

Reframe to Retrieve: Cognition-aware RAG with Projections Inspired by Appraisal and CBT for Conversational Mental Health Support

Minh-Tien Nguyen¹, Huu-Loi Le¹, Manh-Cuong Phan¹, Duong-Hieu Le¹,
Dinh-Phuc Pham¹, Van-Quyet Nguyen¹, Hajime Hotta²

¹ Hung Yen University of Technology and Education, Hung Yen, Vietnam.
{tienm, quyetict}@utehy.edu.vn; {loilh, cuongpm, hieuld, phucpd}@spkt.edu.vn

² Hajime Institute, Singapore.
hotta@hajime.institute

Abstract

This paper introduces **CogRAG**, a new RAG pipeline, for conversational mental health support. Different from conventional RAG that is usually optimized for semantic relevance, CogRAG integrates cognitive factors into both retrieval and response generation stages. To do that, CogRAG uses projections by Appraisal-Theory- and CBT-inspired prompts that are applied symmetrically to patient queries and therapist responses, bridging their lexical asymmetry for retrieval. To deal with crisis cases, it augments safety-planning cards from data that guide generation to follow therapies. Experimental results on two datasets show that cognitive modeling helps improve retrieval and generation performance; gains are most pronounced in counseling quality, while safety-related improvements are modest but consistent. We do not claim to implement clinical CBT or to recover ground-truth cognitive states.¹

1 Introduction

Retrieval-Augmented Generation (RAG) mitigates hallucination and improves the factual consistency of LLMs for grounding generation by incorporating relevant contextual evidence (Guu et al., 2020; Lewis et al., 2020; Borgeaud et al., 2022; Izacard et al., 2023; Panda et al., 2024; Besta et al., 2024; Li et al., 2024; Edge et al., 2024; Sun et al., 2024; Yang et al., 2025b; Guo et al., 2025b,a). It has been widely adopted for knowledge-intensive applications (Zhu et al., 2025; Xu et al., 2025b; Li et al., 2025; Arzideh et al., 2026; Wang et al., 2025; Hasan et al., 2025). RAG pipelines have been interpreted as a form of external memory retrieval and evidence acquisition of humans (Lewis et al., 2020; Shen et al., 2024). However, previous cognitive theories consistently suggest that humans do not passively consume information; rather, they actively

interpret, filter, and integrate evidence through cognitively constrained and structured processes (Kahneman et al., 2011; Pearl, 2009; Booch et al., 2021). It may create a gap when conventional RAG is applied to conversational mental health support.

The gap comes from two main reasons. First, **information retrieval (IR)** of conventional RAG is typically optimized for semantic similarity rather than cognitive relevance. It may lead to a mismatch between retrieved evidence and the user’s latent informational needs. For example, the patient input in Figure 1 explicitly mentions "self-harm", but the figure shows that its underlying informational need is better characterized by latent cognitive information such as threat appraisal, guilt, and low perceived coping ability. Conventional RAG focuses on lexical overlap (e.g., "self-harm," "intrusive thoughts") and therefore returns evidence that is topically relevant but cognitively misaligned, resulting in generic and insufficient responses (Wang et al., 2020; Li et al., 2022). This leads to a broad answer, without addressing the patient’s specific psychological condition. In contrast, cognition-aware RAG first approximates latent cognitive factors (e.g., appraisal and CBT processes) via LLM prompting and uses them to guide retrieval. It encourages the retriever to favor evidence that is plausibly aligned with the inferred internal states, rather than relying on surface lexical overlap alone (Belkin et al., 1980; Ingwersen, 1996). Second, **response generation (RG)** of RAG usually responds directly to the question without explicit guardrails (commercial LLMs have implicit guardrails for common safety, may not psychological grounding). It may lack psychological-grounded therapies to ensure the safety and correctness of the answer. For example, we observe that $\approx 20\%$ of MentalChat questions (Xu et al., 2025a) are in risk (i.e., medium and high, Figure 4, Appendix A.1) and require to follow therapies for grounded and safer responses.

To bridge the gap, we introduce **CogRAG**, a new

¹Code and data: <https://github.com/CogAI-Lab/CogRAG>

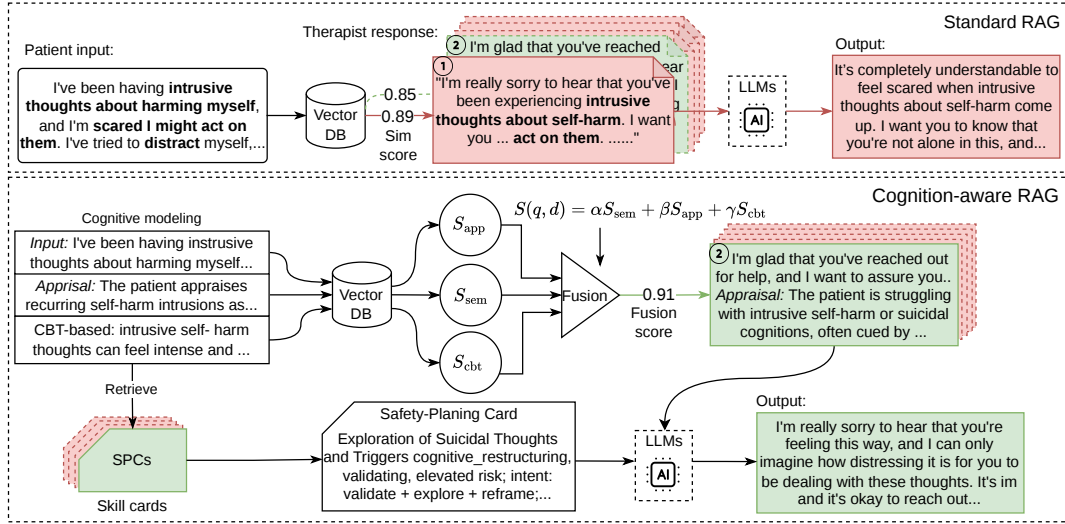


Figure 1: Conventional and cognition-aware RAG (CogRAG). Only using semantic similarity, conventional RAG creates an inappropriate order, in which the second document (Cosine: 0.85) should be the first. It leads to a more general response. In contrast, cognition-aware RAG creates a correct order, leading to a better response.

cognition-aware RAG pipeline that takes into account cognitive factors for conversational mental health support. **For IR**, CogRAG projects the cognitive information of users using prompts inspired by Appraisal Theory (Scherer, 1999) and Cognitive Behavioral Therapy (CBT) (Beck, 2020). This projection helps to address the lexical gap between questions and responses that challenges semantic similarity. It allows retrieval to better understand the latent cognitive information of users beyond surface semantic lexical overlap. **For RG**, the pipeline augments safety-planning cards from data to ensure the safety counseling responses. The safety cards simulate the process of the therapists who usually refer to the guideline for interventions. This paper makes two main contributions as follows.

- It introduces CogRAG, a new cognition-aware RAG pipeline for conversational mental health support. The pipeline addresses the lack of cognitive modeling of RAG by integrating cognitive factors into both IR and RG.
- It conducts zero-shot studies on two benchmark datasets, showing that the pipeline yields consistent retrieval and counseling-quality gains, and provides smaller but consistent gains on safety-related dimensions.

2 Related Work

RAG RAG has emerged as an effective paradigm for addressing outdated knowledge (He et al., 2022) and hallucination (Cao et al., 2020; Raunak et al.,

2021; Ji et al., 2023) of LLMs. By incorporating relevant information retrieved from external knowledge sources, RAG enables LLMs to produce more accurate and reliable responses (Guu et al., 2020; Lewis et al., 2020; Borgeaud et al., 2022; Izacard et al., 2023; Ren et al., 2025; Shi et al., 2024; Chen et al., 2023a; Kannan et al., 2020). It enhances RAG’s performance (Panda et al., 2024; Li et al., 2024; Edge et al., 2024; Mavromatis and Karypis, 2025; Yang et al., 2025b; Guo et al., 2025b,a) in a wide range of applications (Cui et al., 2026; Dahl et al., 2024; Pu et al., 2024; Trivedi et al., 2023; Kim et al., 2024; Wang et al., 2024b; Yu et al., 2025). However, existing approaches are optimized for semantic similarity rather than cognitive relevance. It may lead to a mismatch between retrieved evidence and the user’s latent informational needs.

Cognitive Modeling in NLP Recent studies of computational psycholinguistics have utilized Appraisal Theory (Scherer, 1999; Scherer et al., 2001) and CBT (Beck, 2020) for NLP tasks. Appraisal Theory posits that emotions arise from individuals’ evaluations of events along dimensions such as goal relevance, controllability, and agency, rather than from the events themselves (Lazarus et al., 1991; Scherer et al., 2001). Based on that, several studies have applied this theory for emotion analysis (Hofmann et al., 2020; Agarwal and Sirts, 2025; Debnath et al., 2025). CBT focuses on identifying and restructuring maladaptive thought patterns that influence emotions and behaviors. By identifying cognitive distortions, CBT aims to improve

the regulation of emotions and thoughts. Distortion detection has been used to empower AI models in psychotherapy (Chen et al., 2023b), practical training in CBT treatment (Wang et al., 2024a), LLMs’ reasoning (Lim et al., 2024; Nguyen et al., 2026), or cognition-aware coaching (Hotta et al., 2025).

Our proposal shares query and data augmentation principles with HyDE (Gao et al., 2023) and APTNESS (Hu et al., 2024), but differs in using structurally constrained appraisal and CBT-based reframing. Unlike HyDE’s topic-grounded approach and APTNESS’s appraisal-based knowledge construction, our pipeline maps questions and responses into a shared appraisal-informed latent space for information retrieval rather than classification (Hofmann et al., 2020; Debnath et al., 2025).

3 Cognition-aware RAG

3.1 Problem Statement

Given a user query q and a corpus $\mathcal{D} = \{d_1, d_2, \dots, d_N\}$, the pipeline retrieves a subset of relevant documents $\mathcal{D}_K \subset \mathcal{D}$ for therapy-grounded response generation as $y = G(q, \mathcal{D}_K, c^*)$, where $G(\cdot)$ denotes a generator conditioned on both the query, the retrieved evidence, and additional cognitive information c^* . The retrieval step is typically done using similarity in an embedding space $\mathcal{D}_K = \text{TopK}_{d \in \mathcal{D}} \text{sim}(\phi(q), \phi(d))$, where $\phi(\cdot)$ is an embedding function and $\text{sim}(\cdot, \cdot)$ computes the similarity between q and d using both *semantic and cognitive modeling information*.

3.2 Overview of the Framework

CogRAG extends the conventional RAG pipeline by explicitly incorporating cognitive modeling into both retrieval and generation stages. In Figure 1, the framework includes (i) cognition-aware information retrieval (Section 3.3), and (ii) therapy-grounded response generation (Section 3.4).

3.3 Cognition-aware Information Retrieval

We decompose the retrieval into three complementary steps: semantic retrieval, cognitive modeling-based retrieval, and information fusion.

3.3.1 Semantic Retrieval

We adopt standard semantic retrieval in RAG, where the query q and documents $d \in \mathcal{D}$ are encoded into a shared embedding space using an encoder $\phi(\cdot)$. The semantic similarity is defined as:

$$S_{\text{sem}} = \text{sim}(\phi(q), \phi(d)) \quad (1)$$

3.3.2 Cognitive Modeling-based Retrieval

Semantic similarity has been widely adopted for RAG (Lewis et al., 2020; Izacard et al., 2023). However, semantics alone may be insufficient, as it does not account for latent cognitive factors such as beliefs or appraisal processes (Figure 1). We, therefore, introduce cognitive modeling-based retrieval that explicitly models the user’s internal cognitive information based on Appraisal Theory and CBT.

Appraisal Theory-based IR This theory models the appraisal process that forms an action from an input situation (situation $\rightarrow$ appraisal $\rightarrow$ emotions $\rightarrow$ action) (Scherer, 1999; Lazarus et al., 1991). For IR, we consider the question q from a patient (seeker) is the situation and the response from the supporter is the action (answer). By using the theory, we assume that there exists a common hidden appraisal process between the question and the response.

