Limics at #SMM4H-HeaRD 2026: Uncertainty-Driven Prediction for ADE Detection in Social Media

Nour Allam[1], Marie-Christine Jaulent[1]

[1]Sorbonne Université, INSERM, Université Paris-Nord,
Limics laboratoire de recherche en informatique pour la santé, Paris, France
Correspondence: nour.allam@sorbonne-universite.fr

Abstract

This paper describes our system for the SMM4H-HeaRD 2026 Task 1: *Detection of Adverse Drug Events in Multilingual and Multi-platform Social Media Posts*. We developed a two-stage pipeline combining a fine-tuned XLM-RoBERTa-large encoder-only model with a large language model for final decision on ambiguous cases. To handle complex linguistic boundaries, we explore explicitly training the encoder to treat ambiguity as a discrete third label to delegate those cases to the generative model. Although introducing the third label was associated with lower performance than relying on a binary model, when using the encoder as a preliminary filter for classifying 78.62% of posts as negatives, we achieved a global F_1 score of 0.614 (+0.034 over task median).

1 Introduction

Social media and patient forums can serve as useful sources for detecting signals complementary to spontaneous reporting systems for pharmacovigilance activities (Sarker et al., 2015; Klein et al., 2025). However, automatically extracting Adverse Drug Events (ADEs) from such content remains a challenging open problem. This is particularly difficult in social media, where posts are short, informal, multilingual, and frequently ambiguous.

Task 1 of SMM4H-HeaRD 2026 formalizes this problem as a multilingual binary classification task. Given a social media post, the system must predict 0 (no ADE mentioned) or 1 (at least one ADE mentioned) (Lopez-Garcia et al., 2026). We approach this task by combining the efficiency of masked language models with the reasoning capabilities of generative LLMs. Rather than forcing a hard binary decision at training time, we hypothesize that explicitly representing annotation ambiguity as a discrete third label can improve the encoder's ability to delegate difficult cases to an LLM. To our knowledge, this explicit semantic treatment of ambiguity as a routing signal has not previously been investigated in this domain.

Related work. Although different than our approach, hybrid SLM+LLM pipelines have been explored in the broader natural language processing literature, where cascading architectures delegate uncertain instances to stronger models to improve efficiency and accuracy (Varshney and Baral, 2022; Zhang et al., 2026). In the ADE detection setting specifically, recent systems combine encoder efficiency with LLM reasoning in different ways: for instance, Zheng et al. (2024), participating in SMM4H 2024 Task 1, chain a RoBERTa classifier with GPT-4 for span extraction in a multi-stage pharmacovigilance pipeline. Kondadadi and Ortega (2026) propose a Learn-to-Defer (L2D) framework that trains the encoder to predict its own likely errors, using that to trigger delegation. Both treat the delegation decision as based on the model's output distribution, whereas our approach explicitly attempts to encode ambiguity as a discrete training label. Zero-shot LLM-only approaches have also been explored for the same task series; Guo et al. (2025) apply GPT-4o without fine-tuning to multilingual ADE classification in SMM4H 2025.

Task data. The provided data contains 46,737 posts in the training set and 8,033 posts in the development set. The dataset has six different languages in the training phase: English, Russian, Japanese, German, French, and Mandarin. Additionally, Farsi health forum posts were included in the test set only to evaluate zero-shot cross-lingual transfer. The data is highly variable. Messages originate from different platforms that do not follow the same linguistic patterns. Posts from patient forums or drug review websites (like KEEPHA, CADEC-v2, and RuDReC) are usually structured and long. In contrast, data from X and 120ask are shorter, less formal, and frequently contain sarcasm or in-

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 230–236

ternet slang. Furthermore, the class distribution is highly imbalanced, with positive ADE examples making up only 6.4% of the training set.

2 Methods

2.1 Sliding window preprocessing

We used a sliding-window strategy for posts over 300 characters, as longer texts often contain irrelevant noise. This also standardizes input lengths across diverse sources, given that forum texts and social media posts vary significantly in size. This strategy was chosen to lead the model's focus on the local context around a potential ADE (Jaiswal and Milios, 2023). Texts are first segmented based on punctuation and grouped into overlapping windows to preserve context. Specifically, the window size dynamically adapts to span two consecutive sentences, expanding to a third sentence if the initial pair is shorter than 180 characters. The window strides forward by a single sentence at a time, ensuring a robust overlap of one to two sentences between adjacent windows. For unpunctuated posts, a fallback fixed-length chunking of approximately 150 characters is employed.

