苏格拉底式探询的艺术：一种用于主动式模板引导治疗对话生成的框架

Mingwen Zhang[1], **Minqiang Yang**[1], **Changsheng Ma**[1], **Yang Yu**[1]

Hui Bai[1], **Chen Xu**[2],**Xiangzhen Kong**[1],**Bin Hu**[1]1兰州大学信息科学与工程学院2北京理工大学医学技术学院

arXiv:2602.01598v1 [cs.CL] 2 Feb 2026

摘要

主动提问，即治疗师有意发起结构化、引导认知的询问，是认知行为疗法（认知行为疗法）的基石。然而，当前的心理学大语言模型（大语言模型）仍然极为被动，往往默认给出富有共情却流于表面的回应，无法揭示潜在信念或引导行为改变。为弥合这一差距，我们提出了**苏格拉底式探询框架（SIF）**，这是一种轻量级、即插即用的治疗意图规划器，可将大语言模型从被动倾听者转变为主动的认知引导者。SIF 将**何时提问**（通过策略锚定）与**提问什么**（通过模板检索）分离开来，从而无需端到端重新训练即可实现兼顾上下文、以理论为指导的提问。作为对 SIF 的补充，我们推出了**Socratic-QA**，这是一个高质量数据集，包含与策略对齐的苏格拉底式对话序列，为主动推理提供明确监督。实验表明，SIF 显著提升了主动提问频率、对话深度和治疗契合度，标志着从被动安慰转向主动探索的明确转变。我们的工作为受心理学启发的大语言模型确立了一种新范式：不仅要回应，更要引导。

1 引言

心理健康障碍是全球面临的一项重大健康挑战。由于训练有素的临床医生长期短缺、资源分配不均以及治疗费用高昂，获得专业护理的机会仍然有限（Busch 等，2024年；Hall，2023年；Patel 等，2018；Ophuis 等，2017年）。大语言模型（LLM）已成为提供可规模化心理支持的一条前景广阔的途径，能够开展富有共情、开放式的对话，模拟心理支持的某些方面（Trosper 等，2009；Garber 等，2016）。近期研究进一步探索将大语言模型整合到认知行为疗法（CBT）中，并通过微调、提示工程或增强记忆的架构来构建受启发的系统（Raile，2024年；Stade 等，2024年；Lee 等，2024年；Wang 等，2024年；Na，2024年），这反映出利用大语言模型应对心理健康护理难题的趋势日益增强。

如图 1 所示，现有大语言模型仍主要采取**被动响应**的方式：它们用安慰性陈述回应用户的情绪，却很少主动发起结构化、以认知为导向的探询，而这正是有效干预的关键（Qi 等人，2023年）。尽管适配器（Hu 等人，2025年）已初步尝试赋予心理学大语言模型主动提问的能力，但其策略选择基于事后推理，即利用大语言模型生成解释，为已经选定的策略提供合理化依据，而非驱动决策过程本身。

我们认为，这一局限源于缺乏规划治疗意图的机制。在临床实践中，治疗师会在超过 64% 的话语中提出有目的的问题（Chen 等人，2025年），有意识地引导来访者发现隐性信念并挑战认知扭曲。然而，当前的大语言模型将战略推理与回复生成混为一体，因此难以确保提问始终一致且有理论依据。若没有明确的信号指示**何时**提问，以及**何种**类型的问题符合认知行为疗法的目标，模型往往会默认采用浅层的安慰保证（Raile，2024年；Stade 等人，2024年）。

为解决这一问题，我们提出了**苏格拉底式探究框架（SIF）**——一种轻量、即插即用的治疗意图规划器，可赋予任何大语言模型提出有目的、符合认知行为疗法的问题的能力。SIF 的核心是引入一种解耦的规划架构，由两个协同模块组成：策略锚定，用于确定当前轮次应由何种治疗意

