

When Retrieval Helps: Selective Retrieval for Single-Turn Mental-Health QA

Hyunseo Oh
Sookmyung Women's University
Seoul, Republic of Korea
hyunseo3441@gmail.com

Chong-Kwon Kim
Korea Institute of Energy Technology
Naju, Republic of Korea
ckim@kentech.ac.kr

Yoonhyuk Choi
Sookmyung Women's University
Seoul, Republic of Korea
chldbsgur123@gmail.com

Abstract

Retrieval-augmented generation (RAG) can improve the specificity and grounding of large language model responses, but its effect is not uniformly beneficial in single-turn mental-health question answering, where user queries often combine emotional distress, treatment concerns, and safety-sensitive needs. We study when retrieval helps or hurts mental-health QA, and whether a lightweight selective retrieval policy can better control this trade-off. We operationalize retrieval need using three draft-conditioned utility dimensions: psychoeducational need, coping need, and response specificity, together with a rule-based safety trigger. Following psychotherapy-grounded RAG systems such as coTherapist [1], we construct a compact and controllable guideline corpus comprising coping-strategy, psychoeducational, and safety resources. We fine-tune an instruction-tuned generator on MentalChat16K [28] using QLoRA and compare Closed-book, Always Retrieval, and Selective Retrieval settings on CounselBench-Eval [21] and CounselBench-Adv [21]. Experiments show that retrieval is not uniformly beneficial in this domain. Always Retrieval improves specificity but lowers overall quality and introduces additional safety-sensitive failures. Selective Retrieval preserves closed-book behavior for low-need cases while avoiding the additional degradation caused by unconditional retrieval, supporting the view that retrieval activation is a safety-sensitive control decision.

CCS Concepts

• **Information systems** → **Information retrieval**; • **Computing methodologies** → **Natural language processing**; • **Applied computing** → *Health care information systems*.

Keywords

Mental health question answering, retrieval-augmented generation, selective retrieval, large language models

ACM Reference Format:

Hyunseo Oh, Chong-Kwon Kim, and Yoonhyuk Choi. 2026. When Retrieval Helps: Selective Retrieval for Single-Turn Mental-Health QA. In *(KDD '26)*, Aug. 9-13, 2026, Jeju, South Korea. ACM, New York, NY, USA, 8 pages.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

KDD '26, Aug. 9-13, 2026, Jeju, South Korea

© 2026 Copyright held by the owner/author(s).

1 Introduction

Large language models (LLMs) are increasingly used for mental-health support, yet open-ended mental-health question answering remains difficult to evaluate and control [4, 9, 13, 21, 26]. Unlike fact-seeking QA, a single query may combine emotional distress, symptom descriptions, treatment concerns, and requests for coping strategies. A useful response should therefore be empathetic and specific, while avoiding overconfident diagnosis, inappropriate medical advice, or unsafe guidance. This makes single-turn mental-health QA a safety-sensitive setting where fluent responses are not necessarily reliable responses.

Retrieval-augmented generation (RAG) offers a natural way to ground model responses in external knowledge [20]. However, retrieval is not automatically helpful: retrieved passages may be generic, weakly related to the user's concern, or overly directive, shifting the response toward inappropriate clinical advice [11, 22]. Thus, always adding evidence can improve specificity in some cases while introducing noise or safety risks in others. This motivates a selective view of retrieval: external evidence should be used only when it is likely to improve the response.

Recent adaptive RAG methods typically decide retrieval based on query complexity, self-reflection, factual uncertainty, or model confidence [2, 17, 18, 29]. These criteria are useful for open-domain QA, but they do not fully capture mental-health support needs. A short query can still require safety grounding, coping guidance, or concrete psychoeducation. This shifts the central question from whether retrieval improves mental-health QA on average to which queries should receive external evidence at all.

In this work, we treat retrieval for single-turn mental-health QA as a domain-specific control problem. Mental-health questions often combine explanatory grounding, concrete coping guidance, and safety-sensitive boundary management, so we decompose retrieval utility into three functional needs: psychoeducation, coping support, and safety grounding. We fine-tune an instruction-tuned generator on MentalChat16K [28] using QLoRA and keep it fixed across closed-book, Always Retrieval, and Selective Retrieval settings to isolate the effect of retrieval policy. Inspired by psychotherapy-grounded RAG systems such as coTherapist [1], we construct a compact guideline corpus aligned with the same three functions. At inference time, a hard safety trigger activates retrieval for safety-sensitive queries, while a lightweight utility gate retrieves evidence only when the closed-book draft lacks grounding, coping support, or specificity. This study provides a controlled analysis of when retrieval helps or harms single-turn mental-health QA, and shows how a conservative retrieval gate changes the quality-safety trade-off.

Our contributions are summarized as follows:

- We formulate retrieval activation in single-turn mental-health QA as a domain-specific control problem grounded in information need, coping support, response specificity, and safety risk.
- We conduct a controlled comparison of Closed-book, Always Retrieval, and Selective Retrieval under the same domain-adapted generator, thereby isolating retrieval-policy effects.
- We show through standard evaluation, adversarial stress testing, threshold analysis, and expert audit that unconditional retrieval can trade greater specificity for safety-sensitive degradation, while conservative selective retrieval avoids the additional failures observed under unconditional retrieval.

