Improving RAG for LLM-based Mental Health Risk Detection through Contrastive Embeddings

Anonymous Authors[1]

Abstract

Users can disclose psychological suffering or suicidal intentions during interactions with Large Language Models (LLMs). Therefore, it is important to enable LLMs to recognize suicide and depression risks, which can support early intervention. To avoid costly backbone fine-tuning, retrieval-augmented generation (RAG) has been adopted to enhance prompts with retrieved information. However, in mental health risk detection, texts that are semantically similar but correspond to opposite risk labels can lead RAG to introduce misleading retrieval evidence. In this study, we propose a framework that leverages contrastive learning to reduce the retrieval ambiguity suffered from RAG. Our experiments compare four contrastive embedding variants across two datasets using five evaluation metrics. Results show RAG can improve accuracy over vanilla prompting, with contrastive embeddings providing additional gains. Notably, small LLMs with RAG augmentation achieve comparable performance to larger models without augmentation, demonstrating clear cost-performance trade-offs. Our framework offers a practical guidance for balancing retrieval quality and model capacity in detection.

1. Introduction

As large language models (LLMs) are gradually integrated into daily conversation systems, more and more users may disclose psychological suffering or even suicidal intentions in their interactions with LLMs (Li et al., 2025; Levkovich & Elyoseph, 2023; Hua et al., 2024). If these systems can possess mental health risk recognition capabilities, they can assist in early intervention and bring positive social value. The backbone models of such systems are typically pretrained. If directly fine-tuning them for suicide risk detection, which not only requires substantial computational costs but also modifies the backbone parameters, potentially degrades performances on other tasks. Therefore, a desirable approach is to endow LLMs with additional information, but without modifying the parameters of the pretrained backbone.

The Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Gao et al., 2023; Singhal et al., 2023) has been widely applied in improving the LLM's factual judgements without any extra LLM training process. This method provides relevant retrieved samples based on the semantic similarity to enrich the prompt contextual information for the LLMs, improving both detection accuracy and robustness. Nevertheless, RAG still faces challenges (Barnett et al., 2024) when texts are lexically and semantically similar but correspond to opposite labels. Texts often contain only a tiny fraction of suicide signals in the suicide detection task, while the rest may be neutral or opposing. Due to the overall semantic similarity, RAG may retrieve samples that are highly relevant to original texts but have opposite tendencies, introducing retrieval ambiguity and weakening model accuracy.

To address this challenge, we propose a framework by utilising contrastive learning (Khosla et al., 2020) in the RAG process to enhance the suicide risk detection capability of LLMs. Contrastive learning has been widely applied in text generation (An et al., 2022) and cross-lingual applications (Pan et al., 2021). It can optimise the semantic representation, enabling the model to narrow the embedding distance between 'suicide-suicide' samples and push away 'suicide-non-suicide' samples, thereby producing higher discriminability semantic embeddings, which may reduce the semantic confusion and improve the retrieved examples. Unlike general-purpose RAG frameworks that focus on broad semantic similarity, our method targets a critical failure mode in mental health risk detection: texts with highly similar surface semantics but opposite risk labels.

We conduct experiments on integrating contrastive learning into RAG on two public datasets, systematically evaluating different contrastive objectives to identify which best mitigates such retrieval ambiguity. The results demonstrate the effectiveness of our framework. We also analyse the impact

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

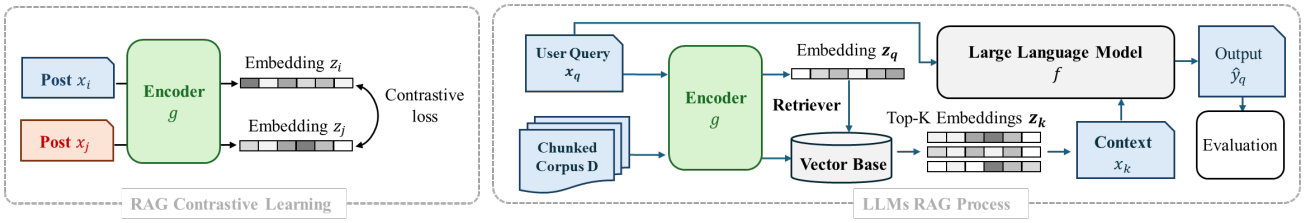

Figure 1. Proposed LLM RAG Framework with Contrastive Embeddings. (left) The encoder $g(\cdot)$ is first trained with contrastive loss on paired posts. (right) During retrieval, the trained $g(\cdot)$ encodes user queries and document chunks into embeddings for similarity matching. Top-K retrieved contexts are fed to the LLM for response generation.

of thinking mode on LLMs under our proposed framework, which shows that the small-sized LLMs benefit from thinking mode, whereas the large-sized LLMs show diminishing or even negative returns.

