Secure and Private Offline Mental-Health Analysis Using Open Lightweight LLMs with RAG

Makoto Koshino* [a] and Yamato Himi [b]

[a] Department of Electronics and Information Engineering,
National Institute of Technology, Ishikawa College (Ishikawa KOSEN), Ishikawa, Japan
[b] Graduate School of Science and Technology,
Nara Institute of Science and Technology, Nara, Japan

*koshino@ishikawa-nct.ac.jp

Large language models (LLMs) have demonstrated remarkable ability to understand and generate human-like text, but relying on cloud-based proprietary services raises significant concerns around data privacy and security—issues that are particularly acute in mental-health applications. To address these challenges, we introduce MentalGemma2, an end-to-end offline pipeline that integrates the open-source Gemma 2-9B-it model with retrieval-augmented generation (RAG) to perform interpretable mental-health analysis entirely on local hardware. MentalGemma2 eliminates all external API calls and ensures that sensitive user data never leaves the secure deployment environment.

We evaluate four system configurations on the Interpretable Mental-Health Instruction (IMHI) dataset, which aggregates five Reddit-derived subtasks: depression detection (DR), stress detection (Dreaddit), interpersonal risk-factor identification (IRF), multi-dimensional wellness assessment (MultiWD), and stress-cause classification (SAD). The configurations are: (1) few-shot prompting using in-context examples, (2) prompting enhanced with RAG retrieval, (3) supervised fine-tuning via 4-bit quantized low-rank adaptation (QLoRA), and (4) fine-tuning combined with RAG. Each configuration is assessed for classification accuracy (weighted F1-score) and explanation quality (BARTScore).

Our experiments reveal that supervised fine-tuning alone delivers the highest overall classification performance, consistently outperforming the few-shot and RAG-augmented variants. While RAG provides a measurable improvement in the few-shot setting—boosting F1-scores by several points—it yields negligible gains once the model has been fine-tuned, indicating that domain-specific knowledge acquired during training subsumes the auxiliary retrieved information. Notably, MentalGemma2's fine-tuned configuration matches or exceeds the accuracy of larger, cloud-based LLMs such as LLaMA2-7B and MentaLLaMA-chat-13B, despite its smaller footprint and offline operation.

By maintaining all computation and data storage locally, MentalGemma2 offers a practical, privacy-first solution for mental-health analytics in environments where network connectivity is limited or data governance is highly constrained. These results demonstrate that compact, open-source LLMs can achieve competitive, interpretable performance on sensitive NLP tasks without sacrificing user confidentiality, making MentalGemma2 a promising candidate for deployment in resource-constrained educational and clinical settings.

Keywords: *mental health analysis, open lightweight language models, retrieval-augmented generation (RAG), privacy-preserving, fine-tuning*

Introduction

Student mental health is increasingly recognized as crucial for a positive learning environment and academic success. Simultaneously, advancements in Artificial Intelligence, particularly large language models (LLMs), present new opportunities for technology to support mental health assessment and intervention in schools. With their strong ability to understand and generate human-like text, LLMs could potentially analyze student communications for early signs of distress or offer tailored wellness resources.

However, deploying general-purpose LLMs, especially powerful proprietary systems like ChatGPT and Gemini, directly in educational contexts raises significant concerns. Data privacy and the confidentiality of sensitive student information are major issues. Reliance on cloud-based APIs, common for many large LLMs, limits offline use, posing challenges for schools with restricted network access or strict data protection policies. Furthermore, the black box character of many large language models prevents any direct inspection of their internal decision processes. Because it is not possible to determine how a given output was generated, this opacity undermines the interpretability and accountability required for responsible use in student mental health support.

To tackle these critical issues surrounding student data, we recognized the need for LLM-based solutions prioritizing security, privacy, and interpretability. Furthermore, such solutions must operate effectively within the often resource-constrained or entirely offline environments found in educational institutions. In response, our research investigates a novel approach centered on an open-source, lightweight language model, Google's Gemma2-9B, augmented with Retrieval-

Augmented Generation (RAG). While we drew inspiration from prior work like MentaLLaMA, which highlighted the potential of LLMs in mental health analysis, we selected Gemma2-9B as our foundational model. We deemed Gemma2 particularly suitable due to its favorable balance of performance and efficiency, its open-source availability, and its demonstrated capacity for handling sensitive content without the overly strict filtering observed in some alternatives. We refer to our proposed model as MentalGemma2.

