored the use of GPT-4o in a prompt-based zero-shot setup to perform ADE classification across German (Raithel et al. 2022), French (Raithel et al. 2024), and Russian (Tutubalina et al. 2020; Magge et al. 2021) social media content. We approached this task with a zero-shot philosophy, intentionally avoiding training on any labeled dataset to prevent domain overfitting and preserve generalization. Instead, we iteratively engineered a multilingual prompt using GPT-4o to predict whether a post mentions an ADE.

We split the test data by language (German, French, Russian) and developed prompts using the validation set solely for prompt refinement—not as labeled examples. Prompt design included sentence fragments, example phrases, and semantic cues (e.g.,"caused", "after taking") indicative of ADEs. We partitioned the posts into batches to accommodate the prompt length constraints: 75 posts for German and French, 100 for Russian (owing to shorter average post lengths). GPT-4o returned binary labels (0/1) for each post ID in a numbered list. We then concatenated batch outputs to generate full prediction tables for each language.

To mitigate the model's tendency to generate false positives, we introduced a second-stage refinement. Posts predicted as ADE-positive were rerun in smaller batches (5 for German and French, 20 for Russian). The new outputs were merged with the initial predictions, retaining only confirmed positives. The final prompt used for ADE classification is provided below.

Final Prompt Used for ADE Classification

Definition of an Adverse Drug Reaction (ADR):
An ADR is a harmful or unintended physical or psychological effect that a user personally experiences after taking a drug, and where there is either a clear statement or a strong implication that the drug caused or contributed to the negative effect.
Label a post as 1 (ADR present) only if all of the following are true: The user states or clearly implies they have taken a drug. The drug may be named or unnamed, but it must be evident that the user personally took it. The user describes one or more negative or harmful symptoms or effects that occurred after taking the drug. These effects can be physical (e.g., nausea, rash, dizziness) or psychological (e.g., anxiety, mood changes, insomnia). There is a causal connection between the drug and the symptoms, either explicitly stated or implied through timing or language. Phrases indicating a link might include: *"after I started," "since taking," "because of,"* or any statement that suggests the drug led to the reaction.
Label a post as 0 (No ADR) if any of the following are true: The user does not mention taking a drug. The user describes symptoms but does not link them to any drug or suggests a different cause. The user discusses a drug in general terms, hypothetical scenarios, or someone else's experience. The user describes mild or expected effects without distress or concern (e.g., "I felt a little sleepy, but that's normal"). The described symptoms are likely due to other causes such as illness, withdrawal, diet, or stress, and there is no clear link to the drug.
Special Considerations:
Uncertainty or speculation (e.g., "not sure if it's the medication") should be counted as ADR if the timing and symptoms align and the user shows reasonable suspicion. Vague or emotional language should be interpreted carefully. Posts that express frustration or distress may describe ADRs even without clinical language. Posts referencing treatments, therapies, or "this medicine" may count as ADRs if it is clear the user is referring to a drug they personally took and describes negative effects.
Instructions:
Your task is to read each post carefully and assign 1 if the post meets all ADR conditions, or 0 if it does not. Each post comes from a user-generated discussion board. Language may be informal or ambiguous—use careful judgment based on the criteria above.

Results

Table 1 shows the results of our first-pass predictions across languages. As anticipated, GPT-4o achieved high recall but lower precision, particularly in the French and German validation sets. This likely stems from the model's cautious strategy to avoid missing true positives, especially when processing large batches of 75–100 posts. The increased prompt complexity in these batches may have introduced ambiguity, contributing to the high false positive rate.

		Precision	**Recall**	**F1 Score**
ru	Validation	0.4528	0.6667	0.5393
	Test	0.4551	0.6237	0.5262
de	Validation	0.2979	0.8750	0.4444
	Test	0.5714	0.8000	0.6667
fr	Validation	0.4286	0.9310	0.5870
	Test	0.4914	0.8190	0.6143

Table 1: Performance of the initial zero-shot classification across Russian (ru), German (de), and French (fr).

To mitigate this, Table 2 presents results from a second-stage refinement, where only ADE-positive posts were rerun in smaller batches (5 for German and French, 20 for Russian). This significantly improved precision and F1 scores, with minimal impact on recall—for example, German precision rose from 0.5714 to 0.7619, and French F1 improved from 0.6143 to 0.7072. Additionally, our approach performed well compared to the mean and median results of all team submissions for Task 1 (see Table 1 in the Supplementary Materials), achieving strong performance in the majority of cases. These results highlight the influence of prompt complexity and batch size on classification accuracy.

		Precision	**Recall**	**F1 Score**
ru	Validation	0.7826	0.5000	0.6102
	Test	0.6336	0.3820	0.4767
de	Validation	0.6667	0.6250	0.6452
	Test	0.7619	0.6095	0.6772
fr	Validation	0.7308	0.6552	0.6909
	Test	0.8421	0.6095	0.7072

Table 2: Performance after applying second-stage refinement, where posts initially labeled as ADE-positive were reprocessed in smaller batches to confirm predictions.