2. Methodology

Figure 1 provides outlines our proposed framework, which integrates contrastive learning and RAG process.

2.1. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is a framework designed to improve the domain adaptability and factual consistency of LLMs by providing external knowledge through information retrieval. In this study, we introduce RAG to enhance LLMs' adaptability to tiny differences in semantics related to suicide and depression posts.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a dataset where x_i represents a social media post and $y_i \in \mathcal{Y}$ its corresponding label. The LLM for the classification task is denoted as $f_\theta : \mathcal{P} \to \mathcal{Y}$ with parameters θ, where $\mathcal{P}$ denotes the prompt constructed from the query and retrieved examples. RAG enhances LLM inference by retrieving informative examples from $\mathcal{D}$ and including them in the prompt. The RAG setting used in this study includes three key components:

(1) Encoder: An encoder g can map an text input x_i to a vector embedding $z_i = g(x_i)$. The goal of an encoder is to keep semantically related texts locally clustered in the embedding space to support further retrieval.

(2) Retriever: A retriever can retrieve the most relevant post from an external knowledge corpus based on the encoded query. For a query x_q, retrieve top-K samples $(x_k, y_k)_{k=1}^{K}$ from the corpus $\mathcal{D}$ based on the similarity score $\mathrm{sim}(g(x_q), g(x_k))$. Similarity is measured by cosine similarity $\mathrm{sim}(z_q, z_k) = \frac{z_q^\top z_k}{\|z_q\| \, \|z_k\|}$.

(3) Prompt: A prompt framework can integrate retrieved samples with a user query as input to the LLM for inference. The model then predicts the label based on the query and the RAG-selected demonstrations $\hat{y} = f_\theta\big(x_q \,\big|\, \{(x_k, y_k)\}_{k=1}^{K}\big)$.

2.2. RAG with Contrastive Learning

To improve retrieval quality for mental health risk detection, we optimize the retriever's embedding encoder with contrastive learning (CL), which pulls semantically relevant posts closer to the query while pushing irrelevant ones apart in the embedding space. We investigate three contrastive objectives that impose progressively stronger constraints. **SimCSE** (Gao et al., 2021) serves as a label-free semantic baseline, using dropout-based augmentations to assess whether generic sentence-level discrimination alone benefits risk-aware retrieval. **Triplet Loss** (Yanga et al., 2025) leverages risk labels with hard negative mining to explicitly separate embeddings that are semantically similar but carry opposite risk labels. **HAR Loss** (Hu et al., 2025) further introduces hardness-aware reweighting as a supplementary variant, adaptively emphasizing difficult positives and negatives beyond the uniform margin used in triplet learning. Together, these objectives let us examine how different contrastive constraints affect retrieval quality for risk detection. Detailed formulations are provided in the Appendix B.2.

3. Experiments

To evaluate LLM prediction ability in classification task, we consider both zero-shot and RAG-based few-shot strategies, as well as both thinking and non-thinking modes in all settings. For few-shot methods, we compare RAG using embeddings directly from a frozen pre-trained encoder, RAG using representations obtained by adding and fine-tuning a lightweight MLP classifier on top of the frozen encoder, and RAG using contrastively trained embeddings. Three contrastive losses, SimCSE, Triplet Loss, and HAR Loss, are explored, and the optimal choice of K shots is discussed.

3.1. Datasets, Models and Evaluations Setting

Datasets. We use two labeled Reddit datasets from Kaggle: one for the binary task and one for the three-class task (Komati, 2021). Each dataset contains a `text` field and a `class` field. For fair comparison, we use fixed, balanced test subsets with equal class counts and a fixed random seed. The default K value in demonstration selection is set to 5 in

few-shot selection strategies. We emphasize that this study uses public, non-clinical datasets. Our goal is about method validation rather than clinical deployment. Any real-world application would require validation with clinical experts and specialized datasets.

Models. The experiments are under 4 large language models with different sizes including Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B, and Qwen3-8B (Yang et al., 2025). We deploy Qwen3 series models to isolate RAG's impact, ensuring fair comparisons within one model family, and keep our research focusing on evaluating our proposed method rather than evaluating different LLMs under the same setting. Results are reported in both LLMs' thinking and non-thinking modes. The sentence encoder $g(\cdot)$ is set to `all-MiniLM-L6-v2` (MiniLM) based on (Wang et al., 2020).

Metrics. The experiments are evaluated under five metrics, including Accuracy (Acc), Precision (Prec), Recall (Rec), F1-score (F1), and Area Under the ROC Curve (AUC).