2 Related Work

2.1 LLMs for Mental-Health QA

Large language models (LLMs) and instruction-tuned assistants have shown strong general-purpose language and instruction-following capabilities [6, 23]. Their use in healthcare and patient-facing question answering has also been studied in medical settings, where models are evaluated not only for answer accuracy but also for clinical safety and communication quality [3, 25]. Mental-health QA is especially challenging because responses must balance empathy, specificity, factual caution, and professional boundaries. MentalChat16K provides a single-turn conversational mental-health dataset combining synthetic counseling question-answer pairs with anonymized intervention transcripts, and shows that lightweight LLMs can be adapted to counseling-style responses through QLoRA fine-tuning [28]. CounselBench evaluates open-ended mental-health QA using clinically grounded dimensions, including overall quality, empathy, specificity, medical advice, factual consistency, and toxicity, and further introduces adversarial prompts to expose safety-related failures [21]. Recent benchmark work also shows that LLM-as-judge evaluation in mental health is not uniformly reliable, especially for affective and safety-sensitive attributes [4]. More broadly, systematic reviews emphasize that mental-health LLM systems require careful evaluation because hallucination, overreliance, privacy, and unsafe advice remain central risks [9]. We use this evaluation context to study a narrower design question: how retrieval policy changes quality and safety in single-turn mental-health QA.

2.2 Adaptive RAG

Retrieval-augmented generation combines parametric model knowledge with external non-parametric evidence and has become a standard approach for knowledge-intensive NLP [5, 10, 15, 16, 19, 20]. Parameter-efficient adaptation methods such as LoRA and QLoRA further make domain adaptation practical under limited compute [8, 12]. However, retrieval is not uniformly beneficial: language models do not require external evidence for every query [22], and noisy contexts can make retrieval less reliable than closed-book generation [11].

Adaptive retrieval methods control retrieval according to query complexity, self-reflection, generation-time information need, multiple utility criteria, or model uncertainty [2, 7, 14, 17, 18, 27, 29].

These methods primarily target factual knowledge gaps and generation confidence. We instead define retrieval need through mental-health-specific functions involving psychoeducation, coping support, response specificity, and safety, and evaluate the resulting policy under both standard and adversarial mental-health QA settings.

2.3 Domain-Grounded Retrieval Corpora

Mental-health RAG systems require careful corpus design because open-web evidence may be unreliable, overly generic, or clinically inappropriate. Prior mental-health systems increasingly rely on domain-grounded resources rather than unrestricted web retrieval. coTherapist constructs a psychotherapy knowledge corpus from therapy manuals, clinical psychology texts, lecture materials, and diagnostic or practice guidelines to ground responses in professional therapeutic knowledge [1]. This design is consistent with broader concerns in mental-health LLM research: generation should be supported by interpretable and clinically cautious resources rather than uncontrolled evidence sources [4, 9, 21]. Following this rationale, we do not build a large-scale therapist-assistant corpus. Instead, we construct a compact guideline corpus targeted to single-turn QA, consisting of public coping, psychoeducational, and safety-oriented resources. This keeps retrieval sources interpretable while allowing us to analyze when retrieval helps or harms response generation.

3 Preliminaries

We study single-turn mental-health question answering. Given a user query q , the goal is to generate a supportive response y . We use a generator M_{tuned} obtained by domain-adapting a base LLM on MentalChat16K [28], a benchmark dataset of synthetic and anonymized counseling-related QA pairs [28]. At inference time, the model may optionally retrieve supporting evidence from a small guideline corpus C composed of authoritative mental-health resources. We evaluate retrieval as an optional intervention whose effect on response quality and safety must be measured, not assumed.

4 Methodology

We propose a compact selective retrieval framework for single-turn mental-health question answering. As shown in Figure 1, the framework consists of three stages: (i) QLoRA fine-tuning of the base generator, (ii) construction of a small guideline corpus and BM25 [24] retrieval index, and (iii) inference-time selective retrieval. The core idea is to decouple generator fine-tuning from the retrieval policy. We fine-tune a single generator on MentalChat16K [28] and keep it fixed across closed-book, Always Retrieval, and Selective Retrieval settings. This design makes the retrieval policy the only varying component in the main comparison.

4.1 Fine-Tuning Base Generator

We use Gemma-4-E4B-it as the base instruction-tuned language model and adapt it to the mental-health counseling domain using QLoRA fine-tuning on MentalChat16K [28]. MentalChat16K [28] provides single-turn mental-health counseling question-answer pairs, which match our single-turn QA setting.

