Bhramastra at #SMM4H-HeaRD 2026: A Multi-Stage Hunter-Judge Pipeline using DSPy-Optimized LLMs for Multilingual ADE Detection

Bhaarat Pachori
Independent Researcher
Garner, North Carolina, USA
bhaarat.pachori85@gmail.com

Abstract

This paper describes the submission by Team Bhramastra for the #SMM4H-HeaRD 2026 Shared Task 1, focused on personal Adverse Drug Event (ADE) detection in multilingual social media. We propose **Hunter-Judge**, a decoupled architecture designed to handle extreme class imbalance and linguistic variance across seven languages, including a surprise zero-shot Farsi set. Our system employs a fine-tuned multilingual mDeBERTa-v3 model as a high-recall filter ("Hunter"), followed by a Gemini-2.5-Flash model ("Judge") optimized via the DSPy framework for precision-oriented agentic adjudication. By implementing a reasoning protocol grounded in clinical RAG evidence and universal ingredient mapping, our pipeline achieved the highest average F_1-score (0.6653) among all teams. Notably, it demonstrated strong zero-shot generalizability on Farsi (F_1: 0.5863), highlighting the effectiveness of medically-grounded adjudication in low-resource contexts.

1 Introduction

Automated Adverse Drug Event (ADE) extraction from social media is fundamentally hindered by informal syntax and extreme class imbalance (Sarker et al., 2016). While transformer-based architectures (Vaswani et al., 2017; Devlin et al., 2019) have improved extraction, distinguishing whether a drug was taken for a symptom or caused it (Indication vs. Reaction) remains the primary bottleneck, especially in multilingual contexts (Liu et al., 2019; Conneau et al., 2019). We propose **Hunter-Judge** (Pachori, 2026), a multi-stage hybrid pipeline building upon the agentic framework established in related prior work. The system employs a fine-tuned multilingual mDeBERTa-v3 encoder (He et al., 2021) for high-recall screening and a Gemini-2.5-Flash (Gemini Team and Google DeepMind, 2025) "Judge" optimized via DSPy (Khattab et al., 2023) for medical adjudication. Our system achieved the highest average F_1-score (0.6653) in #SMM4H-HeaRD 2026 Shared Task 1 (Lopez-Garcia et al., 2026), demonstrating robust generalizability in zero-shot contexts (Nori et al., 2023). Notably, the reasoning-heavy Stage 2 was the primary precision driver, correcting 822 false positives where symptoms were misidentified as drug indications. [1]

2 System Architecture

The Hunter-Judge pipeline decouples detection logic to maximize recall in Stage 1 and precision in Stage 2 (Pachori, 2026).

2.1 Stage 1: The Hunter (Multilingual Screening)

The Hunter serves as a high-recall filter for potential ADE mentions within noisy social media text. We utilize mDeBERTa-v3-base (He et al., 2021), which employs Electra-style pre-training for superior multilingual performance compared to standard architectures (Liu et al., 2019; Conneau et al., 2019). While domain-specific models like BioBERT (Lee et al., 2019) or SciBERT (Beltagy et al., 2019) focus on clinical text, mDeBERTa provides the necessary multilingual breadth for this task.

Fine-tuning utilized the SMM4H training set augmented with translated instances from the CADEC corpus (Karimi et al., 2015). To address class imbalance, the Hunter was optimized via a weighted loss function and Language-Specific Thresholding. Thresholds derived from validation gap analysis (e.g., 0.005 for Russian) prioritize recall, ensuring all suspect cases are passed to Stage 2 for adjudication.

[1] To ensure reproducibility, the source code, DSPy signatures, and model configurations will be made publicly available on GitHub at https://github.com/Team-Bhramastra/smm4h-task1-2026 upon the conclusion of the peer-review process.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 49–55

2.2 Stage 2: The Judge (Adjudication via DSPy)

The internal optimization trace and selected reasoning examples from the DSPy program training are documented in Appendix E. Flagged posts are adjudicated by a Gemini-2.5-Flash (Gemini Team and Google DeepMind, 2025) model optimized via the DSPy framework (Khattab et al., 2023). The Judge applies a five-pillar reasoning protocol to distinguish drug indications from reactions, leveraging the zero-shot medical capabilities of high-parameter models (Brown et al., 2020; Nori et al., 2023). This stage utilizes greedy decoding ($T = 0.0$) and a checkpointed asynchronous batching system for deterministic results and efficiency. The full protocol and DSPy instructions are provided in Appendix A.

