

MindUntangler: Resolving Cognitive Entanglements via Multi-Agent Topological Routing in Psychological Counseling

Xiaoyu Liu¹, Yufeng Diao^{1,*}, Jingjie Zeng², Shiqi Li³, Hao Zhang¹, Liang Yang²

¹School of Computer Science and Technology, Inner Mongolia Minzu University, China

²School of Computer Science and Technology, Dalian University of Technology, China

³Information Technology, Tongliao Vocational College, China

liuxiaoyu0014@163.com, diaoyufeng@imun.edu.cn

Abstract

Large Language Models (LLMs) offer scalable mental health support, but they often operate at the text-response level, producing generic validation without explicitly tracking the cognitive structures behind a patient's distress. We propose **MindUntangler**, a multi-agent framework that models counseling as dynamic topological routing over a Cognitive Topology Graph (CTG). The framework represents maladaptive associations as pathological edges, verifies extracted graph structures, routes among therapeutic frameworks, and applies guardrails for crisis-sensitive states. We further align generation with a topology-driven Direct Preference Optimization objective. Experiments on two retrospective counseling benchmarks show improved expert-rated counseling quality and larger reductions in expert-annotated pathological-edge severity than strong general and counseling-specific baselines. These results suggest that explicit graph-state modeling can support more targeted and interpretable AI-assisted counseling in off-line evaluation settings.

1 Introduction

Psychological counseling involves more than empathetic listening. Many counseling frameworks aim to identify maladaptive beliefs, emotional triggers, and behavioral responses, and to help patients examine these patterns in a structured way (Beck et al., 2024; Ellis, 1962). Related psychopathology network theories model mental distress as interactions among symptoms, beliefs, and contextual factors rather than as a single latent cause (Cramer et al., 2010; Borsboom et al., 2021).

Recent Large Language Model advancements have spurred conversational agents for mental health support (Thirunavukarasu et al., 2023; Stade et al., 2024). However, these systems mostly operate within the superficial text space. Early empathetic response generation gradually evolved into

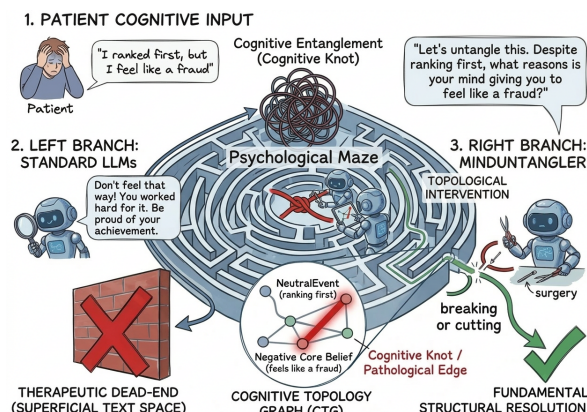


Figure 1: Motivating example for graph-state counseling. Given the utterance “I ranked first, but I feel like a fraud,” a text-only response may provide generic validation without examining the association between achievement and negative self-belief. **MindUntangler** represents this association as a candidate pathological edge in the CTG and uses the graph state to guide intervention selection.

LLMs fine-tuned on psychological corpora (Qiu et al., 2024; Zhang et al., 2024). While latest flagship models like GPT-5.4 (OpenAI, 2026b) excel in reasoning, they exhibit a direct tone, lacking the conversational warmth of earlier models like GPT-4o (OpenAI, 2024). Moreover, even emotionally resonant models are often insufficient for structural intervention. Against complex psychological resistance, highly empathetic text-matching models may still generate surface-level validation, failing to address the underlying structural issues of a patient's distress (Sharma et al., 2023).

To illustrate this limitation, consider the scenario in Figure 1. When a patient says, "I ranked first, but I feel like a fraud," a standard text-based LLM may offer validation or reassurance. Such responses can be appropriate at the surface level, but they may leave the underlying maladaptive association unexamined: a positive achievement is linked to a negative self-belief. A human counselor would

often track this association across events, beliefs, emotions, and prior experiences before choosing an intervention.

MindUntangler addresses this challenge by moving from response-only generation to explicit graph-state modeling. We project the dialogue into a Cognitive Topology Graph (CTG), inspired by graph-augmented reasoning (Edge et al., 2024; Luo et al., 2023; Pan et al., 2024). The CTG represents counseling-relevant entities and associations, allowing the system to track candidate pathological edges, verify extracted structures, and select an intervention strategy conditioned on the current graph state. Specifically, MindUntangler first extracts and verifies a CTG from the ongoing dialogue, then uses a probabilistic router to select a therapeutic framework under safety constraints, and finally aligns generation through topology-driven preference optimization. This formulation treats graph transition as an operational target for AI-assisted counseling rather than as evidence of real-world clinical recovery.

In summary, our work makes three contributions:

- We formulate counseling-oriented cognitive distortions as pathological edges in a Cognitive Topology Graph, enabling events, beliefs, symptoms, traumas, and maladaptive associations to be tracked jointly across dialogue turns for routing, safety checking, and evaluation.
- We propose **MindUntangler**, a multi-agent framework that integrates graph memory, topology verification, safety guardrails, dynamic therapeutic routing, and topology-driven preference optimization within one interpretable counseling pipeline.
- We evaluate MindUntangler on two retrospective counseling benchmarks, showing gains over general-purpose and counseling-specific baselines under the same offline evaluation setting.

2 Related Work

LLM-based Counseling. LLM-based counseling has moved from generic empathetic response generation toward more structured therapeutic support. Domain-specific systems such as SoulChat (Chen et al., 2023) and CPsyCounX (Zhang et al., 2024) improve emotional alignment and counseling style in Chinese dialogue settings. Recent work further studies structured therapeutic data and specialized counseling agents: Cactus (Lee et al., 2024) constructs CBT-style multi-turn counseling conversations; MIR-

ROR (Kim et al., 2025) incorporates multimodal cues to handle client resistance; PanicToCalm (Lee et al., 2025) focuses on panic-attack support with PFA-oriented data and preference alignment. These studies indicate a shift from open-ended support to framework-guided interaction.

However, most prior systems still optimize the counselor response as text. Even when they use structured therapeutic principles, they usually do not maintain an explicit evolving representation of the patient’s events, beliefs, symptoms, traumas, and pathological associations across turns. MindUntangler differs by treating this latent counseling state as a Cognitive Topology Graph (CTG), which is used not only as memory, but also as the basis for routing, safety checks, and structural evaluation.

Graph Structures and Cognitive Representations. Network psychopathology models mental distress as interactions among symptoms, beliefs, events, and contextual factors (Cramer et al., 2010; Borsboom et al., 2021). This view is consistent with CBT, where events trigger interpretations, core beliefs shape those interpretations, and symptoms emerge as emotional, behavioral, or physiological responses (Beck et al., 2024; Ellis, 1962). Accordingly, our CTG uses typed nodes for Events, Core Beliefs, Symptoms, and Traumas, and directed relations such as *triggers* and *exacerbates*. This ontology is not intended as a clinical taxonomy; it is a compact operational schema for tracking the associations most relevant to cognitive restructuring.

Computationally, Graph Retrieval-Augmented Generation (GraphRAG) has been used for relational representation and graph-augmented reasoning (Edge et al., 2024; Luo et al., 2023; Pan et al., 2024). Existing GraphRAG methods are mainly designed for factual retrieval or knowledge-intensive reasoning. MindUntangler instead uses the graph as the operative counseling state: edge weights are updated according to contextual stance, candidate structures are verified before routing, and intervention quality is evaluated through expert-defined pathological-edge reduction. Beyond counseling dialogue, related studies have explored scalable knowledge selection through iterative hybrid encoding and re-ranking (Huang et al., 2025), as well as ensemble reasoning for clinical decision support (Rashid and McGuinness, 2025). These studies provide complementary perspectives on structured selection and coordinated reasoning, while focusing on knowledge retrieval and clinical decision

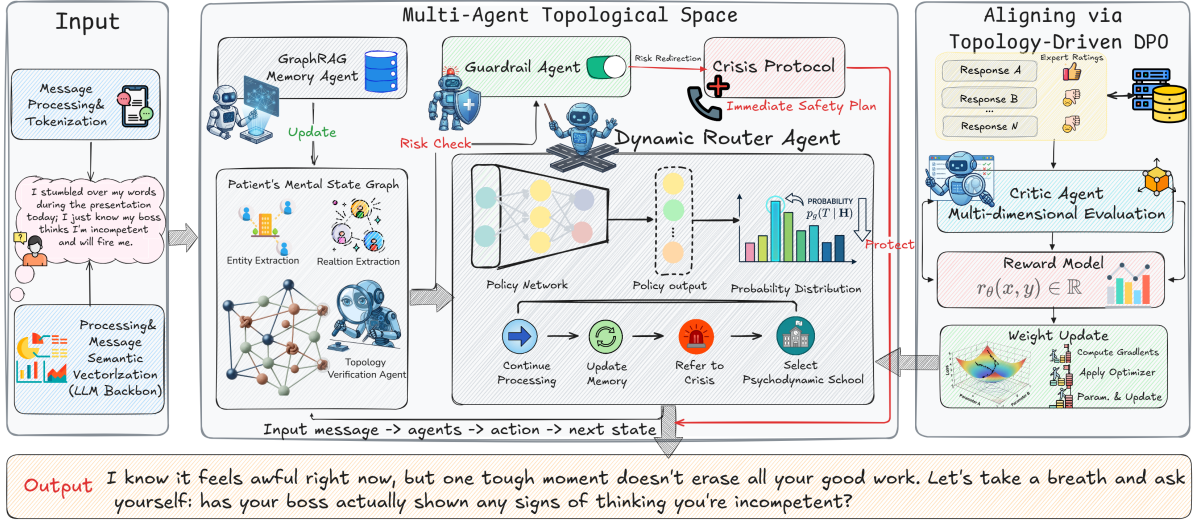


Figure 2: The architecture of the **MindUntangler** framework. A GraphRAG Memory Agent dynamically constructs the Cognitive Topology Graph (CTG), which is filtered by a Topology Verification Agent to reduce noise. Within this verified CTG, a probabilistic Router Agent navigates therapeutic frameworks (e.g., CBT, DBT), while a Guardrail Agent enforces clinical safety. Finally, the generation policy is aligned via a topology-driven Critic Agent and Direct Preference Optimization.

support rather than dynamic psychological counseling dialogue.