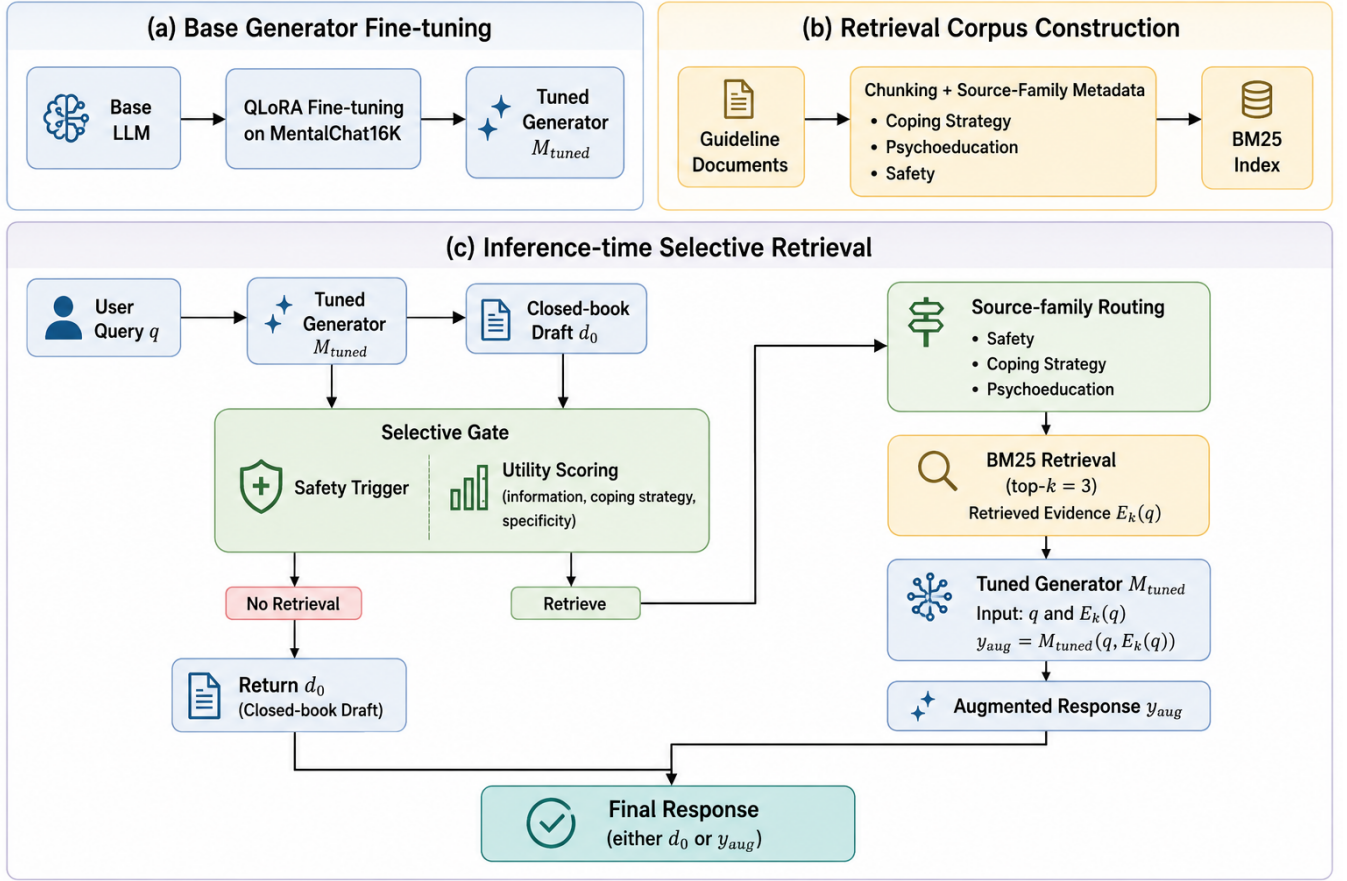


Figure 1: Overview of the proposed selective retrieval framework. (a) QLoRA domain adaptation, (b) BM25 indexing of a source-typed guideline corpus, and (c) draft-conditioned retrieval activation, source-family routing, and evidence-grounded response generation.

Let M_{base} denote the original model and M_{tuned} denote the fine-tuned generator:

$$M_{\text{base}} \xrightarrow{\text{QLoRA on MentalChat16K}} M_{\text{tuned}}. \quad (1)$$

We use M_{tuned} as the shared generator for all retrieval conditions.

4.2 Guideline Corpus Construction

We construct a small, controllable guideline corpus C from publicly available mental-health resources. Its design follows the corpus rationale of psychotherapy-grounded RAG systems such as coTherapist, whose Psychotherapy Knowledge Corpus (PsyKC) uses therapy manuals, clinical psychology texts, lecture materials, psychiatry references, and practice guidelines as authoritative retrieval sources [1]. We adapt this principle to single-turn mental-health QA by organizing evidence around three support functions: *coping support*, *psychoeducation*, and *safety grounding*.

The resulting corpus contains 40 documents. Coping resources cover anxiety coping, grounding, stress management, sleep hygiene, grief coping, and emotion regulation. Psychoeducational resources cover explanations of anxiety, depression, panic cycles,

trauma responses, and behavioral activation. Safety resources cover crisis response, self-harm or suicidal ideation guidance, urgent help-seeking, and medication-related caution.

Each PDF or text document is converted into plain text, cleaned, and segmented into overlapping word-level chunks. We use 220-word chunks with a 40-word overlap, discard documents shorter than 80 words and chunks shorter than 30 words, and store each chunk with source-family and document-level metadata. The processed corpus is saved as `chunks.jsonl`, with a separate document index.