3.2. Results on Zero-shot and RAG-Based Few-Shot

Table 1 compares LLM predictive performances under zero-shot and few-shot settings. The few-shot is extracted under standard RAG settings. RAG yields substantial overall gains compared to the zero-shot baseline, with the largest improvements for smaller models (e.g., Qwen3-0.6B, Qwen3-1.7B) and for binary classifications. Gains on the three-class task are more modest, likely due to semantic overlap among 'depression' and 'suicidewatch' labels, which reduces the discriminative value of retrieved evidence. Notably, the incremental benefits of RAG in the Thinking Mode are smaller than in the Non-Thinking mode, plausibly because the built-in reasoning enhancements of the former leave less headroom for additional improvements from retrieval.

Based on our proposed three contrastive learning objectives strategies (SimCSE, Triplet Loss and HAR loss), we further evaluate the performances of the improved RAGs with standard RAG or Zero-shot. Overall, the performance differences among contrastive-learning variants are marginal. A plausible explanation is twofold: for binary classification, the pretrained encoder operates near saturation, leaving limited headroom for additional gains from contrastive learning; for three-class settings, most contrastive objectives fail to adequately mitigate noise arising from semantic overlap among labels, thereby limiting improvements. However, contrastive learning with triplet loss consistently outperforms the baseline RAG with a pre-trained encoder across most experimental conditions. This advantage is particularly pronounced in 3-class classification tasks, as corroborated by manually curated test datasets. The results suggest that anchor-based learning enables more discriminative semantic representations, facilitating retrieval of more relevant examples. HAR loss is evaluated only on the three-class task, as

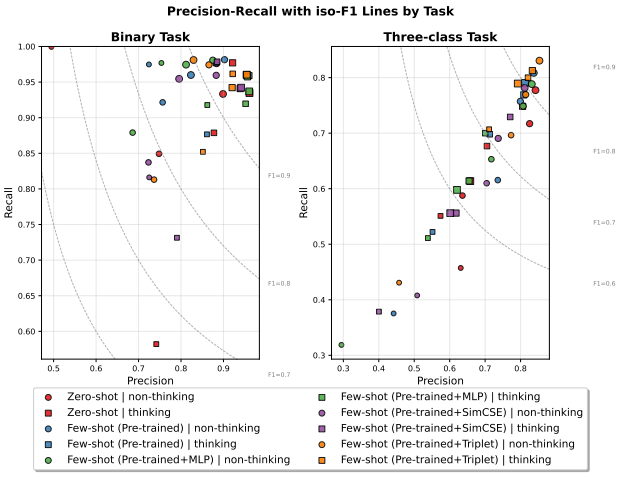

Figure 2. Precision–Recall Analysis with iso-F1 Curves.

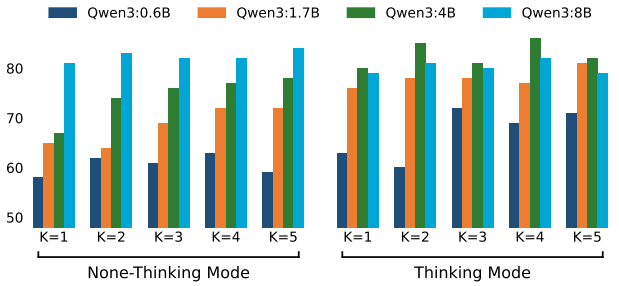

Figure 3. Accuracy across Different K Demonstrations with Triplet Loss on Three-class classification

the three-class setting involves greater semantic overlap between labels, where the reweighting mechanism is designed for. HAR loss shows some improvements in Qwen3-4B and Qwen3-8B (in Table 2), however, performance decreases on small-sized Qwen3 models, which illustrates that the optimization effects of HAR Loss can rely on LLMs size.

3.3. Parameters Sensitivity Analysis

Precision–Recall Analysis with iso-F1 Curves. Figure 2 shows precision values on the x-axis and recall values on the y-axis, which served to evaluate model performance in positive class prediction. The "iso-F1" lines (grey lines) show where precision and recall pairs have the same F1-score values. If the curves is close to the top right corner and in a high F1 area (eg. F1 = 0.9), the model has a better balance between precision and recall.

For binary classification, enabling Thinking mode reliably increases recall, with large gains for small and medium models (Qwen3 0.6–1.7B) and little change for larger models such as Qwen3-8B, suggesting limited headroom. Among methods, Few-shot (Pre-trained + Triplet) and Few-shot (Pre-trained + SimCSE) cluster near the iso-F1 = 0.9 band

Table 1. LLMs Prediction Performance Comparison across Zero-shot and Few-shot RAG with Different Variations. **Bold numbers** indicate each LLM's best score across methods for a given metric within the task.