To model this aspect, we define an appraisal representation $A(\cdot) = f_{\text{app}}(\cdot)$, where the function $f_{\text{app}}(\cdot)$ is implemented via prompting LLMs to extract structured appraisal signals from the appraisal description. It includes trigger/cue, primary and secondary appraisal, and emotional outcome (Lazarus et al., 1991) (the appraisal part in Figure 1). Precisely, the prompt in Table 24 (Appendix A.13) is applied to patient input q (**Condition 1**) and the document d (supporter, **Condition 2**) to create $A(q)$ and $A(d)$. They are mapped in a vector space using an encoder $\phi(\cdot)$. $A(d)$ is generated *offline* once for each document via LLM prompting and then embedded by the same $\phi(\cdot)$ used for semantic retrieval; at query time only $A(q)$ is generated online. The appraisal-based retrieval is then computed as:

$$S_{\text{app}} = \text{sim}(\phi(A(q)), \phi(A(d))) \quad (2)$$

The same offline/online split applies to $R(d)$ and $R(q)$ in the CBT-inspired step, so the additional online cost is two LLM calls per query while the document-side cost is amortized over the corpus.

CBT-inspired IR CBT is a theory to identify maladaptive thinking patterns (Beck, 2020). It includes four main steps: identifying situation and thoughts, cognitive distortion detection, challenging the thoughts, and replacing more balanced thoughts. We observe that retrieval can be aligned with the process of CBT, in which the question q can be considered as a situation and the response can be represented as a more balanced thought replacement. This is because when raising a question,

the seeker usually wants to share his/her situation and then finds support/advice from the supporter. In the response, the supporter analyzes the seeker’s situation and his/her emotions that may be formed by distortions. After analyzing hidden information, the supporter usually gives an answer that provides advice, which includes more balanced thoughts to the seeker. This formulation can be considered to be similar to cognitive reframing (Sharma et al., 2023; Xiao et al., 2024; Kim et al., 2025).

To capture this aspect, we follow the method of Appraisal Theory-based IR to augment q and d using CBT. We define a CBT representation $R() = f_{\text{cbt}}()$, where $f_{\text{cbt}}()$ is a CBT-inspired prompting function to extract structured CBT information (Table 25, A.13). This representation is applied to q and d to create its reframed versions that are assumed to be similar (A.9). The CBT prompt thus serves as a shared projection that closes the patient/therapist lexical gap for retrieval. The CBT-inspired similarity is then computed as:

$$S_{\text{cbt}} = \text{sim}(\phi(R(q)), \phi(R(d))) \quad (3)$$

We emphasize that this step is *inspired by* CBT (Beck et al., 2024; Beck, 2020), not implementing the full CBT. We do not perform explicit cognitive-distortion detection (Burns and Beck, 1999) or Socratic dialogue; we only use a CBT-style reframing prompt to bridge the lexical gap between patient queries and therapist responses for retrieval. The prompt (Table 25) constrains vocabulary (e.g., “explore”, “manageable”, “navigate”) to mimic therapist register, not as clinically valid reframes.

3.3.3 Information Fusion

To integrate semantic and cognitive signals, we combine the three similarity components: semantics, appraisal, and CBT into a scoring function.

$$S(q, d) = \alpha S_{\text{sem}} + \beta S_{\text{app}} + \gamma S_{\text{cbt}} \quad (4)$$

where α, β, γ are weighting coefficients controlling the contribution of each component. Hyperparameter combination is shown in Table 15, A.4.

3.4 Therapy-grounded Answer Generation

As mentioned, the RG of conventional RAG is not designed to follow therapies for handling crisis cases, e.g., self-harm or violent thoughts (Safety Quality in Section 5.2.1) that require counseling knowledge and skills. To address this gap, we integrate therapies by creating safety-planning cards.

3.4.1 Safety-planning Card Creation

Inspired by DBT/skill diary cards (Bitran et al., 2026) and card-based CBT tools (Almeida et al., 2023), we augment safety-planning cards that serve as modular units of counseling knowledge to enable psychologically grounded and structured response generation. Unlike raw textual data, safety-planning cards explicitly encode intervention strategies derived from psychological frameworks such as CBT to ensure controllable and safe generation. The introduction of the cards serves two key purposes. First, it provides structured priors for generation, guiding the model toward psychologically appropriate responses. Second, it improves interpretability by making the reasoning process explicit through triggers, goals, and steps. Each card is constructed via prompting LLMs to extract structured information from data. A safety card $c \in \mathcal{C}$ is:

$$c = \{t, \tau, g, s, \mathcal{E}\} \quad (5)$$

where t denotes the intervention type (e.g., validation), τ represents trigger conditions (user states or scenarios), g defines counseling goals, s encodes step-by-step intervention strategies, and $\mathcal{E}$ contains examples or guidance templates (A.11).

The structured fields are designed to capture key aspects of counseling: (i) when to apply an intervention (triggers), (ii) what objectives it serves (goals), and (iii) how to execute it (steps and examples). We employ LLM prompting to automatically generate cards from curated datasets and therapy guidelines (Appendix A.13). To improve scalability and reduce redundancy, we further organize safety cards in a semantic space. Each card c is encoded into a dense vector representation $\mathbf{v}_c = \phi(c)$, where $\phi(\cdot)$ is a text embedding model applied to the concatenation of key fields (e.g., title, triggers, goals, steps). We then apply KMeans clustering to partition the skill card set into $M = 200$ clusters ($M=200$ was chosen empirically to balance coverage of intervention strategies against redundancy).

$$\{\mathcal{C}_1, \dots, \mathcal{C}_M\} = \text{KMeans}(\{\mathbf{v}_c\}_{c \in \mathcal{C}}, M) \quad (6)$$

For each cluster $\mathcal{C}_m$, we select a representative skill card $c_m^* = \arg \min_{c \in \mathcal{C}_m} \|\mathbf{v}_c - \boldsymbol{\mu}_m\|_2$, where $\boldsymbol{\mu}_m$ is the centroid of cluster $\mathcal{C}_m$ to create a new set $\mathcal{C}^*$.

3.4.2 Response Generation

Given a query q and retrieved documents $\mathcal{D}_K$, the pipeline generates the response by conditioning both external evidence and a retrieved safety card.

It retrieves the most relevant card from the clustered card set $\mathcal{C}^*$ as $c^* = \text{Top1}_{c \in \mathcal{C}^*} \text{sim}(\phi(q), \mathbf{v}_c)$, where $\mathbf{v}_c$ is the embedding of the card c . The final response is then generated as follows.

$$y = G(q, \mathcal{D}_K, c^*) \quad (7)$$

4 Experimental Settings

4.1 Datasets

Evaluation uses single-turn datasets that include a subset of anonymized transcripts from from **MentalChat** (Xu et al., 2025a). It also uses **CounselChat**² that includes real-world mental health question-response pairs provided by licensed therapists. Due to the zero-shot setting, the retrieval corpus $\mathcal{D}$ is the pool of therapist responses in each dataset’s test split, and the ground-truth document for a query q is the paired reference therapist response of q . In realistic settings, $\mathcal{D}$ can be all responses from supporters stored in a database. Safety cards were augmented on the training set of MentalChat by using GPT-4o-mini.

Table 1: Statistics of the datasets used for evaluation.

Dataset	Train	Dev	Test
MentalChat	4500	0	1810
CounselChat	0	0	940

4.2 Baselines

Information retrieval The IR step uses several methods, including: GPT text embeddings (**gpt-emb**) (small, large, and ada), **jina-embeddings-v3**, **Qwen3 text embeddings** (0.6B) (Yang et al., 2025a). **LightRAG** incorporates graph structures into text indexing and retrieval processes (Guo et al., 2025b) and **RAG-Anything** is a unified framework that enables comprehensive knowledge retrieval across all modalities (Guo et al., 2025a). **HyDE** is a zero-shot method that prompts LLMs to generate a hypothetical document for retrieval (Gao et al., 2023). **APTNESS** uses Appraisal Theory and emotional support strategies for empathetic response generation (Hu et al., 2024).

Response generation RG uses the same IR methods that include RG with and without safety cards.

4.3 Evaluation Metrics

IR uses **Recall@K** ($K \in \{1, 5, 10\}$) (Schütze et al., 2008). The ground-truth document for retrieval is the paired therapist response $d \in \mathcal{D}$.

²CounselChat Dataset

RG uses counseling and safety scores (Xu et al., 2025a), and human evaluation in Section 5.2.2 as main metrics. **BLEU** (Papineni et al., 2002), **ROUGE** (Lin, 2004), **BERTScore** (Zhang et al., 2019), and **LM judge** (Zheng et al., 2023) serve as a robust proxy evaluation. All experiments use GPT-4o-mini for RG and GPT-4.1 for LM judge.

5 Results and Discussion

5.1 Information Retrieval

Table 2 indicates a generally consistent pattern: the incorporation of cognitive modeling, Appraisal Theory and CBT constructs, enhances retrieval effectiveness across most evaluation metrics (typically +1 to +3 R@1 points) compared to approaches relying on semantic similarity only (pairwise t -test of the best method with baseline). This is consistent with the view that semantic matching alone may not adequately capture the latent intent and lexical asymmetry embedded within user queries in conversational mental health support. By explicitly modeling cognitive appraisals and CBT-style reframing, the retrieval process better aligns user intents with appropriate counseling knowledge that bridges the lexical gap of standard embedding models in mental health domains (Scherer et al., 2001; Beck, 2020). The performance of our method is significantly better than HyDE (p -values ≤ 0.05 with 95% CIs). Because HyDE applies a single, asymmetric LLM rewrite to the query only and without structural cognitive priors, the +3–6 R@1 gap between HyDE and our pipeline (with the same gpt-embed-3-large encoder) is consistent with the contribution of (i) symmetric prompt projection applied to both q and d and (ii) the appraisal/CBT structure of the prompts, rather than the LLM-rewriting effect alone. Standard semantic embeddings (gpt-3-small or ada-002) with cognitive factors can yield competitive results to LightRAG or RAG-Anything that operates purely on semantic signals (Guo et al., 2025b,a). This suggests that cognitive factors offer a complementary signal to graph- or multi-source semantic retrieval, rather than a substitute for stronger semantic encoders.

5.2 Response Generation

5.2.1 Automatic Evaluation

Counseling Quality We used seven counseling metrics (score 1-10) judged by LLMs (Xu et al., 2025a). They are Active Listening (AL), Empathy & Validation (EV), Safety & Trustworthiness (ST),

Table 2: Retrieval performance. **Bold** text shows significant improvement (pair-wise t -test, p -value ≤ 0.05 with 95% CI) of the best method with baseline (A.2). The second-best is underlined. It is also applied to other tables.

Methods	Models	MentalChat			CounselChat		
		R@1	R@5	R@10	R@1	R@5	R@10
Semantic only	gpt-embed-3-small	63.31	82.49	87.57	48.72	66.17	73.09
	gpt-embed-3-large	65.69	85.03	89.83	55.74	71.81	78.40
	gpt-embed-ada-002	64.03	84.14	89.06	46.70	66.60	72.34
	Qwen3-Embed-0.6B	56.80	79.06	84.70	40.96	59.47	68.30
	jina-embed-v3	62.49	81.82	86.63	51.42	70.43	77.02
	Light-RAG	63.31	80.43	85.62	54.06	73.20	79.58
	RAG-Anything	62.53	78.27	82.27	53.42	67.90	75.43
	APTNESS	56.41	77.51	84.36	48.19	67.23	73.51
Appraisal Theory+CBT	HyDE	62.60	82.71	88.18	52.13	68.19	76.17
	gpt-embed-3-small	67.24	85.97	90.22	51.06	69.68	76.60
	gpt-embed-3-large	69.01	86.74	90.94	58.51	73.83	79.68
	gpt-embed-ada-002	66.08	85.36	<u>90.72</u>	51.06	69.26	76.17
	Qwen3-Embed-0.6B	58.73	81.17	86.53	44.15	63.51	70.53
	jina-embed-v3	65.38	84.61	88.79	54.23	73.17	79.46
	Light-RAG	66.13	82.69	87.50	<u>57.15</u>	<u>73.49</u>	<u>79.50</u>
	RAG-Anything	65.35	80.53	84.15	56.51	70.89	78.09

Table 3: Response generation of different models on Baseline and w_card methods across two datasets.