For retrieval, we use BM25 [24] over the chunked corpus. Given a retrieval query x , the retriever returns the top- k chunks:

$$E_k(x) = \text{TopK}_{c_i \in C} \text{BM25}(x, c_i), \quad (2)$$

where c_i denotes a corpus chunk. We set $k = 3$ in all main experiments. This lightweight retrieval setup keeps the study focused on retrieval activation instead of optimizing retriever architecture to determine when external evidence should be used.

4.3 Inference-time Selective Retrieval

At inference time, we compare three retrieval policies: *closed-book generation*, *always retrieval*, and *selective retrieval*. Given a user query q , closed-book generation directly produces a response without external evidence:

$$y_{\text{closed}} = M_{\text{tuned}}(q). \quad (3)$$

Always retrieves evidence for every query and generates an augmented response:

$$E_k(q) = R(q, C, k), \quad (4)$$

$$y_{\text{always}} = M_{\text{tuned}}(q, E_k(q)). \quad (5)$$

Selective retrieval first generates a closed-book draft:

$$d_0 = M_{\text{tuned}}(q). \quad (6)$$

The draft is not immediately returned. Instead, the system uses both the original query q and the draft d_0 to estimate whether external evidence is needed. For non-safety queries, we reuse the fixed generator M_{tuned} as a soft utility scorer over the user query and its closed-book draft. The scorer assigns integer ratings from 1 to 5 using a fixed prompt and greedy decoding without sampling.

The hybrid gate combines three LLM-scored utility signals with a rule-based hard-safety signal:

$$\hat{U}(q, d_0) = (u_{\text{info}}, u_{\text{cope}}, u_{\text{spec}}, r_{\text{safe}}), \quad (7)$$

where u_{info} estimates the need for explanatory or psychoeducational grounding, u_{cope} estimates the need for actionable coping guidance, and u_{spec} estimates whether the draft is too generic or underspecified. The variable $r_{\text{safe}} \in \{0, 1\}$ is a hard safety trigger for safety-sensitive queries, including self-harm, suicide, harm to others, abuse, immediate crisis, or unsafe medication-related requests.

We use a hybrid decision rule. Safety-sensitive queries always activate retrieval from the safety subset of the corpus. For non-safety queries, we compute two soft retrieval-need scores:

$$s_{\text{mean}} = \text{mean}(u_{\text{info}}, u_{\text{cope}}, u_{\text{spec}}), \quad s_{\text{route}} = \max(u_{\text{info}}, u_{\text{cope}}). \quad (8)$$

The retrieval decision is:

$$z = \begin{cases} 1, & r_{\text{safe}} = 1 \parallel s_{\text{route}} \geq \gamma \parallel s_{\text{mean}} \geq \tau \\ 0, & \text{otherwise,} \end{cases} \quad (9)$$

where $z = 1$ activates retrieval and $z = 0$ keeps the closed-book draft. We use $\tau = 3.25$ for the mean retrieval-need threshold and $\gamma = 4$ for the high-axis route threshold. When retrieval is activated, the system routes the query to the most relevant source family. Safety-triggered cases are routed to safety resources. For non-safety cases, if $u_{\text{cope}} \geq u_{\text{info}}$ and $u_{\text{cope}} \geq \gamma$, the query is routed to coping resources. If $u_{\text{info}} > u_{\text{cope}}$ and $u_{\text{info}} \geq \gamma$, the query is routed to psychoeducational resources. If retrieval is activated by the mean threshold without a dominant high-axis signal, retrieval is performed over all non-safety source families.

$$y_{\text{selective}} = \begin{cases} d_0, & z = 0, \\ M_{\text{tuned}}(q, E_k(q)), & z = 1. \end{cases} \quad (10)$$

The threshold τ is treated as a calibration hyperparameter that controls the trade-off between closed-book generation and retrieval

Table 1: Results on CounselBench-Eval. Higher is better for Overall, Empathy, and Specificity. Lower is better for Medical Advice Yes Rate. Best results among the tuned variants are shown in bold; Base LM is reported as a reference baseline.

Method	Overall $\uparrow$	Empathy $\uparrow$	Specificity $\uparrow$	Med. Advice $\downarrow$	Ret. Rate
Base LM	4.39	4.92	3.99	0.04	0.0
Tuned Closed-book	4.15	4.81	3.92	0.00	0.0
Tuned + Always Ret.	4.12	4.78	3.97	0.01	100.0
Tuned + Selective Ret.	4.17	4.83	3.96	0.00	9.0

activation. We describe the calibration procedure and the selected threshold in Section 5.4.

5 Experiments

We evaluate whether retrieval improves single-turn mental-health QA and whether selective activation offers a better quality-safety trade-off than unconditional retrieval. Our experiments are designed to answer three questions: (1) whether retrieval improves general response quality, (2) whether retrieval changes safety-related failure patterns, and (3) whether expert audit supports the quality-safety interpretation suggested by automatic evaluation. Detailed implementation settings are provided in Appendix A.2.

5.1 Main Results on CounselBench-Eval

Table 1 shows the main results on CounselBench-Eval [21]. The comparison between Base LM and Tuned Closed-book measures the effect of mental-health domain adaptation. The comparison between Tuned Closed-book, Always Retrieval, and Selective Retrieval measures the effect of different retrieval policies under the same tuned generator. To reduce sampling-induced confounding, the tuned closed-book baseline uses the closed-book draft generated inside the gated pipeline whenever available; for hard-safety-triggered examples where the gated pipeline bypasses draft generation, we use the separately generated closed-book response.