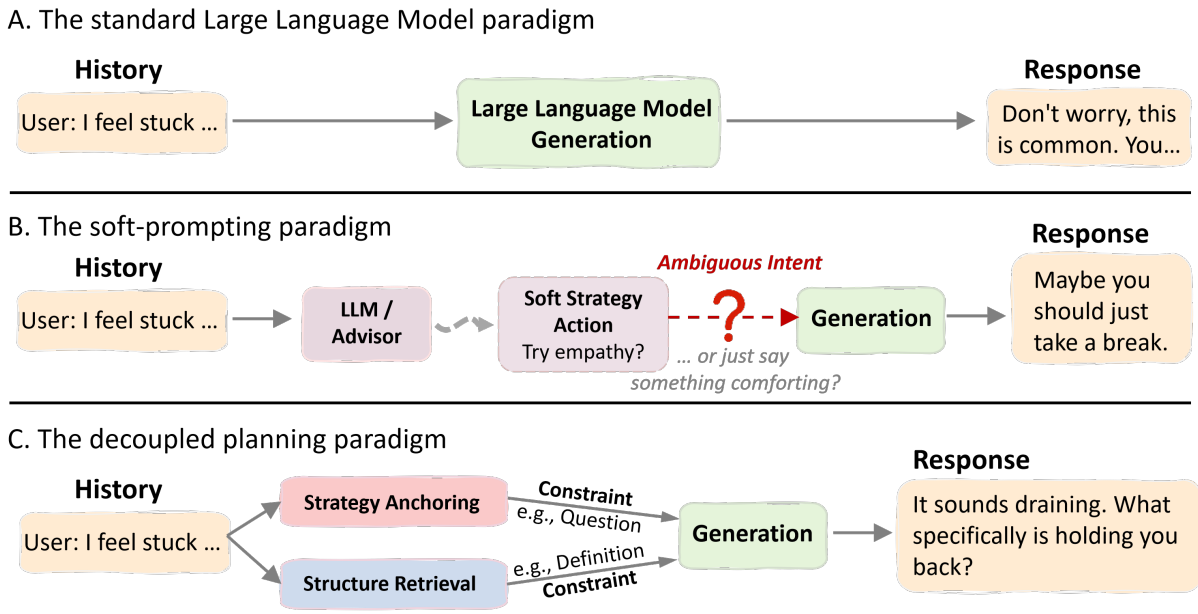

图1：对话控制范式比较。(a) 标准大语言模型范式：基于原始概率进行反应式运作。(b) 软提示范式：依赖事后推理。(c) 解耦规划范式（我们的方法）：引入主动规划。

图引导；模板检索则选择与该意图相符的苏格拉底式提问模式。这些信号共同作用于对话生成器，使其生成自然且贴合上下文的回应，从而无需端到端重新训练，便能实现主动、以理论为依据的对话。为训练这一规划器，我们整理了**Socratic-QA**，这是一个高质量的数据集，包含与策略相符的问题序列，并通过多维临床评估标准进行筛选，优先考虑治疗相关性而非语言流畅度。整合后，SIF 与 Socratic-QA 构成一个统一系统，**将对话范式从被动安慰转向主动探索**。我们的贡献主要有三点：

- 我们将治疗性对话的挑战重新界定为主动意图规划，而不仅仅是共情式回应生成。

- 我们提出了 SIF，这是一种模块化框架，通过解耦架构赋予大语言模型具备临床依据、由模板引导的提问能力。

- 我们通过自动化指标、人工评估和案例研究证明，SIF 显著提升了主动提问频率、对话深度和治疗契合度，使心理学大语言模型更接近主动引导者的角色，而非被动的陪伴者。

2 相关工作

2.1 认知行为疗法中的大语言模型

大语言模型（Xu 等，2024年；Zhang 等，2024a、2023）日益广泛地应用于认知行为疗法，尤其用于动机式访谈和检测认知扭曲等任务，以帮助情感障碍患者（Lai 等，2023年；Do June Min 等，2022）。为提升对话质量，近期方法采用了提示调优和记忆增强技术（Ding 等，2022；Maddela 等，2023），与早期基于规则的系统相比，这些方法能更好地遵循认知行为疗法原则。为缓解领域专用训练数据匮乏的问题，研究者推出了 ExTES 数据集（Zheng 等，2023）和 SMILE 框架（Qiu 等，2023）等资源，用于合成逼真的治疗性对话。此外，CBT-LLM（Na，2024a）和 CoCoA（Lee 等，2024a）等专用模型整合了个性化提示和记忆机制，以提升语境相关性并增强求助者话语的参与度。以往的 CBT 大语言模型将治疗性对话视为端到端生成任务，把策略意图与表层回应生成混为一体，导致互动被动且泛泛。SIF

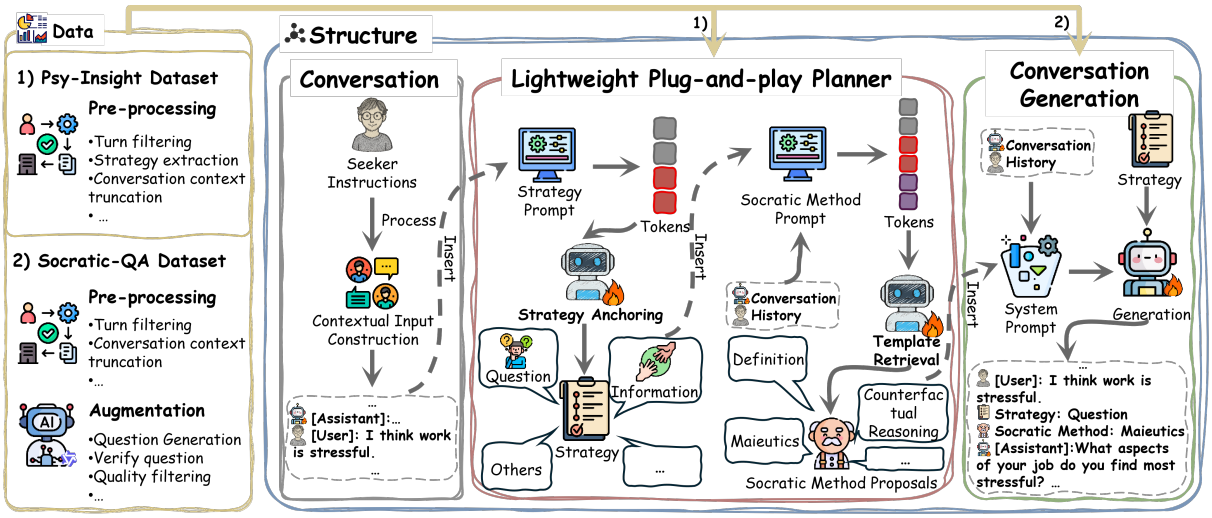

图 2：所提出的 SIF 框架将 LPP 模块与 CG 模块集成在一起。在 LPP 中，SA 组件首先从求助者话语中提取对话策略；随后，这些策略引导 TR 模块选择合适的问题模板。SA 和 TR 的组合输出为 CG 组件提供条件，而 CG 组件在我们领域专属的苏格拉底式问答数据集上进行微调，以生成兼顾上下文、目标明确的回应。

