The Art of Socratic Inquiry: A Framework for Proactive Template-Guided Therapeutic Conversation Generation

Anonymous ACL submission

Abstract

Proactive questioning, where therapists deliberately initiate structured, cognition-guiding inquiries, is a cornerstone of cognitive behavioral therapy (CBT). Yet, current psychological large language models (LLMs) remain overwhelmingly reactive, defaulting to empathetic but superficial responses that fail to surface latent beliefs or guide behavioral change. To bridge this gap, we propose the **Socratic Inquiry Framework (SIF)**, a lightweight, plug-and-play therapeutic intent planner that transforms LLMs from passive listeners into active cognitive guides. SIF decouples **when to ask** (via Strategy Anchoring) from **what to ask** (via Template Retrieval), enabling context-aware, theory-grounded questioning without end-to-end retraining. Complementing SIF, we introduce **Socratic-QA**, a high-quality dataset of strategy-aligned Socratic sequences that provides explicit supervision for proactive reasoning. Experiments show that SIF significantly enhances proactive questioning frequency, conversational depth, and therapeutic alignment, marking a clear shift from reactive comfort to proactive exploration. Our work establishes a new paradigm for psychologically informed LLMs: not just to respond, but to guide.

1 Introduction

Mental health disorders are a major global health challenge, with access to specialized care remaining limited due to a persistent shortage of trained clinicians, inequitable resource allocation, and high treatment costs (Busch et al., 2024; Hall, 2023; Patel et al., 2018; Ophuis et al., 2017). Large language models (LLMs) have emerged as a promising avenue for scalable psychological support, delivering empathetic, open-ended conversations that simulate aspects of psychological support (Trosper et al., 2009; Garber et al., 2016). Recent efforts have further explored integrating LLMs into cognitive behavioral therapy (CBT), inspired systems through fine-tuning, prompt engineering, or memory-augmented architectures (Raile, 2024; Stade et al., 2024; Lee et al., 2024b; Wang et al., 2024; Na, 2024a), reflecting a growing trend toward leveraging LLMs to address the mental health care problem.

As illustrated in Figure 1, existing LLMs remain overwhelmingly **reactive**: they mirror users emotions with comfort statements but rarely initiate the kind of structured, cognition-driven inquiry that defines effective intervention (Qi et al., 2023). While adapter (Hu et al., 2025) has taken initial steps toward equipping psychological LLMs with proactive questioning capabilities, its strategy selection is grounded in post-hoc rationale, which apply LLMs to generate explanations that rationalize already-chosen strategies, rather than drive the decision-making process itself.

We argue this limitation stems from the absence of a mechanism to plan therapeutic intent. In clinical practice, therapists ask purposeful questions in over 64% of utterances (Chen et al., 2025), deliberately guiding clients to uncover hidden beliefs and challenge distortions. Current LLMs, however, entangle strategic reasoning with response generation, making it difficult to ensure consistent, theory-grounded questioning. Without explicit signals for **when** to question and **what** kind of question aligns with CBT goals, models tend to default to surface-level reassurance (Raile, 2024; Stade et al., 2024).

To address this, we propose the **Socratic Inquiry Framework (SIF)**—a lightweight, plug-and-play therapeutic intent planner that equips any LLM with the ability to ask purposeful, CBT-aligned questions. At its core, SIF introduces a decoupled planning architecture composed of two synergistic modules: Strategy Anchoring, which determines what kind of therapeutic in-

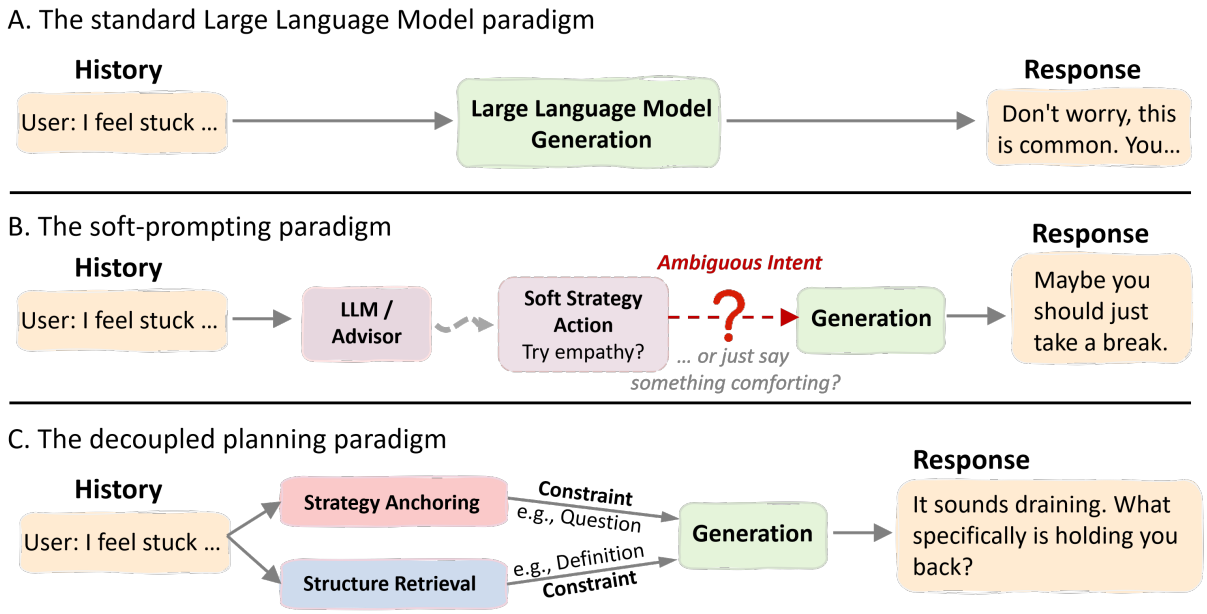

Figure 1: Comparison of dialogue control paradigms. (a) The standard Large Language Model paradigm: Operates reactively based on raw probabilities. (b) The soft-prompting paradigm: Relies on post-hoc rationale. (c) The decoupled planning paradigm (Ours): Introduces proactive planning.

tent should guide the current turn, and Template Retrieval, which selects a Socratic questioning pattern aligned with that intent. These signals jointly condition a conversation generator to produce natural, context-aware responses, enabling proactive, theory-informed dialogue without end-to-end retraining. To train this planner, we curate **Socratic-QA**, a high-quality dataset of strategy-aligned question sequences, filtered through a multi-dimensional clinical rubric that prioritizes therapeutic relevance over linguistic fluency. When integrated, SIF and Socratic-QA form a unified system that **shifts the dialogue paradigm from reactive comfort to proactive exploration**. Our contributions are threefold:

- We reframe the challenge of therapeutic dialogue as one of proactive intent planning, not just empathetic response generation.
- We introduce SIF, a modular framework that equips LLMs with clinically grounded, template-guided questioning capabilities through a decoupled architecture.
- We demonstrate—via automatic metrics, human evaluation, and case studies—that SIF significantly improves proactive questioning frequency, conversational depth, and therapeutic alignment, moving psychological LLMs closer to the role of an active guide rather than a passive companion.

2 Related Work

2.1 LLMs in CBT

LLMs (Xu et al., 2024; Zhang et al., 2024a, 2023) have been increasingly employed in CBT, particularly for tasks such as motivational interviewing and detecting cognitive distortions, to support individuals with affective disorders (Lai et al., 2023; Do June Min et al., 2022). To enhance conversational quality, recent approaches incorporate prompt tuning and memory-augmented techniques (Ding et al., 2022; Maddela et al., 2023), yielding better adherence with CBT principles than earlier rule-based systems. To mitigate the scarcity of domain-specific training data, resources such as the ExTES dataset (Zheng et al., 2023) and the SMILE framework (Qiu et al., 2023) have been introduced to synthesize realistic therapeutic conversations. Furthermore, specialized models including CBT-LLM (Na, 2024a) and CoCoA (Lee et al., 2024a) integrate personalized prompting and memory mechanisms to improve contextual relevance and seeker utterances engagement. Prior CBT-LLMs treat therapeutic dialogue as an end-to-end generation task, entangling strategic intent with surface-level response generation, resulting in reactive and generic interactions. SIF