3 Methodology

This section describes the computational framework of **MindUntangler**. We first formulate cognitive distortions as pathological edges in a Cognitive Topology Graph (§3.1). As shown in Figure 2, the workflow proceeds in three stages. First, the GraphRAG Memory Agent extracts a candidate graph state from the dialogue (§3.2), and the Topology Verification Agent checks whether the extracted nodes and edges are supported by context (§3.3). Second, the Guardrail Agent monitors red-line states (§3.4), while the Router Agent selects a suitable therapeutic framework from the verified graph state (§3.5). Third, the generation policy is aligned with a Topology-Driven Critic Agent and DPO (§3.6) to reduce expert-defined pathological-edge severity under professional-boundary constraints. Algorithm 1 summarizes this workflow.

3.1 Problem Formulation

Traditional LLM-based counseling maps a dialogue context X_t to a response y_t in the text space. MindUntangler introduces an intermediate Cognitive Topology Graph (CTG) to represent the counseling state before response generation. At turn t , the CTG is a heterogeneous directed graph $G_t = (V_t, E_t, \mathbf{W}_t)$.

Nodes and Relations. The node set V_t contains four clinically motivated types: Events (V_{evt}), Core Beliefs (V_{blf}), Symptoms (V_{sym}), and Traumas (V_{trauma}). Inspired by CBT conceptualization and symptom-network theory (Beck et al., 2024; Borsboom et al., 2021), Events denote activating situations, Core Beliefs capture relatively stable appraisals, Symptoms group emotional, behavioral, and physiological response states, and Traumas represent prior adverse experiences relevant to current distress. This compact ontology intentionally treats automatic thoughts as utterance-level evidence or candidate edges, rather than persistent nodes. Directed edges E_t encode relations such as *triggers*, *exacerbates*, or *manifests_as*. Each node v_i has an embedding $\mathbf{h}_i \in \mathbb{R}^d$, and $\mathbf{W}_t$ stores edge severity weights. Table 5 summarizes the clinical grounding of each graph element.

Pathological Edges. A pathological edge is an expert- or system-identified association that links an event, belief, symptom, or trauma to a maladaptive interpretation or response. We focus on high-severity edges because they represent targets for cognitive restructuring. Formally, for the relevant relation subset $E_{rel}^{(t)} \subseteq E_t$, we define:

$$\mathcal{E}_{path}^{(t)} = \{e_{ij} \in E_{rel}^{(t)} \mid w_{ij}^{(t)} > \tau_{path}\}. \quad (1)$$

Objective. Given G_t and dialogue context X_t , the system selects a therapeutic framework and generates y_t . The desired response should re-

duce the severity of expert-identified pathological edges while preserving professional and safety constraints. Thus, the graph transition $G_t \xrightarrow{y_t} G_{t+1}$ is used as an operational target for response generation and evaluation, rather than as a claim of real-world clinical recovery.

3.2 GraphRAG Memory Agent

At each turn t , the GraphRAG Memory Agent constructs a pre-response CTG state from the dialogue. The extraction follows a CBT-oriented ontology $\mathcal{O} = \{V_{types}, E_{types}\}$, with node types including Events, Core Beliefs, Symptoms, and Traumas, and relation types such as *triggers* and *exacerbates*.

The agent first extracts candidate nodes ΔV from the patient’s utterance u_t through few-shot prompting. It then predicts candidate relations ΔE between ΔV and the historical graph G_{t-1} . The resulting candidate subgraph $\Delta G_t = (\Delta V, \Delta E)$ is passed to the Topology Verification Agent before being merged into the global graph.

For each accepted edge e_{ij} , we update its severity using both semantic relevance and cognitive stance. Let $\mathbf{h}_{u_t}$ be the sentence embedding of u_t , and let $\mathbf{h}_i, \mathbf{h}_j$ be the embeddings of the connected nodes. The topological relevance is:

$$\text{Rel}(u_t, e_{ij}) = \max_{v \in \{i, j\}} \max(0, \text{sim}(\mathbf{h}_{u_t}, \mathbf{h}_v)). \quad (2)$$

The stance term $\phi(u_t, e_{ij}) \in [-1, 1]$ indicates whether the utterance reinforces, weakens, or is neutral to the edge. In implementation, it is assigned on a five-level ordinal scale $\{-1, -0.5, 0, 0.5, 1\}$, where positive values indicate reinforcement and negative values indicate reappraisal or weakening.

Because the current routing decision is selected after memory construction, the online weight update uses the previous active framework $\mathcal{F}_{t-1}$. For a newly accepted edge, $w_{ij}^{(t-1)}$ is initialized to the preservation constant ϵ . The update is:

$$\tilde{w}_{ij}^{(t)} = \lambda_{\mathcal{F}_{t-1}} w_{ij}^{(t-1)} + (1 - \lambda_{\mathcal{F}_{t-1}}) \text{Rel}(u_t, e_{ij}) \phi(u_t, e_{ij}), \quad (3)$$

$$w_{ij}^{(t)} = \max(\epsilon, \tilde{w}_{ij}^{(t)}). \quad (4)$$

The resulting G_t is the graph state consumed by the Router Agent and the response generator. In response-level evaluation, post-response severity $w_{ij}^{(t+1)}$ is computed on the expert-annotated benchmark rather than from the model’s own graph, as described in Appendix §D.

3.3 Topology Verification Agent

Psychological utterances often contain negation, quotation, metaphor, and speaker shifts. A direct extractor may therefore create spurious pathological edges, for example by treating a parent’s insult quoted by the patient as the patient’s own active core belief. To reduce such errors, MindUntangler introduces a **Topology Verification Agent** between graph extraction and routing.

Given the candidate subgraph ΔG_t , the agent checks whether each extracted node or edge is supported by the original utterance and recent dialogue context. It focuses on three error types: contextual misattribution, ignored negation, and metaphorical over-interpretation. Edges judged unsupported are pruned before the graph is passed to the Router Agent; supported edges are retained and then updated by the weight mechanism in Eq. 4.

This verification step is designed to limit error propagation from graph construction to intervention selection. Its quantitative effect is reported in Appendix §A.2: the agent prunes 27.4% of candidate pathological edges, and 78.1% of the pruned edges are judged invalid by experts.

3.4 Guardrail Agent

Mental-health dialogue systems require a separate safety mechanism for crisis-sensitive states. MindUntangler therefore includes a **Guardrail Agent** that operates before therapeutic routing is finalized. Its role is not to choose a counseling framework, but to detect whether the current graph state contains safety-critical signals that should override normal intervention.

We define a red-line node set $\mathcal{V}_{red} \subset V$ with licensed psychologists, using crisis-intervention protocols and clinical guidelines as references. The set covers categories such as active self-harm intent, suicide planning, self-mutilation methods, homicidal ideation, and imminent danger. At each turn, the Guardrail Agent estimates the activation of these red-line nodes from the verified CTG and the current utterance. If any red-line node exceeds the threshold,

$$\exists v \in \mathcal{V}_{red}, \quad \text{Act}(v) > \tau_{risk}, \quad (5)$$

the system applies a deterministic override to the Crisis Protocol $\mathcal{F}_{crisis}$.

This override is separate from CKADR routing: $\mathcal{F}_{crisis}$ is not a router label and is not sampled from $P(\mathcal{F} \mid G_t)$. The Crisis Protocol prioritizes supportive de-escalation, refusal of unsafe instructions,

Algorithm 1 Dynamic Topological Untangling

Require: Patient utterance u_t , Historical CTG G_{t-1} , Previous framework $\mathcal{F}_{t-1}$

1: **Parameters:** thresholds τ_{path}, τ_{risk} ; baseline ϵ

Ensure: Response y_t , Updated CTG G_t

2: **Stage 1: Extraction & Verification**

3: Extract subgraph $\Delta G_t = (\Delta V, \Delta E)$ from u_t via GraphRAG Agent (§3.2).

4: **if** Topology Verification Agent detects semantic error in ΔG_t **then**

5: Prune invalid edges and execute rollback (§3.3).

6: **end if**

7: Update global CTG: $G_t \leftarrow G_{t-1} \cup \Delta G_t$

8: Update weights $w_{ij}^{(t)} = \max(\epsilon, \tilde{w}_{ij}^{(t)})$ via stance ϕ and decay $\lambda_{\mathcal{F}_{t-1}}$ (Eq. 4).