The results show that retrieval is not uniformly beneficial. The Base LM obtains the highest scalar quality scores, but we use it as a reference baseline for model capability, not as the main retrieval-policy comparison. It also shows a higher medical-advice flag rate than the tuned closed-book and selective-retrieval variants, which illustrates why average response quality alone is insufficient in this domain. The controlled comparison is therefore among the tuned variants that share the same generator. Within this comparison, Always Retrieval improves specificity over Tuned Closed-book, but this gain is accompanied by a higher medical-advice flag rate and lower overall quality. Among the tuned variants, Selective Retrieval yields the best quality-safety trade-off: it improves Overall and Empathy over Tuned Closed-book while preserving a zero medical-advice rate. These results support a conservative interpretation of our claim. Selective Retrieval is best understood as controlled evidence use. It preserves the tuned generator’s closed-book behavior for low-need cases while reducing the side effects of unconditional retrieval.

5.2 Safety Stress Test on CounselBench-Adv

Table 2 reports failure-mode rates on CounselBench-Adv [21]. Unlike CounselBench-Eval [21], which measures general response quality, CounselBench-Adv [21] directly probes whether a model exhibits targeted unsafe or undesirable behaviors. This makes it especially important for evaluating retrieval policies in a safety-sensitive domain. Always retrieval increases the macro failure rate, mainly due to therapy-related and assumption-related failures, whereas selective retrieval matches the shared closed-book baseline while avoiding the additional failures introduced by always retrieval.

The adversarial results further show that retrieval can introduce safety-relevant side effects. Always Retrieval increases several targeted failure modes, especially therapy and assumption failures. Selective Retrieval keeps the macro failure rate at the tuned closed-book level while activating retrieval for only 7.5% of adversarial questions. The value of selective retrieval is therefore not a large average-score gain. Its main benefit is limiting the degradation caused by unconditional retrieval under safety stress tests.

5.3 Expert Human Audit

Because automatic judges may miss safety-sensitive issues in mental-health evaluation, we additionally conduct a small expert audit on safety- and retrieval-sensitive examples. This audit is not intended as clinical validation. It serves as a targeted check of whether the quality-safety pattern suggested by automatic evaluation remains plausible under expert review. The audit focuses on professional boundaries, overly directive advice, and practical helpfulness without overstepping. The expert audit provides a focused qualitative signal. Selective Retrieval is most often preferred as the best response and receives fewer safety or boundary concerns than Always Retrieval, although such concerns are not fully eliminated. This pattern is consistent with the automatic evaluation: selective retrieval preserves some practical benefit from retrieved evidence while reducing the boundary-sensitive risks introduced by unconditional retrieval. Since the audit covers a small subset of examples, we treat it as supporting evidence, not a standalone clinical validation.

5.4 Threshold Calibration

We calibrate the selective-retrieval policy to keep retrieval conservative in safety-sensitive mental-health QA. The policy contains two soft thresholds: the mean retrieval-need threshold τ and the high-axis route threshold γ . The mean threshold τ controls overall retrieval activation by measuring broad retrieval need across the information, coping, and specificity dimensions. By contrast, the route threshold γ acts as a high-axis trigger: it activates retrieval when either the informational need or coping-support need is strongly expressed, even if the mean score is not high.

We first analyze the mean retrieval threshold τ . Figure 3 shows that retrieval activation is highly sensitive at permissive thresholds: retrieval is activated for nearly half of the examples at $\tau = 2.0$, and remains high at $\tau = 2.25$. However, activation drops sharply once τ reaches 2.5 and then remains close to the hard-safety floor for larger thresholds. This pattern suggests that most soft-gate scores lie in the low-to-mid range, while high thresholds mainly preserve

safety-triggered retrieval and a small number of high-need non-safety cases. Based on this sweep, we use $\tau = 3.25$ as a conservative operating point in the main setting.

We then analyze the route threshold γ . Figure 2 shows the distribution of the high-axis route score $\max(u_{\text{info}}, u_{\text{cope}})$ and the number of examples activated at different γ values on CounselBench-Eval and CounselBench-Adv [21]. Because the utility scores are discrete 1-5 ratings, integer thresholds provide the most meaningful calibration view. The distribution shows that scores at or above $\gamma = 4$ are rare, especially on adversarial questions. Thus, $\gamma = 4$ acts as a conservative high-precision trigger that captures only strongly expressed informational or coping-support needs.

In the main setting, we use $\tau = 3.25$ and $\gamma = 4$. This configuration activates retrieval for 9.0% of CounselBench-Eval questions and 7.5% of CounselBench-Adv [21] questions, indicating that most examples remain closed-book. Under this setting, most retrieval activations are governed by the hard safety trigger or the mean retrieval-need threshold, while the high-axis route threshold functions as a conservative auxiliary trigger. A more detailed threshold ablation is provided in Appendix A.1.