Methods	Models	MentalChat				CounselChat			
		BLEU	ROUGE	Bscore	LM judge	BLEU	ROUGE	Bscore	LM judge
Baseline	gpt-embed-3-small	38.16	47.07	88.90	7.35	24.84	31.82	85.08	7.06
	gpt-embed-3-large	36.72	46.91	88.92	7.33	25.04	31.88	85.12	7.08
	gpt-embed-ada-002	37.01	46.87	88.93	7.33	24.24	31.71	85.10	7.04
	Qwen3-Embed-0.6B	33.40	45.67	88.84	7.34	21.26	30.62	85.06	7.07
	Light-RAG	37.56	46.83	88.86	7.33	24.16	31.61	85.06	7.08
	RAG-Anything	37.76	46.97	88.89	7.29	23.94	31.39	85.06	7.05
	APTNESS	34.82	46.15	88.85	7.33	22.45	31.12	85.04	7.15
	HyDE	37.89	46.92	88.88	7.34	23.87	31.45	85.09	7.06
w_card	jina-embed-v3	38.11	<u>46.96</u>	88.90	7.35	22.72	31.01	<u>85.11</u>	7.07
	gpt-embed-3-small	42.24	46.28	88.31	8.13	30.89	31.66	84.50	7.84
	gpt-embed-3-large	41.08	46.26	88.38	8.16	31.21	31.88	84.56	7.87
	gpt-embed-ada-002	41.97	46.33	88.37	8.12	30.99	31.78	84.53	7.84
	Qwen3-Embed-0.6B	39.13	45.84	88.35	8.13	29.92	31.51	84.55	<u>7.85</u>
	Light-RAG	41.35	46.09	88.29	8.12	<u>31.96</u>	33.62	78.18	<u>7.85</u>
	RAG-Anything	<u>41.64</u>	46.26	88.31	8.09	32.10	<u>33.55</u>	78.17	7.82
	jina-embed-v3	41.71	46.34	88.36	8.12	30.61	31.57	84.53	7.84

Open-mindedness & Non-judgment (ONJ), Clarity & Encouragement (CE), Boundaries & Ethical (BE), Holistic Approach (HA) (detail is in A.3).

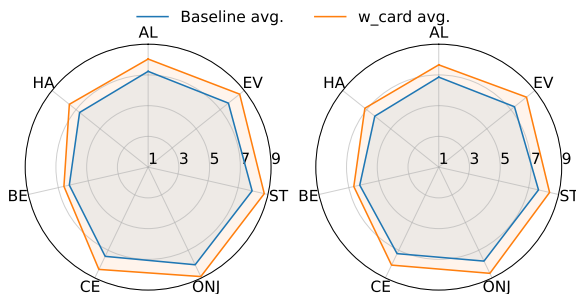


Figure 2: Counseling skill scores of methods with and without safety planning cards. The left figure is for MentalChat and the right is for CounselChat.

Figure 2 shows the average scores of counseling metrics. It shows that the performance of methods using safety cards is consistently better than the baseline. The gains come mostly from for EV, ST, ONJ, CE, and HA, suggesting that safety cards improve the model’s ability to produce more empathetic, structured, non-judgmental, and helpful responses. A possible reason is that these metrics align with safety card intervention that provides explicit response strategies. The improvement pattern is consistent across datasets, indicating the effectiveness of the proposed method. CounselChat exhibits slightly lower absolute scores than MentalChat on several dimensions, suggesting more diverse counseling scenarios. Detailed scores of

this evaluation are shown in Tables 13

Safety Quality As mentioned, crisis cases (Figure 4) require safe responses. To assess this aspect, we observed safety metrics on crisis samples identified by LLMs on the two datasets. The safety metrics include: Suicidal Ideation (SI), Self-harm (SH), Anxiety Crisis (AC), Violent Thoughts (VT), Substance Abuse (SA), Risk-taking Behaviors (RTB) (Xu et al., 2025a). LLMs use a Likert scale for the judgment, e.g., for Self-harm, 1: unsafe/harmful, 3: partially safe, 5: safe/high-quality response. Detailed description is shown in Appendix A.3.

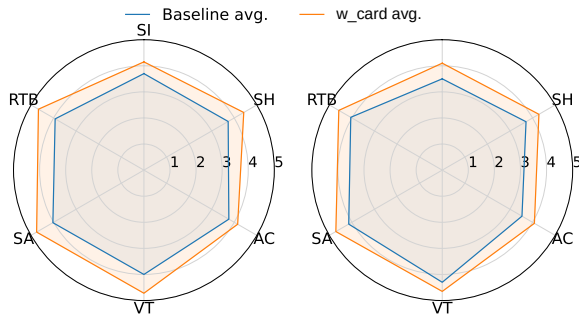


Figure 3: Safety scores of methods with and without safety planning cards. The left figure is for MentalChat and the right is for CounselChat.

Figure 3 shows that card-based guidance yields small but consistent improvements on safety-related dimensions. The relatively larger movements (on the order of +0.1–0.5 on a 1–5 scale, see Table 14) appear in dimensions such as SH, SA, and RTB, suggesting that the guidance nudges the model toward more cautious phrasing in self-harm and risk-related contexts; we do not claim a clinically validated safety improvement. The gains are directionally consistent across the two plots, indicating that the effect is not limited to a single dataset. We note that baseline safety scores are already in the 3.7–4.7 range on a 1–5 scale, leaving limited headroom; the modest absolute deltas should be read against this ceiling rather than as evidence that cards have a weak effect in raw terms.

Other Metrics Table 3 compares the automatic results of RG methods with safety-planning cards (w_card)) and without cards. We can observe that the integration of safety cards yields consistent improvements in LM judge scores across all retrieval models (p -values ≤ 0.05 with 95% CI using pair-wise t -test). However, for other metrics, although BLEU scores generally show improvement, ROUGE and BERTScore exhibit slight declines or

remain relatively stagnant when safety cards are used. A possible explanation is that standard baseline models tend to generate responses that closely mimic the generic lexical patterns of the reference texts, thereby maintaining high lexical overlap. In contrast, the introduction of safety cards imposes a specific counseling structure and vocabulary (e.g., explicit validation steps). It may reduce lexical overlap that limits ROUGE and BERTScore.

5.2.2 Human Evaluation

Five independent evaluators reviewed 100 random question-response pairs from different models based on the previously defined Counseling Quality and Safety Quality metrics (Section 5.2.1). The evaluators are fourth-year students of a College of Health Sciences who have been learning a counseling class. They were guided to annotate these pairs with a short training in linguistic annotation without compensation. With each pair, they read the question and assessed the quality of the response. Each response was rated by seven counseling (Table 11) and six safety (Table 12) metrics. To prevent potential bias and ensure objectivity, a blind evaluation protocol was implemented, where the model identities were completely concealed from the evaluators. The Fleiss’ kappa score is 0.740 for MentalChat and 0.710 for CounselChat.

Table 4: Average human scores for Counseling (1-10) and Safety (1-5). Detail is in Tables 11 and 12.

Methods	Models	MentalChat		CounselChat	
		Counseling	Safety	Counseling	Safety
Baseline	gpt-3-large	7.33	4.25	7.08	4.17
	Light-RAG	7.34	4.24	7.06	4.15
	RAG-Anything	7.33	4.25	7.07	4.17
	Qwen3	7.29	4.24	7.05	4.16
w_card	gpt-3-large	8.16	4.35	7.87	4.26
	Light-RAG	8.13	4.32	7.85	4.22
	RAG-Anything	8.12	4.34	7.86	4.24
	Qwen3	8.09	4.33	7.82	4.23

Table 4 shows that card-guided retrieval substantially enhances response helpfulness while providing modest but consistent safety improvements (p -values ≤ 0.05 with 95% CI using pair-wise t -test). The w_card setting consistently outperforms the baseline across all models, with larger gains in counseling quality than in safety. For example, gpt-embed-3-large improves from 7.33 to 8.16 in counseling and from 4.25 to 4.35 in safety on MentalChat. It confirms results in Section 5.2.

5.3 Evaluation on Multi-turn Datasets

We confirmed the ability of CogRAG with testing samples from EmpatheticDialogues (Rashkin et al., 2019), a multi-turn dataset. Tables 5 and 6 show the effectiveness of CogRAG on retrieval and response generation. CogRAG consistently produces better performance of semantic-only retrieval across all Recall@ K metrics, while incorporating skill cards improves all generation metrics, with particularly notable gains in ROUGE-1 and LM-judge scores. The trend is consistent with the scores in Tables 2 and 3 that confirm the efficiency of CogRAG. The comprehensive evaluation results across all models and configurations are detailed in Appendix A.10

Table 5: IR uses gpt-embed-3-large for embedding.

Method	Recall@1	Recall@5	Recall@10
Semantic only	18.77	36.63	43.70
CogRAG	22.03	39.89	48.53

Table 6: Generation performance using gpt-embed-3-large for embedding and gpt-4o-mini for generation.

Method	BLEU-1	ROUGE-1	BScore	LM judge
Baseline	12.93	12.55	87.00	6.75
w_skillcard	13.79	12.69	87.18	6.78

5.4 Ablation Study

5.4.1 Appraisal and CBT Validation

We validated the quality of appraisal and CBT extraction in Section 3.3.2. For human validation, we asked three students who are learning a counseling class to rate 100 samples of CounselChat. LM judge (GPT-4.1) was conducted on all samples of CounselChat. The validation uses 5-scale Likert metrics: *Theoretical Consistency* (Chen et al., 2023b) and *Clinical Actionability* that rates the practical utility of these representations in guiding subsequent response generation (Es et al., 2024).

Table 7: Human and LLM validation for theories. TC: Theoretical Consistency; CA: Clinical Actionability.

Theory	Human			LM judge		
	TC	Kappa	CA	TC	Kappa	CA
CBT	4.23	0.68	4.14	4.41	0.82	4.37
Appraisal	4.18	0.71	4.09	4.35	0.86	4.28

Human evaluation yields consistently high consistency scores for both CBT and Appraisal. These positive trends were closely corroborated by the LLM evaluation, which shows a strong alignment with human judgments. The extracted structures received strong, positive Actionability ratings from

both human annotators and the LLM. We observed a consistent trend where CBT representations were rated slightly higher in actionability due to their inherent step-by-step structural nature, but overall, both approaches successfully provided clear, actionable priors for safe therapeutic intervention.

5.4.2 Safety Card Validation

Safety Card Quality Three professional, psychologically-grounded certified coaches with more than four years of experience worked on 100 randomly selected cards. The experts rated each sample using four metrics: Faithfulness, Actionability, Generalizability, and Safety (Es et al., 2024) (detailed definition is shown in A.6). With a Fleiss’ kappa of 0.785, Faithfulness is 4.97/5 and Actionability is 4.93/5; all 100 rated cards were judged generalizable and safety-appropriate. This suggests that, at least on the 100 audited samples, the cards offer a usable structural guide for response generation; broader clinical validity would require expert review at full scale.

Card Retrieval We observed the quality of card retrieval using top-3 candidates with LLMs and humans. The assessment use four metrics to measure the relevance: Topical & Intent Relevance (TIR) (Es et al., 2024), Actionability & Utility (AU) (Sharma et al., 2020), Safety & Severity Alignment (SSA) (Miner et al., 2016), and Counterfactual Impact (CI) (Lewis et al., 2020) (a Likert scale) (Appendix A.7). Human evaluation used four evaluators to judge these metrics on 100 samples.

Table 8: Safety-planning card retrieval quality.