		Binary (Thinking)				Binary (Non-thinking)				Three-class (Thinking)				Three-Class (Non-thinking)			
		0.6B	1.7B	4B	4B	0.6B	1.7B	4B	4B	0.6B	1.7B	4B	4B	0.6B	1.7B	4B	4B
Zero-shot	Acc	0.697	0.879	0.949	0.949	0.500	0.700	0.930	**0.920**	0.545	0.677	0.768	**0.828**	0.420	0.580	0.730	0.790
	Prec	0.737	0.875	0.922	0.957	0.400	0.745	0.887	**0.900**	0.574	0.704	0.806	0.655	**0.627**	0.641	**0.822**	0.838
	Rec	0.583	0.875	**0.979**	0.938	**1.000**	0.854	0.879	0.938	0.549	0.676	0.746	0.611	**0.462**	0.589	0.713	0.774
	F1	0.651	0.875	0.949	0.947	0.658	0.760	0.931	**0.918**	0.458	0.675	0.765	0.628	0.351	0.145	0.734	0.795
	AUC	0.694	0.879	0.950	0.949	0.519	0.792	0.932	**0.921**	0.661	0.748	0.808	**0.865**	**0.593**	0.677	0.779	0.828
Few-shot RAG (MiniLM)	Acc	0.870	0.930	**0.960**	**0.960**	0.810	**0.820**	**0.940**	0.880	0.545	0.697	0.778	0.788	0.440	0.630	0.750	0.800
	Prec	0.857	0.887	**0.958**	**0.958**	0.723	**0.769**	**0.904**	0.821	0.552	0.710	0.813	**0.814**	0.444	0.727	0.795	0.833
	Rec	0.875	**0.979**	0.958	**0.958**	0.979	**0.917**	**0.979**	0.958	0.520	0.699	0.768	**0.787**	0.377	0.618	0.753	0.804
	F1	0.866	0.931	**0.958**	**0.958**	0.832	**0.830**	**0.940**	0.885	0.475	0.699	0.784	**0.798**	0.309	0.375	0.769	0.816
	AUC	0.879	0.932	**0.960**	**0.960**	0.817	**0.824**	**0.942**	0.883	0.634	0.798	0.822	0.835	0.531	0.703	0.807	0.847
Few-shot RAG (MiniLM+MLP)	Acc	**0.890**	**0.940**	**0.960**	0.950	**0.840**	0.750	0.920	0.880	0.530	0.680	**0.830**	0.790	0.360	0.650	0.750	0.790
	Prec	**0.863**	**0.957**	**0.958**	0.957	**0.758**	0.689	0.870	0.810	0.536	0.703	0.651	0.616	0.299	0.718	0.804	0.828
	Rec	**0.917**	0.917	0.958	0.938	0.979	0.875	**0.979**	**0.979**	0.507	0.695	0.617	0.594	0.322	0.656	0.744	0.792
	F1	**0.889**	0.936	**0.958**	0.947	0.855	0.771	0.922	0.887	0.460	0.681	0.630	0.601	0.275	0.611	0.758	0.805
	AUC	**0.891**	0.939	**0.960**	0.950	0.845	0.755	0.922	0.884	0.627	0.760	0.768	0.805	0.485	0.729	0.801	0.838
Few-shot RAG (MiniLM+SimCSE)	Acc	0.780	0.930	0.940	0.940	0.760	0.770	0.920	0.860	0.530	0.740	0.770	0.730	0.480	0.620	0.690	0.780
	Prec	0.795	0.887	0.938	0.938	0.722	0.727	0.885	0.793	0.401	0.775	0.614	0.598	0.513	0.704	0.740	0.810
	Rec	0.729	**0.979**	0.938	0.938	0.813	0.833	0.958	0.958	0.383	0.733	0.557	0.553	0.411	0.605	0.691	0.785
	F1	0.761	0.931	0.938	0.938	0.765	0.777	0.920	0.868	0.367	0.748	0.574	0.573	0.377	0.575	0.706	0.794
	AUC	0.778	0.932	0.940	0.940	0.762	0.772	0.921	0.864	0.568	0.795	0.647	0.682	0.561	0.695	0.759	0.832
Few-shot RAG (MiniLM+Triplet)	Acc	0.860	**0.940**	0.930	**0.960**	**0.840**	0.770	0.920	0.890	**0.700**	**0.800**	0.810	0.780	**0.580**	**0.710**	**0.770**	**0.830**
	Prec	0.854	0.920	0.918	**0.958**	0.750	0.736	0.870	0.825	**0.712**	**0.820**	**0.833**	0.793	0.460	**0.769**	0.811	**0.849**
	Rec	0.854	0.958	0.938	**0.958**	**1.000**	0.813	**0.979**	**0.979**	**0.712**	**0.804**	**0.815**	0.786	0.430	**0.700**	**0.772**	**0.833**
	F1	0.854	**0.939**	0.928	**0.958**	**0.857**	0.772	0.922	0.895	**0.698**	**0.810**	**0.821**	0.789	**0.379**	**0.706**	**0.788**	**0.838**
	AUC	0.860	**0.941**	0.930	**0.960**	**0.846**	0.772	0.922	0.893	**0.777**	**0.848**	**0.857**	0.834	0.464	**0.769**	**0.822**	**0.871**