To manage the 42,736 test records under Rate-Per-Day (RPD) limits, we implemented a parallel batching system using `asyncio`. Checkpointing ensured that adjudicated cases were preserved, providing resilience against hardware interruptions and preventing redundant API costs.

3 Experimental Setup

This section details the data, hardware, and hyperparameter configurations used to develop the Hunter-Judge pipeline, focusing on the technical grounding required for multilingual reproducibility.

3.1 Data Utilization

We utilized the official #SMM4H-HeaRD 2026 Task 1 training (47,547 rows) and validation (8,136 rows) sets. To enhance the Hunter's exposure to formal drug-reaction syntax and improve the model's ability to handle extreme class imbalance ($1 : 12.67$), we augmented the training data with translated instances from the CADEC corpus (Karimi et al., 2015), providing a more robust foundation for multilingual ADR detection.

3.2 Hunter Configuration

The Stage 1 Hunter utilized the `mdeberta-v3-base` encoder (He et al., 2021). While domain-specific models like BioBERT (Lee et al., 2019) or SciBERT (Beltagy et al., 2019) focus on clinical text, mDeBERTa-v3 was selected for its superior multilingual breadth and ELECTRA-style pre-training. We implemented a custom `WeightedLoss` trainer with a learning rate of 2×10^{-5} and weight decay of 0.01. Optimization was performed on an Apple M2 Silicon GPU utilizing the Metal Performance Shaders (MPS) backend.

Training dynamics, visualized in Appendix C, indicate that the model converged within the first 100 global steps, with training loss dropping from 1.23 to < 0.20. Occasional volatility in the gradient norm (spiking to 276.97) during the second epoch reflects the heavy penalty applied to rare, high-context ADE mentions. We finalized training at 237 global steps ($\approx$1.99 epochs) upon reaching a validation F_1-score plateau. This early-stopping strategy prevented majority-class overfitting while maintaining the high-recall profile essential for Stage 2.

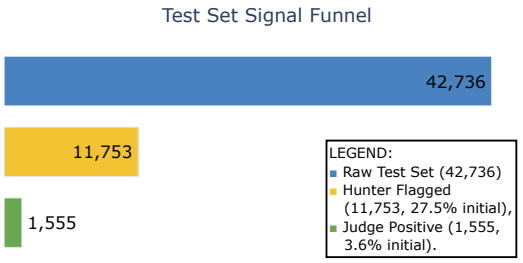

Figure 1: Test set Signal Funnel illustrating the 96.36% noise filtration rate across 42,736 records.

3.3 Judge Configuration

The Stage 2 Judge was implemented as a reasoning agent orchestrated via LangGraph and optimized via the DSPy framework (Khattab et al., 2023). Unlike standard zero-shot prompting, the reasoning logic was compiled using the global training distribution. To resolve cross-lingual variations and brand-name ambiguities, the Judge incorporates an automated INN mapping protocol. For each tokenized mention m_i, a query retrieves the standardized substance profile $P(m_i)$ via a localized RAG framework indexed on FDA-label documentation, explicitly decoupling clinical context from surface-level textual representation. Adjudication was performed on the test records using `asyncio` for parallel batch processing, ensuring that the Judge leveraged the medical reasoning capabilities found in high-parameter language models (Nori et al., 2023) with greedy decoding ($T = 0.0$). The full protocol is detailed in Appendix A.

Team/Metric	en	de	fr	ja	ru	zh	fa	de-cad	fr-cad	Avg F1
Bhramastra	0.7504	0.7866	0.7193	0.5631	0.5921	0.8436	0.5863	0.8846	0.9020	**0.6653**
Mean	0.6845	0.6640	0.6814	0.5342	0.5327	0.8044	0.3670	0.8328	0.8430	0.5465
Median	0.7011	0.6559	0.6961	0.5490	0.5504	0.8210	0.3797	0.8598	0.8829	0.5798

Table 1: System performance results.