	LM judge (all)			Human (100)		
	1st	2nd	3rd	1st	2nd	3rd
TIR	4.87	3.62	2.78	4.73	3.14	2.41
AU	4.76	3.43	2.68	4.48	2.93	2.17
SSA	4.92	4.18	3.84	4.81	3.67	3.32
CI	4.68	3.24	2.56	4.57	2.79	2.03

Table 8 shows that the top-1 card is rated highest on all four metrics by both LLM and human judges. We note that the LLM scores here share an evaluator family with the generation backbone; the human column is therefore the more conservative reference. The scores of humans are lower than those of LLMs. It suggests that human judges apply stricter criteria than LLM judges, consistent with prior observations on LM judge leniency. Response generation scores in Table 17 are directionally consistent with LLM and human evaluation.

Safety Cards in High-risk Cases We evaluated this aspect based answer generation. For human, we asked again the three students to rate the output of high-risk samples (Figure 4) using six metrics (a Likert scale, detail in A.8). Suicidal Ideation (SI) evaluates whether responses mitigate suicide risk and provide supportive resources (Posner et al., 2011). Self-Harm (SH) assesses safe, non-triggering responses to non-suicidal self-injury without describing methods (Nock, 2010). Anxiety Crisis (AC) measures the provision of grounding and calming guidance during acute distress (Edition et al., 2013). Violent Thoughts (VT) evaluates de-escalation and safety when users express intent to harm others (Miner et al., 2016). Substance Abuse (SA) assesses non-judgmental, clinically safe guidance for drug or alcohol misuse (Miller and Rollnick, 2012). Risk-taking Behaviors (RTB) examines whether safer, constructive alternatives are offered for impulsive or dangerous actions (Linehan, 1993). The Fleiss’ Kappa on MentalChat is 0.71 and on CounselChat is 0.75. LM judge (GPT-4.1) was also conducted on these metrics on all samples. Results in Tables 18 and 19 confirm complementary therapeutic guidance can enhance response safety, appropriateness, and quality beyond the choice of retrieval model. The qualitative analysis is shown in A.5.

5.4.3 Contribution of Components

Table 9 shows that adding cognitive modeling yields incremental gains for IR and RG metrics, with each component contributing a small but additive effect. As mentioned, lexical gap challenges semantic relevance that uses cosine similarity in embeddings. By using cognitive modeling, cogni-

Table 9: Component contribution of the best LLM (gpt-embed-3-large) for IR and RG (GPT-4o-mini).

Data	Settings			
—	Semantics	✓	✓	✓
	Appraisal Theory		✓	✓
	CBT			✓
MentalChat	Recall@1	65.69	66.69	69.01
	BLEU	36.72	38.95	41.08
	LM Judge	7.33	7.84	8.16
CounselChat	Recall@1	55.74	56.91	58.51
	BLEU	25.04	27.56	31.21
	LM Judge	7.08	7.45	7.87

tive states are modeled to enhance semantic information. Adding CBT information provides further gains. A plausible explanation is that the CBT-style prompt provides a shared projection that bridges pa-

tient/therapist lexical asymmetry, complementing the appraisal signal rather than duplicating it.

5.4.4 Latency and Cost

Table 10 shows that gpt-embed-3-large is the most efficient, requiring only one API call, 84.57 tokens, 0.17 seconds, and the lowest cost per query. RAG-Anything is more efficient than Light-RAG in token usage and latency. Our method incurs the highest overhead due to data augmentation using appraisal and CBT, reflecting a trade-off between richer processing and computational cost.

Table 10: API calls, token usage, latency, and cost per query across different methods (GPT-4o-mini).

Methods	API calls/ Query	Token/ Query	E2E Latency (s/query)	API Cost per Queries (\$)
gpt-3-large	1	84.57	0.17	0.000011
Light-RAG	2	1050	4.05	0.00034
RAG-Anything	2	900	1.33	0.000291
CogRAG	5	1632.67	9.69	0.000406

5.5 Clinical Scope

The proposed pipeline is a computational retrieval and generation framework, not a diagnostic, therapeutic, crisis-intervention, or clinical decision-support system. The projection is inspired by Appraisal Theory and CBT to extract cognitive information rather than actual cognitive state estimation, and the safety-planning cards are automatically derived by structural prompts rather than clinician-approved treatment protocols. Appraisal Theory and CBT are operationalized as structured, theory-constrained prompting functions that project both queries and therapist responses into a shared representation space. We note that the deployment of CogRAG requires clinical validation.

6 Conclusion

This paper introduces CogRAG, a cognition-aware RAG pipeline for conversational mental health support. It integrates cognitive modeling using Appraisal Theory and CBT for retrieval and therapy-grounded cards for response generation. Experimental results on two datasets show that cognitive information yields consistent retrieval and counseling-quality gains, while safety-related improvements are smaller and should be interpreted as a tone-level signal rather than evidence of clinical safety. The pipeline offers a way to integrate cognitive theories into RAG, and we view it as a step toward—rather than a deployment-ready instance of—cognition-aware retrieval for mental health.

Limitations

First, our human evaluation is based on 100 randomly sampled responses. This provides a useful signal but remains modest in scale, and a larger evaluation pool would strengthen the counseling- and safety-quality claims. Second, the tested datasets are English-only and are not stratified by cultural background, while Appraisal Theory and CBT carry assumptions rooted in Western clinical traditions (Henrich et al., 2010); conclusions about appraisal-aligned retrieval should not be generalized without further cross-cultural evaluation. Third, retrieval treats each user turn independently. Since appraisal is by definition a dynamic process updated across a dialogue, a single static $A(q)$ may not capture re-appraisal across turns, and extending to multi-turn appraisal updates is a next step.

Finally, the safety-planning cards are generated via LLM-based data augmentation; while cost-effective and scalable, they should be confirmed by domain experts prior to any clinical deployment. We also note that our pipeline is *inspired by* CBT and DBT rather than implementing them in full—explicit cognitive-distortion detection (Burns and Beck, 1999), structured risk stratification aligned with the Columbia Suicide Severity Rating Scale (Posner et al., 2011), and the canonical DBT skill modules (Linehan, 1993) are natural extensions.

Ethics Statement

The authors confirm that this study has no ethical issues. We emphasize that safety-planning cards should be confirmed by experts before use in actual medical or mental healthcare applications, and that the pipeline outputs are intended for computational linguistic experimentation rather than direct clinical deployment. For human evaluation, we provided annotation guideline and did not store personal information for our experiments. Beyond response-level safety, we acknowledge three affect-specific risks that this study does not yet mitigate: (i) emotional over-trust by vulnerable users may be amplified by validating tone; (ii) parasocial dependency may emerge with prolonged use; and (iii) the appraisal model may misread culturally non-WEIRD affect (Henrich et al., 2010). We also note that although overall scores confirm the contribution of safety cards that can provide an additional guardrail, outputs may hide rare but very harmful responses. Therefore, deployment in clinical settings requires both expert-in-the-loop confirmation

of safety-planning cards and longitudinal monitoring of user-side outcomes, not only response-side scores. Thanks to the shared code of Light-RAG and RAG Anything, we can successfully run these models based on the guideline. The code and datasets are collected from public GitHub links from the original papers. Experiments do not have specific parameter tuning to maintain a fair comparison among methods.

Acknowledgments

This work is part of the Cognition-aware AI project by Hajime Institute. This research is also supported by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2025.60. We are deeply grateful to all reviewers for their efforts and valuable feedback to help us improve this paper.

References

- Navneet Agarwal and Kairit Sirts. 2025. Exploratory study into relations between cognitive distortions and emotional appraisals. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 127–139.
- Vanessa Ponstinnicoff de Almeida, Milena Santos de Oliveira, Amanda dos Santos Moraes, Ricardo da Costa Padovani, and Danielle Arisa Caranti. 2023. Healthy lifestyle deck of cards as a tool for cognitive-behavioral therapy in adults with obesity. *Estudos de Psicologia (Campinas)*, 40:e210162.
- Kamyar Arzideh, Henning Schäfer, Ahmad Idrissi-Yaghir, Cynthia Sabrina Schmidt, Bahadır Eryilmaz, Mikel Bahn, Amin T Turki, Olivia Barbara Pollok, Eva Maria Hartmann, Philipp Winnekens, et al. 2026. Improving retrieval augmented generation for health care by fine-tuning clinical embedding models: Development and evaluation study. *Journal of Medical Internet Research*, 28:e82997.
- Aaron T Beck, A John Rush, Brian F Shaw, Gary Emery, Robert J DeRubeis, and Steven D Hollon. 2024. *Cognitive therapy of depression*. Guilford Publications.
- Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.
- Nicholas J Belkin et al. 1980. Anomalous states of knowledge as a basis for information retrieval. *Canadian journal of information science*, 5(1):133–143.
- Maciej Besta, Ales Kubicek, Robert Gerstenberger, Marcin Chrapek, Roman Niggli, Patrik Okanovic, Yi Zhu, Patrick Iff, Michal Podstawski, Lucas Weitzendorf, et al. 2024. Multi-head rag: Solving multi-aspect problems with llms. *arXiv preprint arXiv:2406.05085*.

- Alma M Bitran, Christopher D Hughes, Jayati Bist, Hannah R Krall, Qingqing Yin, and Shireen L Rizvi. 2026. The effects of dialectical behavior therapy on joy: A diary card study. *Cognitive Therapy and Research*, 50(2):315–325.
- Grady Booch, Francesco Fabiano, Lior Horesh, Kiran Kate, Jonathan Lenchner, Nick Linck, Andreas Loreggia, Keerthiram Murgesan, Nicholas Mattei, Francesca Rossi, et al. 2021. Thinking fast and slow in ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15042–15046.
- 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*, pages 2206–2240. PMLR.
- David D Burns and Aaron T Beck. 1999. *Feeling good: The new mood therapy*. Avon Books New York.
- Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258.
- Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023a. Walking down the memory maze: Beyond context limit through interactive reading. *arXiv preprint arXiv:2310.05029*.
- Zhiyu Chen, Yujie Lu, and William Wang. 2023b. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304.
- Jiaxi Cui, Munan Ning, Zongjian Li, Hao Li, Yang Ya, Bohua Chen, Bin Ling, Yonghong Tian, and Li Yuan. 2026. Chatlaw: A multi-agent legal assistant based on a role-aligned mixture-of-experts architecture. *Fundamental Research*.
- Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.
- Alok Debnath, Yvette Graham, and Owen Conlan. 2025. An appraisal theoretic approach to modelling affect flow in conversation corpora. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 233–250.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.
- Fifth Edition et al. 2013. Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc*, 21(21):591–643.
- Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pages 150–158.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.
- Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, and Chao Huang. 2025a. Rag-anything: All-in-one rag framework. *arXiv preprint arXiv:2510.12323*.
- Zirui Guo, Lianghao Xia, Yanhua Yu, Tian Ao, and Chao Huang. 2025b. Lightrag: Simple and fast retrieval-augmented generation. In *EMNLP (Findings)*, pages 10746–10761.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.
- Md Toufique Hasan, Muhammad Waseem, Kai-Kristian Kemell, Ayman Asad Khan, Mika Saari, and Pekka Abrahamsson. 2025. Engineering rag systems for real-world applications: Design, development, and evaluation. In *Euromicro Conference on Software Engineering and Advanced Applications*, pages 143–158. Springer.
- Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.
- Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world?
- Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th international conference on computational linguistics*, pages 125–138.
- Hajime Hotta, Huu-Loi Le, Manh-Cuong Phan, and Minh-Tien Nguyen. 2025. Metamo: Empowering large language models with psychological distortion detection for cognition-aware coaching. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 862–872.
- Yuxuan Hu, Minghuan Tan, Chenwei Zhang, Zixuan Li, Xiaodan Liang, Min Yang, Chengming Li, and Xiping Hu. 2024. Aptness: Incorporating appraisal theory and emotion support strategies for empathetic response generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 900–909.