Table 2. Few-shot RAG performance on MiniLM with HAR Loss-based contrastive optimization in Three-class Dataset

Mode	LLMs	Acc	Prec	Rec	F1
Thinking	Qwen3-0.6B	0.520	0.368	0.369	0.364
	Qwen3-1.7B	0.660	0.667	0.676	0.671
	Qwen3-4B	0.840	0.871	0.832	0.846
	Qwen3-4B	0.850	0.662	0.639	0.649
Non-thinking	Qwen3-0.6B	0.470	0.309	0.306	0.299
	Qwen3-1.7B	0.690	0.725	0.667	0.681
	Qwen3-4B	0.790	0.620	0.586	0.599
	Qwen3-4B	0.840	0.643	0.637	0.638

and generally outperform others. Few-shot (Pre-trained + MLP) attains higher recall at a modest precision cost, whereas Zero-shot often maintains high precision only at low recall and can underperform on both axes. For the three-class task, performance depends more on context and Thinking mode. Non-thinking mode particularly harms small models, while Thinking mode consistently lifts recall without destabilizing precision, most notably for the Triplet variant, which shifts operating points toward iso-F1 = 0.7. The distribution moves downward, with recall around 0.5–0.8 and precision around 0.4–0.8, reflecting greater task complexity. Overall, Thinking mode mainly improves recall, the triplet-based few-shot approach is most balanced, and binary classification already sits in a high-precision, high-recall regime with limited room to improve.

Impact of K Values on LLM Performance. We perform the study exploring different values of K (1 to 5) to assess retrieval size's impact on LLM few-shot learning performance. Results in Figure 3 demonstrate an overall monotonic improvement with increasing K, with smaller models (0.7B) showing greater sensitivity and benefiting more substantially from additional demonstrations compared to larger models (8B). Performance gains plateau beyond K=4 for most tasks, indicating the trade-off between information richness and computational efficiency. These trends suggest that smaller models rely more heavily on in-context examples to compensate for parametric knowledge, whereas larger models may have internalized sufficient task-relevant information and thus exhibit reduced sensitivity to additional demonstrations. Future work exploring more strategic few-shot retrieval methods (Jin et al., 2024; Hu et al., 2024; Gabouj et al., 2025) would be beneficial for optimizing performance-efficiency trade-off across different model scales.

4. Conclusion

This study explores how contrastive learning can be leveraged into retrieval-augmented generation to deal with the challenge when samples are semantically similar but do not have the same risk labels, and further to improve LLMs' performance on mental health risk detection. Experiments on two public datasets, four differently sized LLMs, test different contrastive learning losses. The results show that while standard RAG already improves performance over vanilla prompting, contrastive embeddings are particularly beneficial. These findings highlight the potential of contrastive learning to enhance retrieval-augmented methods in sensitive mental health applications. Future work may explore more contrastive signals, such as contextual user information, to further improve fine-grained risk detection.

Impact Statement

This work aims to improve how large language models detect mental health risk in user text. Better detection may support earlier help for people who express depression or suicidal thoughts, especially when using smaller and lower cost models. The method is tested only on public, non clinical Reddit datasets, so it should not be used as a medical tool without expert review, clinical validation, privacy safeguards, and careful checks for bias and errors.

References

An, C., Feng, J., Lv, K., Kong, L., Qiu, X., and Huang, X. Cont: Contrastive neural text generation. *Advances in Neural Information Processing Systems*, 35:2197–2210, 2022.

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., and Abdelrazek, M. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pp. 194–199, 2024.

Gabouj, O., Charaf, K., Zakazov, I., Baldwin, N., and West, R. Grad: Generative retrieval-aligned demonstration sampler for efficient few-shot reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19226–19244, 2025.

