

# Y2K@SMM4H 2025: MUSIC: A Multilingual Social-Media Interpreter with Self-Correction Ability

Yufeng Wang<sup>1\*</sup>, Yifan Gao<sup>2, 1\*</sup>, Yimeng He<sup>1, 3\*</sup>, Debiao Li<sup>1, 4†</sup>, and Xiuzhen Huang<sup>3†</sup>

<sup>1</sup>Biomedical Imaging Research Institute, Cedars-Sinai Medical Center, CA90048, USA

<sup>2</sup>School of Medicine, Tsinghua Medicine, Tsinghua University, Beijing 100084, China

<sup>3</sup>Department of Computational Biomedicine, Cedars-Sinai Medical Center, CA90048, USA

<sup>4</sup>Department of Bioengineering, University of California, Los Angeles, CA 90095, USA

{Yufeng.Wang, Yifan.Gao, Yimeng.He, Xiuzhen.Huang, Debiao.Li}@chshs.org, yf-gao21@mails.tsinghua.edu.cn

## Abstract

SMM4H 2025 Task 1 requires detecting adverse drug-event (ADE) mentions in social-media posts of four languages. We introduce MUSIC, a compact two-stage decoder-only language model to tackle the task. MUSIC combines a Classifier that assigns an initial label with a Judge that confirms or overturns that decision. An ensemble of Classifier and Judge checkpoints achieves a weighted-F1 of 0.7079 on the blind test set, outperforming the task median by 8 percentage points and the mean by 17.

## Introduction

Adverse Drug Events (ADEs)—harmful and unintended responses to medications—pose a significant public health burden, accounting for approximately 6.5% of hospital admissions in the UK (Pirmohamed et al. 2004), 4.93% in China (Zhang et al. 2021), and 0.9–7.9% in India (Patel et al. 2007), while substantially driving up healthcare costs. Traditional pharmacovigilance systems relying on spontaneous reports (e.g., FAERS, Vigibase) suffer from severe under-reporting—estimated at 78%–99% in France between 1997 and 2002 (Hazell and Shakir 2006)—which delays safety-signal detection and prolongs patient risk. Retrospective analyses of user-generated content on social media platforms (e.g., patient forums, blogs, microblogs) have shown that mentions of ADEs often appear weeks to months before formal post-marketing safety reports, highlighting social media mining’s potential as an early-warning pharmacovigilance tool (Lee et al. 2021).

The Social Media Mining for Health-HeaRD workshop (#SMM4H) shared tasks aim to advance the use of user-generated social media data for pharmacovigilance, epidemiology, patient-centered results and tracking the impacts of non-medical substance use (Xu and Gonzalez-Hernandez 2024). The 2025 shared task 1 (Detection of adverse drug events in multilingual and multi-platform social media posts) challenges participants to develop binary classification systems that identify ADE mentions across diverse

languages and platforms. This task offers a unique opportunity to evaluate how modern natural language processing (NLP) techniques adapt to noisy user-generated text across linguistic and platform variations.

Previous work in this field mainly focused on fine-tuning monolingual encoder-only transformers such as *RoBERTa-base* (Liu et al. 2019) or smaller encoder-decoder models such as *T5-small* (Raffel et al. 2020), each requiring language-specific adaptation and offering limited zero-shot or generative reasoning. Recent breakthroughs in autoregressive, decoder-only foundation models, such as *GPT-4* (Achiam et al. 2023), *LLaMA* (Touvron et al. 2023), and *Gemma* (Team et al. 2025), deliver strong prompting potential and cross-lingual transfer out of the box (Xu et al. 2024), yet their use in specialized sequence classification tasks such as ADEmention detection remains scarce.

Based on insights into decoder-only architectures, we introduce MUSIC (Multilingual Social-media Interpreter with Self-Correction), a streamlined pipeline that applies parameter-efficient fine-tuning (PEFT) to *Gemma3* for binary ADE detection on social media posts. To boost both in-language and cross-lingual performance, we introduce a two-step fine-tuning architecture, augmenting our training set with translations into three additional languages and external reasoning assistance from the state-of-the-art Large Language Model (LLM), and finally stabilizing output via a majority-voting ensemble.