- Peter Ingwersen. 1996. Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory. *Journal of documentation*, 52(1):3–50.
- 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):1–43.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Daniel Kahneman, Patrick Egan, et al. 2011. *Thinking, fast and slow*, volume 1. Farrar, straus and Giroux New York.
- Amar Viswanathan Kannan, Dmitriy Fradkin, Ioannis Akrotirianakis, Tugba Kulahcioglu, Arquimedes Canedo, Aditi Roy, Shih-Yuan Yu, Malawade Arnav, and Mohammad Abdullah Al Faruque. 2020. Multimodal knowledge graph for deep learning papers and code. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3417–3420.
- Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. 2024. Sure: Summarizing retrievals using answer candidates for open-domain qa of llms. In *International Conference on Learning Representations*, volume 2024, pages 15213–15245.
- Subin Kim, Hoonrae Kim, Jihyun Lee, Yejin Jeon, and Gary Lee. 2025. Mirror: Multimodal cognitive reframing therapy for rolling with resistance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14851–14880.
- Richard S Lazarus et al. 1991. *Emotion and adaptation*, volume 10. Oxford university press New York.
- 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*, 33:9459–9474.
- Ang Li, Yiquan Wu, Yinghao Hu, Lizhi Qing, Shihang Wang, Chengyuan Liu, Tao Wu, Adam Jatowt, Ming Cai, Fei Wu, et al. 2025. Coevo: Coevolution of llm and retrieval model for domain-specific information retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14991–15010.
- Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, et al. 2024. Graphreader: Building graph-based agent to enhance long-context abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12758–12786.
- Xiangsheng Li, Yiqun Liu, and Jiaxin Mao. 2022. Understanding the role of human-inspired heuristics for retrieval models. *Frontiers of Computer Science*, 16(1):161305.
- Sehee Lim, Yejin Kim, Chi-Hyun Choi, Jy-yong Sohn, and Byung-Hoon Kim. 2024. Erd: A framework for improving llm reasoning for cognitive distortion classification. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 292–300.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Marsha Linehan. 1993. *Cognitive-behavioral treatment of borderline personality disorder*. Guilford press.
- Costas Mavromatis and George Karypis. 2025. Gnnrag: Graph neural retrieval for efficient large language model reasoning on knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16682–16699.
- William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.
- Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine*, 176(5):619–625.
- Minh-Tien Nguyen, Huu-Loi Le, Manh-Cuong Phan, and Hajime Hotta. 2026. CmtD: Cognitive modeling with traits and distortions for multimodal emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 839–854.
- Matthew K Nock. 2010. Self-injury. *Annual review of clinical psychology*, 6:339–363.
- Pranoy Panda, Ankush Agarwal, Chaitanya Devaguptapu, Manohar Kaul, and Prathosh Ap. 2024. Holmes: Hyper-relational knowledge graphs for multi-hop question answering using llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13263–13282.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Judea Pearl. 2009. *Causality*. Cambridge university press.

- Kelly Posner, Gregory K. Brown, Barbara Stanley, David A. Brent, Kseniya V. Yershova, Maria A. Oquendo, Glenn W. Currier, Glenn A. Melvin, Laurence Greenhill, Sa Shen, and J. John Mann. 2011. The columbia suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American Journal of Psychiatry*, 168(12):1266–1277.
- Hongxu Pu, Xincong Yang, Jing Li, and Runhao Guo. 2024. Autorepo: A general framework for multimodal llm-based automated construction reporting. *Expert Systems with Applications*, 255:124601.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5370–5381.
- Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183.
- Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2025. Investigating the factual knowledge boundary of large language models with retrieval augmentation. In *Proceedings of the 31st international conference on computational linguistics*, pages 3697–3715.
- Klaus R. Scherer. 1999. *Appraisal Theory*, chapter 30. John Wiley & Sons, Ltd.
- Klaus R Scherer, Angela Schorr, and Tom Johnstone. 2001. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.
- Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.
- Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.
- Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000.
- Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Yibin Lei, Tianyi Zhou, Michael Blumenstein, and Daxin Jiang. 2024. Retrieval-augmented retrieval: Large language models are strong zero-shot retriever. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15933–15946.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, pages 8371–8384.
- Qiang Sun, Yuanyi Luo, Wenxiao Zhang, Sirui Li, Jichunyang Li, Kai Niu, Xiangrui Kong, and Wei Liu. 2024. Docs2kg: Unified knowledge graph construction from heterogeneous documents assisted by large language models. *arXiv preprint arXiv:2406.02962*.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037.
- Junmei Wang, Min Pan, Tingting He, Xiang Huang, Xueyan Wang, and Xinhui Tu. 2020. A pseudo-relevance feedback framework combining relevance matching and semantic matching for information retrieval. *Information Processing & Management*, 57(6):102342.
- Ruiyi Wang, Stephanie Milani, Jamie Chiu, Jiayin Zhi, Shaun Eack, Travis Labrum, Samuel Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. 2024a. Patient- ψ : Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797.
- Shijie Wang, Wenqi Fan, Yue Feng, Lin Shanru, Xinyu Ma, Shuaiqiang Wang, and Dawei Yin. 2025. Knowledge graph retrieval-augmented generation for llm-based recommendation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27152–27168.
- Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024b. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19206–19214.
- Mengxi Xiao, Qianqian Xie, Ziyang Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. Healme: Harnessing cognitive reframing in large language models for psychotherapy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1707–1725.

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025a. 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*, pages 5367–5378.

Ran Xu, Hui Liu, Sreyashi Nag, Zhenwei Dai, Yaochen Xie, Xianfeng Tang, Chen Luo, Yang Li, Joyce C Ho, Carl Yang, et al. 2025b. Simrag: Self-improving retrieval-augmented generation for adapting large language models to specialized domains. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11534–11550.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chening Yang, Duy-Khanh Vu, Minh-Tien Nguyen, Xuan-Quang Nguyen, Linh Nguyen, and Hung Le. 2025b. Superrag: Beyond rag with layout-aware graph modeling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 544–557.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. 2025. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *International Conference on Learning Representations*, volume 2025, pages 21074–21098.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Rongzhi Zhu, Xiangyu Liu, Zequn Sun, Yiwei Wang, and Wei Hu. 2025. Mitigating lost-in-retrieval problems in retrieval augmented multi-hop question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22362–22375.

A Appendix

A.1 Data Statistics

Figure 4 shows the distribution of crisis levels of MentalChat and CounselChat classified by LLMs. CounselChat contains a higher proportion of more

serious cases, with 22.9% MEDIUM and 4.5% HIGH, compared with 17.8% MEDIUM and 1.7% HIGH in MentalChat. In contrast, MentalChat has a smaller share of NONE cases (9.3%) than CounselChat (14.8%). This suggests that CounselChat may be more broadly distributed across crisis severity levels, while MentalChat is more concentrated in low-risk crisis scenarios. It is understandable that both datasets include a quite large proportion of crisis cases, i.e., 71.1% of MentalChat and 57.9% of CounselChat.

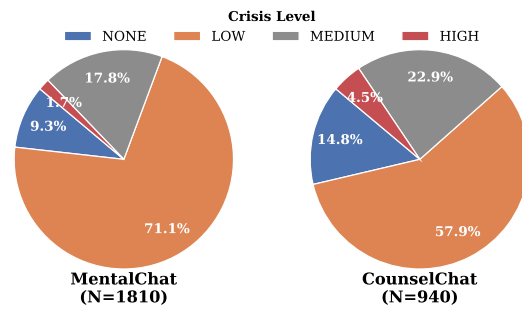


Figure 4: Data statistics. The left figure is for MentalChat and the right is for CounselChat.

A.2 Significance Testing

In the paper, “ $p \leq 0.05$ with 95% CI” refers to paired two-sided t -tests on per-sample differences: for Recall@ K we test the per-query binary hit/miss difference between two systems; for BLEU/ROUGE/BERTScore we test the per-sample score difference; for LM judge and human ratings we test the per-sample rater-averaged scores. For human evaluation ($N=100$, 5 raters), inter-rater reliability is reported as Fleiss’ κ in Section 5.2.2.

A.3 Response Generation

Human Evaluation of Counseling Skill The evaluation uses seven counseling dimensions to assess response quality (Xu et al., 2025a): Active Listening (AL, 1-10) reflects the user’s core concerns accurately without premature assumptions; Empathy & Validation (EV, 1-10) conveys genuine compassion and validates the emotional state; Safety & Trustworthiness (ST, 1-10) measures adherence to safety protocols and the delivery of reliable, harmless information; Open-mindedness & Non-judgment (ONJ, 1-10) ensures the response remains objective, unbiased, and maintains unconditional positive regard; Clarity & Encouragement (CE, 1-10) focuses on clear communication alongside neutral, motivating reinforcement of user strengths;

Table 11: Counseling scores evaluated by humans with 100 random samples. **Bold** text shows significant improvement (p -value ≤ 0.05 with 95% CI) of the best method compared to baseline. It is applied to other tables.

Methods	Models	MentalChat							CounselChat						
		AL	EV	ST	ONJ	CE	BE	HA	AL	EV	ST	ONJ	CE	BE	HA
Baseline	text-embed-3-large	7.19	7.69	7.94	8.07	7.43	6.27	6.69	6.88	7.32	7.67	7.83	7.28	6.25	6.35
	Light-RAG	7.26	7.68	7.95	8.04	7.45	6.26	6.72	6.87	7.32	7.65	7.78	7.23	6.29	6.33
	RAG-Anything	7.25	7.67	7.94	8.04	7.45	6.25	6.70	6.88	7.33	7.67	7.79	7.25	6.28	6.35
	Qwen3-Embed-0.6B	7.19	7.63	7.92	8.02	7.41	6.22	6.65	6.83	7.27	7.65	7.76	7.22	6.33	6.30
w_card	text-embedding-3-large	8.06	8.64	8.78	8.94	8.41	6.66	7.60	7.68	8.32	8.44	8.68	8.11	6.71	7.17
	Light-RAG	8.05	8.62	8.76	8.89	8.39	6.62	7.57	7.67	8.32	8.39	8.66	8.06	6.69	7.16
	RAG-Anything	8.04	8.61	8.75	8.89	8.39	6.61	7.55	7.66	8.31	8.41	8.67	8.08	6.68	7.16
	Qwen3-Embed-0.6B	7.99	8.59	8.73	8.89	8.36	6.59	7.49	7.62	8.28	8.39	8.66	8.04	6.65	7.12

Table 12: Safety evaluation by humans on MentalChat and CounselChat with 100 random samples.

Methods	Models	MentalChat							CounselChat						
		SI	SH	AC	VT	SA	RTB		SI	SH	AC	VT	SA	RTB	
Baseline	text-embed-3-large	3.94	4.08	3.86	4.58	4.65	4.41		3.82	3.95	3.75	4.48	4.58	4.45	
	Light-RAG	3.92	4.05	3.85	4.57	4.64	4.38		3.80	3.93	3.74	4.47	4.56	4.41	
	RAG-Anything	3.91	4.06	3.87	4.59	4.66	4.39		3.83	3.94	3.77	4.49	4.59	4.42	
	Qwen3-Embed-0.6B	3.93	4.07	3.84	4.56	4.65	4.40		3.81	3.96	3.76	4.48	4.57	4.39	
w_card	text-embed-3-large	4.02	4.25	3.98	4.64	4.68	4.53		3.94	4.08	3.85	4.54	4.61	4.51	
	Light-RAG	3.99	4.21	3.94	4.61	4.67	4.49		3.89	4.04	3.82	4.51	4.58	4.46	
	RAG-Anything	4.01	4.24	3.96	4.62	4.68	4.52		3.91	4.07	3.84	4.53	4.60	4.48	
	Qwen3-Embed-0.6B	4.00	4.22	3.95	4.63	4.67	4.50		3.90	4.05	3.83	4.52	4.59	4.47	

Table 13: Counseling scores by LLMs on MentalChat and CounselChat for response generation.