2.2 Exploration & categorization phase

Manual review showed that the boundary between ADE and non-ADE is often ambiguous, and many posts are hard to classify even for humans. Among positive examples, we identified six recurring ADE sub-categories: direct explicit side effects, withdrawal symptoms, allergic reactions, loss of drug efficacy, speculative or uncertain reports, and vague ADEs where the drug–symptom link is unclear. To explore these, we used GPT-5-mini (OpenAI) with a decompositional prompting strategy (Khot et al., 2023). Rather than asking the model to classify a post in one single instruction, we decomposed the task into six independent API calls with independent yes/no questions, each designed to detect one specific category hierarchically (see Appendix A). We applied this scheme to all 2,992 positive training examples.

Extending this categorization to a sampled subset of negatives highlighted a recurring pattern: a noticeable number of posts in the *negative* class received strong or vague ADE flags from the LLM. These posts, despite being formally annotated as non-ADE, were very similar to ADE expressions (see the Appendix B).

2.3 Three-label data construction & training

The exploration phase indicated that the six ADE sub-categories carry varying levels of medical certainty. Because speculative and vague reports form a soft boundary, treating them identically to explicit side effects might blur the model's binary distinction. We hypothesized that isolating these uncertain cases could prevent them from diluting the primary ADE signal. To test this, we collapsed the categories into a three-label taxonomy: Label 1 captures explicit ADEs such as obvious adverse effects and withdrawal symptoms; Label 2 isolates vague or speculative associations and Label 0 handles all clearly negative and unrelated posts.

LLM Model Parameters. All LLM calls used the OpenAI API with ‘max_completion_tokens‘ set to 500 and ‘reasoning_effort‘ set to "low"; ‘temperature‘ and ‘top_p‘ were omitted because this model's reasoning architecture does not support them.

Data preparation. Mapping was applied per window. For multi-window posts, if any window had Label 1, the post was positive. If no window had Label 1 but at least one had Label 2, the post was *Ambiguous*. This procedure yielded a training set of 22,829 windows across 12,230 posts (16,997 Label 0; 5,349 Label 1; 483 Label 2). Since positive examples represent only 6.4% of the total data, we retained all Label 1 and Label 2 instances. To prevent the negative class from dominating the training process while still maintaining the significant presence of non-ADE content typical of social media, we downsampled the negative posts to a 30:70 (Positive:Negative) ratio with the random seed 42. By artificially increasing the prevalence of the minority class, we ensure that the model is exposed to a high enough density of ADE signals to prioritize recall. The negative examples sampling was performed randomly and proportionally to the language distribution of the positive examples.

Model training. We fine-tuned XLM-RoBERTa-large, a 560M-parameter multilingual transformer with strong cross-lingual transfer properties. Models were trained for 10 epochs with a learning rate of 1e-5, batch size of 16, warmup ratio of 0.2, and label smoothing of 0.1, selecting the best checkpoint based on macro F_1 on the development set. We trained two variants: the three-label model and a binary model version of the same dataset, merging Labels 1 and 2 as positive.

2.4 System design

The decision rule relies on the encoder's output. If the encoder predicts Label 0, the post is classified as negative without delegation to the LLM. For the remaining predictions, we evaluated three distinct pipeline configurations, as illustrated in Figure 1. Either the ambiguous cases alone, both explicit and ambiguous cases, or all binary positive predictions are delegated to the LLM arbiter for final classification. For the three-label pipeline, both Label 1 and Label 2 predictions are delegated to the LLM arbiter, as preliminary experiments showed that delegating Label 2 alone did not yield meaningful improvements over the encoder-only baseline.

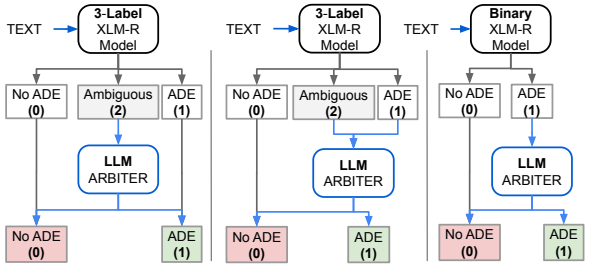

Figure 1: System architecture and inference pipelines evaluated. *Left:* Three-label pipeline sending only *Ambiguous* (Label 2) predictions to the LLM. *Middle:* Three-label pipeline sending both *Ambiguous* and ADE predictions to the LLM. *Right:* Binary pipeline sending all positive ADE predictions to the LLM. In all configurations, negative predictions (Label 0) directly considered as negatives.