3.4 Empirical Threshold Calibration

To ensure the Stage 1 Hunter maximized recall without overwhelming the Stage 2 Judge, we performed a two-phase statistical calibration on the training distribution.

1. **Statistical Distribution Audit:** We extracted the Hunter's probability scores for all gold-standard ADEs to determine language-specific thresholds (τ) levels. We calculated the 1st (P1) and 5th (P5) percentiles to identify the "Hard Tail" of ADEs. As shown in Table 2, Russian ADEs required a P5 threshold of 0.0089 to maintain a 95% recall baseline.

2. **Linguistic Gap Analysis:** To understand the model's uncertainty, we performed a word-frequency analysis on high-uncertainty ADEs (below the 10th percentile). This audit revealed linguistic noise, such as conversational fillers in French or specific Russian causal markers (поэтому), which lowered confidence scores despite the presence of true ADEs.

Combining these statistical percentiles with qualitative insights enabled the Hunter to target the linguistic "blind spots" identified during training. This calibration allowed the system to navigate the Twitter firehose efficiently. Appendix B provides detailed linguistic signals isolated for each language's high-uncertainty regions.

Lang	τ (P1)	τ (P5)	Median Prob
English (en)	0.0009	0.0140	0.9943
Russian (ru)	0.0012	0.0089	0.9924
German (de)	0.0024	0.0072	0.9966
French (fr)	0.0054	0.0190	0.9966
Japanese (ja)	0.0010	0.0061	0.9938
Chinese (zh)	0.0049	0.0341	0.9944

Table 2: Empirical probability thresholds derived from training set ADE distributions to optimize Stage 1 recall.

4 Results and Discussion

We evaluated the pipeline on unified validation and hidden test sets (Lopez-Garcia et al., 2026). Results demonstrate the Hunter-Judge funnel's efficacy (Pachori, 2026): Stage 1 establishes a high-recall ceiling across 42,736 records, while the Stage 2 Judge drives precision through structured causal adjudication grounded in clinical RAG evidence.

4.1 Performance Benchmarks

The transition from validation to the final test environment showed remarkable stability, as illustrated in Table 3, between the internal validation environment ($F_1 = 0.6772$) and the official hidden test set ($F_1 = 0.6653$). As summarized in Table 1, our system achieved the highest average F_1-score in Task 1, significantly outperforming the competition's mean (0.5465) and median (0.5798) (Lopez-Garcia et al., 2026). This consistent performance across the leaderboard confirms that the DSPy-optimized protocol effectively generalizes to the distributional shifts and informal syntax found in raw social media discourse. Notably, the system maintained a $\approx 21 - point$ lead over the task median on the surprise Farsi set (F_1: 0.5863), illustrating the robustness of reasoning-based adjudication in languages where no training data was provided.

Language	Val F1	Test F1
English (en)	0.7612	0.7504
German (de)	0.6400	0.7866
German (de-cad)	0.8727	0.8846
French (fr)	0.7105	0.7193
French (fr-cad)	0.8077	0.9020
Russian (ru)	0.6597	0.5921
Japanese (ja)	N/A	0.5631
Chinese (zh)	N/A	0.8436
Farsi (fa)*	N/A	0.5863
Overall F1	**0.6772**	**0.6653**

Table 3: System performance: Validation vs. Test benchmarks. *Zero-shot surprise set.

4.2 System Synergy and Significance

The Hunter-Judge architecture enables high-recall screening ($R \approx 0.93$) followed by precise medical adjudication (Pachori, 2026). As visualized in Fig-

ID / Type	Text Snippet	Judge's Reasoning
`fa_1` *Indication*	...مراد ی‌نیی ی‌گتفرگ زونهه...	Identified the stuffy nose and loss of smell as persistent attributes of the underlying cold (*indication*) rather than a reaction caused by the Erythromycin.
`fr_95` *Reaction*	"...si la douleur aux jambes est un effet secondaire..."	Correctly adjudicated as **Positive** by identifying the user's explicit questioning of a causal link between new physiological distress and the medication.
`ru_110` *Indication*	...Вентолин останавливает спазмы...	Mention categorized as an *indication* because the user describes a therapeutic effect (stopping spasms) rather than an adverse reaction.