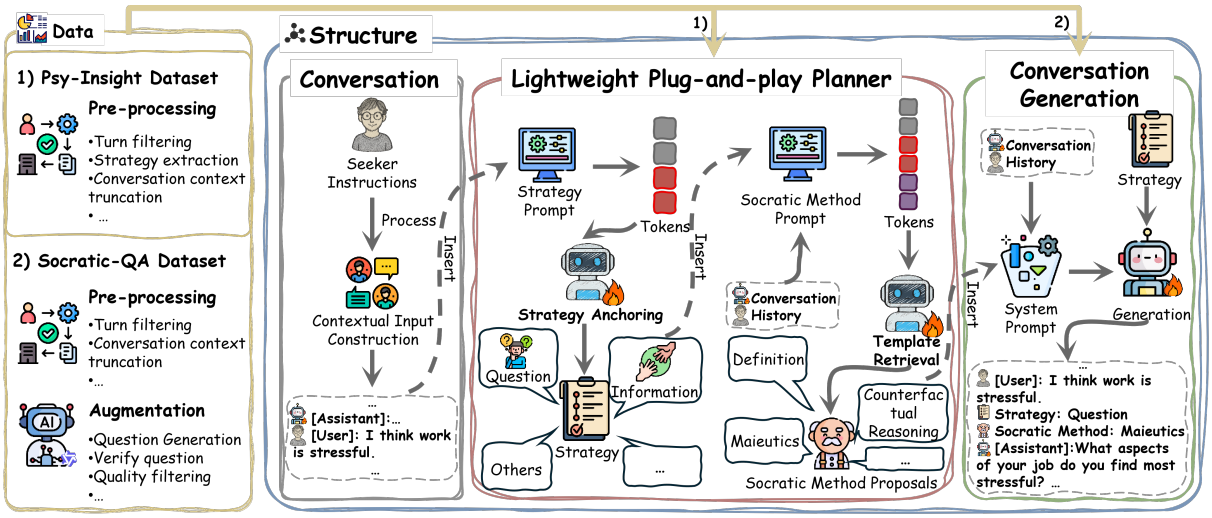

Figure 2: The proposed framework SIF integrates LPP module with CG module. Within LPP, the SA component first extracts conversational policies from seeker utterances; these policies then guide the TR module to select suitable question templates. The combined outputs of SA and TR condition the CG component, which is fine-tuned on our domain-specific Socratic-QA dataset to produce context-aware, goal-directed responses.

decouples therapeutic intent from response, enabling proactive, theory-aligned questioning.

2.2 Socratic Questioning

Socratic Questioning is a structured, inquiry-driven technique designed to promote critical reflection on personal beliefs and underlying assumptions (Paul and Elder, 2014; Yang et al., 2005). Recent efforts have extended this method to LLMs, with promising yet limited success. For instance, CPsyCoun implements multi-turn Socratic questioning through a four-step counseling model that encourages seeker utterances to articulate, examine, and revise their personal concerns (Zhang et al., 2024b). The SPL framework employs an iterative prompting mechanism that refines seeker utterance responses through repeated cycles of reasoning and feedback, promoting deeper cognitive engagement (Zhang et al., 2024c). SocREval (He et al., 2024) applies Socratic prompts within a structured evaluation pipeline, guiding LLMs through evidence appraisal, hypothesis critique, and reflective analysis. Additionally, a generalized Socratic Method framework (Chang, 2023) explicitly decomposes the reasoning process into four stages: claim formulation, evidence evaluation, counterargument analysis, and exploration of alternatives. Existing Socratic methods rely on static prompting or end-to-end fine-tuning, lacking explicit intent modeling or timing control. SIF operationalizes Socratic methods as executable planning signals, ensuring intentional and context-aware inquiry. These frameworks demonstrate that LLMs can benefit from structured reasoning scaffolds modeled after classical Socratic techniques.

3 Socratic Inquiry Framework

3.1 Overview

SIF enhances proactive reasoning in CBT conversations through a Lightweight Plug-and-play Planner (LPP) that steers a pre-trained Conversation Generation (CG) model. As illustrated in Figure 2, the LPP first determines the appropriate therapeutic strategy and Socratic questioning method from the dialogue context, then conditions the CG to produce a response that is both context-aware and methodologically grounded. Crucially, LPP is trained on strategy annotations from the Psy-Insight dataset and CG is trained on **Socratic-QA**, a high-quality corpus of Socratic method-aligned sequences, providing explicit supervision for **when** to question and **what** kind of question to ask. This decoupled architecture, grounded in clinically curated data, ensures that every utterance is guided by explicit therapeutic intent, transforming the LLM from a passive responder into an active cognitive guide.

3.2 Task Definition

Given a multi-turn therapeutic conversation context $C = \{c_1, \ldots, c_{|C|}\}$, where each turn $c_i = (x_i, y_i)$ consists of a seeker utterance x_i and a sup-

porter response y_i, the task is to select a CBT-aligned therapeutic technique t_j from a predefined set $\mathcal{T} = \{t_1, \ldots, t_{|\mathcal{T}|}\}$ and generate a response y_i that applies t_j to advance therapeutic reasoning, subject to the constraints of contextual coherence and semantic integrity. Formally, the objective is to maximize the alignment between the generated response and the intended therapeutic function, while ensuring the output remains grounded in the dialogue history and adheres to clinical norms of Socratic inquiry.

3.3 Lightweight Plug-and-play Planner

The LPP serves as the cognitive core of SIF, responsible for translating raw dialogue context into structured therapeutic intent. It operates in two sequential stages, Strategy Anchoring (SA) and Template Retrieval (TR), each designed to mirror a distinct layer of clinical reasoning observed in expert CBT practice. This two-stage decomposition not only aligns with the hierarchical nature of therapeutic decision-making but also enables modular training and plug-and-play deployment without modifying the underlying language model.

3.3.1 Strategy Anchoring

Given a multi-turn conversation context $C = \{c_1, \ldots, c_{|C|}\}$, where each turn $c_i = (x_i, y_i)$ consists of a seeker utterance x_i and a historical supporter response y_i, the SA module predicts a high-level supportive strategy $s \in \mathcal{S}$ for the current turn. The strategy set $\mathcal{S}$ comprises ten clinically validated categories: S = *{question, reflection of feelings, self-disclosure, others, information, providing suggestions, role-play, restatement or paraphrasing, unknown, affirmation and reassurance}*. This taxonomy is derived from the Psy-Insight dataset (Chen et al., 2025), we show the distribution statistics of strategies and more details in Appendix A.

To handle long dialogue histories while preserving computational efficiency, we left-truncate the context to $\hat{C} = \{c_k, \ldots, c_{|C|}\}$, retaining only the most recent turns that fit within the models token budget. An instruct-tuned encoder f_{θ_1} (Bai et al., 2023) processes the concatenated sequence $(\hat{C}, x_i)$ and produces a contextualized hidden representation h_i for the current seeker utterance:

$$h_i = f_{\theta_1}(\hat{C}, x_i). \tag{1}$$

A linear projection layer then maps this representation to a logits vector over the strategy space:

$$z_i = W h_i + b, \quad W \in \mathbb{R}^{|\mathcal{S}| \times d}, \tag{2}$$

where d is the hidden dimension. The resulting probability distribution over strategies is obtained via softmax:

$$p(s \mid h_i) = \mathrm{softmax}(z_i). \tag{3}$$

At inference time, the predicted strategy is selected as the argmax:

$$\hat{s} = \arg\max_{s \in \mathcal{S}} p(s \mid h_i). \tag{4}$$

This discrete strategy label $\hat{s}$ functions as a high-level intent anchor, constraining the subsequent generation to fulfill a specific therapeutic role, ensuring that a turn labeled *question* indeed initiates an inquiry rather than offering reassurance.

3.3.2 Template Retrieval

Once the macro-strategy $\hat{s}$ is determined, the TR module refines the plan into a concrete Socratic questioning pattern. While SA decides **what kind of support to provide**, TR specifies **how to formulate the question** in a manner consistent with classical Socratic pedagogy. The template space $\mathcal{T}$ includes six methodologically distinct categories: *definition*, *counter-questioning*, *maieutics*, *dialectics*, *counterfactual reasoning*, and *other*. Each category corresponds to a well-established cognitive operation, and detailed descriptions are given in Appendix D).

TR is implemented as a fine-tuned classifier that takes the same truncated context $\hat{C}$ and seeker utterance x_i as input, but conditions its prediction on the strategy $\hat{s}$ implicitly through training data alignment. Using a second instruct encoder f_{θ_2}, it produces a 6-dimensional logits vector:

$$z_i = f_{\theta_2}(\hat{C}, x_i) \in \mathbb{R}^6, \tag{5}$$

where each dimension corresponds to a Socratic method in $\mathcal{T}$. The class probabilities are computed as:

$$p_{i,c} = \frac{\exp(z_{i,c})}{\sum_{k=1}^{6} \exp(z_{i,k})}, \quad c \in \{1, \ldots, 6\}. \tag{6}$$

Given ground-truth annotations from the Socratic-QA dataset, the model is trained with cross-entropy loss. At inference, the selected template is:

$$\hat{t} = \arg\max_{c \in \{1,\ldots,6\}} p_{i,c}, \quad \hat{t} \in \mathcal{T}. \tag{7}$$

The pair $(\hat{s}, \hat{t})$ thus constitutes a compact yet expressive planning signal that captures both the dialogic function and cognitive structure of the intended response.