Posts sent to the LLM are evaluated using GPT-5-mini via an API. To assess the extent to which prompting influences the final outcomes, we evaluated two distinct prompt variants (detailed in Appendix C). The *Short Prompt* provides a concise instruction requiring a simple binary output. The *Long Prompt* uses a structured format with explicit inclusion and exclusion criteria.

3 Results

3.1 Development set performance

Table 1 presents the performance of our system configurations on the development set, comparing standalone encoders against the full two-stage pipelines. To evaluate the standalone three-label encoder without an LLM, we tested three baseline strategies for handling *Ambiguous* predictions: treating them all as positive, treating them all as negative, or excluding them from evaluation entirely.

System	Prompt/Strat.	F_1	P	R	LLM%
Encoder-Only Baselines					
Binary	Standard	0.496	0.349	0.856	0%
3-Label	Ambig → Pos	0.421	0.275	0.900	0%
3-Label	Ambig → Neg	0.426	0.287	0.823	0%
3-Label	Excluded	0.435	0.287	0.891	0%
LLM-Only Baselines					
LLM Only	Long	0.596	0.469	0.819	100%
LLM Only	Short	0.554	0.408	0.864	100%
Two-Stage Pipeline					
Bin + LLM	Long	**0.646**	0.597	0.703	12.84%
Bin + LLM	Short	0.645	0.561	0.758	12.84%
3-Lab + LLM	Long	0.626	0.551	0.724	21.38%
3-Lab + LLM	Short	0.563	0.467	0.707	21.38%

Table 1: Development set performance. P = Precision, R = Recall.

All encoder-only three-label configurations had lower performance than the binary encoder in overall F_1 score. When *Ambiguous* predictions are mapped to the positive class, recall reaches 0.900, but F_1 drops to 0.421. This indicates that the ambiguous class mislabels a large number of true negatives along with positive or uncertain cases. Consequently, the ambiguous label does not cleanly separate the uncertain positives from the negative class. For the Binary + LLM pipeline, the overall F_1 scores for Long and Short prompts are nearly identical. In contrast, for the Three-Label + LLM pipeline, the *Long Prompt* outperforms the *Short Prompt* (0.626 vs. 0.563). This difference likely indicates that because the three-label encoder explicitly sends ambiguous cases to the LLM, a simple instruction prompt is not enough. The detailed inclusion and exclusion criteria provided in the *Long Prompt* were critical for the LLM to better decide on these ambiguous posts.

In the two-stage pipelines, the *Long Prompt* outperforms the *Short Prompt*. Notably, the Binary + LLM configuration with the *Long Prompt* achieves the highest development F_1 (0.646), outperforming its 3-Label counterpart (0.626). Overall, the two-stage pipelines improve results compared to both the encoder-only baselines and the LLM-only zero-shot approach.

3.2 Official test set results

Table 2 reports the official test set results for our submitted system (Three-Label + LLM, Short Prompt) alongside the shared task mean and median across all participating teams. The three-label system was submitted as the three-label approach represented our primary research hypothesis.

Our submitted system achieves a global F_1 of 0.614, above the task mean (0.547) and median (0.580). Global performance is highest on Ger-

System	Global	EN	FR	DE	RU
Ours (Sub.)	**0.614**	0.683	0.657	0.684	0.576
Task Mean	0.547	0.685	0.681	0.664	0.533
Task Median	0.580	0.701	0.696	0.656	0.550
System	**ZH**	**JA**	**FA**	**FR-CADEC**	**DE-CADEC**
Ours (Sub.)	0.840	0.486	0.525	0.844	0.852
Task Mean	0.804	0.534	0.367	0.843	0.833
Task Median	0.821	0.549	0.380	0.883	0.860

Table 2: Official test set F_1 scores for all subsets using the Three-Label + LLM pipeline with the *Short Prompt*.

man CADEC (0.852) and French CADEC (0.844). These subsets correspond to longer, more formally structured forum posts. For Mandarin, despite being only 6% of the training data, the model performed strongly, with results close to both the mean and median scores. This is notable given that all posts come from 120ask, a brief and informal platform, and deserves further examination.