Our primary objective in this study is to assess the performance of MentalGemma2, particularly when enhanced with RAG (denoted MentalGemma2-RAG), for conducting secure and private offline mental health analysis. We also explore its use in specific educational settings, such as teacher-training courses in educational technology, AI modules within computer science curricula, and professional-development workshops on ICT integration, where students and educators can gain practical experience with AI tools that preserve data privacy. Integrating RAG enables the model to leverage external knowledge bases, potentially including curated educational materials or local support resources. This integration offers a pathway to improve the accuracy and grounding of the model's outputs while maintaining data privacy, as sensitive information can remain local. Gemma2's lightweight nature facilitates offline deployment, ensuring the accessibility and confidentiality crucial for mental health tools within schools. To evaluate this approach, we systematically test various configurations of MentalGemma2 and MentalGemma2-RAG using the Interpretable Mental Health Instruction (IMHI) dataset. Our evaluation focuses on comparing their ability to detect mental health states from text and assesses the quality of the explanations they generate.

By demonstrating the feasibility and effectiveness of using open-source, lightweight models like MentalGemma2 for offline mental health analysis, this work contributes to the development of technological tools that inherently respect privacy, offer cost-effectiveness, and are potentially more understandable. We believe advancements like these hold significant implications for technology education, offering concrete examples of responsible AI development while providing potential tools to better support student well-being within the educational environment.

Related Work

This section surveys three areas of prior work relevant to our approach: mental-health analysis with large language models (LLMs), the MentaLLaMA framework for interpretable diagnosis, and retrieval-augmented generation (RAG), as well as advances in lightweight LLM architectures.

1. MentaLLaMA

Yang et al. (2024) introduced MentaLLaMA, which fine-tunes LLaMA2-7B and -13B on the Interpretable Mental-Health Instruction (IMHI) dataset, a collection of social media posts labeled with mental-health conditions. Through instruction fine-tuning, MentaLLaMA produces both diagnostic labels and accompanying rationales in a question-answer format, providing transparent insight into its predictions.

While MentaLLaMA showed good results, being based on the Llama2 architecture, it may have certain limitations compared to newer architectures like Gemma2 in terms of efficiency or specific capabilities. Our work builds upon the inspiration from MentaLLaMA but adopts a different, more recent base model.

2. Retrieval-Augmented Generation (RAG)

Lewis et al. (2020) proposed RAG, which augments pre-trained LLMs with retrieved documents to ground generation in real evidence. By conditioning on both the input and top-k relevant passages, RAG mitigates hallucination and grants access to up-to-date or domain-specific knowledge—properties valuable for mental health applications. Although RAG has proven effective in medical question answering, its integration with lightweight, offline LLMs remains underexplored. In this work, we assess RAG in conjunction with the Gemma 2-9B-it model for secure, interpretable offline mental-health analysis.

3. Lightweight Language Models

Lightweight LLMs enable on-device processing, reduce latency, and lower energy consumption. Examples include LLaMA3 and Gemma2-9B-it, both of which are designed for high performance at a practical scale. Gemma2-9B-it, developed by Google DeepMind (Gemma2 Team, 2024), employs a novel model architecture that balances efficiency and accuracy.

In this study, we use the Gemma2-9B-it variant because it runs entirely offline, addressing the strict privacy and accessibility requirements of educational settings. Moreover, in our initial tests, Gemma2-9B-it did not exhibit any problematic content filtering, making it well suited for sensitive mental-health discussions.

Materials and Methods

1. Dataset

We used the Interpretable Mental-Health Instruction (IMHI) dataset (Yang et al., 2024) for all experiments. IMHI comprises five Reddit-derived subtasks—depression detection, stress detection, interpersonal risk-factor identification, multi-dimensional wellness assessment, and stress-cause classification—each labeled with both a target and a human-written rationale. Although IMHI includes these rationales, our study evaluates only classification accuracy and explanation quality via weighted F1-score and BARTScore. The five subtasks are summarized in Table 1 and together provide a comprehensive evaluation of mental-health analysis performance.

Table 1. Overview of Key Mental Health Tasks in the IMHI Dataset (Adapted from Yang et al., 2024)

Data	Task	Description
DR	depression detection	Binary classification (Yes/No) for the presence or absence of depression.
Dreaddit	stress detection	Binary classification (Yes/No) for the presence or absence of stress.
Irf	interpersonal risk factors detection	Binary classification assessing risks like Thwarted Belongingness or Perceived Burdensomeness.
MultiWD	Wellness dimensions detection	Multi-question binary classification assessing dimensions like Spiritual, Physical, Vocational.
SAD	stress cause detection	Multi-class classification identifying the stress cause (School, Finance, Family, Social Relation, Work, Health, Emotion, Decision, Others).