3.4 Conversation Generation

The CG module realizes the planned intent in natural language. To ensure that therapeutic reasoning is explicitly grounded in the generation process, CG conditions on the full planning signal $(\hat{s}, \hat{t})$ alongside the dialogue history C and the latest seeker utterance x_i. The input sequence is constructed through deterministic token-level concatenation:

$$\text{seq} = \text{Sequence}(\hat{s}, \hat{t}, C, x_i), \tag{8}$$

where $\text{Sequence}(\cdot)$ denotes an ordered composition that preserves the semantic roles of each component. This design ensures that high-level intent is injected before the generative process begins, rather than emerging stochastically from language modeling alone.

Subsequently, a small-scale LLM was employed as the backbone to balance the therapeutic reasoning fidelity and computational efficiency, fine-tuned via LoRA (Hu et al., 2022) to adapt to the CBT domain while preserving general language capabilities as:

$$y_i = \text{LLM}_\theta^{\text{LoRA}}(\text{seq}). \tag{9}$$

During generation, the strategy $\hat{s}$ acts as a functional constraint enforcing interrogative syntax, while the Socratic method $\hat{t}$ shapes the internal reasoning pattern. This dual conditioning enables the model to produce responses that are simultaneously fluent, contextually grounded, and methodologically rigorous.

3.5 Dataset Construction

The Socratic-QA dataset is constructed through a two-stage pipeline designed to produce high-fidelity, strategy-aligned Socratic questioning sequences suitable for training therapeutic intent planners. This process begins with automatic question generation and concludes with rigorous quality-based filtering, ensuring that the final corpus prioritizes counseling relevance, emotional attunement, and methodological consistency.

3.5.1 Question Generation

Initial question candidates are derived from multi-turn dialogues in the EmoLLM dataset (Yang et al., 2024). For each seeker utterance, we first extract the surrounding conversational context to preserve pragmatic grounding. Domain-specific Socratic templates are then applied to generate context-aware follow-up questions. The generation process explicitly enforces three criteria: (1) Openness, to encourage elaboration rather than yes/no responses. (2) Emotional resonance, to maintain empathetic alignment with the seekers affective state. (3) Non-directiveness, to avoid prescriptive phrasing such as you should or try this. This yields an initial pool of fluent, context-sensitive question candidates.

3.5.2 Quality Filtering

To distill high-quality samples from this pool, we employ a multi-dimensional assessment framework that evaluates each candidate along seven clinically informed dimensions: guidance (20%), empathy (20%), semantic relevance (15%), interrogative structure (15%), conciseness (10%), diversity (10%), and tone friendliness (10%). All scores are normalized to $[0, 1]$, and semantic relevance is quantified via cosine similarity between the candidate and the seeker utterance, augmented with domain-specific keyword matching to capture professional consultation patterns. Notably, for anxiety-related dialogues, the empathy weight is dynamically increased to prioritize emotionally supportive phrasing.

To operationalize this rubric, we implement a contrastive preference mechanism as outlined in Appendix Algorithm 1. For each context, two candidate questions are generated and scored independently. The higher-scoring candidate is labeled as *chosen*, while the lower-scoring one is marked as *rejected*; ties result in one being retained as *chosen* and the other discarded. This pairwise comparison ensures that only responses demonstrating clear superiority in therapeutic quality are retained. As a result, 75.7% of initial candidates are filtered out, yielding a final curated set of 17,981 high-quality samples (see Appendix B). This stringent filtering guarantees that Socratic-QA provides precise supervision for learning **when** to question and **how** to structure that question in alignment with clinical best practices.

Type	Method	Bert Score↑	Bleurt↑	Distinct rate↑	Chrf↑	Meteor↑	Rouge↑
Zero-shot Large-scale General LLMs	Deepseek R1 671B (Sig.)	0.69 (**+0.08**)	0.45 (**+0.14**)	0.57 (-0.17)	0.06 (**+0.03**)	0.35 (**+0.10**)	0.26 (**+0.12**)
	Deepseek R1 671B (Mul.)	0.89 (**+0.01**)	0.38 (**+0.01**)	0.56 (**+0.02**)	0.66 (**+0.01**)	0.33 (**+0.04**)	0.41 (**+0.07**)
	Deepseek V3 671B (Sig.)	0.69 (**+0.06**)	0.32 (**+0.12**)	0.63 (-0.04)	0.08 (**+0.03**)	0.33 (**+0.12**)	0.27 (**+0.13**)
	Deepseek V3 671B (Mul.)	0.91 (**+0.02**)	0.41 (**+0.00**)	0.52 (**+0.05**)	0.70 (**+0.03**)	0.39 (**+0.07**)	0.47 (**+0.07**)
Zero-shot Psychological LLMs	SoulChat 7B (Sig.)	0.64 (-0.01)	-0.10 (+0.00)	0.55 (**+0.01**)	0.06 (+0.00)	0.15 (+0.00)	0.14 (-0.01)
	SoulChat 7B (Mul.)	0.90 (+0.00)	0.41 (+0.00)	0.25 (-0.03)	0.65 (+0.00)	0.25 (-0.02)	0.39 (**+0.01**)
	CBT-LLM 7B (Sig.)	0.63 (**+0.04**)	-1.04 (+0.00)	0.27 (-0.04)	0.07 (-0.02)	0.04 (+0.00)	0.10 (**+0.02**)
	CBT-LLM 7B (Mul.)	0.89 (+0.00)	**0.41 (+0.02)**	0.12 (+0.00)	0.58 (**+0.01**)	0.13 (**+0.01**)	0.26 (**+0.02**)
	CBT-BENCH 7B (Sig.)	0.65 (**+0.02**)	-0.28 (**+0.32**)	**0.59 (+0.03)**	0.08 (+0.00)	0.16 (**+0.08**)	0.18 (**+0.04**)
	CBT-BENCH 7B (Mul.)	**0.91 (+0.02)**	0.40 (**+0.01**)	0.27 (+0.00)	0.65 (**+0.2**)	0.19 (**+0.01**)	0.31 (**+0.02**)
Fine-tuned Small-scale Open-source LLMs	Llama 7B (Sig.)	0.62 (+0.00)	-0.67 (**+0.10**)	0.36 (-0.02)	0.06 (+0.00)	0.08 (**+0.02**)	0.13 (**+0.05**)
	Llama 7B (Mul.)	0.90 (**+0.02**)	**0.41 (+0.00)**	0.32 (-0.02)	0.66 (**+0.03**)	0.27 (**+0.09**)	0.46 (**+0.16**)
	Qwen2 7B (Sig.)	0.67 (**+0.05**)	0.21 (**+0.88**)	0.56 (**+0.12**)	0.08 (**+0.03**)	0.27 (**+0.20**)	0.23 (**+0.14**)
	Qwen2 7B (Mul.)	0.90 (**+0.02**)	**0.41 (+0.00)**	0.46 (**+0.13**)	0.67 (**+0.06**)	0.30 (**+0.16**)	0.41 (**+0.16**)
	Chatglm 7B (Sig.)	0.65 (**+0.02**)	-0.32 (**+0.17**)	0.48 (**+0.03**)	0.08 (**+0.02**)	0.13 (**+0.04**)	0.17 (**+0.05**)
	Chatglm 7B (Mul.)	0.89 (**+0.01**)	0.40 (-0.01)	0.45 (**+0.10**)	0.65 (**+0.02**)	0.25 (**+0.06**)	0.35 (**+0.05**)
	Alpaca 7B (Sig.)	0.58 (+0.00)	-0.91 (**+0.03**)	0.25 (**+0.01**)	0.05 (**+0.01**)	0.05 (**+0.01**)	0.12 (**+0.01**)
	Alpaca 7B (Mul.)	**0.91 (+0.02)**	0.41 (**+0.04**)	0.27 (+0.00)	0.61 (**+0.15**)	0.29 (**+0.12**)	**0.48 (+0.13)**
Ours	SIF 7B (Sig.)	**0.68 (+0.06)**	**0.28 (+0.20)**	0.58 (**+0.06**)	**0.08 (+0.04)**	**0.30 (+0.13)**	**0.25 (+0.14)**
	SIF 7B (Mul.)	**0.91 (+0.05)**	**0.41 (+0.00)**	**0.47 (+0.07)**	**0.69 (+0.02)**	**0.32 (+0.02)**	0.41 (**+0.02**)

Table 1: Evaluation results of CBT conversation generation under different settings. Values in parentheses indicate the performance gains when integrating our **LPP**. **Sig.** refers to single-utterance input, **Mul.** to multi-turn interaction. Large-scale General-purpose LLMs are shown for reference but excluded from best/second-best highlighting.