Gao, T., Yao, X., and Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., and Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

Guo, Q., Tang, J., Sun, W., Tang, H., Shang, Y., and Wang, W. Soullmate: An application enhancing diverse mental health support with adaptive llms, prompt engineering, and rag techniques. *arXiv preprint arXiv:2410.16322*, 2024.

Hu, J., Liu, W., and Du, M. Strategic demonstration selection for improved fairness in llm in-context learning. *arXiv preprint arXiv:2408.09757*, 2024.

Hu, J., Bo, H., Hong, J., Liu, X., and Liu, W. Mitigating degree bias adaptively with hard-to-learn nodes in graph contrastive learning. *arXiv preprint arXiv:2506.05214*, 2025.

Hua, Y., Liu, F., Yang, K., Li, Z., Na, H., Sheu, Y.-h., Zhou, P., Moran, L. V., Ananiadou, S., Clifton, D. A., et al. Large language models in mental health care: a scoping review. *arXiv preprint arXiv:2401.02984*, 2024.

Hwang, Y., Yun, H., and Jung, K. Contrastive learning for context-aware neural machine translationusing coreference information. *arXiv preprint arXiv:2109.05712*, 2021.

Jin, B., Yoon, J., Han, J., and Arik, S. O. Long-context llms meet rag: Overcoming challenges for long inputs in rag. *arXiv preprint arXiv:2410.05983*, 2024.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Komati, N. Suicide and depression detection. Kaggle, 2021. URL https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch. Accessed: 2026-05-02.

Levkovich, I. and Elyoseph, Z. Suicide risk assessments through the eyes of chatgpt-3.5 versus chatgpt-4: vignette study. *JMIR mental health*, 10:e51232, 2023.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Li, T., Yang, S., Wu, J., Wei, J., Hu, L., Li, M., Wong, D. F., Oltmanns, J. R., and Wang, D. Can large language models identify implicit suicidal ideation? an empirical evaluation. *arXiv preprint arXiv:2502.17899*, 2025.

Nishikawa, S., Ri, R., Yamada, I., Tsuruoka, Y., and Echizen, I. Ease: Entity-aware contrastive learning of sentence embedding. *arXiv preprint arXiv:2205.04260*, 2022.

Pan, X., Wang, M., Wu, L., and Li, L. Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*, 2021.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Uluslu, A. Y., Michail, A., and Clematide, S. Utilizing large language models to identify evidence of suicidality risk through analysis of emotionally charged posts. Association for Computational Linguistics, 2024.

Waaler, P. N., Hussain, M., Molchanov, I., Bongo, L. A., and Elvevåg, B. Prompt engineering an informational

chatbot for education on mental health using a multiagent approach for enhanced compliance with prompt instructions: Algorithm development and validation. *JMIR AI*, 4(1):e69820, 2024.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33: 5776–5788, 2020.

Wang, Y., Zhou, Z., and Wang, J. 2-tier simcse: Elevating bert for robust sentence embeddings. *arXiv preprint arXiv:2501.13758*, 2025.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yanga, Y.-T., Jiangb, J.-Y., Linc, Y.-T., and Changb, C.-Y. Enhancing retrieval-augmented generation with knowledge graph-based soft-labeling and triplet similarity sbert. *ToG*, 4:1–16, 2025.

Zhang, J., Lan, Z., and He, J. Contrastive learning of sentence embeddings from scratch. *arXiv preprint arXiv:2305.15077*, 2023.

Zhang, X., Liu, H., Zhang, Q., Ahmed, B., and Epps, J. Speecht-rag: Reliable depression detection in llms with retrieval-augmented generation using speech timing information. *arXiv preprint arXiv:2502.10950*, 2025.

A. Related Work

A.1. Retrieval-Augmented Generation (RAG)

In the mental health domain, (Guo et al., 2024) implemented an RAG-driven LLMs answering system for psychological support, which combined "prompt control" to enhance the responsiveness of the answering system to various psychological conditions. (Waaler et al., 2024) integrated the multi-agent frameworks with the Self-RAG architecture to design a psychoeducational chatbot, which aimed to improve the reliability of the generated response. (Zhang et al., 2025) introduced an innovative method in suicide and depression detection by combining speech rhythm features with the retrieval process to construct a robust generative model for emotion recognition. (Uluslu et al., 2024) explored the potential usage of LLMs for detecting suicidal ideation within social media texts and suggests that the RAG could help mitigate LLM hallucinations. These works primarily focus on using RAG to improve response quality, while less regard for similar texts may lead to opposite risk labels, posing a unique challenge for retrieval-based methods.