The IMHI dataset includes two prompt formats: completion data and instruction data. In the completion format, each example consists of the original post and a question ending with "is," and the model is trained to complete the prompt by outputting a classification label followed by "Reasoning:" and its rationale. The instruction format adds an explicit directive—such as "Consider this post:"—before the question, requiring the model to produce both the classification and a full explanatory sentence (e.g., "the poster does not suffer from depression.") in a single response. The completion format evaluates fill-in-the-blank style completion ability, while the instruction format assesses the model's capacity to interpret and execute more complex, conversational instructions.

2. Model Configurations

The backbone of our system is the open-source Gemma2-9B-it model (Gemma Team, 2024), chosen for its balance of accuracy, computational efficiency, offline capability, and robust content handling. We compare four configurations:

(1) **MentalGemma2 (few-shot prompting)**

We prepend one example per class—drawn from the IMHI training set—to each input at inference time. These demonstrations guide the base Gemma2-9B-it model in classifying new instances.

(2) **MentalGemma2-RAG**

We augment the base model with RAG. For each query, we retrieve the top three relevant documents from the IMHI training set and include them in the prompt. We test three retrieval strategies:

BM25: classic keyword matching using the BM25 ranking function.

Vector: semantic search via multilingual-e5-large embeddings, re-ranked by Jina-reranker-v2.

Hybrid: a weighted combination of BM25 and vector scores (3 : 7 ratio).

(3) **MentalGemma2-SFT (supervised fine-tuning)**

We fine-tune Gemma2-9B-it on each task using 4-bit QLoRA (quantized low-rank adaptation) with rank = 64 and learning rate = 1×10^{-6}. Owing to different label schemes, we train separate models for SAD and for the remaining four subtasks (DR, Dreaddit, IRF, MultiWD).

(4)**MentalGemma2-SFT-RAG (fine-tuning + RAG)**

We combine the fine-tuned models from (3) with RAG, using the same hybrid retrieval method described above.

Results and Discussion

This section reports the performance of the four MentalGemma2 configurations on the IMHI dataset. We compare few-shot prompting, retrieval-augmented generation (RAG), and supervised fine-tuning (SFT) under both completion- and instruction-style prompts. Classification accuracy is measured with weighted F1-score, and explanation quality with BARTScore (Yuan et al., 2021).

1. Performance on Completion Data

Table 2 lists weighted F1-scores for all configurations in the completion setting. Supervised fine-tuning (MentalGemma2-SFT) yields the best overall result, averaging 79.72 %, a gain of more than ten points over the few-shot baseline (MentalGemma2, 69.00 %). This large margin underscores the value of adapting the model to the mental-health domain with QLoRA.

Adding RAG to the unfine-tuned model (MentalGemma2-RAG) produces consistent but modest gains—BM25 70.90 %, vector 72.22 %, and hybrid 72.47 %—with the hybrid retriever performing best. In contrast, attaching RAG to the fine-tuned model (MentalGemma2-SFT-RAG) slightly reduces the average score to 76.73 %, indicating that the domain knowledge learned during fine-tuning already covers most of the information supplied by retrieval.

For context, a fine-tuned LLaMA2-7B baseline reported by Yang et al. (2024) reaches only 67.00 %, confirming that MentalGemma2-SFT delivers substantially higher accuracy while running fully offline.

2. Performance on Instruction Data

Table 3 reports weighted F1-scores for the instruction format. Supervised fine-tuning again dominates: MentalGemma2-SFT attains 74.46 %, versus 68.41 % for the few-shot baseline.

RAG improves the unfine-tuned model only modestly, yet, unlike in the completion setting, combining RAG with the fine-tuned model yields a small additional gain:

Table 2. Results on Completion Data (Weighted F1-score (%))

	DR	dreaddit	Irf	MultiWD	SAD	Average
MentalGemma2	85.79	58.94	69.66	75.53	55.08	69.00
MentalGemma2-RAG (BM25)	87.82	63.24	70.74	74.80	57.88	70.90
MentalGemma2-RAG(vector)	86.99	65.91	72.44	74.81	60.94	72.22
MentalGemma2-(hybrid)	88.34	66.84	71.23	75.06	60.89	72.47
MentalGemma2-SFT	**90.09**	**86.23**	**77.91**	**79.97**	**64.42**	**79.72**
MentalGemma2-SFT-RAG(hybrid)	88.34	80.65	74.60	77.31	62.77	76.73
LLaMA2-7B (Yang, et al. 2024)	84.94	61.59	73.50	65.52	49.60	67.00

Table 3. Results on Instruction Data (Weighted F1-score (%))