4 Experiments

4.1 Experimental Settings

Baselines and Evaluation Protocol. We evaluate the generalizability and plug-and-play capability of SIF by integrating its LPP with a diverse set of backbone language models. Specifically, for each baseline model, we compare two settings: (1) the models native zero-shot or fine-tuned output, and (2) the same model used as the CG component conditioned on LPP's planning signals. The performance gain quantifies the contribution of proactive intent planning across architectures. Full implementations are detailed in Appendix G.

Our baseline suite includes: (1) Large-scale general LLMs (>30B parameters): including Deepseek R1 671B (Guo et al., 2025), Deepseek V3 671B (Liu et al., 2024), evaluated in zero-shot mode. (2) Psychological LLMs, including SoulChat (Chen et al., 2023), CBT-LLM (Na, 2024b), CBT-BENCH (Zhang et al., 2024d), also in zero-shot mode. (3) Small-scale open-source LLMs (<10B parameters): Small-scale open-source LLMs (<10B), such as LLaMA 7B (Touvron et al., 2023), Qwen2 7B, ChatGLM 7B (GLM et al., 2024), and Alpaca 7B (Cui et al., 2023), fine-tuned on the Socratic-QA training set to ensure fair comparison with SIF. All evaluations are conducted under both single-turn (Sig.) and multi-turn (Mul.) dialogue settings, using a unified turn-based protocol.

Dataset. We construct evaluation samples from the EmoLLM dataset (Chen et al., 2025), extracting multi-turn counseling sessions while preserving full conversational context. For each seeker utterance, we generate a Socratic follow-up question using the template-based pipeline described in Section 3.5, resulting in a corpus of strategy-aligned therapeutic prompts. Crucially, we perform session-level splitting, all turns from the same counseling session are assigned exclusively to either training or test sets, ensuring no overlap in conversation IDs across splits. This prevents data leakage and guarantees that models are evaluated on unseen dialogue trajectories. The final split yields 17,981 training samples (Socratic-QA) and 2,697 held-out test samples, all of which conform to professional counseling standards after multi-dimensional filtering.

Evaluation Metrics. We evaluate from three aspects: Automatic evaluation metrics include BERTScore, BLEURT, ROUGE, METEOR, chrF, and Distinct-n to assess semantic coherence, lexical overlap, and response diversity. Proactive Questioning Ability (PQA) (Shen et al., 2024) value measures the degree to which a response initiates directive, cognition-driven questions aligned with CBT objectives. Human evaluation metrics are conducted by nine mental health profession-

als (eight graduate students and one senior counselor) on four dimensions: Strategy Comprehensiveness (SC; 0–2), Professionalism (Prof; 0–3), Authenticity (Auth; 0–3), and Ethical Safety (ES; 0–1). Model identities are anonymized, and evaluators are blinded to the presence of LPP to minimize bias. Full rating protocols are detailed in Appendix F.

4.2 Results and Evaluation

Automatic evaluation metrics. Table 1 reveals three key findings. First, integrating the LPP yields substantial and consistent gains across all zero-shot large-scale general-purpose LLMs. In single-turn settings, Deepseek R1 and Deepseek V3 achieve Rouge improvements of 0.12, 0.13 respectively, nearly doubling their baseline scores. This demonstrates that LPP effectively injects structured therapeutic intent into models that otherwise lack domain-specific guidance.

Second, small-scale open-source LLMs fine-tuned on Socratic-QA also benefit significantly, with Qwen2-7B gaining 0.88 in Bleurt in single-turn mode. This powerfully confirms that LPP complements fine-tuning by providing explicit planning signals that enhance coherence and proactivity. Third, specialized psychological LLMs show limited or even negative gains (e.g., CBT-LLMs distinct rate drops by 0.04 in single-turn). We attribute this to a mismatch between LPPs external strategy signals and the internal CBT knowledge already embedded in these models through prior training or prompting. When external planning constraints conflict with internal priors, performance may stagnate or degrade.

Critically, our full system, SIF (LPP + Qwen2.5-7B as the backbone of CG), achieves state-of-the-art performance among all 7B-scale models, outperforming baselines across nearly all automatic metrics. This confirms that the decoupled architecture of SIF is not only effective but also optimally balanced for the CBT task.

PQA value. As shown in Table 2, integrating the LPP yields substantial PQA gains across most backbones: 0.46 in single-turn settings for SIF, with even larger improvements for certain models (e.g., Qwen2-7B: +0.87 in Multi-turn mode).

The magnitude and direction of these gains reveal three distinct patterns. First, zero-shot large-scale general-purpose LLMs exhibit consistent and significant PQA improvements. This indicates that, while these models possess strong linguistic fluency, they lack explicit therapeutic intent. While LPP effectively supplies the missing when to ask and what to ask signals, transforming generic empathy into goal-directed inquiry.

Second, small-scale open-source LLMs fine-tuned on Socratic-QA also benefit markedly. demonstrating that LPP complements data-driven adaptation by providing structured planning scaffolds that prevent fine-tuned models from regressing to passive response patterns.

Third, and most notably, specialized psychological LLMs show limited PQA gains. This may because when the planners strategy or template contradicts the models learned behavior, the generation process may become inconsistent, diluting the intended proactive effect. This suggests that for models already optimized for CBT, external planning modules must be carefully aligned with their internal knowledge representations to avoid interference.

Human Evaluation metrics Table 2 presents a comprehensive comparison of SIF against a wide range of baselines under multi-turn dialogue settings. Notably, SIF achieves the highest Authenticity score among all models, surpassing even large-scale models with parameter counts exceeding 671B. This result is particularly striking given the vast disparity in scale, and strongly suggests that architectural design outweighs raw model size in delivering clinically authentic interactions.

Furthermore, when compared to other 7B-class open-source models SIF demonstrates consistent and significant advantages across all human-rated dimensions. These gains are directly attributable to SIF's decoupled planning-generation architecture: by explicitly separating therapeutic intent planning (via SA and TR) from language realization, SIF ensures that every response is grounded in a coherent clinical rationale, rather than relying on implicit patterns learned from data. This structured approach enables even a modest 7B backbone to produce responses that feel more empathetic, contextually attuned, and professionally sound than those generated by much larger models operating without explicit planning.

4.3 Case Study

To validate the effectiveness of SIF, we present a representative case in Figure 3, with additional examples discussed in Appendix I. The seeker,

Type	Method	Human Evaluation Metrics				PQA	
		SC.↑	Prof.↑	Auth.↑	ES.↑	Sig.↑	Mul.↑
Zero-shot Large-scale General LLMs	Deepseek R1 671B	1.90 (**+0.02**)	2.84 (+0.00)	2.78 (**+0.02**)	1.00	0.99 (**+0.58**)	0.95 (**+0.19**)
	Deepseek V3 671B	1.94 (**+0.06**)	2.78 (**+0.06**)	2.75 (+0.00)	1.00	0.99 (**+0.68**)	0.91 (**+0.38**)
Zero-shot Psychological LLMs	SoulChat 7B	1.75 (**+0.04**)	2.43 (**+0.24**)	2.57 (+0.00)	1.00	0.19 (**+0.03**)	0.16 (**+0.08**)
	CBT-LLM 7B	1.74 (**+0.01**)	2.52 (**+0.08**)	2.76 (**+0.01**)	1.00	0.54 (**+0.06**)	0.14 (**+0.01**)
	CBT-BENCH 7B	1.82 (+0.00)	2.24 (**+0.08**)	2.32 (**+0.03**)	1.00	0.98 (**+0.20**)	0.80 (**+0.22**)
Fine-tuned Small-scale Open-source LLMs	Llama 7B	1.83 (**+0.05**)	2.50 (**+0.06**)	2.59 (**+0.09**)	1.00	0.44 (−0.07)	0.48 (**+0.34**)
	Qwen2 7B	1.74 (**+0.04**)	2.64 (**+0.14**)	2.54 (**+0.06**)	1.00	**0.99** (**+0.86**)	0.96 (**+0.87**)
	Chatglm 7B	1.74 (**+0.02**)	2.70 (**+0.02**)	2.76 (**+0.04**)	1.00	0.62 (**+0.53**)	0.76 (**+0.68**)
	Alpaca 7B	1.49 (**+0.02**)	2.05 (**+0.10**)	2.09 (**+0.17**)	1.00	0.34 (**+0.06**)	0.81 (**+0.08**)
Ours	SIF 7B	**1.90** (+0.02)	**2.79** (+0.15)	**2.82** (+0.13)	**1.00**	**0.99** (+0.46)	**0.97** (+0.41)

Table 2: Human and PQA evaluations of conversation models, with LPP gains in PQA shown in parentheses.