A.2. Contrastive Learning

Prior work on label-aware retrieval for RAG emphasized embedding quality and entity sensitivity. (Gao et al., 2021) leverages dropout-based contrastive learning to build more contrastive sentence embeddings and reports gains over SBERT on STS-B with a simple training recipe, making it a practical retrieval backbone. Entity-aware approaches further highlight core emotional terms to guide retrieval (Nishikawa et al., 2022). For psychologically nuanced inputs with conflicting or multiple emotions, (Wang et al., 2025) proposes a SimCSE-based hierarchical architecture that improves robustness. CorefCL (Hwang et al., 2021) enhance the ability to handle contextual semantic conflicts. (Zhang et al., 2023) use diverse augmentation to broaden positive/negative pairs and mitigate low-resource constraints. In this work, we explore representative contrastive learning methods with differing impacts on retrieval quality, aiming to address the specific challenges of mental health risk detection

B. Experiment Details

B.1. Motivation

As illustrated in Figure 4, sentences embedded using all-MiniLM-L6-v2 and visualized via PCA show that expressions such as 'I am ready to leave for my vacation' and 'I am ready to leave this world' are located close to each other in the embedding space, despite indicating fundamentally different suicide risk levels.

Figure 4. PCA Visualization of Embeddings with Opposite Labels.

B.2. RAG with Contrastive Learning

Contrastive learning (CL) is a representation-learning paradigm that trains an encoder to bring sentences with similar semantic information together and push those with different semantic information farther apart in the embedding space,

thereby better capturing semantic information. We use contrastive learning to optimize the embedding encoder of the retriever in RAG. The goal is to make the 'relevant posts' in the corpus closer to the query and the 'irrelevant posts' farther away, thereby improving the overall RAG performance. The core idea is to use contrastive loss to make the posts of positive sample pairs more similar than those of negative sample pairs. In this study, we primarily focus on SimCSE and Triplet loss as our main contrastive objectives, and additionally include HAR loss as a supplementary hardness-aware variant. These allow us to explore how the different contrastive constraints affect retrieval quality in the mental health risk detection task.

SimCSE Loss. Simple Contrastive Sentence Embedding (SimCSE) (Gao et al., 2021) is a self-supervised CL method that uses dropout noise to create different views of the same sentence. It treats representations of identical sentences under different dropout masks as positive pairs, while other sentences in the batch serve as negatives. The contrastive loss (NT-Xent) pulls positive pairs closer and pushes negatives apart (Wang et al., 2025). SimCSE optimizes embeddings based on generic semantic consistency without using label information. In our setting, it serves as a semantic baseline to assess whether improving general sentence-level discrimination alone is sufficient to enhance retrieval for risk detection. Given a batch of $2N$ samples, the loss for a positive pair (i, j) is defined as equation 1, where τ is temperature parameter, $\mathbb{1}_{[k\neq i]}$ is an indicator function excluding the anchor. The final loss is averaged across all positive pairs in the batch $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{i,j(i)}$.

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp\left(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)} \tag{1}$$

Triplet Loss. Triplet contrastive learning (Yanga et al., 2025) trains with triplets (anchor, positive, negative) to minimize anchor-positive distance while maximizing anchor-negative distance. The Triplet Loss function minimizes anchor-positive distances, maximizes anchor-negative distances, and maintains minimum separation α for discriminative learning. Hard negative mining selects the nearest negatives for each anchor in the batch (Equation 2). Here, $\mathcal{N}$ denotes the set of negative samples within the batch that have different risk labels from the anchor, and hard negatives are selected based on cosine similarity in the embedding space. To our task, this mechanism is well suited, as it directly separate embeddings with similar semantic but opposite risk, which are the primary source of retrieval challenge of standard RAG. The final loss is averaged across all positive pairs in the batch is $\mathcal{L} = \frac{1}{2N}\sum_{i=1}^{2N}\mathcal{L}_i$, where

$$\mathcal{L}_i = \frac{1}{|\mathcal{N}(i)|} \sum_{k\in\mathcal{N}(i)} \max\Big(\left\|\mathbf{z}_i - \mathbf{z}_{j(i)}\right\|_2^2 - \left\|\mathbf{z}_i - \mathbf{z}_k\right\|_2^2 + \alpha,\ 0\Big) \tag{2}$$

HAR Loss. Hard-Aware Reweighted (HAR) (Hu et al., 2025) addresses the challenge of imbalanced learning by dynamically adjusting sample weights based on their learning hardness.

