Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
5(2): 337–347.
DOI: 10.3934/aci.2025019
Received: 13 November 2025
Revised: 12 December 2025
Accepted: 12 December 2025
Published: 24 December 2025

Research article

RECAP: reinforced, explainable LLM classifier for behavioral-health analysis in police narratives

William A. Stigall[1], Francis Nweke[1,*], Hailey N. Walker[1], Md Abdullah Al Hafiz Khan[1], Sharon Perry[1], Yong Pei[1], Dominic Thomas[2] and Monica Nandan[3]

[1] Department of Computer Science, Kennesaw State University, Marietta, GA 30060, USA

[2] Department of Information Systems and Security, Kennesaw State University, Kennesaw, GA 30144, USA

[3] Department of Social Work and Human Services, Kennesaw State University, Kennesaw, GA 30144, USA

* **Correspondence:** Email: fnweke@students.kennesaw.edu.

Academic Editor: Pasi Fränti

Abstract: Police narrative reports from 911 responses contain valuable signs for early behavioral health intervention, but obtaining them manually is time-consuming and tedious. We present RECAP, a human-AI collaboration framework that fine-tunes three large language models: Mistral-7B, Llama-3-8B, and TinyLlama to classify BH cases in narratives and generate short, sentence-level rationales. The models are trained on an annotated corpus using supervised instruction data and then aligned using direct preference optimization (DPO), allowing patrol officer preferences to continually influence system behavior without disrupting existing workflows. On a held-out test set, Mistral-7B achieves 85.2% weighted accuracy and 84.2% F1-score, matching strong prior baselines while improving interpretability by short, text-span-linked explanations; Llama-3-8B performs similarly, while TinyLlama provides competitive accuracy at a lower compute cost. RECAP is designed to reduce manual effort and identify behavioral health signs earlier in public narratives, while providing rationales and maintaining officer control.

Keywords: behavioral health; text classification; large language model; human-in-the-loop; explainable AI

1. Introduction

Police respond to a large number of behavioral-health (BH)-related incidents, and a significant share of 911 calls involve BH concerns [10, 18]. Each incident results in a narrative report; however, extracting BH-relevant signals from these narratives is manual, time-consuming, and inconsistent in practice. We explore whether large language models (LLMs) can reliably help officers by automating multilabel BH classification and producing clear, human-readable rationales during report writing without disrupting existing workflows.

We present RECAP, a human-AI collaboration framework that integrates fine-tuned LLMs directly into the reporting workflow. RECAP uses three backbones trained on an annotated corpus of police narratives: Mistral-7B, Llama-3-8B, and TinyLlama to (i) assign BH categories and subcategories and (ii) generate short, sentence-level explanations based on the narrative text. Models are trained using supervised instruction data and then aligned using direct preference optimization (DPO), allowing patrol officers' preferences to refine system behavior iteratively.

Previous work shows that deep neural models (for example, hierarchical RNNs, CNNs, and transformers such as BERT, RoBERTa, and XLNet) can detect mental-health-related signals in social media [8, 11, 15, 20, 24], but comparable, workflow-integrated solutions for police incident narratives remain limited. Early detection of BH cases can improve response triage, inform co-responder deployment, and reduce paperwork burden, which frees up time for direct service and safety-critical duties. The work is important for the classification of BH from police narratives, considering not only mental health but also domestic, non-domestic, and substance abuse BH indicators [4]. This work creates the "gold standard" benchmark for BH classification from police narratives and introduces a CNN-LSTM reaching an accuracy of 86.67% [3]. Brown et al. [2] build on this work by adopting a novel query strategy, which increases accuracy to 92%. Another study proposed a transformer-based classifier for behavioral health analysis with an accuracy of 53% [17].

In summary, we achieved the following in this work:

- A novel Human-AI teaming framework **[RECAP Framework]** that improves police report BH classification using large language models (LLMs), enhancing accuracy and interpretability.
- To ensure the framework is up to date, we use iterative RLHF updates to keep the models aligned with practitioner judgment while reducing officers' effort. **[Human-in-the-loop learning]**
- We demonstrated that fine-tuned LLMs (Mistral-7B, TinyLlama, and Llama-3-8B) yield promising results: 85.2% weighted accuracy and 84.2% F1 score, for BH classification in police reports and providing rationales for classification.