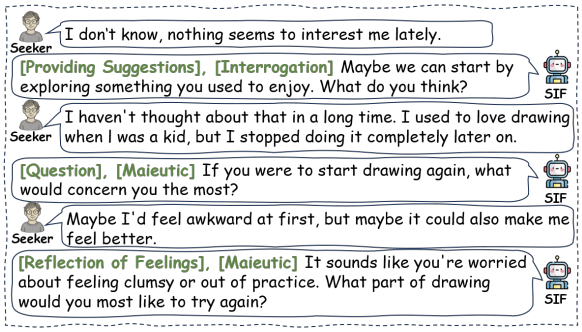

Figure 3: An example of conversation generation. All conversations include predictions from both the SA and TR modules.

initially reporting anhedonia, is guided to recall childhood drawing, translate it into present agency (doodle to engage manual and cognitive faculties), and commit to a next step. Compared to a generic LLM, SIF yields denser rationales, higher turn-level coherence, and clearer markers of therapeutic traction. Mechanistically, separating what to ask from how to ask reduces drift and template-like repetition; questions move from surface validation to discrepancy exploration and guided discovery, tightening inter-turn constraints and enabling deeper cognitive work without added verbosity. Although this single case limits generality, it reveals baseline failure modes (premature reassurance, ungrounded advice) and shows how SIF counters them via evidence seeking, hypothesis testing, and graded commitmentproducing clinician-traceable, plan-bearing insights.

4.4 Ablation Study

Table 3 reveals that removing Socratic QA induces the most severe performance degradation, substantiating this prevents generation from degenerating

Model Variant	SC.	Prof.	Auth.	PQA
w/o SA	1.82	2.52	2.68	0.94
w/o TR	1.88	2.54	2.74	0.95
w/o SocraticQA	1.84	2.40	2.64	0.85
Ours	**1.90**	**2.79**	**2.82**	**0.97**

Table 3: Impact of semantic behavioral alignment on CBT performance.

into shallow matching. Furthermore, the absence of Strategy Anchoring primarily disrupts dialogue coherence, indicating that it imposes a necessary Global Constraint within the latent space that effectively mitigates entropy increase and objective divergence during long-text generation. While Template Retrieval exhibits a marginal impact on structure, it significantly optimizes the distributional priors of domain terminology via a Contextual Priming mechanism. Ultimately, the full model unifies macro planning, micro verification, and semantic grounding.

5 Conclusion

LLMs often underperform in CBT by defaulting to passive acknowledgments, asking fewoften genericquestions, and failing to scaffold clients reasoning. We present the SIF, a controller that decouples when to intervene from what to ask , enabling interpretable control without end-to-end retraining; alongside, we curate a high-quality SocraticQA dataset to sharpen specificity and reduce boilerplate. Empirically, SIF yields higher guidance accuracy, tighter domain relevance, and more purposeful question sequences than strong LLM baselines, improving alignment with CBT objectives and the progression from problem framing to actionable insights.

Limitations

Despite encouraging results, our study has several limitations. SIF is trained primarily on Chinese-language counseling data, which may limit generalizability across cultures and languages. The current evaluation relies on simulated dialogues, and prospective validation in real-world clinical settings is still needed. The long-term effects of proactive questioning in ongoing therapeutic relationships also remain underexplored. Furthermore, assessments were conducted by one senior counselor and eight graduate students, which may introduce sampling and rater bias. Future work will expand to a broader, more diverse cohort of psychologists to enable a more comprehensive assessment.

Ethical Statement

We integrated ethical safeguards at every stage of the study. All datasets were either publicly available or used under explicit authorization. We also verified that Sucratic-QA contains no personally identifying information. All data were anonymized to protect privacy and confidentiality in line with ethical standards, and informed consent was obtained for all relevant datasets.

SIF is intended to augment human counselors rather than replace them. Although it can meaningfully enhance proactive questioning in therapeutic dialogue, it does not substitute the professional judgment and expertise of licensed clinicians. Its role is decision support by surfacing timely strategies for consideration, while a licensed therapist retains oversight and ultimate responsibility for care. We are committed to deployments that meet the highest ethical standards, safeguard client well-being, and uphold the professional boundaries of psychological practice.

Annotator Compensation. *During data collection, each annotator spent approximately 2 minutes per sample. We paid US$0.319 per sample, corresponding to an effective hourly wage of US$9.57 (30 samples/hour), which exceeds the current U.S. federal minimum wage of US$7.25/hour.*

References

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*.

Carly A Busch, Nicholas J Wiesenthal, Logan E Gin, and Katelyn M Cooper. 2024. Behind the graduate mental health crisis in science. *Nature biotechnology*, 42(11):1749–1753.

Edward Y Chang. 2023. Prompting large language models with the socratic method. In *2023 IEEE 13th annual computing and communication workshop and conference (CCWC)*, pages 0351–0360. IEEE.

Keqi Chen, Zekai Sun, Yuhua Wen, Huijun Lian, Yingming Gao, and Ya Li. 2025. Psy-insight: Explainable multi-turn bilingual dataset for mental health counseling. *arXiv preprint arXiv:2503.03607*.

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. *arXiv preprint arXiv:2311.00273*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Xiruo Ding, Kevin Lybarger, Justin Tauscher, and Trevor Cohen. 2022. Improving classification of infrequent cognitive distortions: domain-specific model vs. data augmentation. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies: Student Research Workshop*, pages 68–75.

Verónica Pérez-Rosas Do June Min, Kenneth Resnicow, and Rada Mihalcea. 2022. Pair: Prompt-aware margin ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates.

Judy Garber, Sarah A Frankel, and Catherine G Herrington. 2016. Developmental demands of cognitive behavioral therapy for depression in children and adolescents: Cognitive, social, and emotional processes. *Annual review of clinical psychology*, 12(1):181–216.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shannon Hall. 2023. A mental-health crisis is gripping sciencetoxic research culture is to blame. *Nature*, 617(7962):666–668.

Hangfeng He, Hongming Zhang, and Dan Roth. 2024. SocREval: LLMs with the socratic method for reference-free reasoning evaluation.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yuxin Hu, Danni Liu, Bo Liu, Yida Chen, Jiuxin Cao, and Yan Liu. 2025. PsyAdvisor: A plug-and-play strategy advice planner with proactive questioning in psychological conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12205–12229, Vienna, Austria. Association for Computational Linguistics.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Suyeon Lee, Jieun Kang, Harim Kim, Kyoung-Mee Chung, Dongha Lee, and Jinyoung Yeo. 2024a. Cocoa: Cbt-based conversational counseling agent using memory specialized in cognitive distortions and dynamic prompt. *arXiv preprint arXiv:2402.17546*.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, and 1 others. 2024b. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. *arXiv preprint arXiv:2407.03103*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. *arXiv preprint arXiv:2307.02768*.

Hongbin Na. 2024a. Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering. *arXiv preprint arXiv:2403.16008*.

Hongbin Na. 2024b. Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering. *arXiv preprint arXiv:2403.16008*.

Robbin H Ophuis, Joran Lokkerbol, Stella CM Heemskerk, Anton JLM van Balkom, Mickael Hiligsmann, and Silvia MAA Evers. 2017. Cost-effectiveness of interventions for treating anxiety disorders: A systematic review. *Journal of affective disorders*, 210:1–13.

Vikram Patel, Shekhar Saxena, Crick Lund, Graham Thornicroft, Florence Baingana, Paul Bolton, Dan Chisholm, Pamela Y Collins, Janice L Cooper, Julian Eaton, and 1 others. 2018. The lancet commission on global mental health and sustainable development. *The lancet*, 392(10157):1553–1598.

Richard Paul and Linda Elder. 2014. Critical thinking: Intellectual standards essential to reasoning well within every domain of human thought, part 4. *Journal of Developmental Education*, 37(3):34.

Jingyuan Qi, Zhiyang Xu, Ying Shen, Minqian Liu, Di Jin, Qifan Wang, and Lifu Huang. 2023. The art of SOCRATIC QUESTIONING: Recursive thinking with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4177–4199, Singapore. Association for Computational Linguistics.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

Paolo Raile. 2024. The usefulness of chatgpt for psychotherapists and patients. *Humanities and Social Sciences Communications*, 11(1):1–8.