将治疗意图与回应分离，从而能够主动提出符合理论指导的问题。

2.2 苏格拉底式提问

苏格拉底式提问是一种结构化的探究技术，旨在促进人们批判性地反思个人信念及其潜在假设（Paul 和 Elder，2014年；Yang 等人，2005）。近年来，研究者尝试将这种方法拓展至大语言模型，并取得了虽有潜力但仍有限的成效。例如，CPsyCoun 通过四步咨询模型实施多轮苏格拉底式提问，鼓励求助者话语表达、审视并修正个人困扰（Zhang 等人，2024年）。SPL 框架采用迭代提示机制，通过反复进行推理与反馈来优化对求助者话语的回应，从而促进更深入的认知参与（Zhang 等人，2024c）。SocREval（He 等人，2024年）在结构化评估流程中运用苏格拉底式提示，引导大语言模型进行证据评估、假设批判和反思性分析。此外，通用苏格拉底方法框架（Chang，2023年）将推理过程明确分解为四个阶段：提出主张、证据评估、反驳论证分析和替代方案探索。现有的苏格拉底式方法依赖静态提示或端到端微调，缺乏明确的意图建模或时机控制。SIF 将苏格拉底式方法转化为可执行的规划信号，从而确保探究具有目的性并能适应上下文。这些框架表明，大语言模型能够从借鉴经典苏格拉底式技术构建的结构化推理支架中获益。

3 苏格拉底式探究框架

3.1 概述

SIF 通过轻量级即插即用规划器（LPP）引导预训练的对话生成（CG）模型，从而增强认知行为疗法对话中的主动推理。如图 2 所示，LPP 首先根据对话语境确定合适的治疗策略和苏格拉底式提问方法，随后以此为条件引导 CG 生成兼顾上下文且有方法论依据的回应。重要的是，LPP 在 Psy- Insight 数据集中的策略标注上进行训练，而 CG 则在 **Socratic-QA** 上进行训练；这是一个高质量、遵循苏格拉底式方法的序列语料库，为 **何时** 提问以及 **提出** 哪类问题提供明确监督。这种基于临床精心整理数据的解耦架构，确保每句话语都受到明确治疗意图的引导，使大语言模型从被动回应者转变为主动的认知引导者。

3.2 任务定义

给定一个多轮治疗对话上下文 $C = \{c_1, \ldots, c_{|C|}\}$，其中每一轮 $c_i = (x_i, y_i)$ 包含求助者话语 x_i 以及支持

者回应 y_i，任务是从预定义集合 $\mathcal{T} = \{t_1, \ldots, t_{|\mathcal{T}|}\}$ 中选择一种与认知行为疗法一致的治疗技术 t_j，并生成应用 t_j 的回应 y_i，以推进治疗性推理，同时受上下文连贯性与语义完整性的约束。形式上，目标是最大化所生成回应与预期治疗功能之间的一致性，同时确保输出立足于对话历史，并遵循苏格拉底式探询的临床规范。

3.3 轻量级即插即用规划器

LPP 是 SIF 的认知核心，负责将原始对话语境转化为结构化的治疗意图。它分两个依次进行的阶段运行：策略锚定（SA）和模板检索（TR），分别旨在模拟专家认知行为疗法实践中临床推理的不同层面。这种两阶段分解不仅契合治疗决策的层级特性，也支持模块化训练和即插即用部署，而无需修改底层语言模型。

3.3.1 策略锚定

给定一个多轮对话语境 $C = \{c_1, \ldots, c_{|C|}\}$，其中每轮 $c_i = (x_i,\ y_i)$ 由求助者话语 x_i 和历史支持者回复 y_i构成，SA 模块会为当前轮预测一种高层支持策略 $s \in \mathcal{S}$。策略集合 $\mathcal{S}$ 包含十种经临床验证的类别：$S =$ {提问、情感反映、自我披露、其他、信息提供、提出建议、角色扮演、重述或释义、未知、肯定与安慰}。该分类体系源自 Psy-Insight 数据集（Chen 等人，2025年），我们在附录A中展示了策略的分布统计及更多详情。

为了在保持计算效率的同时处理较长的对话历史，我们对ˆ上下文进行左截断至 $C = \{c_k, \ldots, c_{|C|}\}$，仅保留能纳入模型词元预算的最近若干轮次。指令微调编码器 f_{θ_1} (Bai等人，ˆ2023年)处理拼接序列$(C,\ x_i)$，并生成上下文化的隐藏表示 h_i，用于当前求助者话语：

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

随后，线性投影层将这一表征映射为策略空间上的逻辑值向量 ce:

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

其中 d 为隐藏维度。所得策略概率分布通过Softmax得到：

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

在推理时，预测策略被选取为argmax：

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

该离散策略标签ˆs 充当高层意图锚点，约束后续生成以履行特定治疗角色，确保被标注为提问的轮次确实发起询问，而非提供安慰保证。

3.3.2 模板检索

一旦宏观策略ˆs 确定，TR 模块便将计划细化为具体的苏格拉底式提问模式。策略锚定决定**提供何种支持**，模板检索则规定**如何表述问题**，使其符合古典苏格拉底式教学法。模板空间 $\mathcal{T}$ 包含六个在方法论上彼此不同的类别：定义、反问、助产术、辩证法、反事实推理以及其他。每个类别对应一种成熟的认知操作，详细描述见附录D）。

模板检索由微调分类器实现，该分类器以相同的截断上下文 C 和求助者话语 x_i 作为输入，但通过训练数据对齐，隐式地根据策略s进行预测。它使用第二个指令编码器 f_{θ_2}，生成一个 6 维逻辑值向量：

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

其中，每个维度对应一种苏格拉底式方法 $\mathcal{T}$。类别概率计算如下：

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

给定苏格拉底式问答数据集的真实标注，模型使用交叉熵损失进行训练。推理时，选定的模板为：

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

因此，配对项 ($\hat{s}$, $\hat{t}$) 构成了一种简洁而富有表现力的规划信号，既能捕捉预期回应的对话功能，也能体现其认知结构。

3.4 对话生成

CG 模块以自然语言实现规划好的意图。为确保治疗性推理在生成过程中得到明确体现，CG 会同时以完整的规划信号（$\hat{s}$, $\hat{t}$）、对话历史 C 以及求助者最新的话语 x_i作为条件。输入序列通过确定性的词元级拼接构建：

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

其中，Sequence(·) 表示一种有序组合，可保留各组成部分的语义角色。这种设计确保高层意图在生成过程开始前注入，而非仅由语言建模随机产生。

随后，采用小规模大语言模型作为骨干模型，以兼顾治疗性推理的保真度与计算效率；通过 LoRA（Huet al., 2022）进行微调，使其适应认知行为疗法领域，同时保留通用语言能力，如下所示：

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

生成过程中，策略 $\hat{s}$ 作为功能性约束，确保采用疑问句法；苏格拉底式方法 $\hat{t}$ 则塑造内部推理模式。这种双重条件约束使模型能够生成兼具流畅性、上下文贴合度和方法严谨性的回应。

3.5 数据集构建

Socratic-QA 数据集通过两阶段流水线构建，旨在生成高保真、与策略对齐的苏格拉底式提问序列，以用于训练治疗意图规划器。该流程从自动问题生成开始，以严格的基于质量的过滤结束，确保最终语料优先考虑咨询相关性、情感契合度和方法一致性。

3.5.1 问题生成

初始问题候选项源自 EmoLLM 数据集中的多轮对话（Yanget al., 2024年）。对于每条求助者话语，我们首先提取其周围的对话上下文，以确保语用锚定。随后应用特定领域的苏格拉底式模板，生成符合上下文的追问问题。生成过程明确遵循三项标准：（1）开放性，鼓励求助者展开阐述，而非仅作是或否的回答；（2）情感共鸣，与求助者的情绪状态保持共情上的契合；（3）非指导性，避免使用“你应该”或“试试这个”等指令性措辞。由此形成一组流畅且贴合上下文的初始问题候选项。

3.5.2 质量筛选

为了从这一候选池中筛选出高质量样本，我们采用多维评估框架，从七个基于临床实践的维度对每个候选项进行评估：引导性（20%）、共情（20%）、语义相关性（15%）、疑问结构（15%）、简洁性（10%）、多样性（10%）和语气友好度（10%）。所有分数均归一化至 $[0, 1]$, , 语义相关性则通过候选项与求助者话语之间的余弦相似度进行量化，并结合领域专属的关键词匹配，以捕捉专业咨询模式。值得注意的是，对于焦虑相关对话，我们会动态提高共情维度的权重，以优先保留具有情感支持性的措辞。

为落实这一评分标准，我们按照附录算法 1 中的说明实现了对比偏好机制。对于每个上下文，系统会生成两个候选问题并分别进行评分。得分较高的候选项被标记为选定，而得分较低的候选项则标记为拒绝；若两者得分相同，则保留其中一个作为选定，并丢弃另一个。这种成对比较确保仅保留在治疗质量方面表现出明显优势的回答。因此，初始候选项中有 75.7% 被筛除，最终得到包含 17,981 个高质量样本的精选数据集（见附录B）。这一严格的筛选过程确保 Socratic-QA 能够为学习**何时**提问以及**如何**依据临床最佳实践组织问题提供精准监督。

Type	方法	BERT评分↑	BLEURT↑	Distinct比率↑	chrF↑	METEOR↑	ROUGE↑
零样本大规模通用LLMs	Deepseek R1 671B（显著性）	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 6710亿（多语言）	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（显著性）	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（多语言）	0.91（**+0.02**）	0.41 (**+0.00**)	0.52 (**+0.05**)	0.70 (**+0.03**)	0.39（**+0.07**）	0.47（**+0.07**）
零样本心理学LLMs	SoulChat 70亿参数（单轮）	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（多轮）	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（显著性）	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 70亿参数（多语言）	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 70亿参数（显著性）	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 70亿参数（多语言）	**0.91 (+0.02)**	0.40 (**+0.01**)	0.27 (+0.00)	0.65 (**+0.2**)	0.19（**+0.01**）	0.31（**+0.02**）
微调小规模开源LLMs	Llama 7B（显著性）	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（多语言）	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（显著性）	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（多语言）	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（显著性）	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 70亿参数（显著性）	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（多语言）	**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（单一）	**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（多重）	**0.91 (+0.05)**	**0.41 (+0.00)**	**0.47 (+0.07)**	**0.69 (+0.02)**	**0.32 (+0.02)**	0.41 (**+0.02**)