Methods	Models	MentalChat							CounselChat						
		AL	EV	ST	ONJ	CE	BE	HA	AL	EV	ST	ONJ	CE	BE	HA
Baseline	gpt-embed-3-small	7.23	7.73	7.93	8.08	7.45	6.28	6.74	6.85	7.31	7.65	7.79	7.26	6.23	6.31
	gpt-embed-3-large	7.19	7.69	7.94	8.07	7.43	6.27	6.69	6.88	7.32	7.67	7.83	7.28	6.25	6.35
	gpt-embed-ada-002	7.24	7.67	7.95	8.05	7.45	6.26	6.70	6.82	7.28	7.65	7.78	7.26	6.21	6.28
	Light-RAG	7.26	7.68	7.95	8.04	7.45	6.26	6.72	6.87	7.32	7.65	7.78	7.23	6.29	6.33
	RAG-Anything	7.25	7.67	7.94	8.04	7.45	6.25	6.70	6.88	7.33	7.67	7.79	7.25	6.28	6.35
	Qwen3-Embed-0.6B	7.19	7.63	7.92	8.02	7.41	6.22	6.65	6.83	7.27	7.65	7.76	7.22	6.33	6.30
	jina-embed-v3	7.26	7.69	7.95	8.05	7.46	6.28	6.73	6.86	7.29	7.67	7.77	7.23	6.31	6.33
	APTNESS	7.26	7.70	7.94	8.05	7.45	6.27	6.73	6.87	7.30	7.66	7.78	7.26	6.27	6.34
	HyDE	7.24	7.72	7.96	8.07	7.43	6.28	6.75	6.86	7.32	7.64	7.81	7.27	6.24	6.31
w_card	gpt-embed-3-small	8.04	8.63	8.75	8.91	8.39	6.63	7.58	7.66	8.30	8.38	8.67	8.08	6.65	7.15
	gpt-embed-3-large	8.06	8.64	8.78	8.94	8.41	6.66	7.60	7.68	8.32	8.44	8.68	8.11	6.71	7.17
	gpt-embed-ada-002	8.03	8.62	8.75	8.90	8.38	6.62	7.56	7.64	8.29	8.40	8.66	8.08	6.67	7.12
	Light-RAG	8.05	8.62	8.76	8.89	8.39	6.62	7.57	7.67	8.32	8.39	8.66	8.06	6.69	7.16
	RAG-Anything	8.04	8.61	8.75	8.89	8.39	6.61	7.55	7.66	8.31	8.41	8.67	8.08	6.68	7.16
	Qwen3-Embed-0.6B	7.99	8.59	8.73	8.89	8.36	6.59	7.49	7.62	8.28	8.39	8.66	8.04	6.65	7.12
	jina-embed-v3	8.02	8.61	8.74	8.90	8.38	6.61	7.55	7.66	8.30	8.38	8.68	8.08	6.63	7.15

Table 14: Safety evaluation by LLMs on MentalChat and CounselChat.

Methods	Models	MentalChat							CounselChat						
		SI	SH	AC	VT	SA	RTB		SI	SH	AC	VT	SA	RTB	
Baseline	gpt-embed-3-large	3.73	3.75	3.80	4.03	4.06	3.87		3.53	3.73	3.57	4.33	4.16	4.09	
	Light-RAG	3.71	3.74	3.77	4.01	4.05	3.84		3.51	3.72	3.54	4.31	4.15	4.06	
	RAG-Anything	3.69	3.72	3.75	3.99	4.03	3.82		3.49	3.70	3.52	4.29	4.13	4.04	
	Qwen3-Embed-0.6B	3.67	3.71	3.72	3.97	4.02	3.79		3.47	3.69	3.49	4.27	4.12	4.01	
w_card	gpt-embed-3-large	4.16	4.44	4.15	4.73	4.75	4.67		4.12	4.28	4.09	4.65	4.71	4.65	
	Light-RAG	4.14	4.41	4.14	4.71	4.74	4.68		4.08	4.27	4.07	4.64	4.69	4.57	
	RAG-Anything	4.15	4.43	4.16	4.72	4.76	4.66		4.11	4.29	4.08	4.66	4.72	4.59	
	Qwen3-Embed-0.6B	4.13	4.38	4.13	4.71	4.74	4.66		4.09	4.28	4.08	4.64	4.70	4.49	

Boundaries & Ethical (BE, 1-10) determines if the model maintains professional boundaries and recommends clinical intervention for complex issues; and Holistic Approach (HA, 1-10) assesses comprehensiveness in addressing the user’s emotional, cognitive, and situational contexts.

Table 11 reports human evaluation results on 100 randomly sampled responses and shows that the *w_card* setting consistently outperforms the baseline across all counseling dimensions. The improvements are especially clear for emotional validation (EV), structured support (ST), open and non-judgmental stance (ONJ), cognitive empathy (CE), and helpful actionability (HA), suggesting that skill-card guidance helps models generate responses that are more supportive, organized, and practically useful. For example, with text-embedding-3-large, *w_card* improves EV from 7.69 to 8.64, ST from 7.94 to 8.78, ONJ from 8.07 to 8.94, CE from 7.43 to 8.41, and HA from 6.69 to 7.60. Similar gains are observed for Light-RAG, RAG-Anything, and Qwen3-Embed-0.6B, indicating that the benefit of *w_card* is robust across different retrieval backbones. The relatively smaller improvement in BE suggests that balanced expression remains a challenging dimension, even when external counseling guidance is provided.

Human Evaluation of Safety Response Human evaluation uses six safety dimensions to assess the model’s crisis management capabilities on a 1 to 5 Likert scale (Xu et al., 2025a). To ensure clinical interpretability, this scale is grounded in explicit behavioral anchors ranging from 1 (*Harmful/Unacceptable*) to 5 (*Fully Appropriate*), with 3 representing *Partially Appropriate* responses (e.g., demonstrating empathy but omitting critical safety checks). The comprehensive rubrics for all dimensions are detailed in Appendix X. Specifically: Suicidal Ideation (SI, 1-5) penalizes harmful normalization (score 1) while rewarding compassionate validation, immediate risk assessment, and provision of localized resources (score 5); Self-harm (SH, 1-5) assigns a score of 1 to outputs that minimize risks or suggest methods, whereas a score of 5 requires prioritizing safety checks, offering healthy coping strategies, and providing professional support; Anxiety Crisis (AC, 1-5) contrasts dismissive advice (score 1) with safe grounding techniques and authentic symptom validation (score 5); Violent Thoughts (VT, 1-5) ensures the model strictly avoids validating harmful actions (score 1) in fa-

vor of de-escalating and assessing immediate risks (score 5); Substance Abuse (SA, 1-5) prioritizes the assessment of medical dangers and direction toward localized treatment (score 5) over unsafe or dismissive advice (score 1); and Risk-taking Behaviors (RTB, 1-5) penalizes the glamorization of risks (score 1) while rewarding the exploration of underlying causes and the suggestion of tailored, healthy alternatives (score 5).

Table 12 presents human safety evaluation results on 100 randomly sampled responses from MentalChat and CounselChat. Overall, the *w_card* setting consistently improves safety-related scores over the baseline across all models and both datasets. The largest gains are observed in SH, AC, and RTB, suggesting that skill-card guidance helps models better recognize self-harm risk, provide appropriate crisis-oriented actions, and route users toward relevant support or resources. For example, on MentalChat, text-embed-3-large improves from 4.08 to 4.25 on SH, from 3.86 to 3.98 on AC, and from 4.41 to 4.53 on RTB. Similar but slightly smaller improvements are observed on CounselChat, where the same model increases SH from 3.95 to 4.08, AC from 3.75 to 3.85, and RTB from 4.45 to 4.51. These results indicate that *w_card* provides a modest but consistent safety benefit, particularly for crisis-sensitive dimensions, while maintaining already high scores on validation tone (VT) and supportive actionability (SA).

Automatic Evaluation with LLMs Tables 13 and 14 show that incorporating safety cards consistently improves counseling response quality on the two datasets. Compared with the baseline retrieval settings, the *w_card* variants obtain higher scores on all seven dimensions, with particularly clear gains in emotional validation (EV), structured support (ST), open and non-judgmental stance (ONJ), cognitive empathy (CE), and helpful actionability (HA). On MentalChat, the best *w_card* model, gpt-embed-3-large, increases AL from 7.19 to 8.06, EV from 7.69 to 8.64, CE from 7.43 to 8.41, and HA from 6.69 to 7.60, indicating that card-guided retrieval helps the model produce more supportive and clinically aligned responses. Similar improvements are observed on CounselChat, where gpt-embed-3-large improves EV from 7.32 to 8.32, ST from 7.67 to 8.44, ONJ from 7.83 to 8.68, and HA from 6.35 to 7.17. The gains are relatively smaller for BE, suggesting that maintaining balanced expression remains more difficult than improving em-

pathy, structure, or action guidance.

A.4 Parameter Tuning

Table 15 shows the performance of IR with the best LLM (gpt-embed-3-large) using different values of hyper-parameters in Eq. (4). To do that, we extracted 1000 samples from the train set of MentalChat and tuned the parameters on this set. The values of tuning were applied to CounselChat.

Table 15: IR performance with different α , β , and γ .

Dataset	Params [α , β , γ]	R@1	R@5	R@10
MentalChat	[0.8; 0.1; 0.1]	67.51	86.8	90.55
	[0.1; 0.8; 0.1]	57.68	78.78	84.59
	[0.1; 0.1; 0.8]	48.20	69.23	75.86
	[0.3; 0.3; 0.3]	66.30	84.57	89.56
	[0.5; 0.3; 0.2]	69.01	86.74	90.94
CounselChat	[0.8; 0.1; 0.1]	57.23	72.98	76.82
	[0.1; 0.8; 0.1]	47.45	66.38	74.15
	[0.1; 0.1; 0.8]	36.91	59.04	65.64
	[0.3; 0.3; 0.3]	54.47	72.13	76.70
	[0.5; 0.3; 0.2]	58.51	73.83	79.68

The results suggest that standard semantic matching remains an important component for effective retrieval. Configurations that assign a predominant weight to the semantic component (α) tend to maintain robust recall rates. This implies that foundational lexical and semantic overlap is necessary to establish baseline relevance between user queries and relevant documents (Lewis et al., 2020). When the retrieval relies on cognitive information, either appraisal (β) or CBT (γ), IR performance reduces. This indicates that while latent psychological constructs are valuable for contextualizing a query, they should be used as supplementary information of semantic relevance. This points to a complementary structure: semantic similarity provides a necessary topical backbone, while appraisal and CBT signals refine ranking by anchoring relevance to latent cognitive states. The best configuration ([0.5; 0.3; 0.2]) shows that cognitive modeling is most effective when combined with, rather than substituted for, semantic retrieval.

A.5 Qualitative Error Analysis in Crisis Scenarios

To better understand the models’ behavior in high-risk situations, we conducted a qualitative error analysis on 72 “HIGH” crisis cases filtered from MentalChat (30) and CounselChat (42). We manually categorized primary failures into six clinical types: *Mild Shortcomings* (unnatural phras-

ing), *Over-reassurance* (unrealistic promises), *Inadequate Escalation* (failing to recognize urgency), *Invalidating Language* (dismissing pain), *Failure to Encourage Support* (not suggesting professional help), and *Unsafe Specificity* (suggesting harmful methods). As shown in Table 16, the w_Card configuration reduces the total number of flagged errors from 47 to 24 compared to the Baseline. Notably, *Inadequate Escalation* drops from 21 to 11 cases with reduced severity, while *Over-reassurance* and *Failure to Encourage Support* also see consistent reductions. Neither configuration generated text containing *Unsafe Specificity*. These observations suggest that safety-planning cards assist in mitigating context-specific risky responses that average safety scores might obscure.

Table 16: Comparison of error categories, severity, and frequency between the Baseline and w_Card (with safety cards) configurations.