Our contributions are threefold:

1. PEFT on a decoder-only LLM for ADE detection. We demonstrate that low-rank adapter (LoRA) tuning of *Gemma3* yields strong classification performance, obviating the need for separate encoder or encoder-decoder architectures.
2. Multilingual data augmentation based on cross-lingual reasoning. We show that augmenting with multi-lingual reasoning expansions can boost both monolingual and cross-lingual performance.
3. Lightweight ensembling for robustness. A nine-model majority vote ensemble further improves stability and pushes our final positive  $F_1$  score to 0.7079, outperforming the SMM4H 2025 Task 1 median (0.6268) by more than 8 points and the mean (0.5394) by over 17 points.

\*These authors contributed equally.

†corresponding author

## Task Review

The shared task requires systems to detect mentions of adverse drug events (ADEs) in multilingual, multi-platform social-media posts with limited and biased labeled groundtruth (see Appendix A for detailed dataset review).

Specifically, each input to the system is the tuple

$$(L, R, P),$$

where

- $L$  is a fixed leading prompt,
- $R$  is a reminder prompt indicating the target language,
- $P$  is a single post drawn from different social-media platforms or patient forums, microblogs, or health-related blogs. All posts are written in one of the four languages  $\mathcal{L} = \{\text{English, French, German, Russian}\}$ .

The system  $\mathbf{S}$  must assign a binary label

$$c = \mathbf{S}(L, R, P) \in \{0, 1\},$$

where

$$c = \begin{cases} 1, & \text{if } P \text{ contains a mention of an ADE,} \\ 0, & \text{otherwise.} \end{cases}$$

This setup evaluates how well modern NLP approaches learn features of ADE-related social media expressions, and handle noisy, user-generated text across both linguistic and platform variation.

## System Description

Our end-to-end system  $\mathbf{S} : (L, R, P) \rightarrow c$  is composed of two decoder-only LLMs: the *Classifier*  $\mathbf{C}$  and the *Judge*  $\mathbf{J}$ . See Fig. 1 for detailed system architecture.

**Classifier (C)** :  $(L, R, P) \rightarrow c'$  is fine-tuned on the official SMM4H-2025 Task 1 training set (Magge et al. 2021; Tutubalina et al. 2020), which contains social-media posts in English, French, German, and Russian with binary ADE labels (see Appendix A). Given the prompt  $L$  and language reminder  $R$ , it outputs a provisional label  $c' \in \{0, 1\}$ .

**Judge (J)** :  $(L', R, P, c') \rightarrow c_j$  receives the original post  $P$ , the language prompt  $R$ , a new leading prompt  $L'$  indicating judgment task, and the provisional label  $c'$ . It produces a final judgment  $c_j \in \{0, 1\}$ , indicating whether  $c'$  is correct.

**Two-Step Training Strategy.** To endow  $\mathbf{J}$  with self-corrective and cross-lingual capabilities, we fine-tune it in two phases:

- **Reasoning Step:** We construct a balanced reasoning dataset  $\{(L', R, P, c') \rightarrow (l_c, l_r)\}$ , where  $l_c$  is the true judgment and  $l_r$  is a natural-language rationale generated by *Claude 3.7-Sonnet (v20250219)*. Positive, negative, false-positive, and false-negative cases are equally represented (see Appendix A).
- **Generalization Step:** We translate all training posts into the other 3 target languages, run  $\mathbf{C}$  on posts in these languages to obtain  $(L', R', P', c')$ , and fine-tune  $\mathbf{J}$  to predict only  $l_c$  (setting  $l_r = \text{NaN}$ ) for faster inference.

**Majority Vote.** To boost the complementary nature of  $\mathbf{C}$  and  $\mathbf{J}$ , let  $\{\mathbf{C}_i\}_{i=1}^{N_C}$  and  $\{\mathbf{J}_i\}_{i=1}^{N_J}$  be ensembles of Classifier and Judge instances, each fine-tuned with different random seed and language distribution. For a given input  $(L, R, P)$ , each Classifier  $\mathbf{C}_i$  produces a provisional label  $c'_i = \mathbf{C}_i(L, R, P)$ , and each Judge  $\mathbf{J}_i$  evaluates its corresponding provisional label on  $c'_i$  to yield a final judgment  $c_j^i = \mathbf{J}_i(L', R, P, c'_i)$ . The system’s overall output  $c$  is determined by the majority vote over all  $N_C + N_J$  judgments.