Japanese (0.486) and Farsi (0.525) represent the lowest performance among the evaluated languages. Farsi was absent from the training corpus, yet our system (0.525) exceeds both the task mean (0.367) and median (0.380). Nonetheless, the statistics regarding the number of posts sent to the LLM indicate the encoder routed 5,202 posts (34.3% of the Farsi subset) to the LLM, which reclassified 82.76% of these instances as negative which is the highest rate across all languages (Appendix D). This suggests that the zero-shot language transfer for Farsi was only partially successful. In the case of Japanese, the data is drawn from a corpus limited to misuse (Nishiyama et al., 2025). This mismatch likely contributes to the weaker performance observed for Japanese. We provide a brief manual examination of system errors in Appendix E.

As detailed in Section 3.1, experiments on the development set revealed that a simpler Binary + LLM pipeline (using the *Long Prompt*) outperformed our officially submitted three-label system. Furthermore, we observed a notable increase in the submitted system's performance between the development set (F_1 0.56) and the test set (F_1 0.61). This performance bump is likely explained by underlying distributional differences between the two sets. Specifically, the encoder delegated 12.84% of development set messages to the LLM, compared to 18.54% in the test set. This difference may partly reflect the inclusion of Farsi in the test set or other distributional shifts between development and test.

4 Discussion

This work is an exploratory attempt to capture ambiguous posts as a discrete third semantic label. The ambiguity class was constructed from the *vague_AE* and *doubt* sub-categories identified during our LLM-based exploration phase. Although speculative and doubtful mentions of ADEs tend to be ambiguous, they form their own distinct category and do not fully define linguistic ambiguity. Future work should develop explicit annotation criteria targeting what we consider genuinely ambiguous, for example: irony or sarcasm, incomplete or missing context, double meaning, speculative language, and second-hand or hypothetical reports.

We also note that ambiguity may not necessarily carry consistent semantic patterns, which may partly explain why the binary pipeline outperformed the three-label variant in our experiments. An alternative approach would be to construct the ambiguity class empirically by collecting systematic misclassification patterns across multiple models with different architectures.

Finally, it is worth noting that what we consider ambiguous as researchers may not be treated as ambiguous by annotators. For instance, if annotation guidelines classify ironic or joking ADE mentions as positive, then those posts are unambiguous from the label's perspective, regardless of the interpretive uncertainty we could have. This is a broader issue in social media pharmacovigilance, as there is a degree of irreducible uncertainty, not only in ambiguous cases but in positive ones as well. This reflects a broader issue in annotation methodology: what Plank (2022) calls "human label variation", the idea that annotator disagreement is not noise but genuine signal about instance difficulty, suggesting that in tasks like ours, a single gold label may obscure meaningful uncertainty.

5 Conclusion

Using a fine-tuned encoder as a preliminary filter for negative cases showed benefits: by delegating only 12% to 21% of posts to the LLM, this two-stage pipeline systematically outperformed both standalone XLM-R and LLM-only baselines, offering a highly efficient, low-cost alternative to fully LLM-based systems. While our attempt in formalizing ambiguity as a discrete third label did not yield significant gains over a binary pipeline in our current setup, it remains a promising direction that warrants further investigation with human-validated annotations and more principled ambiguity criteria.

6 Limitations

We used GPT-5-mini for both the exploration phase and the LLM arbiter, which introduces a potential confirmation bias (Pangakis et al., 2023). The model's underlying tendencies remain constant across both roles, meaning some degree of systematic bias cannot be fully ruled out. We did not perform a direct comparison between the sliding-window strategy and full-text processing, nor an exhaustive hyperparameter optimization for fine-tuning XLM-RoBERTa-large. Future work should consider evaluating alternative encoder architectures. We did not evaluate confidence-threshold-based delegation as an alternative to the three-label approach. Finally, our manual data exploration was limited to French and English, which may have introduced bias in our characterization of ambiguous posts across other languages.

References

Ziqi Guo, Robert Palermo, and Luis M. Rocha. 2025. Zero-shot multilingual ADE detection with GPT-4o for #SMM4H-HeaRD 2025 task 1. In *Proceedings of the 10th Social Media Mining for Health and Health Real-World Data Workshop and Shared Tasks*.

Aman Jaiswal and Evangelos Milios. 2023. Breaking the token barrier: Chunking and convolution for efficient long text classification with BERT. *Preprint*, arXiv:2310.20558.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Ari Z. Klein, Titas Dasgupta, Ismael Flores Amaro, Sanya Jana, Shounak Khademi, Graciela Lopez-Garcia, Takashi Onishi, J. Powell, Lisa Raithel, S. Rajwal, and 1 others. 2025. Overview of the 10th Social Media Mining for Health (#SMM4H) and Health RealWorld Data (HeaRD) Shared Tasks at ICWSM 2025. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. AAAI Press.