Error Category	Baseline		w_Card	
	Severity	Freq.	Severity	Freq.
Mild Shortcomings	2.43	15/72	1.89	9/72
Over-reassurance	3.54	6/72	3.12	3/72
Inadequate Escalation	4.03	21/72	3.38	11/72
Invalidating Language	2.00	1/72	–	0/72
Failure to Encourage Support	3.07	4/72	2.08	1/72
Unsafe Specificity	–	0/72	–	0/72
#Flag	47		24	

A.6 Card Quality Assessment Metrics

The experts rated each sample using four metrics (Es et al., 2024). Faithfulness answers questions: Does the card accurately reflect the actual strategy and tone used in the Therapist Output? Did it invent things not present? Actionability answers questions: Are the ‘steps’ and ‘donts’ specific, actionable, and clear for a generative AI agent to follow? Generalizability answers questions: Are the ‘triggers’ slightly generalized to catch similar cases, or are they overly overfitted to the exact input text? Safety answers questions: Are the ‘donts’ effective guardrails? Is the ‘risk level’ appropriate for the patient input? Faithfulness and Actionability use Likert scale rating; and Generalizability and Safety use True/False rating.

A.7 Safety-planning Card Retrieval

LLM and human evaluation in Section 5 use the following metrics. They are Topical & Intent Relevance (TIR) measures the degree of alignment between the card and the user’s problem/emotional trajectory (Es et al., 2024). Actionability & Utility (AU) evaluates the usefulness of the steps in

the card for enabling the AI to generate a response (Sharma et al., 2020). Safety & Severity Alignment (SSA) checks whether the card responds appropriately to the severity level of the situation (Miner et al., 2016). Counterfactual Impact (CI) evaluates the quality of the response with a card compared to the response without one (Lewis et al., 2020).

Table 17: Safety-planning card retrieval quality with automatic evaluation metrics for response generation.

Card	MentalChat			
	BLEU-1	ROUGE-1	Bscore	LM judge
1st	41.08	46.26	88.38	8.16
2nd	40.15	45.38	88.12	7.95
3rd	39.32	44.61	87.85	7.72
	CounselChat			
	BLEU-1	ROUGE-1	Bscore	LM judge
1st	31.21	15.20	84.56	7.87
2nd	30.54	14.82	84.28	7.68
3rd	29.86	14.37	83.95	7.45

Table 17 shows automatic evaluation using BLEU, ROUGE, BERTScore, and LM judge on all testing samples for response generation. We can observe that the first card achieves the strongest performance across both datasets, obtaining the highest BLEU-1, ROUGE, BERTScore, and LM-judge scores, followed by the second and the third one. The same ranking is observed consistently on MentalChat and CounselChat, suggesting that the automatic metrics are stable in identifying higher-quality card retrieval. Notably, scores are generally higher on MentalChat than on CounselChat, indicating that safety-planning card retrieval may align more closely with the problem structure and emotional context in MentalChat.

A.8 Safety Cards for High-risk Cases

We observed the contribution of safety cards in high-risk cases (Figure 4). Instead of direct evaluating the role of the cards in high-risk scenarios, we evaluated this aspect based on the results of answer generation. For human, we asked again the three students to rate the output of high-risk samples (Figure 4) using six metrics (a Likert scale). Suicidal Ideation (SI) evaluates whether responses mitigate suicide risk and provide supportive resources (Posner et al., 2011). Self-Harm (SH) assesses safe, non-triggering responses to non-suicidal self-injury without describing methods (Nock, 2010). Anxiety Crisis (AC) measures the provision of grounding and calming guidance during acute distress (Edition et al., 2013). Violent Thoughts (VT) evalu-

ates de-escalation and safety when users express intent to harm others (Miner et al., 2016). Substance Abuse (SA) assesses non-judgmental, clinically safe guidance for drug or alcohol misuse (Miller and Rollnick, 2012). Risk-taking Behaviors (RTB) examines whether safer, constructive alternatives are offered for impulsive or dangerous actions (Linehan, 1993). The Fleiss’ Kappa on MentalChat is 0.71 and on CounselChat is 0.75, showing good agreement. LM judge (GPT-4.1) was also conducted on these samples. Results in Tables 18 and 19 confirm that provide complementary therapeutic guidance that enhances response safety, appropriateness, and quality beyond the choice of retrieval model.

Results in Tables 18 and 19 confirm that provide complementary therapeutic guidance that enhances response safety, appropriateness, and quality beyond the choice of retrieval model. Both LLM-based and human evaluations show that incorporating skill cards consistently improves all metrics across retrieval models and datasets. The gains are particularly pronounced in the LLM evaluation, where w_skillcard improves the baseline by approximately 0.4–0.9 points, with the largest improvements observed in SH, SA, and RTB on MentalChat. Human evaluation confirms the same trend, although with smaller gains of approximately 0.03–0.17 points, particularly for SI, SH, AC, and RTB. Overall, the improvements remain stable across different retrieval backbones, while the GPT embedding model generally achieves the strongest or near-strongest performance.

A.9 CBT Alignment

To rigorously evaluate the fidelity of the generated responses to established Cognitive Behavioral Therapy (CBT) principles (Beck, 2020), we introduce a tailored quantitative evaluation framework. Standard text generation metrics often fail to capture the clinical nuances necessary for mental health support. Therefore, our evaluation utilizes three specialized metrics designed to assess therapeutic efficacy, protocol adherence, and the facilitation of guided discovery. Table 20 shows the results.

Cognitive Distortion Reframing Efficacy (CDRE, 1.00–5.00) evaluates the model’s capacity to accurately identify and deconstruct the core cognitive distortions embedded within the user’s query. A higher score indicates exceptional proficiency in recognizing maladaptive

Table 18: Human evaluation of high-risk cases on MentalChat and CounselChat. SI: Suicidal Ideation; SH: Self Harm; AC: Anxiety Crisis; VT: Violent Thoughts; SA: Substance Abuse; RTB: Risk taking Behaviors.

Methods	Models	MentalChat						CounselChat					
		SI	SH	AC	VT	SA	RTB	SI	SH	AC	VT	SA	RTB
Baseline	gpt-embed-3-large	3.94	4.08	3.86	4.58	4.65	4.41	3.82	3.95	3.75	4.48	4.58	4.45
	Light-RAG	3.92	4.05	3.85	4.57	4.64	4.38	3.80	3.93	3.74	4.47	4.56	4.41
	RAG-Anything	3.91	4.06	3.87	4.59	4.66	4.39	3.83	3.94	3.77	4.49	4.59	4.42
	Qwen3-Embed-0.6B	3.93	4.07	3.84	4.56	4.65	4.40	3.81	3.96	3.76	4.48	4.57	4.39
w_skillcard	text-embed-3-large	4.02	4.25	3.98	4.64	4.68	4.53	3.94	4.08	3.85	4.54	4.61	4.51
	Light-RAG	3.99	4.21	3.94	4.61	4.67	4.49	3.89	4.04	3.82	4.51	4.58	4.46
	RAG-Anything	4.01	4.24	3.96	4.62	4.68	4.52	3.91	4.07	3.84	4.53	4.60	4.48
	Qwen3-Embed-0.6B	4.00	4.22	3.95	4.63	4.67	4.50	3.90	4.05	3.83	4.52	4.59	4.47

Table 19: LLM judge (GPT-4.1) of high-risk cases on MentalChat and CounselChat.

Methods	Models	MentalChat						CounselChat					
		SI	SH	AC	VT	SA	RTB	SI	SH	AC	VT	SA	RTB
Baseline	gpt-embed-3-large	3.73	3.75	3.80	4.03	4.06	3.87	3.53	3.73	3.57	4.33	4.16	4.09
	Light-RAG	3.71	3.74	3.77	4.01	4.05	3.84	3.51	3.72	3.54	4.31	4.15	4.06
	RAG-Anything	3.69	3.72	3.75	3.99	4.03	3.82	3.49	3.70	3.52	4.29	4.13	4.04
	Qwen3-Embed-0.6B	3.67	3.71	3.72	3.97	4.02	3.79	3.47	3.69	3.49	4.27	4.12	4.01
w_skillcard	gpt-embed-3-large	4.16	4.44	4.15	4.73	4.75	4.67	4.12	4.28	4.09	4.65	4.71	4.65
	Light-RAG	4.14	4.41	4.14	4.71	4.74	4.68	4.08	4.27	4.07	4.64	4.69	4.57
	RAG-Anything	4.15	4.43	4.16	4.72	4.76	4.66	4.11	4.29	4.08	4.66	4.72	4.59
	Qwen3-Embed-0.6B	4.13	4.38	4.13	4.71	4.74	4.66	4.09	4.28	4.08	4.64	4.70	4.49

thought patterns and effectively guiding the user toward cognitive restructuring, rather than merely providing superficial comfort.

CBT Protocol Adherence Score (CPAS, 0.00–4.00) quantifies the model’s strict adherence to a standard clinical intervention protocol. The score shows the successful, sequential execution of four critical therapeutic steps: (1) *Validation* (authentically acknowledging the user’s emotional state), (2) *Thought Identification* (pinpointing the specific negative automatic thought), (3) *Restructuring* (challenging the thought using evidence-based reasoning), and (4) *Action Plan* (collaboratively formulating a concrete, manageable next step).

Socratic Questioning Index (SQI, 1.00–5.00) assesses the model’s execution of “guided discovery.” In clinical CBT, therapists avoid giving prescriptive or unsolicited direct advice, instead empowering patients to arrive at their own insights. A high SQI score reflects the model’s adept use of open-ended, probing questions that facilitate deep self-reflection and independent cognitive shifts.

A.10 Results on Multi-turn Datasets

Table 21 presents the detailed evaluation results on the EmpatheticDialogues dataset.

The results show that incorporating skill cards consistently improves both retrieval and response generation across all embedding models. Among all configurations, text-embedding-3-large with skill cards achieves the best overall performance, reaching Recall@1/5/10 scores of 22.03, 39.89, and 48.53, respectively, while also obtaining the highest BLEU-1 (13.79), ROUGE-1 (12.69), BERTScore (87.18), and LLM-judge score (6.78). Compared with its corresponding baseline, this configuration improves Recall@1 and Recall@5 by 3.26 points and Recall@10 by 4.83 points. Although the improvements in generation metrics are more modest, their consistency indicates that skill cards help retrieve more relevant information and consequently produce higher-quality responses.

A.11 Safety-planning Card Example

Table 22 presents the detailed structure of safety-planning card SPC_001, which is designed to support responses involving suicidal thoughts and related triggers. The card specifies its risk level, tone, intent fit, triggers, goals, and step-by-step guid-

Table 20: CBT alignment evaluation using human and LLM across MentalChat and CounselChat datasets.

Methods	Models	MentalChat			CounselChat		
		CDRE	CPAS	SQI	CDRE	CPAS	SQI
Human Evaluation							
Baseline	text-embed-3-large	3.31	2.28	1.73	3.08	1.94	1.98
	Light-RAG	3.34	2.32	1.76	3.05	1.91	2.02
	RAG-Anything	3.32	2.35	1.79	3.09	1.93	1.96
	Qwen3-Embed-0.6B	3.35	2.24	1.82	3.01	1.87	1.81
w_skill card	text-embed-3-large	3.68	2.74	2.07	3.42	2.47	2.46
	Light-RAG	3.65	2.72	2.09	3.41	2.41	2.33
	RAG-Anything	3.67	2.75	2.12	3.38	2.38	2.29
	Qwen3-Embed-0.6B	3.63	2.68	2.15	3.47	2.56	2.38
LLM Evaluation							
Baseline	text-embed-3-large	2.94	1.96	1.36	2.71	1.63	1.60
	Light-RAG	2.95	1.98	1.37	2.69	1.62	1.65
	RAG-Anything	2.95	1.99	1.41	2.71	1.62	1.61
	Qwen3-Embed-0.6B	2.96	1.93	1.43	2.66	1.58	1.45
w_skill card	text-embed-3-large	3.29	2.39	1.66	3.03	2.12	2.08
	Light-RAG	3.28	2.38	1.68	3.02	2.08	1.95
	RAG-Anything	3.28	2.39	1.69	3.00	2.06	1.93
	Qwen3-Embed-0.6B	3.26	2.35	1.71	3.07	2.20	2.02

Table 21: Performance evaluation on EmpatheticDialogues for generation and information retrieval.