The core formulation of HAR loss for sample v_i is defined as below, where POS_i and NEG_i represent the adaptively weighted positive and negative terms, respectively. Given an input sample, we generate two stochastic views via independent dropout masks, producing embeddings z' and z''.

$$\ell_i(z_i', z_i'') = -\log \frac{\mathrm{POS}_i}{\mathrm{POS}_i + \mathrm{NEG}_i} \tag{3}$$

$$\mathcal{L}_i = \frac{1}{2N}\sum_{i=1}^{N}\left[\ell_i(z_i', z_i'') + \ell_i(z_i'', z_i')\right] \tag{4}$$

The positive POS_i incorporates both inter-relationships and self-augmentation emphasis through $\mathrm{POS}_i = \sum_{j=1}^{N}(W_{ij}^{+}\cdot S_{ij}^{+})$, where $W_{ij}^{+} = \alpha\cdot\bar{s}_{ij} + \delta_{ij}$ combines scaled similarity weights with an identity matrix to emphasize hard positives (sample with the same augmentation origin). The negative NEG_i applies debiased reweighting to focus on challenging negative samples, and assigns higher weights to hard negatives (samples with high similarity but different labels):

$$\mathrm{NEG}_i = \max\left(\frac{\sum_{j=1}^{N}(W_{ij}^{-}\cdot S_{ij}^{-})}{1-\tau^{+}}, e^{-1/\tau}\right) \tag{5}$$

$$W_{ij}^{-} = \frac{\beta \cdot S_{ij}^{-}}{\frac{1}{N}\sum_{j=1}^{N} S_{ij}^{-}} \tag{6}$$

The similarity matrices S^{+} and S^{-} are constructed by applying positive and negative masks to the base similarity matrix S, which captures both intra-view and inter-view associations:

$$f_\tau(z_i, z_j) = \exp(\text{sim}(z_i, z_j)/\tau) \tag{7}$$

$$\text{sim}(z_i, z_j) = \frac{z_i \cdot z_j}{\|z_i\|\|z_j\|} \tag{8}$$

$$s_{ij} = f_\tau(z_i', z_j') + f_\tau(z_i', z_j'') \tag{9}$$

The re-weighting can help the retriever to focus on cases that have highly similar semantics but are difficult to separate by labels beyond uniform triplet constraints.

B.3. Evaluation Metrics

Evaluation metrics used in the experiments are derived from the confusion matrix, which includes True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). AUC is calculated by integrating the area under the ROC curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR). The equations for calculating Acc, Prec, Rec and F1 metrics are as below. For all these metrics, higher values indicate better model performance.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}, \qquad \text{Prec} = \frac{TP}{TP + FP},$$

$$\text{Rec} = \frac{TP}{TP + FN}, \qquad \text{F1} = \frac{2\,\text{Prec} * \text{Rec}}{\text{Prec} + \text{Rec}}.$$

B.4. Results on the Impact of Thinking Mode on LLM Performance

Figure 5 compares accuracy changes between using and not using thinking mode. For a given LLM and dataset, we define accuracy gain as the difference between thinking mode and non-thinking mode accuracy. Binary classification results show thinking mode delivers higher gains for smaller models. Qwen3-0.6B achieves a substantial $+0.197$ improvement in zero-shot settings, while larger models (Qwen3-4B, Qwen3-8B) show minimal gains ($\leq +0.08$). Medium-sized Qwen3-1.7B shows optimal cost-performance balance with $+0.190$ gains using few-shot methods. Three-Class Classification presents more complex dynamics but maintains the core advantage for smaller models. Qwen3-0.6B still achieves significant gains ($+0.170$), while larger models become unstable with alternating positive and negative effects. Although overall improvements are reduced compared to binary tasks, small models consistently benefit from thinking mode, whereas large models show diminishing or negative returns. This pattern suggests smaller models have greater optimization space, offering a viable path to achieve comparable performance to larger models while reducing inference costs.

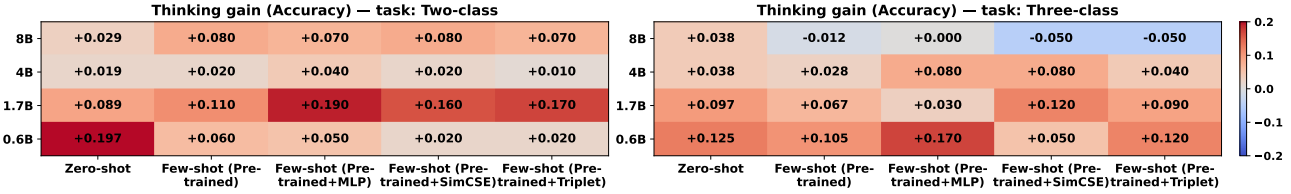

Figure 5. Accuracy Gains of Thinking Mode across Different LLMs