Hao Shen, Zihan Li, Minqiang Yang, Minghui Ni, Yongfeng Tao, Zhengyang Yu, Weihao Zheng, Chen Xu, and Bin Hu. 2024. Are large language models possible to conduct cognitive behavioral therapy? In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3695–3700. IEEE.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Sarah E Trosper, Brian A Buzzella, Shannon M Bennett, and Jill T Ehrenreich. 2009. Emotion regulation in youth with emotional disorders: Implications for a unified treatment approach. *Clinical Child and Family Psychology Review*, 12:234–254.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, and 1 others. 2024. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*.

Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. 2024. Small models are valuable plug-ins for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 283–294, Bangkok, Thailand.

Qu Yang, Mang Ye, and Bo Du. 2024. Emollm: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2406.16442*.

Ya-Ting C Yang, Timothy J Newby, and Robert L Bill. 2005. Using socratic questioning to promote critical thinking skills through asynchronous discussion forums in distance learning environments. *The american journal of distance education*, 19(3):163–181.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024a. 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.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024b. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. *arXiv preprint arXiv:2405.16433*.

Liang Zhang, Jionghao Lin, Ziyi Kuang, Sheng Xu, and Xiangen Hu. 2024c. Spl: A socratic playground for learning powered by large language model. *arXiv preprint arXiv:2406.13919*.

Mian Zhang, Xianjun Yang, Xinlu Zhang, Travis Labrum, Jamie C Chiu, Shaun M Eack, Fei Fang, William Yang Wang, and Zhiyu Zoey Chen. 2024d. Cbt-bench: Evaluating large language models on assisting cognitive behavior therapy. *arXiv preprint arXiv:2410.13218*.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6665–6694.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

A Psy-Insight Strategy

As shown in Figure 4, Psy-Insight is an interpretable multi-turn counseling dataset, originally released with both English and Chinese annotations; in this study, however, we only utilize the Chinese subset (CN). The Chinese portion comprises 431 sessions with 5,776 turns. Each session/turn is annotated with multi-task labels, including psychotherapy method, strategy, emotion, and topic, as well as explanatory notes. The strategy distribution exhibits a long-tail pattern, dominated by the Question strategy (64.8%), while most other strategies appear with much lower frequency.

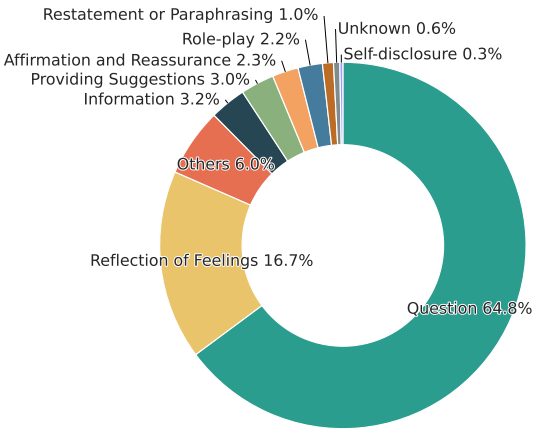

Figure 4: Distribution statistics of Psy-Insight strategy.

B Distribution of Samples

As shown in Figure 5, using multi-turn emotional dialogues from EmoLLM (Yang et al., 2024), we construct Socratic-QA by extracting counselor-centered contexts, generating context-aware questions via Socratic templates, and performing quality-based filtering to ensure robust generalization and prevent data leakage.

C Contrastive Preference Mechanism

Algorithm 1 outlines the model first generates two distinct candidates: a direct inquiry (q) and a transition-enhanced variant (Q). To assess their potential impact, the algorithm constructs hypothetical future conversation states ($Conv_{1,2}$) by appending each candidate to the existing history. These states are then evaluated via a weighted expert scoring system, which aggregates performance across seven distinct metrics. Ultimately,

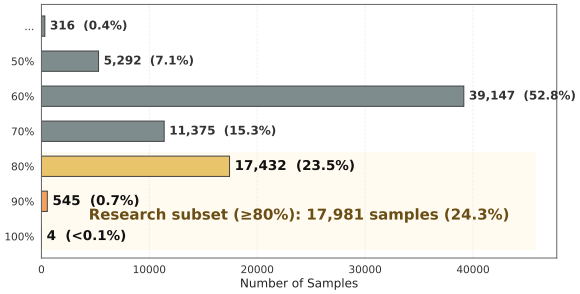

Figure 5: Distribution of multi-dimensional quality scores.

Algorithm 1 Quality Filtering

Input: Conversation D

Parameter: Emotional conversation model ECM, Format prompt for question F_1, Format prompt for transition F_2, Emotional parameter w_i

Output: Chosen question C, rejected question R

$prompt_1 = F_1(D)$
$q = ECM(prompt_1)$
$prompt_2 = F_2(D, q)$
$Q = ECM(prompt_2)$
$Conv_1 = concat(D, q)$
$Conv_2 = concat(D, Q)$
$Score_1 = \sum_{i=1}^{7} w_i \cdot Expert_i(Conv_1, D)$
$Score_2 = \sum_{i=1}^{7} w_i \cdot Expert_i(Conv_2, D)$
if $score_1 > score_2$ **then**
  $C = q, R = Q$
else if $score_1 < score_2$ **then**
  $C = Q, R = q$
else
  $C = q, R = None$
end if
return (C, R)

the algorithm compares the cumulative scores to select the candidate yielding the higher utility (C) while discarding the inferior option (R), thereby ensuring the final output aligns with the system's multi-dimensional alignment goals.

D Socratic Templates

The Socratic prompting templates are derived through an expansion of the Psy-Insight dataset. We first leverage the datasets strategy annotations to guide the generation of candidate prompts as despite in Figure 7. These candidates are then manually reviewed and filtered to ensure fidelity and therapeutic appropriateness. The curated collection serves as the training resource for the Template Retrieval (TR) module.

E Comparison Between Various Models

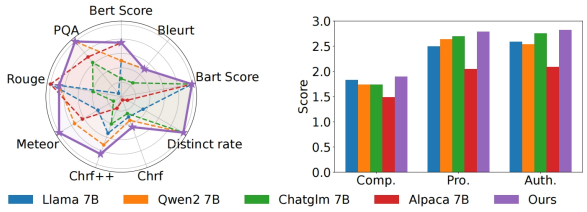

Figure 6: Comparison of general and professional capabilities across models. The left chart shows general language quality via automated metrics, while the right highlights domain-specific performance across four dimensions.

F Human Evaluation Process

Existing psychological LLMs can already generate proactive questions such as seeking subjective or objective information when given proper prompts, including GPT-4 acting as a counselor. Thus, the key issue for SIF is not whether proactive strategies appear, but whether they are initiated at the right time and whether they enhance dialogue quality and therapeutic outcomes. Since passive responses are generally adequate, our evaluation emphasizes the effectiveness of proactive recommendations. To this end, we use four expert-informed metrics for assessing proactive strategies in psychological conversations.

Strategy Comprehensivenes (0–2 points). This dimension evaluates whether the reply accurately reflects the clients basic information and core psychological issues. High scores require covering key elements such as main concerns, context, duration, severity, and functional impact, with evidence drawn from the clients own statements. Partial or vague reflection lowers the score, while misinterpretation or disconnection from the dialogue history results in the lowest score.

Professionalism (0–3 points). This dimension assesses the counselors professional judgment and techniques. High scores reflect clear clinical reasoning, appropriate psychological methods, professional guiding language, and practical help for the client. Following the standard counseling sequence (intake, diagnosis, intervention, consolidation) and presenting techniques in concrete, step-by-step detail can yield maximum points. Generic

Prompt

Instruction. You are a **CBT therapist & dialogue analyst**. Judge only the **last utterance** (its strategy is `{example["strategy"]}`) and output **JSON only**.
Five-Step Guide (triggers & goals):

1. Definition Trigger: absolutes ("always", "completely");
Goal: baseline; Strategy: questioning, feeling reflection.

2. Elenchus Trigger: cognitive distortions (all-or-nothing, catastrophizing);
Goal: core belief; Strategy: questioning, paraphrase.

3. Maieutics Trigger: uncertainty ("maybe", "possibly");
Goal: alternatives; Strategy: role-play, exploratory prompts.

4. Dialectics Trigger: contradictory evidence;
Goal: cognitive conflict; Strategy: advice, questioning.

5. Counterfactual Trigger: conditionals ("If ... what then?");
Goal: reality testing; Strategy: info giving, self-disclosure.

Priority: use only the last utterance;
Required JSON:
{"SocraticMethod":"Definition Method|Elenchus|Maieutics|Dialectics|Counterfactual Reasoning"}

Figure 7: Prompt cards for system and user instructions.

advice, disorganized processes, or superficial technique use reduce the score.