The remainder of this paper is organized as follows: Section 2 provides existing works in LLMs and their fine-tuning. Section 3 explains the framework. Section 4 describes the experimental setup and evaluation metrics used in this work. Section 5 shows the results and outputs of my framework. Section 6 examines the study's limitations and proposes potential avenues for future research.

2. Background: transformers, parameter-efficient fine-tuning, and preference alignment

2.1. Transformers and LLMs

Since the introduction of the transformer architecture, it has quickly become the dominant architecture in numerous natural language processing tasks [22]. The impact of the transformer architecture has been enhanced by architectures such as BERT [7] and GPT [4, 19]. However, LLMs like GPT 3.5 and GPT 4 are costly to train and infer; with the addition of these models being closed-source, companies and research teams developed open-source solutions that can challenge the performance of much larger language models. Open-source LLMs such as Mistral and Llama provide advanced NLP capabilities to a wider community while keeping costs low and allowing task-specific fine-tuning [1, 13, 25].

2.2. *Parameter-efficient fine-tuning (PEFT)*

Even "small" LLMs like Mistral-7B or Llama-3-8B still hold billions of parameters, making full fine-tuning slow and memory-intensive on consumer hardware. PEFT mitigates this by updating only a tiny weight subset, retaining accuracy while slashing memory use [23]. A leading PEFT method, LoRA [9], freezes the base network and adds trainable low-rank adapters to each transformer layer, achieving near full-model performance with far fewer learnable parameters. LoRA makes it easy to compress the model further, enabling affordable fine-tuning on commodity GPUs [6].

2.3. *Reinforcement learning from human feedback (RLHF) and DPO*

RLHF trains a reward model based on direct human judgments, allowing language models to be optimized for human preference [5]. However, classical RLHF demands a distinct reward network as well as a complex feedback loop. Stanford's DPO avoids these costs by learning an implicit reward. It compares the preferred and rejected answers, then updates the main model against a fixed reference model using a binary-cross-entropy loss [21]. The updated conservative DPO (cDPO) reduces noisy or contradicting annotations by providing a debiased loss optimized for imperfect preference labels [16].

3. Reinforced, explainable LLM classifier for BH analysis (RECAP)

Our framework allows us to continuously improve our LLMs by integrating data accumulation, annotation, and deployment activities.

Figure 1 illustrates our Human-AI teaming workflow for behavioral health classification in police systems. For each unlabeled report, the model provides two candidate predictions, which are batched at moderate temperatures to probe the decision boundary and provide self-contained DPO preference pairings. Figure 2 shows the candidates side by side. A single click accepts the response (green), whereas a double-click rejects it (red). If both are refused, a manual entry box appears, and the annotation is considered the "accepted" answer for the pair. This method lowers the need for ongoing human oversight while still collecting high-quality feedback for DPO.

All runs use 4-bit BitsAndBytes quantization with QLoRA adapters and NEFTune ($\alpha = 10$) to optimize instruction tuning and reduce synthetic-data overfitting [12]. We trained on 7600 SFT examples and 1300 DPO pairs, with an 85/15 split. Additional parameters include cDPO label smoothing, Adam optimizer, and 10 dropout.

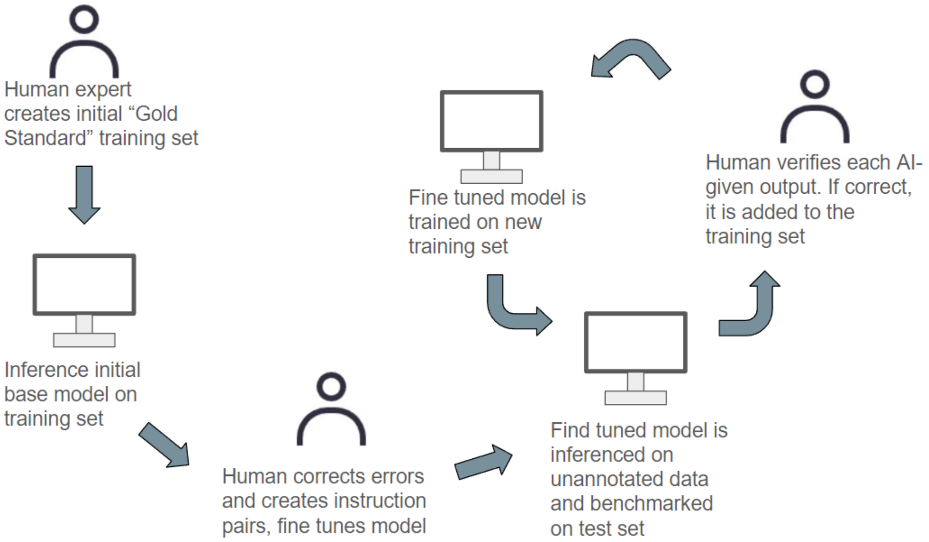
Human expert creates initial “Gold Standard” training set
Inference initial base model on training set
Human corrects errors and creates instruction pairs, fine tunes model
Find tuned model is inferenced on unannotated data and benchmarked on test set
Human verifies each AI-given output. If correct, it is added to the training set
Fine tuned model is trained on new training set