表 1：不同设置下认知行为疗法对话生成的评估结果。括号中的数值表示集成我们的 **LPP** 后的性能提升。**显著性** 指单轮话语输入，**多语言** 指多轮交互。此处列出大规模通用大语言模型作为参考，但不参与最佳/次佳结果的突出显示。

4 实验

4.1 实验设置

基线模型与评估协议。 我们通过将 SIF 的 LPP 与多种骨干语言模型相结合，评估 SIF 的泛化能力和即插即用能力。具体而言，对于每个基线模型，我们比较两种设置：（1）模型原生的零样本或微调输出；（2）将同一模型作为 CG 组件，并以 LPP 的规划信号为条件。性能增益量化了跨架构的主动意图规划所作出的贡献。完整实现详见附录G。

我们的基线套件包括：(1) 大规模通用大语言模型（>30B 参数）：包括 Deepseek R1 671B（Guo 等人，2025年）、Deepseek V3 671B（Liu 等人，2024年），以零样本模式进行评估。(2) 心理学大语言模型，包括 SoulChat（Chen 等人，2023年）、CBT-LLM（Na，2024年）、CBT-BENCH（Zhang 等人，2024d），同样以零样本模式进行评估。(3) 小规模开源大语言模型（<10 B 参数）：小规模开源大语言模型（<10B），例如 LLaMA 7B（Tou- vron 等人，2023年）、Qwen27B、ChatGLM7B（GLM 等人，2024年）和 Alpaca 7B（Cui 等人，2023年），在 Socratic-QA 训练集上进行微调，以确保与 SIF 进行公平比较。所有评估均在单轮（多语言）和多轮（多语言）对话设置下进行，并采用统一的轮次基于协议。

数据集。 我们从 EmoLLM 数据集（Chen 等人，2025年）中构建评估样本，提取多轮咨询会话，同时保留完整的对话上下文。针对每条求助者话语，我们使用第 3.5 节所述的基于模板的流程生成一个苏格拉底式追问，从而形成与策略对齐的治疗提示语料库。重要的是，我们按会话级划分数据，同一咨询会话中的所有轮次仅分配至训练集或测试集之一，确保不同划分之间的对话 ID 不重叠。这样可以防止数据泄漏，并确保模型在未见过的对话轨迹上接受评估。最终划分得到 17,981 个训练样本（Socratic-QA）和 2,697 个留出测试样本；这些样本经多维过滤后均符合专业咨询标准。

评估指标。 我们从三个方面进行评估：自动评估指标包括 BERTScore、BLEURT、ROUGE、METEOR、chrF 和 Distinct-n，用于评估语义连贯性、词汇重合度和回复多样性。主动提问能力（PQA）（Shen 等人，2024年）衡量回复发起与认知行为疗法目标相符的指令性、认知驱动型问题的程度。人工评估指标由九位心理健康专业人士（

八名研究生和一名资深咨询师）从四个维度进行评估：策略全面性（SC；0–2）、专业性（Prof；0–3）、真实性（Auth；0–3）和伦理安全性（ES；0–1）。模型身份均已匿名，且评估者不知道是否使用了 LPP，以尽量减少偏差。完整的评分流程详见附录 F。

4.2 结果与评估

自动评估指标。表1 揭示了三项关键发现。首先，在所有零样本大规模通用大语言模型中，整合 LPP 都带来了显著且稳定的提升。在单轮设置下，Deepseek R1 和 Deepseek V3 的 ROUGE 分别提升了 0.12 和 0.13，几乎达到基线分数的两倍。这表明，LPP 能够有效地向原本缺乏领域特定指导的模型中注入结构化的治疗意图。

其次，在 Socratic-QA 上微调的小规模开源大语言模型也显著受益：Qwen2-7B 在单轮模式下的 BLEURT 提高了 0.88。这有力地证实，LPP 通过提供明确的规划信号增强连贯性和主动性，从而与微调形成互补。第三，专业心理学大语言模型的收益有限，甚至出现负面收益（例如，认知行为疗法大语言模型在单轮设置下的 Distinct比率下降了 0.04）。我们认为，这是由于 LPP 的外部策略信号与这些模型通过先前训练或提示已内化的认知行为疗法知识不匹配。当外部规划约束与内部先验相冲突时，性能可能停滞或下降。

值得注意的是，我们的完整系统 SIF（LPP + Qwen2.5-7B 作为 CG 的骨干模型）在所有 7B 规模模型中取得了最先进的性能，在几乎所有自动指标上都优于基线模型。这证实了 SIF 的解耦架构不仅有效，而且针对认知行为疗法任务实现了最佳平衡。

PQA 值。如表2所示，集成 LPP 后，大多数骨干模型的 PQA 值都有显著提升：SIF 在单轮设置下提升了 0.46，某些模型的提升幅度更大（例如，Qwen2-7B 在多轮模式下提升了 +0.87）。