Rishik Kondadadi and John E. Ortega. 2026. L2D-Clinical: Learning to Defer for Adaptive Model Selection in Clinical Text Classification. *Preprint*, arXiv:2604.13285.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Ganesh Chandrasekar, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z. Klein, Farnoush Zeidi Kolehparcheh, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Amirali Rezaie Mianroodi, Roland Roller, Judith Rosell, and 10 others. 2026. Overview of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ACL 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Tomohiro Nishiyama, Shuntaro Yada, Shoko Wakamiya, Satoko Hori, and Eiji Aramaki. 2025. Monitoring over-the-counter drug misuse in japanese user-generated data. In *MEDINFO 2025 — Healthcare Smart × Medicine Deep*, volume 329 of *Studies in Health Technology and Informatics*, pages 733–737. IOS Press.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. *arXiv preprint arXiv:2306.00176*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212.

Neeraj Varshney and Chitta Baral. 2022. Model cascading: Towards jointly improving efficiency and accuracy of NLP systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Chuang Zhang, Zizhen Zhu, Yihao Wei, Bing Tian, Junyi Liu, Henan Wang, Wang Xavier, and Yaxiao Liu. 2026. Confidence-calibrated small-large language model collaboration for cost-efficient reasoning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4480–4501.

Yifan Zheng, Jun Gong, Shushun Ren, Dalton Simancek, and V.G.Vinod Vydiswaran. 2024. LHS712_ADENotGood at #SMM4H 2024 task 1: Deep-LLMADEminer: A deep learning and LLM pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. In *Proceedings of the 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*.

A Category Detection Prompts

The following prompts were used during the decompositional prompting phase to categorize posi-

tive ADE reports. Example inputs are drawn from the positive-labeled posts from the training dataset.

Vague_AE

> "Does this sentence mention a drug or substance AND any symptom or negative effect, even if the relationship is unclear? Answer only yes or no."

Example input: "despite a difficult meeting with hr today i feel good, i feel chilled, i feel happy but i still feel fat! #olanzapine"

Strong_AE

> "Does this text explicitly describe a drug causing an adverse effect? Answer only yes or no."

Example input: "my increase of quetiapine is making me so tired #sleepy"

Withdrawal

> "Does this sentence talk about withdrawal or addiction related to a drug or molecule? Answer only yes or no."

Example input: "thank god for vyvanse #addicted"

Efficacy

> "Does this sentence talk about a drug or molecule that stopped working? Answer only yes or no."

Example input: "not sure the paxil is working any more"

Reaction

> "Does this sentence talk about an allergic reaction or interaction related to a drug or molecule? Answer only yes or no."

Example input: "the only things i know i'm allergic to are sulfa drugs and ciprofloxacin."

Doubt

> "Does this sentence suggest a possible adverse effect of a drug or substance, but with uncertainty (e.g., question, doubt, speculation, or second-hand report)? Answer only yes or no."

Example input: "or maybe that's the prozac ugh i need a nap always"

B Examples of Borderline Negative Posts

The following are illustrative examples of posts formally annotated as non-ADE (Label 0) in the training data that nonetheless received strong ADE flags from the LLM during our category discovery phase. These cases illustrate the inherent ambiguity in social media pharmacovigilance:

- *"if i run out of prozac i'ma be pissed man ughhh"*
- *"I've been sleeping so much since I haven't taken my vyvanse since last Thursday!"*
- *"04.40 just taken Flecainide and 2 paracetamol to dull Pain side effect of Rivaroxaban."*
- *"@USER_____ the drugs are what cause calcification. Especially flouride. Prozac is half fluoride by weight."*
- *"@USER____ Humira. I'm fine 98% of the time but once every few months I just have a bad week."*
- *"I absolutely adore what Cymbalta does for my fibro, but sometimes the side effects are a bit Mac truck-ish, like today."*
- *"@USER____ lamictal is just making the side effects of the carbatrol worse. can't wait to get off the carbatrol. haven't heard from destiny"*
- *"however i do love my #paxil my dreams have been all over the place since i've been on it!"*

C Prompts Used by the LLM Arbiter

Short Prompt. *You are a pharmacovigilance expert. Your task is to classify social media text. Answer ONLY '1' if the text contains an Adverse Drug Event (ADE). Answer '0' if it does not. CRITICAL RULES: - Do NOT flag the 'Indication' (the disease or symptom the patient took the drug to cure). - Do NOT flag general negative emotions unless explicitly caused by a drug. - If unsure, answer '0'.*

Long Prompt. *### Task Classify whether the social media text describes a clear Adverse Drug Event (ADE). Output ONLY '1' (Positive) or '0' (Negative).*

ADE Inclusion Criteria (Label 1) Label '1' ONLY if there is a clear or strongly implied causal link between a drug and a negative medical effect.