Figure 1. Human-AI teaming.

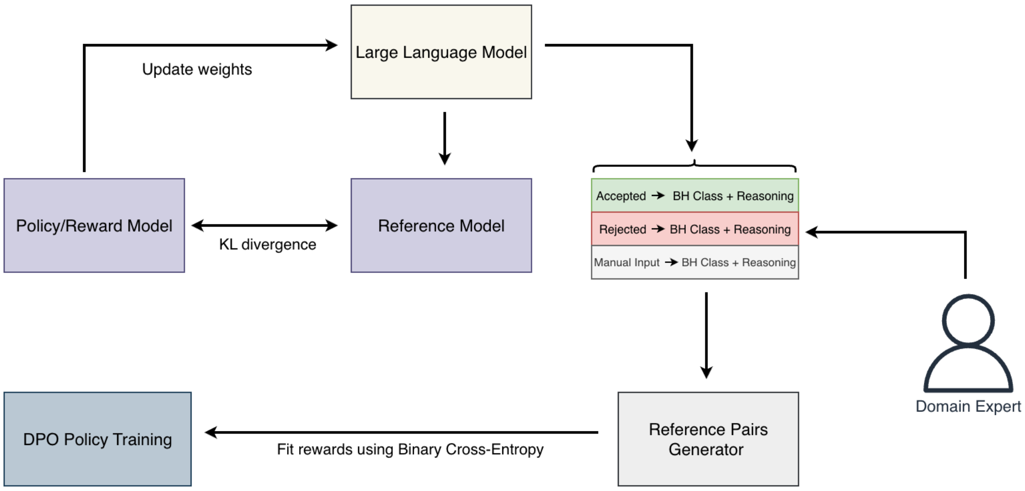
Large Language Model
Update weights
Policy/Reward Model
KL divergence
Reference Model
Accepted → BH Class + Reasoning
Rejected → BH Class + Reasoning
Manual Input → BH Class + Reasoning
Domain Expert
DPO Policy Training
Fit rewards using Binary Cross-Entropy
Reference Pairs Generator

Figure 2. RECAP architecture. The model outputs two generations of the behavioral health class, along with the rationale behind the classification.

4. Experimental setup and evaluation

In this section, we provide detailed information about the experimental setup, including the description of the data, pre-processing, and the evaluation metrics employed to assess the performance of the models.

Table 1 presents the behavioral health classes employed in this study, highlighting significant issues.

Table 1. BH classes.

BH Class	Example
Mental Health	Involves an individual with a diagnosed mental disorder, like schizophrenia or suicidal ideations.
Domestic/Social	Involving multiple individuals in a home setting, like husband/wife or parent/children domestic disputes.
Non-Domestic/Social	Involving multiple individuals not in a home setting, like committing crimes on those not related to the perpetrator.
Substance Abuse	Individuals with persistent drug/alcohol abuse problems.

4.1. Data pre-processing, generation and augmentation

Our corpus contains 8900 police narrative incidents. These include 7600 instruction prompts for supervised fine-tuning (SFT) and 1300 preference pairings for RLHF. The Gold Standard set is maintained completely separate from training and is only used for benchmarking. We used the Mistral model [14] to generate synthetic data from the annotated 1050 samples in a recursive manner. To avoid repetition, each subsequent prompt incorporates the preceding generation's results.

(1) Synthetic data: We generated 7000 instruction examples from the Gold Standard reports via the pipeline in Figure 3.
(2) Human-AI teaming data: An additional 600 instructions and 250 preference pairs were gathered across successive Human-AI teaming rounds.
(3) Public narrative set: We also utilized a 1050 annotated samples that include classes and expert rationales. In Figure 3, we replace the original "synthetic text generation" block with an informed reasoning step, where prompts are augmented by expert annotations to strengthen the generated rationales.