这些提升的幅度和方向揭示了三种截然不同的模式。首先，零样本大规模通用大语言模型的 PQA 表现持续且显著改善。这表明，尽管这些模型具备较强的语言流畅度，却缺乏明确的治疗意图。LPP 则有效补足了何时提问和提什么问题这两类信号，将泛化的共情转化为目标导向的询问。

其次，在 Socratic-QA 上微调的小规模开源大语言模型也显著受益。这表明，LPP 通过提供结构化的规划脚手架来补充数据驱动的适配，防止微调后的模型退回到被动响应模式。

第三，也是最值得注意的一点，专门面向心理学的大语言模型在 PQA 方面的提升有限。这可能是因为，当规划器的策略或模板与模型习得的行为相矛盾时，生成过程可能变得前后不一致，从而削弱预期的主动干预效果。这表明，对于已经针对认知行为疗法进行优化的模型，外部规划模块必须与其内部知识表示仔细对齐，以避免产生干扰。

人工评估指标 表2 全面比较了 SIF 与多种基线模型在多轮对话设置下的表现。值得注意的是，SIF 在所有模型中取得了最高的真实性得分，甚至超过了参数量超过 6710亿 的大规模模型。考虑到模型规模存在巨大差异，这一结果尤为显著，并有力表明，与原始模型规模相比，架构设计对实现具有临床真实性的交互更为重要。

此外，与其他 7B级开源模型 相比，SIF 在所有人工评定维度上均表现出稳定且显著的优势。这些提升直接归功于 SIF 将规划与生成解耦的架构：通过将治疗意图规划（借助 SA 和 TR 实现）与语言表达明确分离，SIF 确保每个回答都以连贯的临床依据为基础，而非依赖从数据中习得的隐式模式。这种结构化方法使得即便是规模适中的 70亿参数 骨干模型，也能生成比许多未采用显式规划的大型模型更富同理心、更贴合上下文且更专业可靠的回答。

4.3 案例研究

为验证 SIF 的有效性，我们在图3中展示了一个具有代表性的案例，更多示例见附录I。来访者

Type	方法	人工评估指标				PQA	
		SC.↑	专业性↑	真实性↑	ES.↑	显著性↑	多语言↑
零样本大规模通用大语言模型	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）
零样本心理学LLMs	SoulChat 70亿参数	1.75（+0.04）	2.43 (+0.24)	2.57 (+0.00)	1.00	0.19 (+0.03)	0.16 (+0.08)
	CBT-LLM 70亿参数	1.74 (+0.01)	2.52 (+0.08)	2.76（+0.01）	1.00	0.54 (+0.06)	0.14 (+0.01)
	CBT-BENCH 70亿参数	1.82 (+0.00)	2.24（+0.08）	2.32 (+0.03)	1.00	0.98 (+0.20)	0.80 (+0.22)
微调小规模开源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)

表2： 对话模型的人工评估和 PQA 评估，括号中为 PQA 中的 LPP增益。

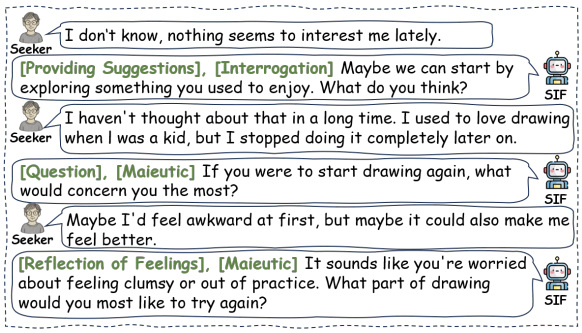

图3：对话生成示例。所有对话均包含策略锚定和模板检索模块的预测。

起初报告有快感缺失，随后在引导下回忆童年绘画，将其转化为当下的行动力（通过涂鸦调动手动与认知能力），并承诺采取下一步行动。与通用大语言模型相比，SIF 生成的推理依据更为丰富，轮次间连贯性更高，也更清晰地呈现出治疗进展的迹象。从机制上看，将“问什么”与“如何问”分开，可以减少偏离和模板化重复；问题也从表层确认转向差异探索与引导式发现，从而强化轮次间约束，并在不增加冗长程度的情况下推动更深入的认知工作。尽管单一个案限制了结论的普适性，但它揭示了基线失败模式（过早给予安慰保证、提出无依据建议），并展示了 SIF 如何通过证据寻求、假设检验和分级承诺来应对这些问题，进而生成临床医生可追溯、带有计划的洞见。

4.4 消融研究

表3 表明，移除苏格拉底式问答会导致最严重的性能下降，证实了它能防止生成退化

模型变体	SC.	专业性	真实性	PQA
不含策略锚定	1.82	2.52	2.68	0.94
不含模板检索	1.88	2.54	2.74	0.95
不含苏格拉底式问答	1.84	2.40	2.64	0.85
Ours	**1.90**	**2.79**	**2.82**	**0.97**

表3：语义行为对齐对认知行为疗法性能的影响。

为浅层匹配。此外，缺少策略锚定主要会破坏对话连贯性，这表明它在潜在空间中施加了必要的全局约束，能有效缓解长文本生成过程中的熵增和目标偏离。虽然模板检索对结构的影响微乎其微，但它通过上下文启动机制显著优化了领域术语的分布先验。最终，完整模型将宏观规划、微观验证和语义落地融为一体。