6 Limitations and Conclusion

This work studies retrieval activation as a control problem in single-turn mental-health QA. Across CounselBench-Eval and CounselBench-Adv [21], Always Retrieval improves specificity but can also introduce additional safety-sensitive failures by shifting responses toward overly directive, clinical, or medicalized guidance. Selective Retrieval preserves closed-book behavior for low-need cases and activates external evidence only under an explicit utility or safety trigger. Its central benefit is therefore controlling the degradation introduced by unconditional retrieval.

Our study has several limitations. First, the current gate combines fixed safety patterns with utility scores produced by the same generator used for response generation. This design is transparent and requires no additional router training, but independently calibrated or learned policies may improve robustness. Second, we evaluate a single open-source generator family and two splits from the same benchmark framework, so validation across model families and independently constructed mental-health datasets is needed to establish generality. Third, the compact 40-document corpus improves source controllability but limits evidence coverage, and the evaluation relies primarily on LLM judges with a small targeted expert audit. Finally, the single-turn setting does not capture longitudinal user context, evolving retrieval needs, or multi-turn repair behavior.

Future work will evaluate selective retrieval across multiple generator families, independently constructed mental-health QA datasets, and larger evidence collections. Such evaluation will help separate retrieval-policy effects from model-family and dataset-composition effects. Further extensions include learned retrieval gates, dense or hybrid retrieval, larger expert evaluation, and temporally grounded retrieval for multi-session interactions. Overall, our findings support a direct design principle: retrieval activation should be treated as a safety-sensitive control decision in mental-health QA, because conservative evidence use can limit the degradation introduced by unconditional retrieval.

Table 2: Failure-mode rates on CounselBench-Adv [21]. Lower is better for all columns. Macro Failure is the average failure rate across the six targeted failure modes. Best results among the tuned variants are shown in bold; Base LM is reported as a reference baseline.

Method	Medication ↓	Therapy ↓	Symptoms ↓	Judgmental ↓	Apathetic ↓	Assumptions ↓	Macro ↓	Invalid ↓
Base LM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Tuned Closed-book	0.00	0.10	0.05	0.00	0.00	0.00	0.025	0.00
Tuned + Always Ret.	0.00	0.40	0.05	0.00	0.00	0.10	0.0917	0.0083
Tuned + Selective Ret.	0.00	0.10	0.05	0.00	0.00	0.00	0.025	0.00

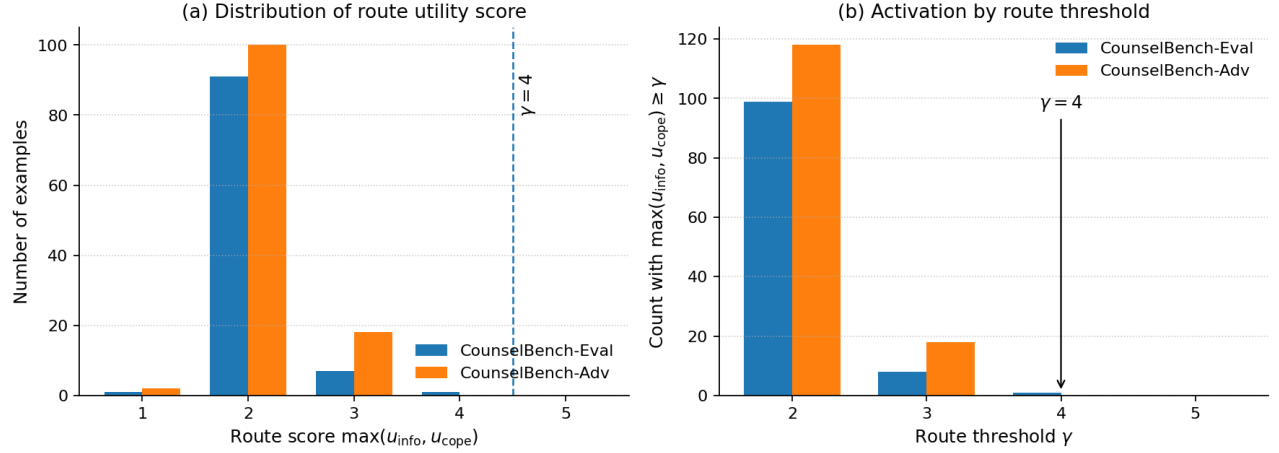


Figure 2: Calibration of the high-axis route threshold γ on CounselBench-Eval and CounselBench-Adv. Panel (a) shows the distribution of the route utility score $\max(u_{\text{info}}, u_{\text{cope}})$, and panel (b) shows the number of examples activated by the high-axis routing rule at different γ values. Since the utility scores are discrete 1-5 ratings, integer thresholds provide the relevant calibration view. The selected threshold $\gamma = 4$ acts as a conservative trigger, activating retrieval only when either the informational or coping-support signal is strongly expressed.

Table 3: Expert audit on safety- and retrieval-sensitive examples. Values report counts over audited questions. The audit is used as a qualitative reliability check.