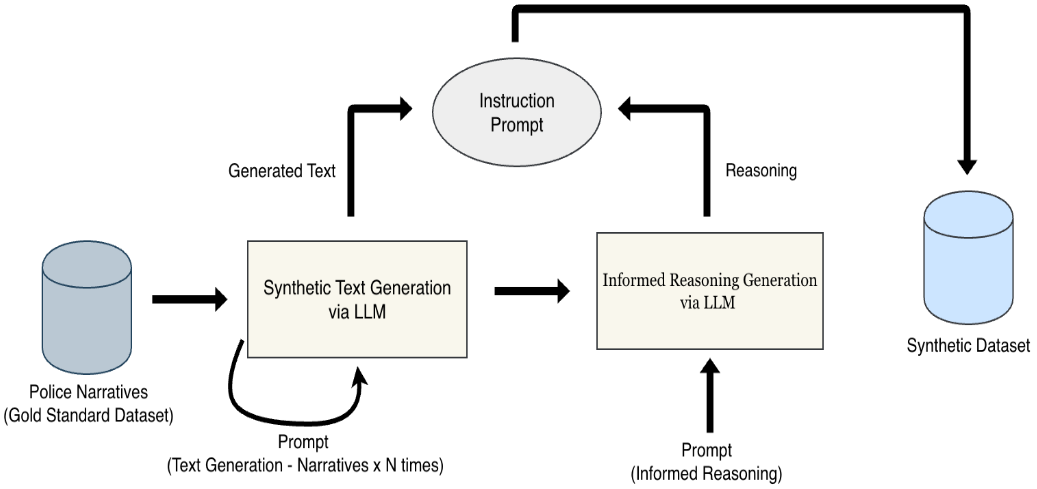
Instruction Prompt
Generated Text
Reasoning
Police Narratives (Gold Standard Dataset)
Synthetic Text Generation via LLM
Informed Reasoning Generation via LLM
Synthetic Dataset
Prompt (Text Generation - Narratives x N times)
Prompt (Informed Reasoning)

Figure 3. Synthetic dataset generation. The first instance, the LLM is assigned a generation task and the new sample is restricted by the bounds of the system prompt and the generation prompt. Recursive generation is used to guarantee diversity.

4.2. Prompt engineering workflow

A single **system prompt** is shared across tasks. It contains (i) a header, (ii) the BH taxonomy, and (iii) explicit instructions for how to tag cases with multiple classes. We have additional prompts to guide the models through each task. These are:

(1) Classification prompt: casts the model as a high-stakes triage aide, requests a class label, and then a one-line rationale.
(2) Informed-reasoning prompt: mirrors the classification prompt but centers on why the tag fits, supporting synthetic-data creation and expert-label conversion.
(3) Generation prompt: runs at high temperature, edits only mutable text, preserves the original tag, forbids out-of-scope additions, and never admits the sample is synthetic.

4.3. Evaluation

Our assessment combines conventional classification metrics with a blinded human-subject study that probes the quality of the rationales for the models.

(1) Teaming cycles and hold-out set (Round 1): the base Mistral-7B model labels every Gold-Standard record; two experts fix any errors. From Round 2 onward, these items are frozen as a test set, while newly generated or corrected samples are added to the training pool. We report accuracy, macro-F1, and weighted-F1 on the held-out set after each snapshot.
(2) Data-flow control: a lightweight CLI drives RECAP. Each narrative carries a persistent ID logged to disk, ensuring that no example is revisited in later teaming sessions.
(3) Iterative RLHF fine-tuning: four DPO rounds are run, each adding 50 new preference pairs. Training always resumes from the previous checkpoint; DPO uses no separate validation split.
(4) Blinded subjectivity test: two experts independently (i) assign a class after reading raw text and (ii) judge whether a model's rationale is plausible, even if its label differs. The protocol is applied to Mistral-7B, Llama-3-8B, and TinyLlama, providing qualitative insight alongside the automatic metrics.

5. Results

In this section, we will discuss the results of the LLMs in the RECAP framework and evaluate the effectiveness of the framework on Mistral-7B after 4 rounds of teaming, BH class, and binary classification, and a model alignment test.