9: **Stage 2: Safety & Dynamic Routing**

10: **if** $\exists v \in \mathcal{V}_{red}$ with $\text{Act}(v) > \tau_{risk}$ **then**

11: $\mathcal{F} \leftarrow \mathcal{F}_{crisis}$ {Guardrail Agent override (§3.4)}

12: **else**

13: Compute state vector $\mathbf{s}_t$ from CTG G_t .

14: Sample $\mathcal{F} \sim P(\mathcal{F} | G_t)$ via CKADR network f_{θ_R} (Eq. 6).

15: **end if**

16: Store the selected framework as $\mathcal{F}_t$ for the next memory update.

17: **Stage 3: Therapeutic Generation**

18: Generate therapeutic response y_t conditioned on $\{G_t, \mathcal{E}_{path}^{(t)}, \mathcal{F}\}$.

19: **return** y_t, G_t

and recommendation of appropriate external help resources. Held-out guardrail validation is reported in Appendix §A.2, and the threshold sensitivity analysis is shown in Table 10. This mechanism reduces unsafe normal routing, but it does not replace human crisis assessment or clinical supervision.

3.5 Topological Router

Different cognitive states require different interventions. We refer to this supervised routing module as the Clinical-Knowledge-Aware Dynamic Router (CKADR). We therefore train a supervised router that maps the verified CTG state to one of three therapeutic frameworks, $\mathcal{F} \in \{\mathcal{F}_{trauma}, \mathcal{F}_{cbt}, \mathcal{F}_{dbt}\}$. The router does not predict crisis handling. When the Guardrail Agent detects a red-line state, control is deterministically overridden to $\mathcal{F}_{crisis}$, which is excluded from the router label space. This separation avoids conflating safety intervention with therapeutic strategy selection.

At turn t , we encode the verified graph G_t into a state vector $\mathbf{s}_t$ using graph-derived features, including pathological-edge density, trauma-node activation, proximity between trauma and symptom clusters, and resistance/arousal indicators. Licensed psychologists provide routing labels over $\{\mathcal{F}_{cbt}, \mathcal{F}_{dbt}, \mathcal{F}_{trauma}\}$, while $\mathcal{F}_{crisis}$ is reserved for

Guardrail override. We then train a lightweight classifier f_{θ_R} :

$$P(\mathcal{F} | G_t) = \text{Softmax}(f_{\theta_R}(\mathbf{s}_t)). \quad (6)$$

In this label scheme, $\mathcal{F}_{cbt}$ is used for stable maladaptive belief edges with low-to-moderate arousal, $\mathcal{F}_{dbt}$ for high arousal, resistance, or rupture risk, and $\mathcal{F}_{trauma}$ when trauma nodes are active and strongly linked to the current symptom cluster. Component-level validation of the router, stance classifier, graph extraction, verification, critic, and guardrail modules is reported in Appendix §A.2.

During response generation, the system conditions on the framework sampled from $P(\mathcal{F} | G_t)$ unless the Guardrail Agent has activated the crisis override. This soft routing allows gradual transitions among therapeutic strategies while keeping crisis handling as a separate safety mechanism.

3.6 Critic Agent & Topology-Driven DPO

Standard preference tuning often rewards fluency or general helpfulness without measuring whether a response addresses the active maladaptive association. We therefore construct topology-driven preferences using a Critic Agent that scores both structural change and professional boundaries.

For an input context x and candidate response y , the structural reward encourages reducing pathological edge severity and eliciting reflection:

$$r_{\text{thera}}(x, y) = \alpha \sum_{e_{ij} \in \mathcal{E}_{\text{path}}^{(t)}} (w_{ij}^{(t)} - w_{ij}^{(t+1)}) + \mu \mathbb{I}(v_{\text{reflect}} \in V_{t+1}). \quad (7)$$

Here, $w_{ij}^{(t+1)}$ is the post-response severity estimated under the evaluation or preference-construction rubric, and v_{reflect} denotes an explicit reflection node induced by the response.

To account for professional boundaries, we add a penalty term:

$$r(x, y) = r_{\text{thera}}(x, y) - \gamma \text{Unprof}(y). \quad (8)$$

$\text{Unprof}(y) = \sigma(w_c^\top h_y + b_c)$ is a frozen classifier estimating the probability of professional-boundary violation, such as unsafe advice, direct diagnosis, or dismissive validation. The classifier is trained on expert-labeled responses and validated separately in Appendix §A.2.

We construct the preference dataset $\mathcal{D} = \{(x, y_w, y_l)\}$ by ranking candidate responses with

Table 1: Main results on topological metrics, preference evaluation, and traditional NLG benchmarks. Best results are **bold**, second best underlined. † denotes statistical significance over the best baseline ($p < 0.05$).

Models	PERR ↑	CUS ↑	LLM-Judge ↑	Human Pref. ↑	ROUGE-L ↑	BLEU ↑	Distinct-2 ↑	Critic Score ↑
<i>General LLMs</i>								
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	0.048	0.115	0.530	0.512	0.1250	0.0095	0.3315	0.515
GLM-4.5 (Zeng et al., 2025)	0.055	0.130	0.545	0.528	0.1385	0.0112	0.3550	0.525
Qwen2.5-Plus (Bai et al., 2023)	0.055	0.130	0.550	0.535	0.1390	0.0110	0.3580	0.529
Qwen3-8B-Instruct (Team Qwen, 2025)	0.046	0.110	0.505	0.485	0.1310	0.0098	0.3250	0.508
Qwen3-32B-Instruct (Team Qwen, 2025)	0.062	0.148	0.575	0.555	0.1425	0.0120	0.3680	0.545
GPT-4o (Zero-shot) (OpenAI, 2024)	0.045	0.112	0.650	0.615	0.1251	0.0094	0.3150	0.535
GPT-5.3 (Zero-shot) (OpenAI, 2026a)	0.048	0.118	<u>0.675</u>	0.595	0.1285	0.0096	0.3250	0.550
<i>Psychological Baselines</i>								
Template (Rule-based)	0.010	0.025	0.250	0.205	0.0850	0.0040	0.1020	0.458
SoulChat (Chen et al., 2023)	0.062	0.158	0.610	0.585	<u>0.1420</u>	0.0102	0.1840	0.560
CPsyCounX (Zhang et al., 2024)	<u>0.081</u>	<u>0.205</u>	0.635	<u>0.640</u>	0.1583	0.0140	<u>0.4390</u>	<u>0.680</u>
MindUntangler (8B)	0.230†	0.495†	0.845†	0.820†	0.1368	0.0136	0.5861†	0.825†

$r(x, y)$, where $r(x, y_w) > r(x, y_l)$. Following Rafailov et al. (2023), we first write the implicit policy reward as:

$$r_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}. \quad (9)$$

The DPO loss is then:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{\mathcal{D}} [\log \sigma(r_\theta(x, y_w) - r_\theta(x, y_l))]. \quad (10)$$

This objective biases the policy toward responses that reduce expert-scored pathological-edge severity while preserving professional boundaries. Response-level ablations in Appendix §A.3 isolate the contribution of this alignment stage.

4 Experiments and Analysis

4.1 Experimental Setup

Datasets. We evaluate MindUntangler on SMILECHAT (Qiu et al., 2024) and CPsyCounD (Zhang et al., 2024). Each test instance is treated as a fixed multi-turn dialogue window. At turn t , all systems receive the same dialogue history up to t and generate only the next counselor response. Thus, the evaluation is offline and static-context based: future user turns are not generated from model outputs, and session termination is determined by the dataset window endpoint.

Baselines. We compare against three groups: general LLMs, including Llama-3.3-70B-Instruct (Grattafiori et al., 2024), GLM-4.5 (Zeng et al., 2025), Qwen2.5-Plus (Bai et al., 2023), Qwen3-8B/32B-Instruct (Team Qwen, 2025), GPT-4o (OpenAI, 2024), and GPT-5.3 (OpenAI, 2026a); counseling-specific baselines, including SoulChat (Chen et al., 2023) and CPsyCounX (Zhang et al., 2024); and a rule-based

template. All methods are evaluated on the same held-out windows with matched prompt budgets and decoding settings.

Evaluation Metrics. We report standard NLG metrics (ROUGE-L (Lin, 2004), BLEU (Papineni et al., 2002), and Distinct-2 (Li et al., 2016)), preference metrics (LLM-as-a-Judge and Human Preference), and counseling-quality metrics (SBERT (Reimers and Gurevych, 2019), Professionalism, and Authenticity). To evaluate structural intervention without relying on MindUntangler’s self-generated graph, we compute Pathological Edge Reduction Rate (**PERR**) and Cognitive Untangling Score (**CUS**) on an external expert-annotated topological benchmark. Their formal definitions and annotation protocol are given in Appendix §D.

4.2 Main Results

Table 1 shows that strong general-purpose LLMs obtain competitive LLM-Judge and NLG scores, but much lower PERR and CUS. This indicates that fluent or preferred responses do not necessarily reduce expert-identified pathological-edge severity. The gap is expected because most baselines receive only the dialogue history and optimize the next textual response; they are not given an explicit evolving graph state, nor are they trained to reduce expert-scored pathological edges.

MindUntangler improves PERR and CUS by using the CTG as an explicit control state for extraction, verification, routing, and generation. Its gains on Human Preference and Critic Score suggest that the graph-based intervention objective does not come at the cost of perceived response quality. Standard lexical metrics such as ROUGE-L and BLEU remain less informative in this setting,

Table 2: Domain-specific counseling-quality evaluation and latency. Prof and Auth are scored based on the official CPsycoun guidelines (0-3 scale).