5 结论

大语言模型默认倾向于被动回应，常常提出很少且泛化的问题，并且无法为来访者的推理提供支架，因此在认知行为疗法中往往表现不佳。我们提出了 SIF，这是一种将“何时介入”与“询问什么”解耦的控制器，无需端到端重新训练即可实现可解释的控制；与此同时，我们整理了一个高质量的苏格拉底式问答数据集，以提升问题的具体性并减少模板化表达。实证结果表明，与强大的大语言模型基线相比，SIF 具有更高的引导准确率、更紧密的领域相关性，以及更具目的性的问题序列，从而改善了与认知行为疗法目标的对齐，并促进了从问题界定到可操作洞见的推进。

局限性

尽管取得了令人鼓舞的结果，我们的研究仍存在一些局限。SIF 主要基于中文咨询数据进行训练，这可能限制其在不同文化和语言之间的普适性。目前的评估依赖于模拟对话，而在真实世界临床环境中的前瞻性验证仍有待开展。在持续的治疗关系中，主动提问的长期影响也尚未得到充分探究。此外，评估由一名资深咨询师和八名研究生完成，这可能引入抽样偏差和评定者偏差。未来的工作将扩展至规模更大、构成更多元的心理学家群体，以开展更全面的评估。

伦理声明

我们在研究的每个阶段都融入了伦理保障措施。所有数据集均为公开数据，或是在获得明确授权后使用的。我们还核实了 Sucratic-QA 不包含任何个人身份识别信息。我们依照伦理标准对所有数据进行了匿名化处理，以保护隐私和保密性，并为所有相关数据集取得了知情同意。

SIF 旨在辅助人类咨询师，而非取代他们。尽管它能够显著增强治疗性对话中的主动提问，但不能替代持证临床工作者的专业判断和专业知识。SIF 的作用是提供决策支持，及时呈现可供考虑的策略；持证治疗师则负责监督并对护理工作承担最终责任。我们致力于确保系统的部署符合最高伦理标准、保障来访者的福祉，并维护心理实践的专业界限。

标注员报酬。 在数据收集期间，每位标注员为每个样本耗时约 *2* 分钟。我们按每个样本支付 *0.319* 美元，相当于时薪 *9.57* 美元（每小时 *30* 个样本），高于目前每小时 *7.25* 美元的美国联邦最低工资。

参考文献

Jinze Bai、Shuai Bai、Yunfei Chu、Zeyu Cui、Kai Dang、Xiaodong Deng、YangFan、Wenbin Ge、YuHan、Fei Huang 等共 12 人。2023年。Qwen 技术报告。*arXiv* 预印本 *arXiv:2309.16609*。

Carly A Busch、Nicholas J Wiesenthal、Logan E Gin 和 Katelyn M Cooper。2024年。科学领域研究生心理健康危机背后的原因。《自然生物技术》，42(11):1749–1753。

Edward Y Chang. 2023年. 使用苏格拉底方法提示大语言模型。载于*2023*年 *IEEE* 第 *13* 届年度计算与通信研讨会及会议（*CCWC*），第 0351–0360 页。IEEE。

Keqi Chen、Zekai Sun、Yuhua Wen、Huijun Lian、Ying- ming Gao 和 Ya Li. 2025年. Psy-Insight：用于心理健康咨询的可解释多轮双语数据集。*arXiv* 预印本 *arXiv:2503.03607*。

Yirong Chen、Xiaofen Xing、Jingkai Lin、Huimin Zheng、Zhenyu Wang、Qi Liu 和 Xiangmin Xu. 2023年. SoulChat：通过使用多轮共情对话进行微调，提升大语言模型的共情、倾听和安慰能力。*arXiv* 预印本 *arXiv:2311.00273*。

Yiming Cui、Ziqing Yang 和 Xin Yao。2023年。 适用于中文LLaMA和 Alpaca 的高效文本编码。 *arXiv* 预印本 *arXiv:2304.08177*。

Xiruo Ding、Kevin Lybarger、Justin Tauscher 和 Trevor Cohen。2022。改进低频认知扭曲的分类：领域专用模型与数据增强。载于 *2022* 年计算语言学协会北美分会会议论文集：人类语言技术：学生研究研讨会，第 68–75 页。

Verónica Pérez-Rosas、Do June Min、Kenneth Resnicow 和 Rada Mihalcea。2022。PAIR：用于动机式访谈中咨询师反思评分的提示感知间隔排序。载于*2022* 年自然语言处理实证方法会议论文集，第 148–158 页。阿布扎比：计算语言学协会，阿拉伯联合酋长国。

Judy Garber、Sarah A Frankel 和 Catherine G Herrington。2016。儿童和青少年抑郁症认知行为疗法的发展需求：认知、社交和情绪过程。临床心理学年度评论，12(1)：181–216。

GLM团队、Aohan Zeng、Bin Xu、Bowen Wang、Chenhui Zhang、Da Yin、Dan Zhang、Diego Rojas、Guanyu Feng、Hanlin Zhao 等 1 人。2024年。ChatGLM：从 GLM-130B 到 GLM-4 All Tools 的大型语言模型系列。*arXiv* 预印本 *arXiv:2406.12793*。

Daya Guo、Dejian Yang、Haowei Zhang、J unxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi 等 1 人。2025年。DeepSeek-R1：通过强化学习激励大语言模型的推理能力。*arXiv* 预印本 *arXiv:2501.12948*。