5.1. Model-wise performance

Tables 2–5 show the performance of the models in different settings for BH classification and evaluation.

Tables 3 and 4 show that our models perform strongly on BH classification from police narratives, with Mistral-7B exceeding Llama-3 by 4 percentage points in overall binary BH classification. For clarity, we define the positive class as non-BH cases, and the negative class as BH cases (i.e., true positives = non-BH, true negatives = BH). DPO achieves accuracy comparable to SFT with 84% fewer labeled samples. Both QLoRA and DPO improve accuracy and F1 over their base models.

Table 2. TinyLlama BH classification results.

Model	Accuracy	Precision	Recall	F1-score
TinyLlama-QLoRA	68.0%	48.0%	85.0%	48.0%
TinyLlama-DPO	76.0%	51.0%	62.0%	51.0%
TinyLlama-RECAP	**77.0%**	**53.0%**	**64.0%**	**52.0%**

Table 3. Mistral BH classification results.

Model	Accuracy	Precision	Recall	F1-score
Mistral-7B-Pretrained	82.0%	72.0%	78.0%	75.0%
Mistral-7B-QLoRA	88.0%	82.0%	81.0%	82.0%
Mistral-7B-DPO	88.0%	85.0%	79.0%	82.0%
Mistral-7B-RECAP	**90.0%**	**90.0%**	**79.0%**	**84.0%**

Table 4. Llama-3-8B BH classification results.

Model	Accuracy	Precision	Recall	F1-score
Llama-3-8B-Instruct	82.0%	66.0%	96.0%	78.0%
Llama-3-8B-QLoRA	87.5%	80.0%	85.0%	82.0%
Llama-3-8B-DPO	84.0%	68.0%	97.0%	80.0%
Llama-3-8B-RECAP	**86.0%**	**71.0%**	**97.0%**	**82.0%**

Table 5. RECAP compared to other models.

Model	Accuracy	F1-score
CNN-LSTM [4]	86.7%	82.0%
CNN-LSTM+Query [2]	92.0%	91.1%
Mixtral-8x7B (0-shot)	80.1%	74.2%
Mistral-7B-RECAP (Ours)	**90.0%**	**84.0%**
Llama-3-8B-RECAP (Ours)	**87.5%**	**82.0%**

Precision-recall characteristics vary between backbones. Mistral-7B outperforms Llama-3 (71%) in predicting the BH class with 90% accuracy. Mistral's false-negative rate (1 - recall) is 21%, whereas Llama-3's is 3%. This suggests that Llama-3 rarely misses BH instances, but is less exact when it does.

Mistakes rarely escape the BH classification. Both models frequently switch one class for another as depicted in Table 6. Mistral's main weakness is the domestic-social class, where accuracy suffers. The majority of these errors occur in narratives with no personal or family ties, such as neighbor disagreements or store fights, implying that the model relies too heavily on relationship cues. C-Miss (confusion miss) indicates the most common incorrect label assigned by the model to each true class, as well as how frequently it happens (for example, "Other (4/4)" meaning all four were misclassified as Other). Adjusted Accuracy is the accuracy after using predefined assessment rules (for example, validator acceptability or equivalence mappings) to count operationally acceptable near-misses as accurate.

Table 6. Class analysis for two models.

		Mistral-7B			Llama-3		
Class	**Samples**	**Acc.**	**C-Miss**	**Adj. Acc.**	**Acc.**	**C-Miss**	**Adj. Acc.**
Other	196	91.3%	—	—	91.3%	—	—
Domestic Social	62	62.0%	—	—	95.8%	—	95.8%
Non-Domestic Social	4	4.0%	Other (4/4)	—	4.0%	Other (4/4)	—

Additionally, subsequent RLHF fine-tuning after SFT results in models that perform better, but similarly to RLHF alone. This indicates that RLHF adds value to the fine-tuning process, enhancing model performance. Overall, these results suggest that both QLoRA SFT and DPO are effective methods for improving BH classification in police reports, with Mistral-7B showing the highest overall performance, but to detect the highest percentage of BH cases, Llama-3-8B should be used.

5.2. *Class subjective agreement analysis*

The experiments were conducted on our last fine-tuning checkpoint for all models. Twenty-five samples are generated randomly from a dataset of unseen samples, and annotators are not exposed to the classification output until after they have selected a class.

Figure 4 shows the agreement between annotator and model for classification and explanation.