Table 4: Enhanced qualitative analysis of Stage 2 adjudication logic across diverse scripts.

ure 1, the system processes 42,736 raw records through a multi-stage funnel: Stage 1 (Hunter) isolates 11,753 suspect mentions (27.5%), which Stage 2 (Judge) refines into 1,555 high-precision detections. This achieves a 96.36% total noise filtration rate, successfully isolating sparse signals from massive background noise.

Standalone ablation confirms that while the Stage 1 Hunter maintains high recall (0.93), it suffers from excessive false positives (Precision: 0.32) by misidentifying symptoms as indications (Appendix D). A McNemar's test (McNemar, 1947) on discordant pairs (Table 5) confirms that Stage 2 adjudication is statistically significant ($p < 0.001$), successfully correcting 822 Hunter errors while introducing only 87. This synergy proves that semantic adjudication leveraging medically-grounded RAG components is essential for high-fidelity multilingual pharmacovigilance.

Hunter → Judge	Correct	Wrong
Hunter Correct	6904 (Both)	87
Hunter Wrong	822	323 (Both)

Table 5: McNemar's contingency counts ($p < 0.001$) verifying Stage 2 precision gains.

5 Qualitative Analysis

A primary strength of the architecture is its ability to handle linguistically distant languages through universal medical logic (Pachori, 2026). Table 4 illustrates the Judge's adjudication process.

5.1 The Farsi "Surprise" Case Study

Despite Farsi (`fa`) being an unseen language during fine-tuning, the system achieved a zero-shot F_1 of 0.5863. This suggests that the Hunter's multilingual embeddings (Conneau et al., 2019) were sufficient to flag suspect clusters, while the Judge applied clinical RAG reasoning (Nori et al., 2023). In case `fa_1`, the Hunter flagged the post ($P = 0.99$), but the Judge correctly adjudicated it as Negative, noting the symptoms were persistent attributes of the underlying cold rather than reactions caused by Erythromycin. This demonstrates the Judge's ability to utilize clinical evidence to distinguish Indication from Reaction in low-resource settings.

5.2 Medical Reasoning Logic

The reasoning protocol excelled in cases where symptoms were temporally linked but linguistically complex. In the French case **`fr_95`**, the user describes severe pain after starting medication. The Judge identified a Positive ADE by analyzing the user's explicit questioning of whether the pain was a "side effect" (*effet secondaire*). This ability to pivot based on user intent and temporal proximity—rather than just keywords—was the deciding factor in precision gains. The internal optimization trace for these patterns is documented in Appendix E.

5.3 Categorization of Failures

Most false negatives occurred in Japanese (`ja`), identifying a "Recall Ceiling" in the screening pipeline. Because Stage 2 adjudication only occurs for posts flagged by the Hunter, any case filtered out during screening is permanently lost. In high-context languages, informal slang often fell below the screening threshold (e.g., 0.006), preventing the Judge from applying its reasoning logic.

6 Conclusion

The Hunter-Judge pipeline decouples multilingual screening from semantic reasoning, achieving a #1 rank (F_1: 0.6653) in #SMM4H-HeaRD 2026. This hybrid approach validates the efficacy of agentic medical reasoning protocols in both high-resource and zero-shot contexts. Future work will investigate adaptive screening thresholds to address recall limitations in high-context languages like Japanese.

7 Ethical Considerations

Research adhered to SMM4H-2026 data-use agreements. All examples (Appendix A) were paraphrased to protect user privacy and prevent reverse-identification while preserving clinical semantics.

Limitations

1. **The Recall Ceiling (Fundamental)**: The most significant limitation is that the pipeline is Recall-Capped by the Stage 1 Hunter. If the transformer model fails to flag a post due to extreme linguistic variance or novel slang (especially in JA or FA), the Judge never has the opportunity to adjudicate it. Future research must focus on "Looser" screening that doesn't exponentially increase noise.