Valid cases: - Side Effects: Physical or mental harm caused by a drug (e.g., "This med gave me a rash", "It caused panic attacks"). - Allergic Reactions: Immune responses triggered by a drug. - Withdrawal: Symptoms occurring after stopping or reducing a drug. - Drug Interaction: Harm caused by combining substances. - Second-hand Reports: ADE experienced by another identifiable person. - Sarcasm: Only if the negative effect is clearly attributable to the drug.

Exclusion Criteria (Label 0) Label '0' in all other cases, including:

- Indications: The condition the drug is meant to treat. - No Causality: Mentions of symptoms and drugs without a clear link. (e.g., "I have a headache and took ibuprofen") - Vague Symptoms: Non-specific complaints without a clear medical effect (e.g., "This med makes me feel weird", "I feel off") - General Sentiment: Opinions or complaints without a specific adverse effect. - Effectiveness Only: Drug "doesn't work" or "never worked" (no harm described). - Negation: Explicitly denying side effects. - Intentional Misuse/Overdose: Recreational use or self-harm.

Decision Rules - A valid ADE requires a clear or strongly implied cause -> effect relationship. - Temporal or causal language ("after taking", "made me", "caused") strengthens evidence. - Mental/emotional effects count ONLY if clearly drug-induced. - If the link is weak, ambiguous, or unclear -> output '0'.

Output Return ONLY: '1' or '0'

D Per-Language Proportion of Posts Sent to the LLM and Reclassified on the Test Set

Lang	Total	Deleg.	Neg.	Rate(%)	Tot.(%)
FA	15,184	5,202	4,305	82.76	28.35
RU	9,293	3,333	1,942	58.27	20.90
EN	11,689	1,993	725	36.38	6.20
JA	3,045	482	299	62.03	9.82
ZH	1,144	261	134	51.34	11.71
DE	1,105	190	64	33.68	5.79
FR	1,102	169	58	34.32	5.26
FR_cadec	87	71	14	19.72	16.09
DE_cadec	87	70	14	20.00	16.09

Table 3: Per-language proportion of posts sent to the LLM for the Three-Label + Short Prompt system on the test set. **Total**: total number of posts; **Deleg.**: number of posts delegated to the LLM; **Neg.**: number of posts reclassified as negative; **Rate (%)**: proportion of sent posts reclassified; **Tot. (%)**: total subset reclassified as negative.

E Qualitative Error Analysis

We present a few English examples from the development set to illustrate error patterns in our system (Three-Label + LLM, Short Prompt).

False Positives. The most frequent source of false positives is posts that mention a drug and a negative effect without a genuine causal link. This includes posts describing the drug's indication rather than an ADE, and sarcastic framing:

- *"with all those fatal side effects its got to be good #xarelto"* –> sarcasm.
- *"no sleep last night. due to running out of seroquel."* –> absence of medication.
- *"I've lost eight pounds thank you vyvanse"* –> Ambiguity between adverse and desired effects.

False Negatives. The system missed several true ADEs. These failures occurred at both stages of the pipeline: either the encoder failed to detect the signal entirely, or the LLM arbiter incorrectly dismissed them:

- *"I might have gained 30 lbs in 6 weeks on Zyprexa"* –> **Encoder Failure:** A clear ADE expressed with doubt ("might have").
- *"does anyone happen to know if lamictal causes you to feel hot? #cantstopsweating"* –> **LLM Arbiter Failure:** The encoder correctly classified this, but the ADE is expressed as a question which likely caused the LLM arbiter to reject strict causality.
- *"21y.o. w/ sickle-cell anemia and taking trazodone presents w/ priapism. what's the cause?"* –> **LLM Arbiter Failure:** The LLM arbiter rejected this because it is formatted as a clinical exam question.

LLM Arbiter Corrections. The LLM arbiter correctly reclassified several encoder false positives as negative:

- *"pradaxa is bad for you the tv told me tho"*
- *"Oh darn! i havent woke up in this much pain in a while...right back 2 bed i go #crohns #ibd #humira"*
- *"Warning: Trazodone will show a false positive for ecstasy."*