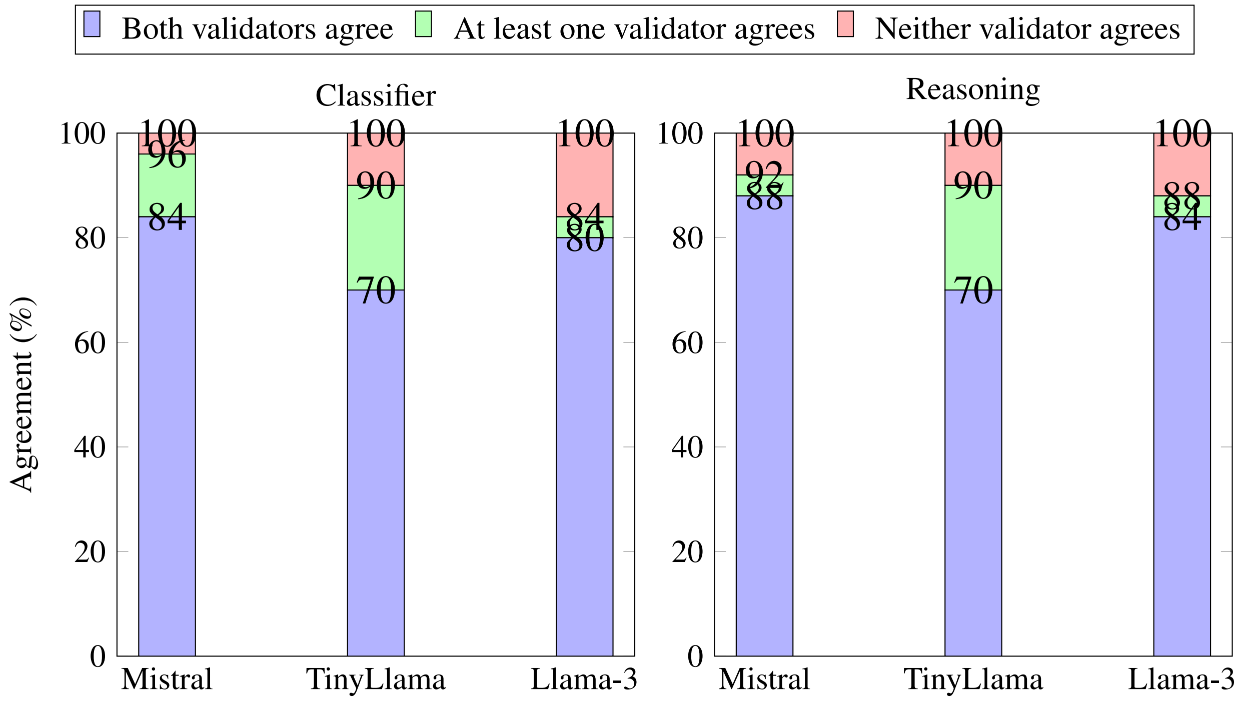

Figure 4. Agreement for the *Classifier* (left) and *Reasoning* (right) evaluations.

6. Limitations, future work, and conclusions

BH labels remain partly subjective. Even with formal definitions, experts can disagree, so the gold standard set may understate true model skill. Second, our DPO loop is model-specific: we replace the backbone (for example, Mistral-7B - Llama-3-8B), the stored preference pairs still mirror the old model's style, leaving the newcomer at a short-term disadvantage until new feedback accrues and prompts are retuned. The improvement could be:

- Prompt maintenance. An architecture that either (i) refreshes training data (with human inputs) with updated system prompts or (ii) mixes legacy and new prompts to encourage more generalizable learning.
- Full-precision models. Phase out heavy quantization to recover reasoning accuracy, accepting higher memory and latency costs; full-precision weights also eliminate the need to de-quantize during training, resulting in a marginal speedup of updates.
- Synthetic-data pipeline. Automate duplicate removal and decontamination, apply low-rank RLHF to curb tag leakage, and run additional synthesis rounds to balance underrepresented behavioral health classes.

Mistral's accuracy is good and also excels at rationale generation, reaching 92% inter-annotator and 96% intra-annotator agreement in our subjectivity test.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

Md Abdullah Al Hafiz Khan is an editorial board member for Applied Computing and Intelligence and was not involved in the editorial review and the decision to publish this article.