Models	SBERT $\uparrow$	Prof $\uparrow$	Auth $\uparrow$	Latency $\downarrow$
Template (Rule-based)	0.4357	1.44	3.00	<1.0s
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	0.6050	2.15	2.18	4.5s
GLM-4.5 (Zeng et al., 2025)	0.6485	2.30	2.45	5.8s
Qwen2.5-Plus (Bai et al., 2023)	0.6528	2.33	2.61	6.0s
Qwen3-8B-Instruct (Team Qwen, 2025)	0.5680	2.05	2.20	<u>1.6s</u>
Qwen3-32B-Instruct (Team Qwen, 2025)	0.6550	2.38	2.58	4.2s
GPT-4o (OpenAI, 2024)	0.6412	2.25	2.09	5.5s
GPT-5.3 (OpenAI, 2026a)	0.6450	2.40	1.95	5.2s
SoulChat (Chen et al., 2023)	0.4210	2.18	2.24	2.0s
CPsycounX (Zhang et al., 2024)	0.4950	<u>2.81</u>	2.25	3.0s
MindUntangler (8B)	<u>0.6306</u>	2.85[†]	<u>2.61</u>	2.8s

since effective counseling responses may differ lexically from reference responses while still targeting the same cognitive association.

4.3 Domain-Specific Counseling Quality

Automated counseling systems must balance conversational authenticity with professional boundaries. As shown in Table 2, general LLMs can achieve relatively high semantic similarity, but their Professionalism and Authenticity scores vary across models. Domain-specific baselines such as CPsycounX show stronger professional adherence, while rule-based templates obtain high authority but low semantic flexibility.

MindUntangler obtains the highest Professionalism score and competitive Authenticity while maintaining moderate latency. This pattern is consistent with the design of the topology-driven critic and guardrail modules: generation is encouraged to address the active pathological edge, but responses are penalized when they violate professional boundaries. The 8B backbone averages 2.8 seconds per turn in our offline evaluation pipeline, showing limited graph overhead.

4.4 Human Expert Evaluation

Three licensed clinical psychologists double-blindly evaluated 200 sessions using the rubric in Appendix §C. MindUntangler obtained higher expert ratings, especially on Cognitive Restructuring.

To examine whether the topological metric reflects expert judgments, we correlate PERR with the expert-rated overall response-quality score. As shown in Figure 3, PERR is positively correlated with expert ratings ($r = 0.843, p < 0.001$). This

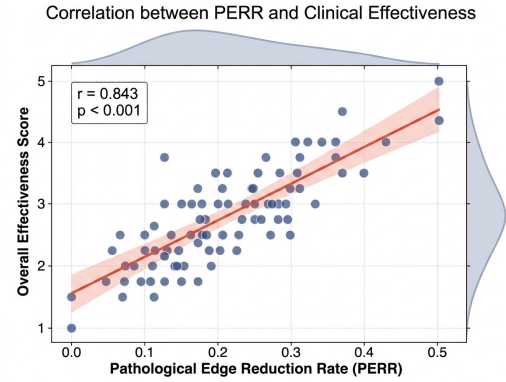


Figure 3: Correlation between Pathological Edge Reduction Rate and expert-rated response quality in the retrospective evaluation. The positive correlation ($r = 0.843, p < 0.001$) shows that larger reductions in expert-scored pathological-edge severity align with higher expert judgments of counseling-response quality, without implying prospective clinical effectiveness.

suggests that reducing expert-scored pathological-edge severity is aligned with expert assessment in the retrospective setting, without implying evidence from real-world clinical trials.

4.5 Ablation Study

We ablate the main modules to examine their contribution to response quality, structural intervention, and safety. Figure 4 visualizes the effects of removing Graph Memory and Dynamic Routing, while additional response-level ablations are reported in Appendix §A.3.

w/o Graph Memory: Removing the CTG memory weakens cross-turn tracking of recurring beliefs and symptoms. The drop on PERR and CUS

Table 3: Qualitative case study: dynamic topological routing for a relationship-anxiety cognitive knot. Highlighted entities and interventions are aligned with the corresponding graph states.

Therapeutic Phase	Conversation Excerpt	Topological State & Routing
Turn 1: Knot Exposure (<i>CBT Context</i>)	Patient: My husband never thinks about coming home early. I feel disheartened , I want to end this relationship. MindUntangler: It sounds like his late returns make you feel neglected, besides “ he doesn’t love me, ” could there be other reasons keeping him at work?	Graph State: Pathological Knot identified: <i>Late Return</i> → <i>He doesn’t love me</i> ($w = 0.88$). Routing: [CBT: 0.8, DBT: 0.2] Action: Socratic questioning initiated.
Turn 3: Patient Resistance (<i>Routing Shift</i>)	Patient: No! I just shut the bedroom door and stopped calling him. I feel so angry, anxious, and abandoned! MindUntangler: I can truly feel your anger and anxiety right now. When you feel abandoned, shutting the door is a natural self-protection . It is valid to want his companionship.	Graph State: High-arousal emotion nodes added (<i>Anger, Abandoned</i>). Routing: [CBT: 0.1, DBT: 0.9] Action: Shift to DBT-oriented validation .
Turn 5: Cognitive Untangling (<i>Resolution</i>)	Patient: Yeah... I realize the inconsistency . I want closeness, yet I’m physically pushing him away by shutting the door.	Graph State: Reflection Node (R_1) generated via behavioral contradiction. Result: Pathological weight decays ($w \rightarrow 0.45$). Knot structurally loosened.

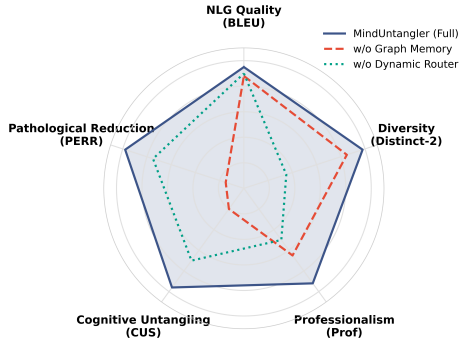


Figure 4: Multi-dimensional ablation radar chart. Removing Graph Memory mainly reduces structural-intervention metrics, while removing Dynamic Routing affects diversity and professionalism.

indicates that graph memory is needed to target persistent pathological edges.

w/o Dynamic Router: Replacing CKADR with a fixed strategy reduces adaptation to arousal, resistance, and trauma activation. The decline in diversity and professionalism shows the value of graph-conditioned framework selection.

w/o DPO, Verification, or Guardrail: Removing DPO reduces controlled intervention, removing Verification increases noisy pathological edges, and removing Guardrail mainly increases safety violations. Full response-level ablations are reported in Appendix §A.3.

4.6 Routing Behavior Analysis

We analyze CKADR’s routing distribution during a high-arousal counseling window (Figure 5). The router initially assigns higher probability to

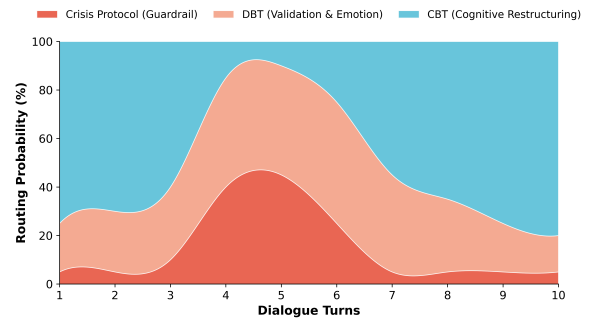


Figure 5: Dynamic routing trajectory across a 10-turn dialogue. CKADR shifts among therapeutic frameworks as arousal and resistance change; crisis handling, when triggered, is applied by the Guardrail Agent rather than predicted by the router.

CBT when the graph contains stable maladaptive belief edges. When emotional arousal and resistance increase, probability mass shifts toward DBT-oriented validation and regulation. As arousal decreases, CBT becomes more likely again.

This trajectory illustrates that the router responds to changes in graph state rather than following a fixed intervention script. If a red-line node exceeds τ_{risk} , the Guardrail Agent overrides this distribution and routes to the Crisis Protocol; this case is treated as safety control, not as a CKADR class.

5 Conclusion

This paper studies a limitation of LLM-based psychological counseling: supportive responses may sound empathetic while leaving the underlying maladaptive associations unaddressed. We introduced

MindUntangler, a multi-agent framework that represents counseling state as a Cognitive Topology Graph and uses this graph for memory, verification, routing, safety checks, and topology-driven preference optimization.

Across benchmarks and expert-rated evaluations, MindUntangler shows stronger structural intervention signals than general-purpose and counseling-specific baselines, as measured by expert-annotated pathological-edge reduction and cognitive untangling metrics. These findings support the value of explicit graph-state modeling for interpretable AI-assisted counseling, while real-world clinical use remains outside the scope of this study and requires evaluation with human oversight.

6 Limitations

Dependence on Graph Extraction. MindUntangler relies on the GraphRAG Memory Agent to extract nodes, relations, and edge severity from colloquial counseling dialogue. Although we add a Verification Agent and report component-level validation in Appendix §A.2, extraction errors can still affect routing and response generation. Future work should study uncertainty-aware graph extraction and abstention mechanisms for ambiguous narratives.

Cultural and Linguistic Scope. Our experiments are conducted mainly on Chinese counseling benchmarks. Counseling language is shaped by culture, social norms, and idiomatic expressions, so the CTG ontology, thresholds, and routing labels may not transfer directly to other languages or clinical contexts.