	DR	dreaddit	Irf	MultiWD	SAD	Average
MentalGemma2	86.23	59.00	68.74	73.64	54.43	68.41
MentalGemma2-RAG (BM25)	86.23	62.81	66.11	74.01	60.25	69.88
MentalGemma2-RAG(vector)	86.77	64.73	66.94	73.96	60.20	70.52
MentalGemma2-(hybrid)	**87.73**	63.69	66.71	73.96	60.08	70.43
MentalGemma2-SFT	86.47	**80.90**	72.54	71.98	60.41	74.46
MentalGemma2-SFT-RAG(hybrid)	85.55	75.71	73.73	**76.26**	62.96	74.84
MentaLLaMA-chat-13B (Yang, et al. 2024)	85.68	75.79	**76.49**	75.11	**63.62**	**75.34**

MentalGemma2-SFT-RAG reaches 74.84 %, edging past SFT alone. This suggests that retrieved context can still provide useful grounding for the more conversational instruction prompts.

All configurations perform several points lower on instruction than on completion inputs, reflecting the higher linguistic complexity of multi-turn prompts. Nevertheless, MentalGemma2-SFT remains competitive with the larger MentaLLaMA-chat-13B (75.34 %; Yang et al., 2024) while running entirely offline on the smaller Gemma2-9B architecture. These findings underscore the influence of prompt style: direct completion prompts appear better suited to short classification tasks, whereas instruction prompts may benefit from supplementary retrieval even after fine-tuning.

3. Comparison with Discriminative Models

Table 4 compares our interpretable generative approach with domain-specific discriminative models MentalBERT and MentalRoBERTa (Ji et al., 2022), as reported by Yang et al. (2024). On completion data, MentalBERT and MentalRoBERTa achieve average F1-scores of 78.98 % and 80.38 %, respectively, slightly above MentalGemma2-SFT's 79.72 %. Although these discriminative models deliver marginally higher classification accuracy, they lack the ability to generate explanatory rationales. In contrast, MentalGemma2-SFT matches their performance while providing interpretable explanations—an essential capability for responsible deployment in sensitive domains such as mental health and education.

Table 4. Performance of MentalBERT and MentalRoBERTa on the IMHI dataset (Weighted F1-score (%)) (Yang, et al. 2024)

	DR	dreaddit	Irf	MultiWD	SAD	Average
MentalBERT(Ji, et al. 2022)	94.62	80.04	76.73	76.19	67.34	78.98
MentalRoBERTa(Ji, et al. 2022)	94.23	81.76	85.33	68.44	72.16	80.38

Table 5. BARTScore Results on Completion Data (−∞~0)

	DR	dreaddit	Irf	MultiWD	SAD	Average
MentalGemma2	-2.90	-2.72	-2.91	-2.87	-2.82	-2.84
MentalGemma2-RAG (BM25)	-2.85	-2.67	-2.83	-2.72	-2.85	-2.78
MentalGemma2-RAG(vector)	-2.75	-2.68	-2.84	-2.75	-2.84	-2.77
MentalGemma2-(hybrid)	-2.90	-2.72	-2.91	-2.87	-2.82	-2.84
MentalGemma2-SFT	-2.42	-2.34	-2.41	-2.03	-2.11	**-2.26**
MentalGemma2-SFT-RAG(hybrid)	-2.51	-2.43	-2.51	-2.17	-2.18	-2.36
LLaMA2-7B (Yang, et al. 2024)	-2.80	-3.00	-3.00	-2.35	-2.65	-2.76

Table 6. BARTScore Results on Instruction Data (−∞~0)

	DR	dreaddit	Irf	MultiWD	SAD	Average
MentalGemma2	-2.73	-2.47	-2.74	-2.44	-2.90	-2.66
MentalGemma2-RAG (BM25)	-2.90	-2.47	-2.90	-2.67	-3.17	-2.82
MentalGemma2-RAG(vector)	-2.85	-2.46	-2.79	-2.55	-3.06	-2.74
MentalGemma2-(hybrid)	-2.85	-2.70	-2.80	-2.55	-3.01	-2.78
MentalGemma2-SFT	-2.43	-2.25	-2.40	-1.89	-2.00	-2.19
MentalGemma2-SFT-RAG(hybrid)	-2.39	-2.22	-2.35	-1.86	-1.95	**-2.15**
MentaLLaMA-chat-13B (Yang, et al. 2024)	-2.85	-2.80	-2.65	-2.15	-3.00	-2.69

4. Quality of Explanations (BARTScore)

Tables 5 and 6 report BARTScore evaluations of each configuration's generated explanations; scores closer to zero indicate greater similarity to the reference rationales. For context, we also include estimated scores for LLaMA2-7B and MentaLLaMA-chat-13B (Yang et al., 2024), whose exact values were inferred from published figures.