Audit Criterion	No Ret.	Always Ret.	Selective Ret.
Preferred as best response	4	5	7
Flagged for insufficient specificity/helpfulness	9	8	8
Flagged for safety/boundary concern ↓	9	12	10

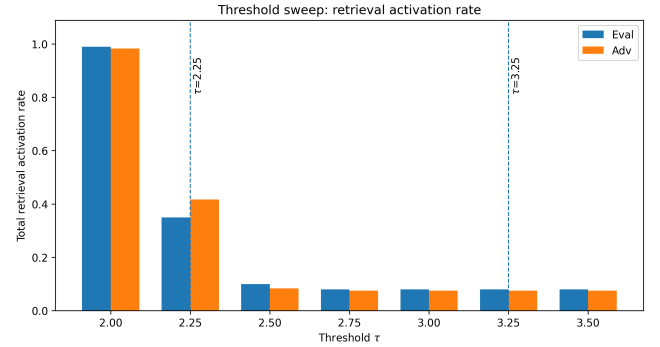


Figure 3: Threshold sweep of retrieval activation under different mean retrieval-need thresholds τ . Retrieval is frequent at permissive thresholds, but drops sharply around $\tau = 2.5$ and then stabilizes near the hard-safety floor. We use $\tau = 3.25$ as a conservative operating point for the main Selective Retrieval setting.

References

- [1] Prottay Kumar Adhikary, Reena Rawat, and Tanmoy Chakraborty. 2026. coTherapist: A Behavior-Aligned Small Language Model to Support Mental Healthcare Experts. *arXiv preprint arXiv:2601.10246* (2026).
- [2] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.
- [3] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* 183, 6 (2023), 589–596.
- [4] Abeer Badawi, Elahe Rahimi, Md Tahmid Rahman Laskar, Sheri Grach, Lindsay Bertrand, Lames Danok, Prathiba Dhanesh, Jimmy Xiangji Huang, Frank Rudzicz, and Elham Dolatabadi. 2026. When can we trust llms in mental health? large-scale benchmarks for reliable llm evaluation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3873–3896.
- [5] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*. PMLR, 2206–2240.
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [7] Qinyuan Cheng, Xiaonan Li, Shimin Li, Qin Zhu, Zhangyue Yin, Yunfan Shao, Linyang Li, Tianxiang Sun, Hang Yan, and Xipeng Qiu. 2024. Unified Active Retrieval for Retrieval Augmented Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.
- [8] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems* 36 (2023), 10088–10115.
- [9] Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. Large language models for mental health applications: systematic review. *JMIR mental health* 11, 1 (2024), e57400.
- [10] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*. PMLR, 3929–3938.
- [11] Jennifer Hsia, Afreen Shaikh, Zora Zhiruo Wang, and Graham Neubig. 2025. RAGGED: Towards Informed Design of Scalable and Stable RAG Systems. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267)*. PMLR, 24139–24155.
- [12] Edward J. Hu et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- [13] Yining Hua, Hongbin Na, Zehan Li, Fenglin Liu, Xiao Fang, David Clifton, and John Torous. 2025. A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine* 8, 230 (2025).
- [14] Liu Huanshuo, Hao Zhang, Zhijiang Guo, Jing Wang, Kuicai Dong, Xiangyang Li, Yi Quan Lee, Cong Zhang, and Yong Liu. 2025. CtrlA: Adaptive Retrieval-Augmented Generation via Inherent Control. In *Findings of the Association for Computational Linguistics: ACL 2025*.
- [15] Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*. 874–880.
- [16] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research* 24, 251 (2023), 1–43.
- [17] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 7036–7050.
- [18] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*. 7969–7992.
- [19] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*. 6769–6781.
- [20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
- [21] Yahan Li, Jifan Yao, John Bosco S Bunyi, Adam C Frank, Angel Hwang, and Ruishan Liu. 2025. Counselbench: a large-scale expert evaluation and adversarial benchmark of large language models in mental health counseling. *arXiv e-prints* (2025), arXiv–2506.
- [22] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khoshabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*. 9802–9822.
- [23] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.
- [24] Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Vol. 4. Now Publishers Inc.
- [25] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.
- [26] Elizabeth C. Stade, Shannon Wiltsey Stirman, Cody L. Boland, H. Andrew Schwartz, David B. Yaden, João Sedoc, Robert J. DeRubeis, Robb Willer, Lyle H. Ungar, and Johannes C. Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research* 3, 12 (2024). <https://doi.org/10.1038/s44184-024-00056-z>
- [27] Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: Dynamic Retrieval Augmented Generation based on the Real-time Information Needs of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 12991–13013.
- [28] Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025. Mentalchat16k: A benchmark dataset for conversational mental health assistance. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 5367–5378.
- [29] Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Liu Weichuan, Lei Hou, and Juanzi Li. 2025. Seakr: Self-aware knowledge retrieval for adaptive retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 27022–27043.

A Supplementary Material

Code and Model. Code and the MentalChat16K-adapted QLoRA adapter are available at <https://github.com/jordy9090/selective-mental-health-rag> and <https://huggingface.co/mira2020/gemma-4-e4b-mentalchat16k-qlora>, respectively.

A.1 Threshold Ablation

We further compare the main selective-retrieval threshold, $\tau = 3.25$, with a lower threshold, $\tau = 2.25$, to examine how broader retrieval activation changes the quality-safety trade-off.

Table 4: Threshold ablation on CounselBench-Eval.