## Experiments and Results

**Experiments.** Each training post was translated by *Gemini-2.0-Flash* into the three other task languages, and the resulting translations were merged with the original post to produce four language versions per instance. We fine-tuned two decoder-only LLMs, *Gemma3-12B* and *Gemma3-27B*, on both original and augmented dataset. 5 best-performing Classifier checkpoints based on macro-F1 were retained for downstream integration. Two Judge variants were trained upon the best Classifier model. The final majority-vote ensemble model consists of  $N_C = 5$  best Classifier checkpoints and  $N_J = 2$  Judge models. We use positive F1-score for evaluation. Implementation and evaluation details can be found in Appendix B.

**Results.** Performance of our approach across both the validation and test sets is presented in Table 2. We conducted ablation study for the best single Classifier model, the best Judge model trained from it, and the full ensemble voted model. Results are summarized in Table 1. Our best model achieved the highest overall performance on the test set, reaching a weighted-F1 of 0.7079—surpassing the task median (0.6268) by more than eight percentage points and the task mean (0.5394) by nearly seventeen.

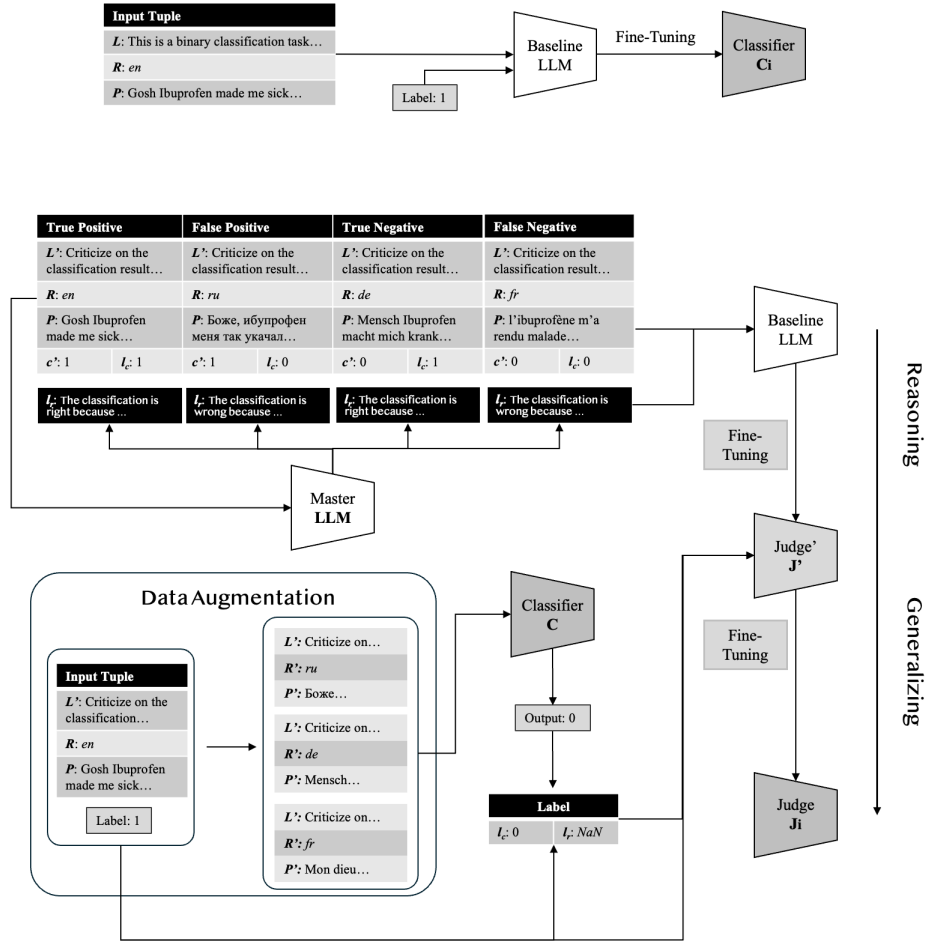
**Discussion.** The Classifier Model demonstrated strong classification capability when trained on languages with abundant and reliable annotations, such as English and Russian. In contrast, the Augmented Judge Model effectively leveraged cross-lingual complementary information and implicit reasoning to enhance performance on lower-resource languages like German and French. Finally, the Majority Voting approach successfully integrated the strengths of both models, yielding improved overall performance through model-level ensembling.

**Code Availability.** Open-source code and our reasoning dataset can be found at <https://github.com/yfngao76/MUSIC>.

## Conclusion

In this work, we introduced MUSIC, a multi-step, self-corrective PEFT framework for detecting adverse drug events in multilingual social-media posts. Our experiments show that by fine-tuning only decoder-only foundation models, MUSIC not only matches but outperforms traditional encoder-decoder architectures. This success is driven by a novel self-correction module that uses cross-lingual reasoning to refine predictions. These findings underscore the potential of self-autoregressive models for robust, multilingual information extraction in social-media mining.