References

1. A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, et al., The Llama 3 herd of models, arXiv: 2407.21783. https://doi.org/10.48550/arXiv.2407.21783
2. M. Brown, A. Azmee, M. Khan, D. Thomas, Y. Pei, M. Nandan, Adaptive attention-aware fusion for human-in-the-loop behavioral health detection, *Smart Health*, **32** (2024), 100475. https://doi.org/10.1016/j.smhl.2024.100475
3. M. Brown, M. Khan, D. Thomas, Y. Pei, M. Nandan, Detection of behavioral health cases from sensitive police officer narratives, *Proceedings of IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2023, 1398–1403. https://doi.org/10.1109/COMPSAC57700.2023.00213
4. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., Language models are few-shot learners, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, 1877–1901.
5. P. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, *Proceedings of 31st Conference on Neural Information Processing Systems*, 2023, 1–9.

6. T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: efficient finetuning of quantized LLMs, *Proceedings of 37th Conference on Neural Information Processing Systems*, 2023, 1–28.
7. J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423
8. A. Dinu, A. Moldovan, Automatic detection and classification of mental illnesses from general social media texts, *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 2021, 358–366.
9. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, et al., Lora: low-rank adaptation of large language models, arXiv: 2106.09685. https://doi.org/10.48550/arXiv.2106.09685
10. *A. Irwin, B. Pearl, The community responder model: how cities can send the right responder to every 911 call*, Center for American Progress, 2020. Available from: `https://www.americanprogress.org/wp-content/uploads/sites/2/2020/10/Alternatives911-report.pdf`.
11. J. Ive, G. Gkotsis, R. Dutta, R. Stewart, S. Velupillai, Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health, *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018, 69–77. https://doi.org/10.18653/v1/W18-0607
12. N. Jain, P. Chiang, Y. Wen, J. Kirchenbauer, H. Chu, G. Somepalli, et al., Neftune: noisy embeddings improve instruction finetuning, arXiv: 2310.05914. https://doi.org/10.48550/arXiv.2310.05914
13. A. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. Chaplot, D. de las Casas, et al., Mistral 7B, arXiv: 2310.06825. https://doi.org/10.48550/arXiv.2310.06825
14. A. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, et al., Mixtral of experts, arXiv: 2401.04088. https://doi.org/10.48550/arXiv.2401.04088
15. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, Roberta: a robustly optimized bert pretraining approach, arXiv: 1907.11692. https://doi.org/10.48550/arXiv.1907.11692
16. E. Mitchell, A note on DPO with noisy preferences and relationship to IPO, 2023. Available from: `https://ericmitchell.ai/cdpo.pdf`.
17. F. Nweke, A. Azmee, M. Khan, Y. Pei, D. Thomas, M. Nandan, A transformer-driven framework for multi-label behavioral health classification in police narratives, *Applied Computing and Intelligence*, **4** (2024), 234–252. https://doi.org/10.3934/aci.2024014
18. *NIH, Mental illness*, National Institute of Mental Health, 2023. Available from: `https://www.nimh.nih.gov/health/statistics/mental-illness`.
19. OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, et al., GPT-4 technical report, arXiv: 2303.08774. https://doi.org/10.48550/arXiv.2303.08774
20. K. O'shea, R. Nash, An introduction to convolutional neural networks, arXiv: 1511.08458. https://doi.org/10.48550/arXiv.1511.08458

21. R. Rafailov, A. Sharma, E. Mitchell, C. Manning, S. Ermon, C. Finn, Direct preference optimization: your language model is secretly a reward model, *Proceedings of 37th Conference on Neural Information Processing Systems*, 2023, 1–14.

22. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, et al., Attention is all you need, *Proceedings of 31st Conference on Neural Information Processing Systems*, 2023, 1–11.

23. L. Xu, H. Xie, S. Qin, X. Tao, F. Wang, Parameter-efficient fine-tuning methods for pretrained language models: a critical review and assessment, arXiv: 2312.12148. https://doi.org/10.48550/arXiv.2312.12148

24. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. Le, Xlnet: generalized autoregressive pretraining for language understanding, *Proceedings of the 33rd International Conference on Neural Information Processing Systems* , 2019, 5753–5763.

25. P. Zhang, G. Zeng, T. Wang, W. Lu, Tinyllama: an open-source small language model, arXiv: 2401.02385. https://doi.org/10.48550/arXiv.2401.02385

AIMS Press

©2025 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)