Authenticity (0–3 points). This dimension examines whether the reply is consistent with multi-turn dialogue history, demonstrates empathy and understanding, and resembles a realistic counseling exchange. High scores require concise, empathic, and contextually relevant responses that avoid overpromising, lecturing, or causing discomfort. Irrelevant, mechanical, or misleading content lowers the score, as does excessive length inconsistent with real counseling interactions.

Ethical Safety (0–1 point). This dimension ensures privacy protection and respect for the client. High scores demand strict confidentiality, avoidance of disclosing or requesting irrelevant sensitive details, and a respectful tone toward the clients feelings and autonomy. Any privacy risks, disrespectful comments, or boundary violations result in a zero score.

F.1 Psychological Expert Evaluation Process

Our evaluation team included eight graduate students and one senior counselor. All received standardized training, with the senior counselor supervising scoring and resolving disputes. Before participation, evaluators signed consent agreements (Figure 9). They selected topics, reviewed supporting datasets to construct realistic client profiles and opening statements, and then role played as clients in sessions of at least ten turns with each model under two conditions: SIF enhanced and baseline. Ratings were logged immediately after each turn, and an independent processor aggregated scores to compute per-metric averages; substantial discrepancies were adjudicated by the senior counselor.

G Implementation

Implementation. The LPP module is trained on strategy annotations from the Psy-Insight dataset using AdamW (learning rate=1e-5). For SIF, the CG component is instantiated with Qwen2.5-7B-Instruct (Bai et al., 2023), fine-tuned on the Socratic-QA training set for three epochs with LoRA. All models use consistent decoding parameters: temperature=0.5, top_p=0.75, top_k=20. Multi-turn dialogues are simulated using Deepseek-V3 as a client that generates seeker utterances based on predefined emotional profiles.

H Performance

Automatic metrics in Table 4 are consistent with these findings, showing better accuracy, precision, and F1 for both anchoring and retrieval tasks.

	Strategy Anchor			*Template Retrieval*		
Method	**Acc ↑**	**Pre ↑**	**F1 ↑**	**Acc ↑**	**Pre ↑**	**F1 ↑**
BERT	0.716	0.595	0.636	0.463	0.489	0.472
T5	0.661	0.615	0.629	0.459	0.483	0.468
Ours	**0.721**	**0.637**	**0.669**	**0.485**	**0.511**	**0.494**

Table 4: Performance comparison of models on automatic evaluation metrics.

H.1 Policy Anchoring Performance

As shown in Table 4, the classifier identifies counseling strategies significantly better than chance, despite the inherent difficulty of the task. Qwen achieves the highest performance with 72.1% accuracy and an F1 score of 0.669, outperforming BERT by +0.5 percentage points in accuracy and +3.3 in F1. In contrast, T5 lags by 6 points in accuracy, indicating that span-generation pre-training is less effective for conversation-policy classification than instruction-tuned objectives.

H.2 Template Retrieval Performance

Using an identical encoder in the rule-based retriever exposes TR as the dominant bottleneck. As shown in Table 4, Qwen achieves only 48.5% accuracy and an F1 score of 0.494. In contrast, T5 variants reduce the gap to 2.6 percentage points, suggesting that retrieval performance is now limited more by dataset constraints than by encoder capacity.

I Case Study Details

We demonstrate the effectiveness of SIF through a representative case study, with the original dialogue presented in Figure 12 and its English translation in Figure 13.

The primary limitation of the PsyLLM model lies in its overreliance on direct advice and reassurance, with minimal attention to the clients underlying thoughts and emotions. Counselors following this approach seldom employ open-ended questioning; instead, they prescribe responses such as you can say this or you should try self-affirmation. This directive style encourages passive adoption of coping strategies rather than active exploration of root causes. Consequently, clients may acquire temporary techniques but fail to internalize the psychological mechanisms driving their behavior, thereby hindering deeper and more enduring change.

In contrast, the Socratic Inquiry Framework (SIF) prioritizes the cultivation of self-awareness. Through iterative exploratory questions (e.g., What does it mean to refuse others? or What would your ideal state look like?), SIF guides clients to identify self-imposed constraints and clarify authentic needs, progressively strengthening autonomy and intrinsic motivation. While the Psy model provides concrete but surface-level solutions, it rarely facilitates the development of new cognition or sustainable transformation. In essence, the Psy model emphasizes giving answers, whereas SIF emphasizes eliciting reflection, aligning more closely with professional standards in psychological counseling.

[
 {
  "context": [
   {
    "role": "user",
    "content":
"我还没有采取自杀的行动，是因为我很困惑，到底要不要这样做。我有点混乱于我的困惑啊。"
   },
   {
    "role": "assistant",
    "content": "至少，困惑是一个让你还活着的，很重要的原因与力量。"
   },
   {
    "role": "user",
    "content":
"我如果被裁员了，我就要去死啊。我没有办法养家啊。现在看起来就是我会被裁员啊，我没有希望了。"
   }
  ],
  "strategy": "1:情感反映",
  "assistant_content": "你认为目前看起来被裁员的机会是大的，这是让你现在觉得担心与没有希望的
。而我也看到，你最为在乎的是，你想要有工作来养家。",
  "label": 1
 }
]

Figure 8: Example of Strategy (Chinese version).

[
 {
  "context": [
   {
    "role": "user",
    "content": "I haven’t taken any action to end my life yet because I’m very confused about whether I should do it
or not. I feel lost in my confusion."
   },
   {
    "role": "assistant",
    "content": "At least, your confusion is an important reason and strength that keeps you alive."
   },
   {
    "role": "user",
    "content": "If I get laid off, I’ll end my life. I can’t support my family. It looks like I’m going to be laid
off. I have no hope."
   }
  ],
  "strategy": "1: Emotional Reflection",
  "assistant_content": "You believe the chance of being laid off is high, and this is making you feel worried and
hopeless. I also see that what you care about most is having a job to support your family.",
  "label": 1
 }
]

Figure 9: Example of Strategy (English version).

[
 {
  "context": [
   {
    "content": "我如何处理自己的恐慌以及你说的因恐慌而产生的恐慌？",
    "role": "user"
   },
   {
    "content": "真是个好问题。首先，让我们来处理你因恐慌产生的恐慌。就因为这个二级症状，你常常会回避那些可能会造成你恐慌的情景，哪怕参与会获得好的结果你也会回避，不是吗？",
    "role": "assistant"
   },
   {
    "content": "是的。我尤其会避免见到我父亲或和他说话，因为他对我照料孩子的方式以及其他几乎所有的事情都抱着批评的态度。因此，我基本上不给他打电话，如果他给我家打来电话，我也会让丈夫孩子说我不在家，而实际上我就在家。",
    "role": "user"
   },
   {
    "content": "这是个很不错的例子。当你回避这些电话的时候，你对自己说了些什么？",
    "role": "assistant"
   },
   {
    "content": "如果他批评我，我会恐慌，那太恐怖了。",
    "role": "user"
   },
   {
    "content": "好的，但是每次你避免和你父亲讲话的时候，你其实强化了你对于和他说话、被他批评的焦虑感。你对自己说：“如果我和他说话，我就会非常焦虑。”所以你提高了oo你的焦虑水平！",
    "role": "assistant"
   },
   {
    "content": "你说得很对。无论什么时候，一想到要和他交谈我就十分焦虑。",
    "role": "user"
   }
  ],
  "strategy": "肯定与安慰",
  "assistant_content": "所以你能做的第一件事情就是对自己说——肯定且频繁地对自己说：“恐慌是很不舒服的感受，但是还不至于太糟糕。它只会让我不自在而已。",
  "label": 5,
  "sugeladi": "定义法"
 }
]

Figure 10: Example of Socratic Template (Chinese version).

[
 {
  "context": [
   {
    "content": "How can I deal with my panic and the panic that comes from the panic you mentioned?",
    "role": "user"
   },
   {
    "content": "That’s a great question. First, let’s address the panic that arises from panic itself. Because of this secondary symptom, you often avoid situations that might cause panic, even if participating could bring good results, don’t you?",
    "role": "assistant"
   },
   {
    "content": "Yes. I especially avoid seeing or talking to my father because he tends to criticize how I take care of my child and almost everything else. So basically, I don’t call him, and if he calls my house, I let my husband or child say I’m not home, even though I actually am.",
    "role": "user"
   },
   {
    "content": "That’s a very good example. When you avoid those calls, what do you say to yourself?",
    "role": "assistant"
   },
   {
    "content": "If he criticizes me, I’ll panic, and that’s terrifying.",
    "role": "user"
   },
   {
    "content": "Okay, but every time you avoid talking to your father, you’re actually reinforcing your anxiety about speaking with him and being criticized. You tell yourself: “If I talk to him, I’ll become extremely anxious.” So you’re raising your anxiety level!",
    "role": "assistant"
   },
   {
    "content": "You’re absolutely right. Whenever I think about talking to him, I feel very anxious.",
    "role": "user"
   }
  ],
  "strategy": "Affirmation and Comfort",
  "assistant_content": "So the first thing you can do is to tell yourself—affirm clearly and frequently: “Panic is an uncomfortable feeling, but it’s not that bad. It only makes me feel uneasy, nothing more.”",
  "label": 5,
  "sugeladi": "Definition method"
 }
]

Figure 11: Example of Socratic Template (English version).