(a) Training Phase



(b) Testing Phase

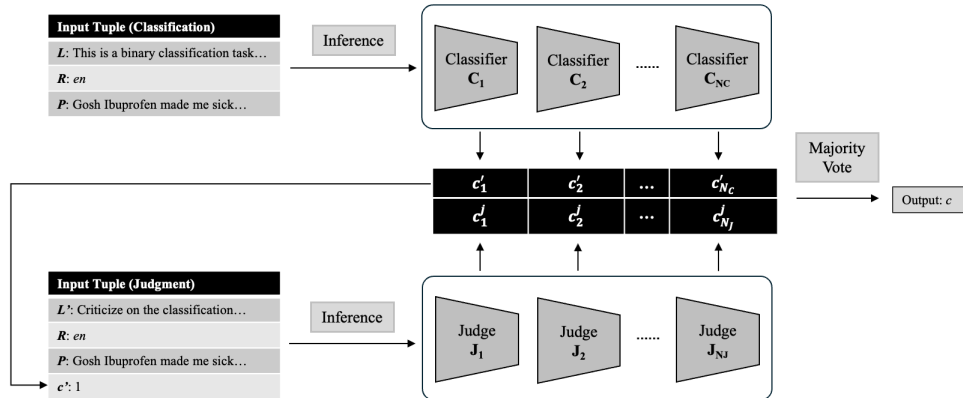


Figure 1: Overall System Architecture.

Table 1: Performance of MUSIC on the SMM4H 2025 Task 1 blind test set. Model C refers to the best Classifier model; Model C + J denotes the best Judge model trained from it; Model C + J + Voting represents the full ensemble. F1-score for each language, an average across 4 languages (Weighted-F1) and a Macro-F1 score considering language distribution are shown.

System	F1-en	F1-de	F1-fr	F1-ru	Weighted-F1	Macro-F1
Model C	0.6933	0.7320	0.7200	0.6247	0.6637	0.6925
Model C + J	0.7258	0.7411	<b>0.7739</b>	0.6070	0.6760	0.7119
Model C + J + Voting	<b>0.7593</b>	<b>0.7660</b>	0.7538	<b>0.6448</b>	<b>0.7079</b>	<b>0.7309</b>

Table 2: Performance on Validation(Dev) and Test sets. The baseline of the Dev set is the result from the baseline model, while the result in the Test set are presented in (median/mean) format. F1 indicates a simple average F1-score across all languages evaluated.

	Dev		Test	
	Macro-F1	F1	Macro-F1	F1
Our Best	<b>0.7330</b>	<b>0.7474</b>	<b>0.7309</b>	<b>0.7079</b>
Baseline	0.6919	0.7048	-	-
Mean	-	-	0.6460	0.6268
Median	-	-	0.5520	0.5394

## Appendices

### A. Training and Evaluation Datasets

Prior to data augmentation, the training dataset exhibited significant imbalance in both language and label distributions. English and Russian together accounted for over 92% of the data, while German and French made up less than 8%. The dataset was also heavily skewed toward the negative class (label 0), which comprised 92.14% of samples in most languages. The positive class (label 1) was particularly underrepresented in German (5.60%) and French (7.16%), while slightly higher in Russian (10.08%).

To address this imbalance, we applied multilingual data augmentation using *Gemini-2.0-Flash*, translating each training instance into the three other task languages. This procedure expanded the dataset fourfold, resulting in 124,748 samples uniformly distributed across the four languages. Each language contributed exactly 31,187 instances (25%), including 2,451 positive and 28,736 negative samples. After augmentation, both the language and label distributions were balanced, with the global positive class ratio preserved at 7.86%.

We randomly selected 2,000 samples following a language distribution 4(en):4(ru):1(de):1(fr) with true and false label, respectively and split them into four evenly distributed group named TP, FP, TN and FN. We mimicked the ADE prediction result from each group. (e.g. In the FP group, the true label is 0 (False) but the mimicked result returned 1 (True)) We provided the post, true label and the mimicked label to *Claude 3.7-Sonnet* (v20250219) and let the LLM generate reasoning and critique. The prompt templated used for generating reasoning is shown below:

```
[User]: The following social media post is
label as (NOT) containing a mention of
an Adverse Drug Event (ADE) by an llm.
The label is correct but lacks
```

```
reasoning. / However the llm made a
mistake in labeling. Please give a
reasoning on how the model understood
the post and give a critique on the
reasoning.
```

```
[User]: Social Media Post: {post}
```

### B. Implementation and Evaluation Details

We fine-tune both the Classifier and Judge models using structured prompt templates of the form  $(L, R, P)$ , where:

- $L$  is the leading instruction,
- $R$  is the language reminder,
- $P$  is the raw social-media post.