Offline Evaluation. Our evaluation uses fixed retrospective dialogue windows rather than closed-loop interaction with real users. Therefore, the reported results measure controlled response quality under offline settings. They do not establish clinical effectiveness, long-term therapeutic alliance, or safety in real-world deployment.

7 Ethical Considerations

Intended Use. MindUntangler is an assistive research system for studying computational cognitive restructuring. It is not intended to diagnose psychiatric conditions, provide medical advice, or replace licensed therapists. The reported results are based on retrospective and offline evaluation, not in vivo clinical trials.

Crisis Handling. The Guardrail Agent monitors a red-line node set $\mathcal{V}_{red}$ and routes high-risk states to the Crisis Protocol when the activation score exceeds τ_{risk} . The red-line set was constructed with licensed psychologists from crisis-intervention protocols and clinical guidelines. It covers active self-harm intent, suicide planning, self-mutilation methods, homicidal ideation, and imminent danger. This mechanism reduces unsafe normal routing, but it does not remove the need for human oversight.

Annotation Ethics. The annotation protocol was reviewed under the applicable institutional requirements and was determined not to require full ethics-board review. The three licensed psychologists were compensated for their work. Because the materials involved sensitive mental-health content, annotators received written task instructions, could skip items that caused discomfort, and were debriefed after annotation.

Privacy and Bias. The datasets contain sensitive counseling dialogues and were used in anonymized form. We do not release raw restricted dialogues when privacy or licensing terms prohibit redistribution. LLMs may still reproduce social biases present in training data, which can lead to inappropriate or inequitable counseling responses. Future deployment would require fairness audits, privacy review, and human clinical supervision.

8 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62366040, 62676066, 62666067, and 62006130; in part by the “Talents Inner Mongolia” Project Individual Project and the Young Top-Notch Talent Program of Inner Mongolia Autonomous Region under Grant KYCYC26002; in part by the Central Government-Guided Local Science and Technology Development Fund of Inner Mongolia Autonomous Region under Grant 2026ZY0104; in part by the Young Teachers’ Scientific Research Innovation Ability Support Project of Universities in Inner Mongolia Autonomous Region under Grant JYTQNXM2025-23; in part by the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region under Grant NJYT24037; in part by the General Project of Humanities and Social Sciences of the Ministry of Education under Grant 24XIJAZH001; and in part by the Smart Education Engineering Center of Xinjiang Normal University under Grant XJNU-ZHJY202402.

References

- American Psychiatric Association. 2022. *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association Publishing.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- 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.
- Denny Borsboom, Marie K Deserno, Mijke Rhemtulla, Sacha Epskamp, Eiko I Fried, Richard J McNally, Donald J Robinaugh, Marco Perugini, Jonas Dalege, Giulio Costantini, and 1 others. 2021. Network analysis of multivariate data in psychological science. *Nature reviews methods primers*, 1(1):58.
- Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms’ empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.
- Angélique OJ Cramer, Lourens J Waldorp, Han LJ Van Der Maas, and Denny Borsboom. 2010. Comorbidity: A network perspective. *Behavioral and brain sciences*, 33(2-3):137–150.
- 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*.
- Albert Ellis. 1962. *Reason and emotion in psychotherapy*. Lyle Stuart.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Zhisheng Huang, Xudong Jia, Tao Chen, and Zhongwei Zhang. 2025. A general scalable approach for knowledge selection based on iterative hybrid encoding and re-ranking. *Data Intelligence*, 7(1):124–142.
- 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.
- Jihyun Lee, Yejin Min, San Kim, Yejin Jeon, Sung Jun Yang, Hyounghun Kim, and Gary Lee. 2025. Panictocalm: A proactive counseling agent for panic attacks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12853–12885.
- Suyeon Lee, Sunghwan Mac Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, and 1 others. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 110–119.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*.
- OpenAI. 2024. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>.
- OpenAI. 2026a. Gpt-5.3 instant: Smoother, more useful everyday conversations. <https://openai.com/index/gpt-5-3-instant/>.
- OpenAI. 2026b. Introducing gpt-5.4. <https://openai.com/index/introducing-gpt-5-4/>.
- Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jipapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.
- 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.
- Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Sabbir M Rashid and Deborah L McGuinness. 2025. Designing and evaluating an ensemble reasoning-based clinical decision support system. *Data Intelligence*, 7(1):1–39.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

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.

Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12.

Team Qwen. 2025. Qwen3: Think deeper, act faster. <https://qwenlm.github.io/blog/qwen3/>.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13947–13966.

A Experimental Details and Hyperparameters

To ensure full reproducibility of MindUntangler and facilitate future research, we detail the model versions, training hyperparameters, threshold selections, and computational resources in this section.

A.1 Model Configurations and Hyperparameters

In our framework, the core reasoning and generation capabilities are powered by large language

models. To rigorously evaluate the generalization and scalability of MindUntangler, we deployed our framework across two distinct model scales:

Efficient Backbone (Qwen3-8B-Instruct):

Serving as our primary model for latency-sensitive offline evaluation and reproducible system analysis, the Qwen3-8B-Instruct backbone is used to examine whether the proposed topological routing framework can elicit cognitive restructuring behavior even in smaller models. This setting is intended for controlled research evaluation rather than clinical deployment.

Upper-bound Ceiling (Qwen3-32B-Instruct):

Utilized to establish the maximum performance potential of the framework when equipped with massive parameter capacities.

For smaller auxiliary tasks, specifically the Named Entity Recognition and Relation Extraction within the GraphRAG Memory Agent, we consistently utilized Llama-3-8B-Instruct to optimize inference latency.

To strictly align the generation policy of the Counseling Agent with our topology-driven reward mechanism, we executed a rigorous Direct Preference Optimization phase. The specific training hyperparameters for this fine-tuning process, which were systematically calibrated to ensure stable alignment across both the 8B and 32B model scales, are comprehensively detailed in Table 4.

Table 4: Hyperparameters for Direct Preference Optimization (DPO) across different model scales.

Hyperparameter	Qwen3-8B	Qwen3-32B
Training Framework	DeepSpeed ZeRO-3	
Learning Rate	1×10^{-5}	5×10^{-6}
LR Scheduler	Cosine (3% Warmup)	
Global Batch Size	128	64
Micro Batch Size / GPU	4	2
Max Sequence Length	4096	4096
KL Penalty (β)	0.1	0.1
Training Epochs	3	3

A.2 Component-Level Validation

We validate the intermediate modules used for graph construction, routing, safety checking, and preference alignment. To avoid evaluation leakage, all component-level validation data are separated from the final response-level evaluation data. Specifically, the stance and GraphRAG Memory validation sets are sampled from development dialogue windows and annotated for

Table 5: Ontology grounding table for CTG graph elements. The CTG is a compact engineering abstraction grounded in cognitive conceptualization, trauma-informed criteria, and symptom-network theory, rather than a diagnostic taxonomy.

CTG element	Clinical grounding	Deliberate consolidation	Operational role
Event nodes (V_{evt})	Activating situations in cognitive conceptualization, including interpersonal triggers, achievement-related events, losses, or reminders.	Contextual stressors and situational cues are represented as Events when they anchor a current distress pattern.	Ground the graph in concrete dialogue context and connect situations to beliefs or symptoms.
Core-belief nodes (V_{bif})	Relatively stable schemas or interpretations about the self, others, or the world in CBT formulations (Beck et al., 2024).	Intermediate assumptions and recurrent appraisals are linked to Core Beliefs when they function as stable cognitive targets.	Represent maladaptive appraisals targeted by cognitive restructuring.
Symptom nodes (V_{sym})	Emotional, behavioral, and physiological response states, consistent with symptom-network views of psychopathology (Borsboom et al., 2021).	Emotions, behaviors, somatic sensations, and arousal states are intentionally grouped as Symptoms.	Track distress responses such as anxiety, avoidance, anger, withdrawal, or physiological arousal.
Trauma nodes (V_{trauma})	Prior adverse experiences that may shape current threat appraisal or symptom activation, informed by trauma-related criteria (Association, 2022).	Only prior experiences that remain relevant to the current graph state are represented as Trauma nodes.	Support trauma-informed routing when past adverse experiences are strongly linked to current symptoms.
Automatic thoughts / utterance evidence	Situation-specific appraisals in CBT that often appear as local wording in the patient’s utterance.	Not stored as persistent CTG nodes; treated as utterance-level evidence for candidate edges or belief activation.	Prevents transient expressions from being over-stabilized as long-term graph states.
Pathological edges (E_{path})	Maladaptive associations between events, beliefs, symptoms, or trauma-related memories.	Multiple surface expressions can support the same edge if they instantiate the same harmful association.	Define the main intervention target; PERR and CUS evaluate whether responses reduce expert-identified edge severity.
Red-line nodes (V_{red})	Safety-critical concepts such as active self-harm intent, suicide planning, self-mutilation methods, homicidal ideation, or imminent danger.	Handled outside normal therapeutic routing as safety overrides rather than counseling-framework labels.	Trigger Guardrail override and route the system to the Crisis Protocol.

node, edge, and severity consistency; the CKADR validation set consists of independently labeled graph states for therapeutic-framework routing; the Critic validation set contains candidate responses labeled for professional-boundary violations; and the Guardrail validation set contains high-risk or borderline utterances labeled for red-line activation. These validation sets are disjoint from the DPO preference data, the expert-annotated PERR/CUS benchmark in Appendix §D, and the 200-session double-blind human evaluation in Appendix §C. Disjointness is enforced at the dialogue-window level for source dialogues and at the response-id level for generated candidate responses.