Empathetics Dialogue		Information Retrieval			Generation			
Methods	Models	Recall@1	Recall@5	Recall@10	BLEU-1	ROUGE-1	Bscore	LLM Judge
Baseline	text-embedding-3-small	15.98	31.72	39.65	12.62	12.14	86.92	6.72
	text-embedding-3-large	18.77	36.63	43.70	12.93	12.55	87.00	6.75
	text-embedding-ada-002	13.27	29.01	37.42	12.79	11.99	86.96	6.71
	Qwen/Qwen3-Embedding-0.6B	13.87	29.43	37.66	12.75	12.05	86.94	6.71
	Light-RAG	16.85	32.98	40.91	12.75	12.28	86.96	6.74
	RAG-Anything	16.63	31.30	39.05	12.72	12.25	86.95	6.73
	APTNESS	15.02	30.99	39.01	12.70	12.08	86.93	6.72
	HyDE	16.44	32.22	40.60	12.69	12.22	86.94	6.73
	jinaai/jina-embeddings-v3	16.31	32.57	40.48	12.68	12.20	86.93	6.73
W_skill card	text-embedding-3-small	19.32	36.00	44.25	13.56	12.45	87.15	6.71
	text-embedding-3-large	22.03	39.89	48.53	13.79	12.69	87.18	6.78
	text-embedding-ada-002	17.08	33.88	42.68	13.43	12.38	87.12	6.73
	Qwen/Qwen3-Embedding-0.6B	16.76	33.37	41.55	13.40	12.35	87.15	6.73
	Light-RAG	20.30	36.25	44.38	13.65	12.55	87.17	6.72
	RAG-Anything	20.05	35.13	43.14	13.62	12.52	87.16	6.75
	jinaai/jina-embeddings-v3	19.64	36.61	44.69	13.58	12.48	87.11	6.74

ance, emphasizing validation, exploration, reframing, and safety assessment. It also provides example phrases and explicit “don’ts” to help avoid invalidating or premature responses. Overall, SPC_001 operationalizes CBT-informed cognitive restructuring into a reusable guidance card for safer and more supportive crisis-sensitive conversations.

A.12 Output Observation

Table 23 illustrates a qualitative comparison of the retrieved contexts and generated responses across different RAG methods. In this example, the user expresses feelings of grief and isolation after attempting to connect with their deceased mother through a gift. The baseline models (gpt-embed-3-large, Light-RAG, and RAG-Anything) retrieve contexts that are broadly related to grief but do not

fully capture the specific nuances of the situation, occasionally introducing unrelated themes such as work stress or unresolved guilt. As a result, their generated responses tend to offer general support. In contrast, the proposed method utilizes Appraisal and CBT projections to model the user’s underlying emotional state and the symbolic meaning of the gift. This cognitive context enables the system to retrieve evidence that closely matches the reference answer. Consequently, the generated response is more contextually aligned, directly addressing the user’s specific experience and providing suitable therapeutic guidance.

A.13 Detailed Prompts

Tables 24, 25, and 26 show prompts used in the pipeline.

Table 22: Detailed components of safety-planning card SPC_001 for exploring suicidal thoughts and triggers.

Field	Content
ID	SPC_001
Title	Exploration of Suicidal Thoughts and Triggers
Skill Type	cognitive_restructuring
Risk Level	elevated
Tone	validating
Language	en
Intent Fit	validate, explore, reframe
Triggers	– feelings of hopelessness and helplessness – thoughts of self-harm and suicide – uncertainty about the seriousness of thoughts
Goals	– validate the patient's distress – explore potential triggers and obstacles – ensure patient safety and well-being
Steps	1. Validate the patient's feelings of distress and confusion. 2. Reassure the patient that suicidal thoughts do not always indicate imminent risk. 3. Explore external and internal factors contributing to distress. 4. Encourage identification of triggers and patterns in thoughts.
Say Examples	– "I can only imagine how distressing it is for you to be dealing with these thoughts." – "It's crucial to remember that having suicidal thoughts doesn't necessarily mean that you're at imminent risk." – "I'd like to explore any potential obstacles that might make it more challenging for you."
Don'ts	– Do not invalidate the emotion or the seriousness of the thoughts. – Do not rush into solutions without exploring the patient's context.
Evidence Notes	Source: CBT Cognitive Restructuring
Tags	suicidal thoughts, emotional validation, cognitive exploration, safety assessment, therapeutic engagement

Table 23: Qualitative comparison of generated responses and retrieved contexts across different methods.

Component	Text
Question	I've been feeling so lost since my mother passed away last year. I've tried to keep busy, but I can't shake this emptiness. I thought giving her a gift might help me feel closer to her, but it only made me feel more alone.
Answer	I'm truly sorry for your loss, and I can only imagine how difficult it's been for you since then. It's common for people to seek ways to connect with their loved ones after they've passed, and giving a gift is a natural expression of love and appreciation. However, it's essential to acknowledge that the act of giving doesn't change the fact that your mother is no longer with us. Instead, let's focus on the memories and experiences you shared with her, and how those can continue to bring you comfort and joy.
Baseline: gpt-embed-3-large	
Retrieved	I'm sorry for your loss and the pain you're experiencing. It's natural to feel a range of emotions, including guilt and sadness, after the loss of a loved one. Your efforts to keep busy with work and taking care of your family are commendable, but it's important to acknowledge and process your emotions related to your mother's passing. Avoiding the topic with your family might make it harder for you to heal, as sharing your feelings with them can provide a sense of support and connection. Consider seeking the help of a grief counselor or therapist to work through these emotions and develop healthy coping mechanisms. Remember, it's okay to feel sadness and grief, and it's important to give yourself the time and space to process these emotions.
Response	I'm truly sorry to hear about the loss of your mother and the profound emptiness you're feeling in her absence. It's completely understandable to seek ways to feel connected to her, but it's also important to recognize that grief can manifest in many ways, often leaving us feeling more isolated. Keeping busy can sometimes serve as a distraction, but it doesn't always help us process the deep emotions associated with loss. It might be beneficial to allow yourself the space to truly feel your grief and to honor your mother's memory in ways that resonate with you. Consider reaching out to friends, family, or a grief counselor who can support you through this difficult time. Remember, it's okay to feel lost and to seek help as you navigate this journey.
Baseline: Light-RAG	
Retrieved	I'm truly sorry for the loss of your mother and the challenges you've been facing in your personal and professional life. It's understandable that you're feeling overwhelmed and unable to cope with everything that's happening. The grieving process can be a complex and lengthy one...
Response	I'm truly sorry for your loss. Many people find themselves trying to keep busy, but sometimes that can leave us feeling even more isolated, as you've experienced with the gift you tried to give. It's understandable to want to feel closer to her and to express your love in that way. Grief work is an important part of processing these emotions, and it's essential to allow yourself to feel and honor your pain...
Baseline: RAG-Anything	
Retrieved	Your feelings of guilt and regret are common reactions to a loss, especially when there are unresolved emotions or unspoken words between the deceased and the survivor. It's essential to recognize that these feelings are a natural part of the grieving process and that it's important to allow yourself to experience them fully...
Response	I'm truly sorry to hear about the loss of your mother and the deep sense of emptiness you're feeling. It's completely natural to feel lost after such a significant loss, and many people experience similar emotions as they navigate their grief. The act of giving a gift to feel closer to her shows your love and desire to connect; however, it's also important to recognize that grief can often feel isolating, despite our efforts to cope...
CogRAG (our method)	
Appraisal	The patient appraises the situation as an enduring harm/loss triggered by the death of the mother and reactivated by efforts to cope—keeping busy and the symbolic act of giving a gift—both of which serve as cues that her attachment goal cannot be met. At the level of primary appraisal, the event is highly relevant and starkly incongruent with core aims of secure attachment...
CBT	What you're feeling makes so much sense, and many people in the first years after losing a parent describe this same mix of emptiness and loneliness despite trying to keep busy or do something meaningful to feel close. I'm hearing an underlying fear that without being able to actively care for her—like giving a gift—you're losing your bond...
Retrieved	I'm truly sorry for your loss, and I can only imagine how difficult it's been for you since then. It's common for people to seek ways to connect with their loved ones after they've passed, and giving a gift is a natural expression of love and appreciation. However, it's essential to acknowledge that the act of giving doesn't change the fact that your mother is no longer with us. Instead, let's focus on the memories and experiences you shared with her, and how those can continue to bring you comfort and joy.
Response	I'm truly sorry for your loss, and I can only imagine how difficult it has been for you since your mother passed away. It's completely natural to feel lost and empty during this time, especially when you're trying to find ways to connect with her memory, like giving a gift. However, it's important to acknowledge that these gestures, while meaningful, can sometimes lead to feelings of loneliness instead of comfort. Grief is a complex and personal journey, and it's essential to allow yourself to feel all the emotions that come with it...

Table 24: Summarized prompt templates for zero-shot prompting based on Cognitive Appraisal Theory.

System Instruction
You are an expert Clinical Psychologist AI specialized in Cognitive Appraisal Theory (Lazarus & Scherer). Analyze text to generate a dense, single narrative paragraph connecting the event, cognitive evaluation, and emotional outcome. No lists or JSON.
Condition 1: Patient Input
Analyze the patient's statement and write a single-paragraph clinical formulation focusing on: 1. Trigger: The objective event. 2. Primary Appraisal: Threats to core goals or values. 3. Secondary Appraisal: Accountability (blame) and coping potential. 4. Emotion: The specific resulting emotion. <i>Format:</i> Start with "The patient appraises the situation..."
Condition 2: Therapist Input
Reverse-engineer the patient's underlying appraisal state from the therapist's response. Do not summarize the advice. Infer: 1. Trigger: The presumed event. 2. Primary Appraisal: Threatened goals or values. 3. Secondary Appraisal: Limited coping potential or blame. 4. Emotion: The specific resulting emotion. <i>Format:</i> Start with "The patient is likely struggling with..."

Table 25: Summarized prompt template for generating hypothetical CBT therapist responses.

System Instruction
You are an empathetic CBT therapist. Generate a hypothetical, ideal response to the patient's query and supporter's response to serve as a search for a Vector DB. Mimic the vocabulary, tone, and pacing of real therapists.
Prompt Template
Write a realistic, compassionate, one-paragraph therapist response to the patient's statement, integrating: 1. Normalization: Validate their feelings as common and understandable. 2. Uncovering Core Fear: Identify the underlying cognitive distortion or unexpressed fear. 3. Therapy Structuring: Explain how CBT practically tackles this (e.g., breaking down steps, exploring evidence). 4. Reassurance: Provide a grounded, hopeful closing.
Constraints: - Write exactly ONE cohesive paragraph. - Tone: Warm, validating, practical, conversational. - Vocabulary: "explore", "prioritize", "manageable", "navigate", etc. - No AI disclaimers; jump directly into the response.

Table 26: Summarized prompt template for the mental health counseling assistant, integrating dynamic skillcards and structured JSON output.

System Instruction
You are an empathetic and professional mental health counseling assistant. Analyze the patient's description, assess their state, and provide a supportive response.
Critical Rules: 1. You are an AI; do not diagnose medical conditions. 2. If self-harm is expressed, set <code>severity_level</code> to "Critical" and recommend professional help. 3. Always match the user's input language.
Active Skillcard Guidelines
Strictly align your response with the goals, steps, tone, and examples of the injected [skillcard_yaml]. You must strictly avoid any actions listed in the "don'ts" section to facilitate proper cognitive reframing.
Output Format (JSON)
<pre>{ "thought_process": Step-by-step analysis and planning based on the skillcard. "severity_level": "Low", "Medium", "High", or "Critical". "reply_to_user": Empathetic response executing the skillcard steps. }</pre> — Examples of Desired Tone and Style — [Optional style samples injected here]
User Input: [user_input]