2. **Linguistic Asymmetry**: While the Judge is zero-shot capable, the reasoning protocol was primarily refined on Latin and Cyrillic scripts. The lower performance in Japanese suggests that "Logical Adjudication" may require cultural context that even high-parameter LLMs struggle to generalize without specific tuning.

3. **Inference Latency (Practical)**: While the hybrid approach is more accurate than a standalone transformer, the multi-stage nature and reliance on LLM APIs introduce significant latency compared to single-pass models, making it less suitable for ultra-high-frequency real-time streams.

4. **Numerical non-determinism**: Furthermore, we acknowledge that while the Hunter's architecture is deterministic, slight variations in floating-point calculations across different hardware backends (e.g., MPS vs. CUDA) may marginally influence the exact calibration of the probability thresholds, a known factor in deep learning reproducibility.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Conference on Empirical Methods in Natural Language Processing*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Thomas Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Jiatao Ji, Roberta Raileanu, and 1 others. 2019. Unsupervised cross-lingual representation learning at scale. *ArXiv*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Gemini Team and Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Google DeepMind Technical Report.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

O. Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *ArXiv*, abs/2310.03714.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

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, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, Abeed Sarker, and 8 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.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Robert Flavell, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *ArXiv*, abs/2303.13375.

Bhaarat Pachori. 2026. Agentic ai in pharmacovigilance: A position paper on opportunities, challenges, and implementation. *Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies*.

Abeed Sarker, Karen O'Connor, Rachel Ginn, Matthew Scotch, Karen Glen Smith, Daniel Malone, and Graciela Gonzalez. 2016. Social media mining for toxicovigilance: Automatic monitoring of prescription medication abuse from twitter. *Drug Safety*, 39:231 – 240.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

A Appendix: Adjudication Protocol Implementation

The following instructions constitute the literal Statistical Protocol utilized by the Stage 2 Judge.

System Prompt: Agentic Judge Protocol

Adjudicate if a social media post describes a personal Adverse Drug Event (ADE).
STATISTICAL PROTOCOL:

1. **IND/REAC:** Distinguish if drug taken FOR (Indication) or CAUSED (Reaction) symptom.
2. **NOISE:** In 'de' and 'fr', ignore conversational etiquette (e.g., 'hallo', 'grüße', 'salut', 'bises'). Focus on the physical narrative.
3. **REASON:** In 'ru', words like 'поэтому' (therefore) or 'помогает' (helps) often signal a cost-benefit analysis of an ADE. Look for the 'cost'.
4. **IMPLICIT DRUGS:** If 'clinical_evidence' is 'Unknown Drug', check for pronouns ('it', 'them') or generic terms ('the pill', 'the dose'). If distress is linked to these, mark as True.
5. **DISTRESS OVERRIDE:** Intense physiological distress (e.g., 'shaking', 'cramps', 'hallucinations') temporally linked to a dose is a True ADE signal.

B Appendix: Linguistic Signals of Model Uncertainty

Statistical analysis of the Hunter's confidence thresholds identified specific linguistic tokens associated with high uncertainty (falling within the P1 to P5 percentile range).

Distribution Audit: Linguistic Uncertainty Signals

Tokens identified during the linguistic gap analysis that correlated with lower model confidence across diverse scripts:

1. **English (en):** conversational noise and causal markers: 'your', 'fuck', 'thats', 'can', 'better', 'user_____', 'an', 'humira', 'taking', 'if', 'just', 'cipro', 'more', 'vyvanse'.
2. **Russian (ru):** pharmaceutical terms and causal connectors: 'препаратов', 'нас', 'применять', 'одной', 'конечно', 'можно', 'чтобы', 'или', 'флуоксетин', 'поэтому', 'эффект', 'просто', 'для', 'более', 'есть'.
3. **German (de):** fillers and treatment-specific tokens: 'ja', 'hab', 'pi', 'grüße', 'u', 'eigentlich', 'östrogen', 'utrogest', 'wj', 'chemo', 'ca', 'euch', 'dadurch', 'natürlich', 'eben'.
4. **French (fr):** clinical context and forum identifiers: 'progestérone', 'sepia', 'bisphosphonates', 'tes', 'cancer', 'lépoque', 'résultat', 'ta', 'chimiothérapie', 'forum', 'naturopathe', 'dentiste', 'sa', 'retraite', 'salive'.
5. **Chinese (zh):** No significant token-level outliers isolated; uncertainty was distributed across diverse syntactic structures rather than specific keywords.
6. **Japanese (ja):** High-uncertainty tokens did not pass frequency thresholds for statistical isolation, suggesting model uncertainty in ja is driven by context density rather than lexical triggers.