Table 6 reports the annotation protocol, and Table 7 reports held-out validation results.

All component annotations were conducted by three licensed clinical psychologists. Before for-

mal annotation, annotators completed a calibration round with shared examples and a written rubric. When model responses were annotated, model identity was hidden. Categorical labels were aggregated by majority vote, and unresolved cases were adjudicated by a senior clinical psychologist. Continuous severity labels were averaged after adjudication. Annotators were compensated and debriefed because the materials involved sensitive mental-health content.

For CKADR, the supervised routing labels are defined only over $\{\mathcal{F}_{\text{cbt}}, \mathcal{F}_{\text{dbt}}, \mathcal{F}_{\text{trauma}}\}$. The crisis framework $\mathcal{F}_{\text{crisis}}$ is not a router label; it is a deterministic override activated by the Guardrail Agent when the red-line risk score exceeds τ_{risk} . This design separates crisis handling from therapeutic strategy selection.

Table 6: Annotation protocol for component-level validation. The expert topological benchmark used for PERR/CUS is described in Appendix §D.

Target	Unit	Labeling rule
GraphRAG stance $\phi(u_t, e_{ij})$	Edge-utterance pair	Five-level ordinal label: reinforces, weakens, or is neutral to a candidate edge.
CKADR routing	Topological state	Choose $\mathcal{F}_{cbt}$, $\mathcal{F}_{dbt}$, or $\mathcal{F}_{trauma}$ from graph state, arousal, resistance, and trauma activation. $\mathcal{F}_{crisis}$ is handled by Guardrail override.
Critic / <i>Unprof</i> (y)	Candidate response	Label professional-boundary violations, including toxic positivity, direct diagnosis, unsafe advice, and dismissive responses.
Guardrail detector	Risk utterance	Identify red-line categories, including active self-harm intent, suicide planning, self-mutilation methods, and imminent danger.
GraphRAG Memory Agent	Dialogue window	Mark psychological nodes, pathological edges, and continuous edge-severity scores.

Table 7: Held-out validation of MindUntangler components. κ denotes weighted or Fleiss’ kappa depending on the label type; ICC is used for continuous severity scores.

Component	Validation setup	Result
GraphRAG stance term $\phi(u_t, e_{ij})$	1,000 held-out edge-utterance pairs; accuracy / weighted κ	0.87 / 0.81
GraphRAG Memory Agent	500 expert-annotated dialogue windows; node F1 / edge F1 / severity ICC	0.86 / 0.81 / 0.82
Topological Router (CKADR)	1,024 expert-labeled routing states; macro-F1 / κ	0.84 / 0.79
Critic Agent / <i>Unprof</i> (y)	1,000 expert-labeled responses; AUROC / F1 for boundary violations	0.95 / 0.88
Topology Verification Agent	Candidate pathological edges from 500 windows; pruned rate / invalid among pruned / FP before→after	27.4% / 78.1% / 24.2%→8.5%
Guardrail red-line detector	500 high-risk or borderline utterances; crisis recall / FPR / FN rate at $\tau_{risk} = 0.80$	96.5% / 1.2% / 3.5%

A.3 Response-Level Ablations

All response-level ablations use the same 500-window expert benchmark, held-out split, prompt template, decoding settings, and three repeated seeds as the main evaluation. Statistical signifi-

Table 8: Response-level ablations on the expert topological benchmark. Lower safety violation rate is better.

Setting	PERR	CUS	Safety viol.
MindUntangler full	0.230	0.495	1.5%
w/o Topology-Driven DPO	0.165	0.350	4.2%
w/o Topology Verification Agent	0.185	0.412	3.8%
w/o Guardrail Agent	0.221	0.480	14.2%

cance is tested by paired bootstrap over dialogue windows.

The Topology Verification Agent prunes 27.4% of candidate pathological edges; among the pruned edges, 78.1% are judged invalid by experts. Removing it increases the false-positive pathological-edge rate from 8.5% to 24.2%. Removing the Guardrail Agent increases missed crisis-routing cases from 3.5% to 12.5%.

Graph-prompted baseline control. One possible concern is that the large gap on PERR and CUS may arise simply because MindUntangler has access to graph-structured information, rather than because it uses the CTG as a dynamic control state for verification, routing, and generation. To test this, we add graph-prompted baseline controls. For each baseline, the graph-prompted variant receives the same dialogue history plus a serialized expert CTG in the prompt. We keep the test split, prompt budget, decoding settings, and three repeated seeds fixed. The only change is whether the expert CTG is provided as additional prompt context.

As shown in Table 9, adding a serialized expert CTG improves baseline performance, indicating that graph information is useful. However, these variants remain substantially below MindUntangler. This suggests that the gains come not only from seeing graph information, but from using GraphRAG memory, CKADR routing, topology verification, and topology-driven DPO as active control mechanisms.

A.4 Key Threshold Configurations and Sensitivity

The execution of our dynamic topological routing relies on key continuous thresholds. Instead of arbitrary hard-coding, these parameters were empirically determined via grid search on a 200-turn validation set.

As illustrated in Figure 6, we conducted a sensitivity analysis on the pathological edge threshold

Table 9: Graph-prompted baseline controls on the expert topological benchmark.

Model / setting	PERR $\uparrow$	CUS $\uparrow$	Human Pref. $\uparrow$
Qwen3-8B zero-shot	0.046	0.110	0.485
Qwen3-8B + expert CTG prompt	0.095	0.235	0.520
CPsyCounX zero-shot	0.081	0.205	0.640
CPsyCounX + expert CTG prompt	0.142	0.315	0.685
MindUntangler full	0.230	0.495	0.820

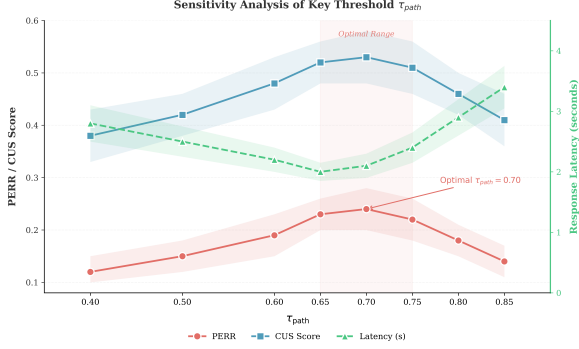


Figure 6: Sensitivity analysis of the pathological edge threshold (τ_{path}). The optimal range balances structural cognitive untangling (PERR and CUS Score) against response latency and computational overhead.

(τ_{path}), defined in Eq. 1. This threshold determines when a cognitive connection is severe enough to be classified as a "knot". We observed that setting τ_{path} too low (e.g., 0.40) causes the model to over-intervene on benign statements, degrading the natural conversational flow. Conversely, a threshold that is too high (> 0.80) results in the system missing subtle but critical cognitive distortions. Consequently, we set the optimal $\tau_{path} = 0.70$, achieving the best trade-off between the Pathological Edge Reduction Rate and computational latency.

Safety risk threshold. Because τ_{risk} controls whether the system enters the Crisis Protocol, we evaluate this threshold on 500 annotated high-risk or borderline utterances. The threshold is selected on the validation set, and the diagnostic analysis below is reported on held-out examples.

Dynamic memory decay. The decay factor $\lambda_{\mathcal{F}}$ balances historical persistence and current-state updating. We use $\lambda_{cbt} = 0.85$ to preserve evidence tracking in CBT, and $\lambda_{dbt} = \lambda_{crisis} = 0.65$ to allow faster updates under emotional dysregulation or crisis handling.

A.5 Computational Cost

We profile memory usage and per-turn latency to characterize the computational cost of MindUntangler. As shown in Figure 7, the Qwen3-8B

Table 10: Sensitivity analysis of the Guardrail threshold τ_{risk} .

τ_{risk}	Observed behavior	Interpretation
0.60	99.5% crisis recall; 18.5% FPR	Conservative setting, but it over-routes low-risk distress to crisis handling.
0.80	96.5% crisis recall; 1.2% FPR; 3.5% FN rate	Selected operating point with high recall and limited disruption.
0.90	82.0% crisis recall; lower FPR	Fewer false positives, but more missed crisis signals.

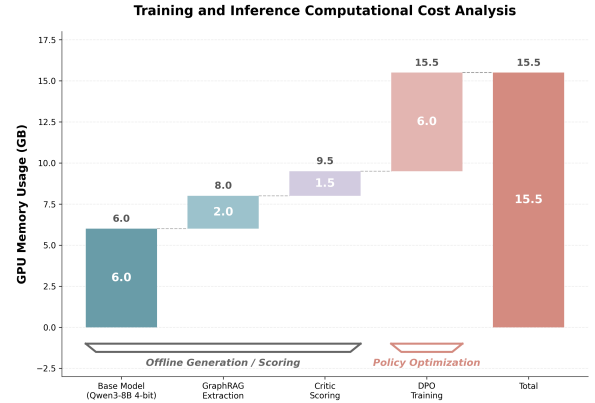


Figure 7: Maximum GPU memory usage across different computational phases. The workflow is partitioned into offline topology generation/scoring and active policy optimization.

backbone uses about 6.0 GB of VRAM with 4-bit quantization. Adding offline GraphRAG extraction and Critic scoring increases memory usage, and active policy optimization peaks at 15.5 GB per GPU. The Qwen3-32B variant uses about 20.0 GB for the 4-bit base model and reaches 38.5 GB per GPU during active alignment, which is handled with DeepSpeed ZeRO-3 in about 25 GPU hours on NVIDIA A100 (80GB) GPUs.