(1) **Completion data (Table 5)**

MentalGemma2-SFT achieves the best average BARTScore (–2.26), consistent with its superior F1-score. This demonstrates that fine-tuning improves not only classification accuracy but also explanation quality. In contrast, adding RAG to the base model yields less consistent gains in explanation alignment. MentalGemma2-SFT also outperforms the estimated LLaMA2-7B score (–2.76).

(2) **Instruction data (Table 6)**

MentalGemma2-SFT-RAG attains the top average BARTScore (–2.15), slightly ahead of MentalGemma2-SFT (–2.19). This pattern mirrors the F1 results, suggesting that retrieval can marginally boost rationale quality for the more complex, conversational prompts. Both fine-tuned configurations deliver explanations substantially closer to zero than the untuned models, and match or exceed the estimated MentaLLaMA-chat-13B score (–2.69), confirming the effectiveness of our QLoRA fine-tuning on Gemma2 for producing high-quality, relevant explanations.

Conclusions

We investigated the use of open-source, lightweight language models—specifically Gemma2-9B-it—combined with retrieval-augmented generation (RAG) and supervised fine-tuning (SFT) for secure, private offline mental-health analysis. Using the IMHI dataset, we compared four configurations (few-shot prompting, prompting + RAG, fine-tuning, and fine-tuning + RAG) under both completion and instruction prompts.

Our experiments demonstrate that supervised fine-tuning alone (MentalGemma2-SFT) consistently achieves the highest classification accuracy and explanation quality, matching or surpassing larger, cloud-based baselines (LLaMA2-7B, MentaLLaMA-chat-13B) and domain-specific discriminative models, while running entirely offline. RAG yields modest gains for few-shot prompting but offers negligible benefit after fine-tuning on completion data and only slight improvements on instruction data. These results indicate that QLoRA-based fine-tuning sufficiently imbues the model with domain knowledge, and that retrieval adds little additional value in most cases.

By combining competitive accuracy, interpretable rationales, and full offline operation, MentalGemma2 offers a practical blueprint for deploying mental-health analytics in environments with strict data-privacy requirements—such as schools and clinics.

Limitations and Future Work

Our study relies on the IMHI dataset and automated metrics such as weighted F1-score and BARTScore, which may not capture all aspects of real-world performance. Future work should validate these findings on more diverse, education-derived datasets and incorporate human evaluation of explanation quality. Exploring advanced fine-tuning techniques, more robust retrieval strategies, and improved automatic metrics will further enhance performance.

To bridge our findings with real-world application, we discuss potential deployment scenarios within an educational context like KOSEN. Future work could explore these applications and their associated challenges.

First, the system could be developed into a self-care tool for students. By analyzing their own texts (e.g., journal entries or draft reports) in a fully private, offline environment, students could engage in self-reflection on their well-being. A key challenge for this application would be designing an intuitive user interface for non-expert users that encourages reflection without causing undue anxiety.

Second, the system could function as an assistive tool for academic advisors and counselors. Crucially, such a tool must be positioned not as a diagnostic instrument or a replacement for professional judgment, but as a supplementary tool to enhance human-led support. With a student's freely given informed consent, it could analyze consultation records to provide insights into emotional nuances that might otherwise be missed. The successful and ethical implementation would require establishing clear institutional guidelines for data

handling and interpretation to support, not automate, the counseling process. The system's offline capability is critical in both scenarios, ensuring the security and confidentiality required when handling sensitive personal information.

Finally, comparative studies with next-generation models such as Gemma 3 (Gemma Team, 2025) and specialized architectures such as MentalQLM (Shi et al., 2024) will continue to advance the development of secure, explainable AI tools for mental-health support.

References

Gemma Team, Google DeepMind. (2024). Gemma 2: Improving Open Language Models at a Practical Size. arXiv preprint arXiv:2408.00118.

Gemma Team, Google DeepMind. (2025). Gemma 3 Technical Report. arXiv preprint arXiv:2503.19786.

Ji, S., et al. (2022). MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022), pp. 7184-7190.

Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems 33, pp. 9459-9474.

Shi, J., et al. (2024). MentalQLM: A lightweight large language model for mental healthcare based on instruction tuning and dual LoRA modules. medRxiv preprint.

Yang, K., et al. (2024). Mentallama: Interpretable mental health analysis on social media with large language models. Proceedings of the ACM Web Conference 2024 (WWW '24) , pp. 4489–4500.

Yuan, W., et al. (2021). BARTSScore: Evaluating generated text as text generation. Advances in Neural Information Processing Systems 34, pp. 27263-27277.