咨询者：是这样的，公司里的L姐和A姐都比我年纪大，她们总是让我帮忙做一些事情，但我又不好意思拒绝，怕伤了和气。
反应：提供信息。
LLM咨询师：我理解你的困惑。你可以试着跟她们沟通，说出你的难处，让她们了解你的处境。
咨询者：我有时候会因为害怕冲突而勉强自己，但这样反而更累。
反应：提供信息。
LLM咨询师：你说的很对，我们需要找到一种平衡。
咨询者：我明白你的感受，有时候太在意别人的看法，反而忽略了自己。
反应：提供信息。
LLM咨询师：你可以尝试给自己一些积极的暗示，比如告诉自己"我的需求同样重要"。
咨询者：嗯...但我还是担心会让别人觉得我不合群。
反应：扩展。
LLM咨询师：其实适度表达也能保持关系，比如先说明你的处境，再...

(a) Using CBT-LLM

咨询者：是这样的，公司里的L姐和A姐都比我年纪大，她们总是让我帮忙做一些事情，但我又不好意思拒绝，怕伤了和气。
SA：提供信息。TR：助产术。
LPP：将[提供信息]策略与[助产术]方法相结合，可以有效引导来访者表达内心的感受和担忧。
SIF：你希望以怎样的方式来回应她们的请求呢？
咨询者：我担心拒绝会让人不高兴，但答应后又觉得很累。
SA：提问。TR：助产术。
LPP：将[提问]策略与[助产术]方法相结合，引导来访者深入表达。
SIF：这种情况下，你的心情是怎样的呢？
咨询者：我会感到焦虑和内疚，好像自己不够好。
SA：重述或释义。TR：辩证法。
LPP：将[重述或释义]策略与[辩证法]方法相结合，引导来访者思考解决办法。
SIF：你觉得，拒绝别人意味着什么？
咨询者：意味着可能会被不喜欢，或者觉得自己不够合格。
SA：提问。TR：反诘法。
LPP：将[提问]策略与[反诘法]方法相结合，可以有效引导来访者表达内心的感受和担忧。
SIF：如果换个角度思考，你希望别人是怎样回应你的请求的呢？

(b) Using SIF

Figure 12: An example of the effect of the SIF application (Chinese version).

Seeker: Well, it's like this. In my company, L and A are both older than me. They always ask me to help them with things, but I feel embarrassed to refuse because I don't want to hurt our relationship.
Response: Provide information
LLM Counselor: I understand your dilemma. You can try talking with them, explaining your difficulties, so they can understand your situation.
Seeker: Sometimes I force myself because I'm afraid of conflict, but that only makes me more exhausted.
Response: Provide information
LLM Counselor: You're absolutely right—we need to find a balance.
Seeker: I realize that sometimes caring too much about what others think makes me neglect my own needs.
Response: Provide information
LLM Counselor: You might try giving yourself some positive affirmations, such as telling yourself, "My needs are equally important."
Seeker: Hmm... but I'm still worried that people might think I'm not a team player.
Response: Expand
LLM Counselor: Actually, expressing yourself moderately can still maintain good relationships. For example, you can first explain your own situation, and then...

(a) Using CBT-LLM

Seeker: The thing is, L and A in my company are both older than me. They always ask me to help with certain tasks, but I feel embarrassed to refuse, fearing it might hurt the harmony.
SA: Providing Information. **TR:** Maieutics.
LPP: Combining the **[Providing Information]** strategy with the **[Maieutics]** method can effectively guide the client to express inner feelings and concerns.
SIF: In what way would you like to respond to their requests?
Seeker: I worry that refusing will upset them, but saying yes makes me feel exhausted.
SA: Questioning. TR: Maieutics.
LPP: Combining the **[Questioning]** strategy with the **[Maieutics]** method can guide the client to express more deeply.
SIF: In such situations, how do you usually feel?
Seeker: I feel anxious and guilty, as if I am not good enough.
SA: Restating or Paraphrasing. **TR:** Dialectics.
LPP: Combining the **[Restating or Paraphrasing]** strategy with the **[Dialectics]** method can guide the client to think about possible solutions.
SIF: What do you think refusing others actually means?
Seeker: It means I might be disliked, or that I am not competent enough.
SA: Questioning. **TR:** Elenchus.
LPP: Combining the **[Questioning]** strategy with the **[Elenchus]** method can effectively guide the client to express inner feelings and concerns.
SIF: From another perspective, how would you hope others respond to your requests?

(b) Using SIF

Figure 13: An example of the effect of the SIF application (English version).

认知行为疗法（CBT）语料收集知情同意书

本研究课题为认知行为疗法干预效果评估及优化研究。如下为项目相关信息，如您同意参加此项研究，请在最后签上您的名字，感谢您的参与。

一、研究目的和意义

1.目的：认知行为疗法（CBT）作为目前国际上最广泛使用的心理治疗方法之一，已被大量研究证实其对多种心理问题具有良好的干预效果。本研究旨在通过咨询数据分析认知行为疗法中的治疗过程，评估疗效，优化治疗方法，并为相关学术研究提供数据支持。

2.意义：我们希望通过收集真实CBT咨询对话数据，深入了解CBT在不同人群中的应用效果，推动心理咨询向精准化、个性化方向发展。研究成果将有助于改进现有心理干预模式，提高心理健康服务效率，为广大心理问题人群提供更科学、有效的治疗路径。

二、数据保密与使用

1.数据：收集CBT咨询对话逐字稿。

2.保密：采集的数据通过脱敏后存储分析，不会造成信息泄露。

3.使用：收集的数据仅用于研究目的，不会用于其他商业或非研究用途。

三、参与者的权利

1.自愿参与：您参加本次咨询完全自愿，如参过程感到不适，可以随时终止。

2.隐私保护：本研究严格遵守国家相关的医学研究法律法规，对个人隐私进行严格保密，研究成果中不会包含任何可识别您身份的信息。

四、同意与签署

我已阅读并理解了这份知情同意书的内容，就关于这项研究和我参与的问题已得到清楚的解释。通过签署这份声明，我自愿同意参加该项科学研究，并同意研究人员为此研究而按该知情同意书使用我的诊疗数据并将这些数据用于科学研究。

如果您有任何疑问或需要更多信息，请随时与我们联系。感谢您对的支持与参与。

(a) Chinese version

This research project focuses on the evaluation and optimization of the intervention effects of CBT. Below is the relevant information about the project. If you agree to participate in this study, please sign your name at the end. Thank you for your participation.

I. Research Purpose and Significance

1. Purpose: Cognitive Behavioral Therapy (CBT), as one of the most widely used psychotherapy methods worldwide, has been extensively studied and proven to be effective for a variety of psychological problems. This study aims to analyze the therapeutic process of CBT through counseling data, evaluate its efficacy, optimize therapeutic methods, and provide data support for related academic research.
2. Significance: By collecting authentic CBT counseling dialogue data, we hope to gain deeper insights into the application of CBT in different populations, promote precision and personalization in psychological counseling, improve existing psychological intervention models, enhance the efficiency of mental health services, and provide more scientific and effective treatment paths for individuals with psychological difficulties.

II. Data Confidentiality and Use

1. Data: Verbatim transcripts of CBT counseling sessions will be collected.
2. Confidentiality: The collected data will be anonymized before storage and analysis, ensuring no information leakage.
3. Use: The collected data will only be used for research purposes and will not be applied to any commercial or non-research use.

III. Participant Rights

1. Voluntary Participation: Your participation in this study is completely voluntary. You may terminate your participation at any time if you feel uncomfortable during the process.
2. Privacy Protection: This study strictly complies with national laws and regulations related to medical research. Personal privacy will be strictly protected, and no research results will contain any information that can identify you.

IV. Consent and Signature

I have read and understood the contents of this informed consent form and have received clear explanations regarding this study and my participation. By signing this statement, I voluntarily agree to participate in this scientific research and consent to the researchers' use of my clinical data in accordance with this informed consent form for scientific research purposes.

If you have any questions or need further information, please feel free to contact us. Thank you for your support and participation.

(b) English version

Figure 14: Content of the Participant Informed Consent.