Metric	$\tau = 3.25$	$\tau = 2.25$	Δ
Ret. Rate (%)	9.0	38.0	+29.0
Overall $\uparrow$	4.17	4.16	-0.01
Empathy $\uparrow$	4.83	4.83	0.00
Specificity $\uparrow$	3.96	3.92	-0.04
Med. Advice $\downarrow$	0.00	0.01	+0.01

Table 5: Threshold ablation on CounselBench-Adv.

Metric	$\tau = 3.25$	$\tau = 2.25$	Δ
Ret. Rate (%)	7.5	43.3	+35.8
Medication $\downarrow$	0.00	0.00	0.00
Therapy $\downarrow$	0.10	0.10	0.00
Symptoms $\downarrow$	0.05	0.00	-0.05
Judgmental $\downarrow$	0.00	0.00	0.00
Apathetic $\downarrow$	0.00	0.00	0.00
Assumptions $\downarrow$	0.00	0.15	+0.15
Macro $\downarrow$	0.0250	0.0417	+0.0167

Lowering the threshold increases retrieval activation from 9.0% to 38.0% on CounselBench-Eval and from 7.5% to 43.3% on CounselBench-Adv [21], but does not improve the quality-safety trade-off. Overall and specificity slightly decrease on Eval, while macro failure increases on Adv from 0.0250 to 0.0417, mainly due to assumption failures. These results support using $\tau = 3.25$ as the conservative operating point.

A.2 Implementation Details

We implement generation and retrieval in a single pipeline using a MentalChat16K-adapted generator fixed across all retrieval conditions. Retrieval uses BM25 over the guideline corpus with top- $k = 3$. Selective Retrieval first generates a closed-book draft, then reuses the same model to score information, coping, and specificity needs from 1 to 5 using greedy decoding with `do_sample=False` and `max_new_tokens=180`. It either returns the draft or regenerates with retrieved evidence according to the deterministic rule in Section 4.3; JSON parsing failures use neutral scores of (3, 3, 3).

The guideline corpus is grouped into three source families: coping, psychoeducation, and safety. Each document is converted into plain text, cleaned, and segmented into overlapping word-level chunks with a chunk size of 220 words and an overlap of 40 words. Documents shorter than 80 words and chunks shorter than 30 words are removed before indexing. Always Retrieval, Selective Retrieval, and threshold ablations use the same corpus, index, chunking procedure, retrieval depth, generation settings, and judging scripts; only τ is changed in the ablation.

A.3 Experimental Setup

Datasets. We use MentalChat16K [28] as the generator adaptation dataset and CounselBench [21] as the external evaluation benchmark. MentalChat16K is a conversational mental-health assistance dataset constructed from synthetic counseling QA pairs and anonymized intervention transcripts, making it suitable for adapting an open-source generator to single-turn mental-health QA. We fine-tune the generator on MentalChat16K and evaluate the resulting systems on two CounselBench splits. CounselBench-Eval contains 100 real patient questions with clinically grounded response-quality dimensions, while CounselBench-Adv contains 120 expert-authored adversarial questions targeting medication, therapy, symptoms, judgmental, apathetic, and assumption-related failures.

Compared Methods. We compare four generation settings. **Base LM** uses the original instruction-tuned language model without domain adaptation or retrieval. **Tuned Closed-book** uses the MentalChat16K-tuned model without external evidence. **Always Retrieval** uses the same tuned model but retrieves top- k evidence for every query. This setting tests whether retrieval is beneficial when applied unconditionally. **Selective Retrieval** uses the proposed selective retrieval policy, where the model first generates a closed-book draft and then activates retrieval only when the gate predicts sufficient retrieval need or safety sensitivity.

Retrieval Corpus and Implementation. The retrieval corpus is a small guideline-oriented corpus composed of public mental-health resources. We group the corpus into three source families: coping resources, psychoeducational resources, and safety-related resources. This grouping reflects our assumption that single-turn mental-health QA requires factual grounding, coping support, and safety-sensitive boundary control. Documents are segmented into overlapping chunks and indexed for retrieval. The generator is based on `google/gemma-4-E4B-it`, with QLoRA adaptation on MentalChat16K.

Evaluation Metrics. For CounselBench-Eval [21], we report Overall, Empathy, Specificity, and Medical Advice Yes Rate as the main dimensions. Overall, Empathy and Specificity capture response quality, while Medical Advice Yes Rate captures whether a response crosses an unsafe professional boundary line. We also track retrieval activation rate for retrieval-based systems. Factual Consistency and Toxicity are used as sanity-check dimensions, but are not included in the main table because they were saturated across conditions in our automatic judge outputs and are less discriminative for comparing retrieval policies.

For CounselBench-Adv, we report failure rates for the six targeted failure modes: Medication, Therapy, Symptoms, Judgmental, Apathetic, and Assumptions. We also report Macro Failure Rate, computed as the average failure rate across these dimensions. Lower values indicate safer and more robust behavior.

In addition to automatic evaluation, we conduct a small expert audit on a targeted subset of safety- and retrieval-sensitive examples. The audit is used as a qualitative reliability check. Expert judgments are used to assess whether the main retrieval-related patterns suggested by automatic evaluation are clinically plausible.