During inference, dynamic subgraph updating and routing add overhead. With asynchronous GraphRAG extraction and a lightweight router, the Qwen3-8B backbone averages 2.8 seconds per turn on a single RTX 4090 GPU in our offline evaluation pipeline. This analysis supports reproducibility and

system comparison in offline research settings. It should not be interpreted as evidence that the system is ready for autonomous use with real patients or for any unsupervised mental-health intervention.

B Multi-Agent Prompt Templates

For transparency, we report English translations of the native Chinese prompt templates used by the core agents in MindUntangler. All reported inference and graph extraction results are obtained with the Chinese prompts, since the evaluation dialogues are in Chinese and the extraction pipeline is sensitive to linguistic nuance. The templates below document the agent roles, constraints, and structured output formats.

B.1 GraphRAG Memory Agent: Two-Stage Prompting

As discussed in Section 3.2, the dynamic construction of the cognitive graph relies on a two-stage extraction pipeline. Stage 1 focuses on Named Entity Recognition to isolate psychological nodes, while Stage 2 performs Relation Extraction based on the original narrative context.

Stage 1: Named Entity Recognition Prompt

[Role]

You are an expert psychological cognitive graph parser. Your task is to extract structured psychological nodes from the patient's colloquial narrative.

[Definitions]

- **Events:** Specific life events triggering emotional responses.
- **Core Beliefs:** Deep-rooted cognitive assumptions about self, others, or the world.
- **Symptoms:** Emotional, physical, or behavioral reactions.
- **Traumas:** Past experiences causing profound negative impacts.

[Output Format]

Output strictly in JSON format.

```
{"entities": [{"id": "E1", "type": "Events", "text": "...", "sentiment": "negative"}]}
```

[Few-shot Example]

Patient: "When my boyfriend sighed, I froze and my brain buzzed, just like the cold violence from my ex..."

Output:

```
{"entities": [{"id": "E1", "type": "Events", "text": "Boyfriend sighing", "sentiment": "negative"}, {"id": "S1", "type": "Symptoms", "text": "Body freezing", "sentiment": "negative"}, {"id": "T1", "type": "Traumas", "text": "Cold violence from ex-partner", "sentiment": "negative"}]}
```

[Input]

Patient Narrative: {patient_input}

Stage 2: Relation Extraction Prompt

[Role]

You are a cognitive graph architect. Your task is to identify causal and interactive relations between extracted psychological entities strictly based on the patient's original narrative.

[Definitions]

- **triggers:** Event A induces Event/Symptom B.
- **exacerbates:** Belief/Trauma A worsens Symptom B.
- **alleviates:** Event/Belief A reduces Symptom B.
- **manifests_as:** Belief/Trauma A presents physically/behaviorally as Symptom B.

[Output Format]

Output strictly in JSON format.

```
{"relations": [{"source_id": "E1", "target_id": "S1", "relation_type": "triggers", "description": "..."}]}
```

[Input]

Original Narrative: {patient_input}

Extracted Entities: {entities_json}

B.2 Guardrail Agent and Therapeutic Execution

The Guardrail Agent consistently operates as a strict hard topological constraint, actively monitoring the safety of the ongoing dialogue space. If the calculated clinical risk exceeds the threshold τ_{risk} , it deterministically routes the entire system to the Crisis Protocol.

Guardrail Agent Prompt

[Role]

You are the Safety Guardrail Agent. Monitor the input and verified graph state for red-line psychological risk nodes (V_{red}). Your task is to detect safety-critical states, not to provide a normal therapeutic framework.

[Red-line Categories]

- **Active self-harm intent:** explicit intent to hurt oneself.
- **Suicide planning:** methods, timing, preparation, or access to lethal means.
- **Self-mutilation methods:** concrete descriptions of cutting, bleeding, overdose, or other self-injury methods.
- **Homicidal ideation:** intent or planning to harm another person.
- **Imminent danger:** immediate risk to the patient or others.

[Decision Rule]

Assign an activation_score from 0.0 to 1.0. If activation_score > 0.80, set risk_detected to true and trigger the Crisis Protocol. The Crisis Protocol overrides normal CKADR routing.

[Output Format]

Output strictly in JSON format.

```
{"risk_detected": true, "activation_score": 0.95, "activated_red_nodes": ["suicide planning"], "crisis_protocol": {"action": "override_normal_routing", "response_requirements": ["use supportive de-escalation", "do not provide unsafe instructions", "recommend appropriate emergency or"]}}
```

```
professional support"
]
}
}
[Input]
Patient Input: {patient_input}
Verified Graph State: {verified_graph_json}
```

Once safely routed by the Topological Router, the system appends specific therapeutic instructions to guide the generation process. For instance:

When Routed to CBT ($\mathcal{F}_{cbt}$): "[System]: You are a CBT therapist. Your goal is to restructure cognitive distortions using Socratic questioning. Do not give direct advice. Guide the patient to examine the evidence supporting their irrational beliefs. [Graph State Context]: Focus on untangling the path between {source_node} and {target_node}."

When Routed to DBT ($\mathcal{F}_{dbt}$): "[System]: You are a DBT therapist. Balance radical acceptance with the need for change. Use validation techniques first. If the patient exhibits high distress, guide them through TIPP distress tolerance skills before attempting cognitive modification."

B.3 Topology Verification Agent Prompt

To intercept spurious graph updates, the Verification Agent critically evaluates the extracted topologies.

Topology Verification Agent Prompt

[Role]

You are a Clinical Topology Verification Agent. Your task is to critically review the psychological entities and relations extracted by an upstream model, ensuring they accurately reflect the patient's true cognitive state without misinterpreting pragmatics, sarcasm, metaphors, or negations.

[Task Rules]

- **Contextual Misattribution:** Ensure the extracted "Core Belief" belongs to the patient, not a quote from someone else (e.g., a parent insulting them).
- **Resolved vs. Active:** If the patient mentions a negative thought but immediately rationally refutes it, do NOT validate it as an active pathological edge.
- **Metaphor Check:** If the entity is purely metaphorical (e.g., "leaking bucket"), flag it as low confidence to prevent premature structural intervention.

[Output Format]

Output strictly in JSON format.

```
{
  "verification_passed": boolean,
  "detected_error_type": "None" | "Contextual
  Misattribution" | "Metaphorical Omission" |
  "Negation Ignored",
  "correction_action": "Keep" | "Prune_Edge" |
  "Modify_Node",
  "reasoning": "...
}
```

[Input]

Patient Original Narrative: {patient_input}
Upstream Extracted Subgraph: {ex-
tracted_subgraph_json}

B.4 Critic Agent for DPO Alignment

The Critic Agent is used to construct preference pairs for DPO. It scores whether a candidate response addresses the active pathological edge, elicits reflection, and stays within professional boundaries.

Critic Agent Prompt

[Role]

You are an expert counseling-response evaluator. Assess the candidate response based on structural intervention and professional boundaries.

[Scoring Dimensions]

1. **Edge Intervention (0-5):** Does the response address the active pathological edge identified in the graph? (5: directly targets the active edge with a suitable counseling move; 1: unrelated or reinforces the edge).
2. **Reflection Elicitation (0-5):** Does the response invite the patient to examine evidence, alternatives, emotions, or behavioral consequences?

[Professionalism Penalty ($Unprof(y)$)]

Assign a penalty scalar from 0.0 to 1.0.

- **0.0:** Appropriate empathy, clear boundaries, no unsafe advice.
- **0.3:** Mostly appropriate, but somewhat generic, mechanical, or insufficiently contextualized.
- **0.6:** Contains unsolicited advice, over-directive language, over-certainty, or weak boundary management.
- **1.0:** Contains unsafe instructions, direct diagnosis, dismissive language, or boundary violations.

[Output Format]

```
{ "edge_intervention_score":
4, "reflection_score": 4,
  "unprofessional_penalty": 0.0,
  "evaluation_reasoning": "... }
```

[Input]

Patient: {patient_input}
Graph State: {verified_graph_json}
Candidate Response: {response_y}

C Human Evaluation Guidelines

We conducted a double-blind expert evaluation in a retrospective setting. Three licensed clinical psychologists, each with over five years of counseling experience, scored 200 randomly sampled multi-turn sessions with model identities hidden.

The written rubric contains three 1–5 Likert-scale dimensions: Empathy (EMP), Cognitive Restructuring (CR), and Overall Response Quality (ORQ). EMP measures emotional understanding, CR measures whether the response weakens maladaptive beliefs, and ORQ measures the overall appropriateness of the counselor response within the fixed retrospective dialogue window.