Below, we give the exact prompt templates used for each model.

#### Classifier

```
[L]: This is a binary classification task.
Given the following social media post,
predict whether the post contains a
mention of an Adverse Drug Event (ADE).
Only answer 0 for not mentioning ADE
and 1 for mentioning ADE.
[R]: The post is in {lang}.
[P]: {post}
```

#### Judge

```
[L]: This is a binary classification task.
Given the following social media post,
predict whether the post contains a
mention of an Adverse Drug Event (ADE).
Only answer 0 for not mentioning ADE
and 1 for mentioning ADE.
```

```
[R]: The post is in {lang}.
[P]: {post}
```

All fine-tuning steps were trained on unsloth (Daniel Han and team 2023), an open-source platform for flexible and efficient fine-tuning. The best classifier model utilizes  $r = 8$  and was trained for 5 epoches, while the judge model uses  $r = 32$ , trained for 3 epochs in reasoning step and another 1 in generalizing step. All training was done on a nvidia 140s GPU with 40GB RAM and cost about 30 minutes per epoch. Model Temperature is set to 0.1 when conducting inference. See source code for more details.

Positive Balanced F-score (F1-score) is used as evaluation metrics. Posts with ADE mention are labeled positive (1) and those without ADE are labeled negative (0). For each language  $lang$  in  $\mathcal{L}$ , we have:

$$F_{1,lang} = \frac{2|TP_{lang}|}{2|TP_{lang}| + |FP_{lang}| + |FN_{lang}|}$$

$$\text{Weighted-}F_1 = \sum_{lang \in \mathcal{L}} F_{1,lang} / |\mathcal{L}|$$

$$\text{Macro-}F_1 = \frac{2 \sum_{lang \in \mathcal{L}} |TP_{lang}|}{\sum_{lang \in \mathcal{L}} (2|TP_{lang}| + |FP_{lang}| + |FN_{lang}|)}$$

## References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Daniel Han, M. H.; and team, U. 2023. Unsloth.
- Hazell, L.; and Shakir, S. A. 2006. Under-reporting of adverse drug reactions. *Drug safety*, 29(5): 385–396.
- Lee, J.-Y.; Lee, Y.-S.; Kim, D. H.; Lee, H. S.; Yang, B. R.; and Kim, M. G. 2021. The use of social media in detecting drug safety-related new black box warnings, labeling changes, or withdrawals: scoping review. *JMIR public health and surveillance*, 7(6): e30137.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- 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.
- Patel, K.; Kedia, M.; Bajpai, D.; Mehta, S.; Kshirsagar, N.; and Gogtay, N. 2007. Evaluation of the prevalence and economic burden of adverse drug reactions presenting to the medical emergency department of a tertiary referral centre: a prospective study. *BMC clinical pharmacology*, 7: 1–5.
- Pirmohamed, M.; James, S.; Meakin, S.; Green, C.; Scott, A. K.; Walley, T. J.; Farrar, K.; Park, B. K.; and Breckenridge, A. M. 2004. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *Bmj*, 329(7456): 15–19.
- Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.
- Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- 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.
- Xu, D.; and Gonzalez-Hernandez, G., eds. 2024. *Proceedings of the 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*. Bangkok, Thailand: Association for Computational Linguistics.
- Xu, D.; Lopez-Garcia, G.; Raithel, L.; Roller, R.; Thomas, P.; Aramaki, E.; Wakamiya, S.; Yada, S.; Zweigenbaum, P.; O’connor, K.; et al. 2024. Overview of the 9th Social Media Mining for Health Applications (# SMM4H) Shared Tasks at ACL 2024—Large Language Models and Generalizability for Social Media NLP. In *The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, 183–195.
- Zhang, Y.; Jin, L.; Zhang, X.; Bai, R.; Chen, D.; Ma, Y.; and Zhai, X. 2021. Emergency hospitalizations for adverse drug events in China: Clinical pharmacists’ approach to assessment and categorization. *Pharmacoepidemiology and Drug Safety*, 30(5): 636–643.