C Appendix: Stage 1 Training Dynamics

Figure 2 illustrates the convergence behavior of the mDeBERTa-v3 Hunter during the fine-tuning phase on the unified global training set (47,547 records).

D Appendix: Stage 1 Hunter Standalone Performance

To evaluate the efficacy of the Stage 2 Judge, we performed a standalone ablation of the mDeBERTa-v3 Hunter on the validation set ($N = 8,136$). The classification report 6 and confusion matrix 7 below illustrate the high-recall, low-precision profile that necessitates agentic adjudication.

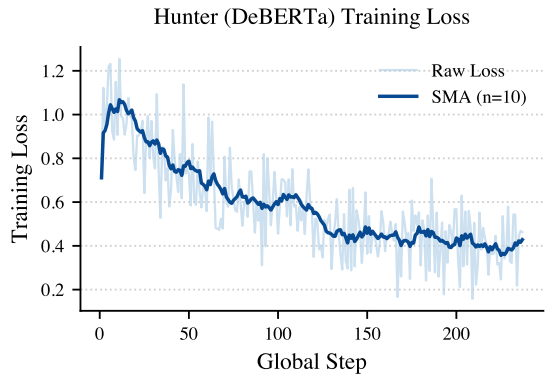

Figure 2: Training loss over 237 global steps for the Stage 1 Hunter. Convergence below 0.2 indicates successful learning of the minority ADE-positive class.

Class	Precision	Recall	F1-score	Support
Negative	0.99	0.85	0.92	7572
ADE	0.32	0.93	0.48	564
Accuracy			**0.86**	8136

Table 6: Stage 1 Hunter standalone classification report.

True \Pred	Negative	ADE
Negative	6465	1107
ADE	38	526

Table 7: Stage 1 Hunter standalone confusion matrix.

E Appendix E: DSPy Program Training Logs

This appendix provides the optimized program signature and qualitative reasoning examples extracted from the DSPy `BootstrapFewShotWithRandomSearch` optimization trace. These logs illustrate the system's ability to resolve medical ambiguities across diverse scripts using the programmatically derived "Statistical Protocol."

5.1 Optimized Program Signature

The following instructions were compiled into the elite program configuration to govern the Judge's adjudication logic:

1. **INDICATION vs. REACTION**: Distinguish if a drug was taken FOR a symptom (Indication) or if it CAUSED a symptom (Reaction).

2. **IMPLICIT DRUGS**: If clinical evidence is absent, check for pronouns such as "it" or "the pill" to identify implied drug mentions.

3. **DISTRESS OVERRIDE**: Prioritize intense physiological distress signals temporally linked to a dose as a True ADE.

5.2 Adjudication Reasoning Examples

The following examples illustrate the Chain-of-Thought (CoT) reasoning patterns generated during the validation trace:

- **German (batch_305_1)**: The user reports having a "Hormonsprirale ziehen lassen" (hormone spiral removed) because symptoms (dizziness) could be caused by it. The Judge adjudicated this as **Positive**, reasoning: "This directly links the hormone spiral to potential adverse symptoms, leading to its removal."

- **Japanese (batch_305_4)**: The user states 糖衣がきもちわるい (the sugar coating feels bad/sickening). The Judge adjudicated this as **Positive**, identifying a direct physical adverse reaction to a component of the drug.

- **Russian (batch_345_4)**: The user describes counterfeit drugs as фальсификат, от которого плохо (fake, from which it's bad). The Judge adjudicated this as **Positive**, identifying that adverse physical effects from ingested substances constitute an ADE, regardless of the drug's authenticity.

- **German (batch_345_3)**: The user discontinued Arimidex due to "sehr schlimme" (very bad) side effects including joint problems. The Judge categorized this as a clear ADE leading to distress and discontinuation.