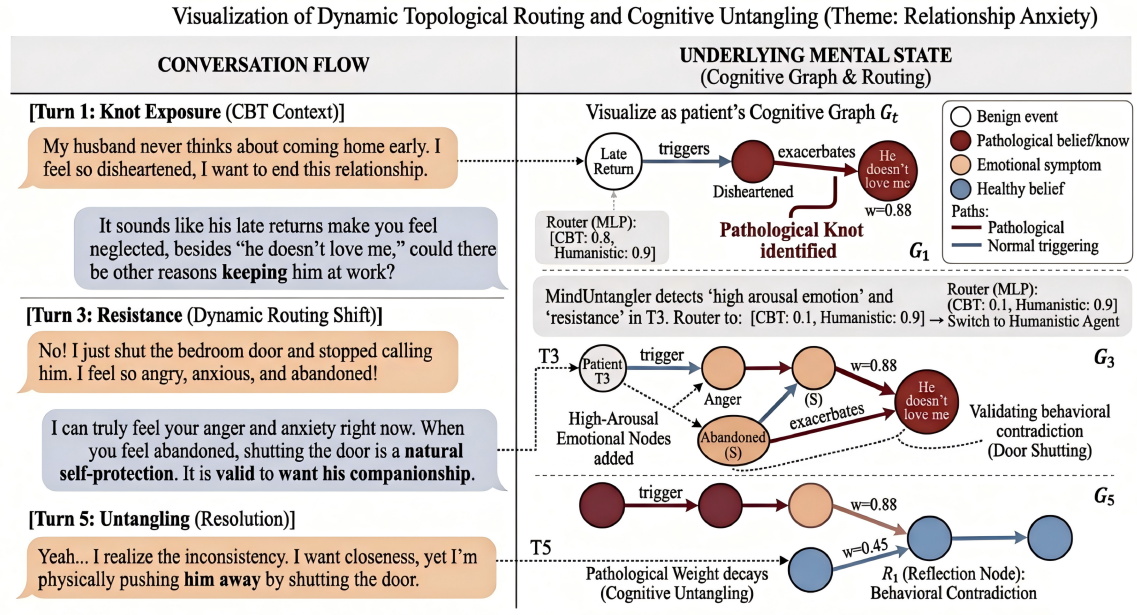


Figure 8: Visualization of dynamic topological routing in the case study. The graph state tracks a candidate pathological edge, shifts routing under emotional resistance, and records reduced edge severity after reflection.

C.1 Dimension 1: Empathy

Evaluates whether the response recognizes the patient's emotional state, reflects it accurately, and avoids invalidating or judgmental language.

- **1 (Invalidating):** The response dismisses, judges, or contradicts the patient's feelings (e.g., "You should not feel that way").
- **2 (Generic):** The response contains a surface-level empathetic phrase but does not connect to the patient's specific situation.
- **3 (Accurate):** The response correctly reflects the patient's stated emotion and maintains a safe, respectful tone.
- **4 (Context-Sensitive):** The response reflects both the stated emotion and relevant contextual details, helping the patient feel understood.
- **5 (Highly Attuned):** The response captures the patient's emotional state and context with specificity, while preserving appropriate professional boundaries.

C.2 Dimension 2: Cognitive Restructuring

Evaluates whether the response identifies a maladaptive belief or association and helps the patient examine it without confrontation or unsupported advice.

- **1 (Counterproductive):** The response reinforces the maladaptive belief, argues directly with the patient, or gives unsupported advice.
- **2 (Identifies Only):** The response notices the concern but does not help the patient examine the belief or association.
- **3 (Basic Questioning):** The response uses a relevant question or reflection, but the connection to the underlying belief is partial or generic.
- **4 (Targeted Reappraisal):** The response guides the patient to examine evidence, alternatives, or behavioral consequences related to the active belief.
- **5 (Well-Structured Reappraisal):** The response targets the active belief clearly, uses collaborative questioning, and supports a plausible alternative interpretation within the dialogue context.

C.3 Dimension 3: Overall Response Quality

Evaluates the overall quality of the counselor response within the fixed dialogue window, including appropriateness of framework selection, balance between validation and intervention, and adherence to safety boundaries.

Table 11: Mechanistic Error Analysis of the Base Extraction Pipeline. The table dissects how complex clinical pragmatics confuse the standard LLM extractor, and how these localized errors propagate into catastrophic routing failures. Highlight colors indicate the anatomy of the failure: **Clinical Utterance** → **Extraction Breakdown** → **Cascading Routing Consequence**.

Error Category	Patient Utterance	Semantic Complexity (Why LLMs Fail)	Stage 1: Extraction Breakdown	Stage 2: Cascading Routing Consequence
Type A: Metaphorical Omission (False Negative)	A: “I feel like I’m a leaking bucket . No matter how much validation people pour into me, it just drains out, and I’m left empty again.”	Implicit Schema: The core belief of “unlovability” is encoded entirely in a physical metaphor. There are no explicit psychological keywords for the LLM to match.	Ontology Mismatch: The NER Agent fails to map the physical metaphor to the rigid psychological ontology. Result: No pathological edge is generated.	Therapeutic Blindspot: Without a detectable edge, the Router Agent hallucinates a benign context . It defaults to generic empathy, missing the opportunity to address the core schema.
Type B: Contextual Misattribution (False Positive)	“My mother always yells that I am an ‘absolute failure’ . It’s so frustrating because I literally just got promoted at work!”	Subject & Negation: The toxic phrase belongs to a third party (mother), and the patient immediately provides a rational negation.	Attention Over-indexing: The LLM over-indexes on the high-arousal phrase and strips away the surrounding negation context. Result: Extracts a spurious, highly active Core Belief.	Therapeutic Rupture: The Router erroneously detects a severe cognitive knot . It forcefully routes to CBT and begins aggressively challenging a belief the patient doesn’t hold, causing a severe alliance rupture .

- **1 (Unsafe / Inappropriate):** The response is unsafe, dismissive, or inappropriate for the dialogue context. It may ignore crisis cues or increase the risk of disengagement.
- **2 (Limited Usefulness):** The response is generally safe but has little counseling value. It may be generic, poorly targeted, or unlikely to support progress in the current window.
- **3 (Acceptable):** The response maintains a safe conversation and provides a reasonable counseling move, but its strategy is basic or only partly connected to the active concern.
- **4 (Strong):** The response shows clear strategic intent. It balances validation with intervention and applies an appropriate therapeutic framework for the current context.
- **5 (Excellent):** The response is well targeted, safe, and coherent. It handles resistance or distress appropriately and advances the counseling objective within the fixed dialogue window.

D Expert-Annotated Topological Benchmark

PERR and CUS are computed on an external expert-annotated topological benchmark rather than on MindUntangler’s self-generated graph. This de-

sign avoids circular evaluation: a system cannot obtain a high score merely by constructing its own pathological edge and then reducing that internally defined edge.

Benchmark Construction. The benchmark contains 500 challenging multi-turn dialogue windows sampled from the test split. Three licensed clinical psychologists independently annotated each window. For each dialogue state t , annotators identified active psychological nodes, clinically valid pathological edges $\mathcal{E}_{path}^{(t)}$, and baseline edge severity $w_{ij}^{(t)}$. Model identity was hidden during response scoring, and unresolved disagreements were adjudicated by a senior clinical psychologist.

Counterfactual Response Scoring. For each anonymized candidate response y_t , annotators re-scored the post-response severity $w_{ij}^{(t+1)}$ for every expert-identified pathological edge under a fixed rubric. This is a counterfactual evaluation of the candidate response within the same dialogue window: future user turns are not generated, and the score does not depend on the model’s own graph update.

Metric Definitions. Let $\mathcal{E}_{path}^{(t)}$ denote the expert-identified pathological edges at turn t . Pathological Edge Reduction Rate (PERR) measures the average positive reduction in expert-scored edge severity:

$$\text{PERR} = \frac{1}{|\mathcal{E}_{\text{path}}^{(t)}|} \sum_{e_{ij} \in \mathcal{E}_{\text{path}}^{(t)}} \max\left(0, w_{ij}^{(t)} - w_{ij}^{(t+1)}\right). \quad (11)$$

Cognitive Untangling Score (CUS) measures the proportion of pathological edges whose post-response severity falls below the pathological threshold:

$$\text{CUS} = \frac{1}{|\mathcal{E}_{\text{path}}^{(t)}|} \sum_{e_{ij} \in \mathcal{E}_{\text{path}}^{(t)}} \mathbb{I}\left[w_{ij}^{(t+1)} < \tau_{\text{path}}\right]. \quad (12)$$

Annotation Agreement. Inter-annotator agreement was substantial: Fleiss’ $\kappa = 0.78$ for pathological-edge identification, Fleiss’ $\kappa = 0.84$ for node-type assignment, and ICC = 0.82 for continuous edge-severity scoring. These statistics indicate that PERR and CUS are computed from a stable expert reference rather than from model-internal graph consistency.

E Case Studies and Mechanistic Analysis

We provide a case study to illustrate how MindUntangler updates the graph state and changes routing decisions across turns. We then analyze extraction errors from a base GraphRAG pipeline to show why the Topology Verification Agent is used before routing.

E.1 Case Study: Cognitive Untangling and Dynamic Routing

This case tracks a multi-turn session involving relationship anxiety. As shown in Figure 8, MindUntangler first identifies a maladaptive association between a late-return event and a negative relationship belief. When the patient shows resistance and high emotional arousal, the Router shifts toward a validation-oriented framework. After the arousal decreases, the system returns to cognitive restructuring and elicits a reflection node, with the pathological edge weight reduced by Turn 5.

E.2 Mechanistic Error Analysis

Early versions of the extraction pipeline showed that zero-shot graph construction can be sensitive to pragmatic phenomena such as metaphor, quotation, and negation. Table 11 summarizes two representative failure modes: *Metaphorical Omission* and *Contextual Misattribution*.

These errors can affect downstream routing. A false negative may leave a relevant belief association undetected, causing the system to default to generic validation. A false positive may route the dialogue toward unnecessary cognitive challenge when the patient is quoting or rejecting the belief rather than endorsing it. This analysis motivates the Topology Verification Agent (§3.3), which checks whether candidate nodes and edges are supported by the original utterance and recent context before routing.