Discussion

This study demonstrates that GPT-4o, when guided by a well-designed prompt, can perform multilingual ADE classification effectively in a zero-shot setting. The model achieved high recall but tended to overpredict ADEs, especially when processing large batches. Our two-stage refinement—re-evaluating initial positives in smaller batches—substantially improved precision and F1 scores, highlighting the impact of prompt complexity and batch size on performance.

Future work should explore the relationship between batch size and sentence length to better understand how input structure affects accuracy. Rerunning negative or mixed cases could offer insights into the model's consistency, and experimenting with different prompt styles may reveal biases or sensitivities to phrasing. Although the zero-shot approach offers scalability, its dependence on prompt design and lack of interpretability remain challenges. Enhancing prompt strategies or incorporating lightweight supervision could further improve reliability for real-world applications.

References

Conneau, A. e. a. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*.

Gururangan, S. e. a. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of ACL*.

Klein, A. Z.; Dasgupta, T.; Flores Amaro, I.; Jana, S.; Khademi, S.; Lopez-Garcia, G.; Onishi, T.; Powell, J.; Raithel, L.; Rajwal, S.; Roller, R.; Sarker, A.; Sinha, M.; Thomas, P.; Tutubalina, E.; Xu, D.; Zweigenbaum, P.; and Gonzalez-Hernandez, G. 2025. Overview of the 10th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ICWSM 2025. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. AAAI Press.

Klein, A. Z. e. a. 2020. Overview of the fifth social media mining for health applications shared task. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop*.

Lee, J. e. a. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4).

Magge, A.; Klein, A.; Miranda-Escalada, A.; Ali Al-Garadi, M.; Alimova, I.; Miftahutdinov, Z.; Farre, E.; Lima López, S.; Flores, I.; O'Connor, K.; Weissenbacher, D.; Tutubalina, E.; Sarker, A.; Banda, J.; Krallinger, M.; and Gonzalez-Hernandez, G. 2021. Overview of the Sixth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at NAACL 2021. In Magge, A.; Klein, A.; Miranda-Escalada, A.; Al-garadi, M. A.; Alimova, I.; Miftahutdinov, Z.; Farre-Maduell, E.; Lopez, S. L.; Flores, I.; O'Connor, K.; Weissenbacher, D.; Tutubalina, E.; Sarker, A.; Banda, J. M.; Krallinger, M.; and Gonzalez-Hernandez, G., eds., *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, 21–32. Mexico City, Mexico: Association for Computational Linguistics.

Nguyen, D. Q. e. a. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9–14.

Nikfarjam, A. e. a. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3).

OpenAI. 2024. GPT-4o: OpenAI's New Multimodal Model. https://openai.com/index/gpt-4o. Accessed: 2025-04-19.

Raithel, L.; Thomas, P.; Roller, R.; Sapina, O.; Möller, S.; and Zweigenbaum, P. 2022. Cross-lingual Approaches for the Detection of Adverse Drug Reactions in German from a Patient‘s Perspective. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3637–3649. Marseille, France: European Language Resources Association.

Raithel, L.; Yeh, H.-S.; Yada, S.; Grouin, C.; Lavergne, T.; Névéol, A.; Paroubek, P.; Thomas, P.; Nishiyama, T.; Möller, S.; Aramaki, E.; Matsumoto, Y.; Roller, R.; and Zweigenbaum, P. 2024. A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions across Languages. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 395–414. Torino, Italia: ELRA and ICCL.

Sakhovskiy, A.; and Tutubalina, E. 2022. Multimodal Adverse Drug Event Detection Using Text and Drug Representations. *Journal of Biomedical Informatics*, 135.

Tutubalina, E.; Alimova, I.; Miftahutdinov, Z.; Sakhovskiy, A.; Malykh, V.; and Nikolenko, S. 2020. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics*, 37(2): 243–249.

Supplementary Material for “BEARCAT at #SMM4H-HeaRD 2025 Task 1: No Training, No Problem - Multilingual ADE Detection ”

Ziqi Guo[1], **Robert Palermo**[2], **Luis M. Rocha***[1,3]
rocha@binghamton.edu

[1]School of Systems Science and Industrial Engineering, Binghamton University, Binghamton, NY, USA.
[2]Department of Mathematics and Statistics, Binghamton University, Binghamton, NY, USA.
[3]Universidade Católica Portuguesa, Católica Biomedical Research Centre, Lisbon, Portugal.

		Precision	**Recall**	**F1 Score**
ru	Validation	0.7826	0.5000	0.6102
	Test	0.6336	0.3820	0.4767
	All Median	0.4951	0.5869	0.5366
	All Mean	0.4614	0.5232	0.4721
de	Validation	0.6667	0.6250	0.6452
	Test	0.7619	0.6095	0.6772
	All Median	0.7236	0.6190	0.6686
	All Mean	0.6359	0.5424	0.5567
fr	Validation	0.7308	0.6552	0.6909
	Test	0.8421	0.6095	0.7072
	All Median	0.7054	0.6857	0.6865
	All Mean	0.6144	0.6064	0.5876

Table 1: Performance after applying second-stage refinement. For each language, median and mean scores from all teams submissions for Task 1 are provided for comparison.