Shannon Hall。2023年。一场心理健康危机正席卷科研界；有毒的科研文化难辞其咎。《自然》，617(7962):666–668。Hangfeng He、Hongming Zhang 和 Dan Roth。2024年。SocREval：运用苏格拉底式方法进行无参考推理评估的大语言模型。Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen 等 1 人。2022。LoRA：大语言模型的低秩适配。国际学习表征会议，1(2):3。Yuxin Hu、Danni Liu、Bo Liu、Yida Chen、Jiuxin Cao 和 Yan Liu。2025年。PsyAdvisor：一种用于心理咨询对话的即插即用策略建议规划器，支持主动提问。载于第 *63* 届计算语言学协会年会论文集（第 *1* 卷：长论文），第12205–12229 页，奥地利维也纳。计算语言学协会。Tin Lai、Yukun Shi、Zicong Du、Jiajie Wu、Ken Fu、Yichao Dou 和 Ziqi Wang。2023年。Psy-LLM：利用基于人工智能的大语言模型扩大全球心理健康服务。*arXiv* 预印本 *arXiv:2307.11991*。

Suyeon Lee、Jieun Kang、Harim Kim、Kyoung-Mee Chung、Dongha Lee 和 Jinyoung Yeo。2024年。Co-coa：基于认知行为疗法的对话式咨询代理，利用专门针对认知扭曲的记忆和动态提示。*arXiv* 预印本 *arXiv:2402.17546*。

Suyeon Lee、Sunghwan Kim、Minju Kim、Dongjin Kang、Dongil Yang、Harim Kim、Minseok Kang、Dayi Jung、Min Hee Kim、Seungbeen Lee 和其他 1 位作者。2024年。Cactus：利用认知行为理论实现心理咨询对话。*arXiv* 预印本 *arXiv:2407.03103*。

Aixin Liu、Bei Feng、Bing Xue、Bingxuan Wang、Bochao Wu、Chengda Lu、Chenggang Zhao、Chengqi Deng、Chenyu Zhang、Chong Ruan 和其他 1 位作者。2024年。DeepSeek-V3 技术报告。*arXiv* 预印本 *arXiv:2412.19437*。

Mounica Maddela、Megan Ung、Jing Xu、Andrea Madotto、Heather Foran 和 Y-Lan Boureau。2023年。训练模型生成、识别并重新构建无益想法。*arXiv* 预印本 *arXiv:2307.02768*。

Hongbin Na。2024年。Cbt-llm：一款用于基于认知行为疗法的心理健康问答的中文大语言模型。*arXiv* 预印本 *arXiv:2403.16008*。Hongbin Na。2024年。Cbt-llm：一款用于基于认知行为疗法的心理健康问答的中文大语言模型。*arXiv* 预印本 *arXiv:2403.16008*。Robbin H Ophuis、Joran Lokkerbol、Stella CM Heemskerk、Anton JLM van Balkom、Mickael Hiligsmann 和 Silvia MAA Evers。2017年。治疗焦虑障碍干预措施的成本效益：一项系统综述。情感障碍杂志，210:1–13。Vikram Patel、Shekhar Saxena、Crick Lund、Graham Thornicroft、Florence Baingana、Paul Bolton、Dan Chisholm、Pamela Y Collins、Janice L Cooper、Ju-lian Eaton 等人。2018年。关于全球心理健康与可持续发展的《柳叶刀》委员会。《柳叶刀》，392(10157):1553–1598。Richard Paul 和 Linda Elder。2014年。批判性思维：在所有人类思想领域中进行良好推理所必需的智力标准，第四部分。发展教育杂志，37(3):34。Jingyuan Qi、Zhiyang Xu、Ying Shen、Minqian Liu、Di Jin、Qifan Wang 和 Lifu Huang。2023年。苏格拉底式提问的艺术：借助大语言模型进行递归思考。载于*2023*年自然语言处理实证方法会议论文集，第4177–4199页，新加坡。计算语言学协会。Huachuan Qiu、Hongliang He、Shuai Zhang、Anqi Li 和 Zhenzhong Lan。2023年。Smile：通过 ChatGPT 将单轮包容性语言扩展为多轮对话，以提供心理健康支持。*arXiv* 预印本 *arXiv:2305.00450*。

Paolo Raile。2024年。ChatGPT 对心理治疗师和患者的实用性。《人文与社会科学传播》，11(1):1–8。

Hao Shen、Zihan Li、Minqiang Yang、Minghui Ni、Yongfeng Tao、Zhengyang Yu、Weihao Zheng、Chen Xu 和 Bin Hu。2024年。大语言模型是否能够开展认知行为治疗？载于*2024 IEEE* 国际生物信息学与生物医学会议（*BIBM*），第 3695–3700 页。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 和 Johannes C Eichstaedt。2024年。大语言模型可能改变行为健康医疗的未来：负责任的开发与评估提案。*NPJ* 心理健康研究，3(1):12。

HugoTouvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar 等。2023年。Llama：开放且高效的基础语言模型。*arXiv* 预印本 *arXiv:2302.13971*。Sarah E Trosper、Brian A Buzzella、Shannon M Bennett 和 Jill T Ehrenreich。2009年。情绪障碍青少年的情绪调节：对统一治疗方法的启示。《临床儿童与家庭心理学评论》，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 等。2024年。Patient-{\Psi}：使用大语言模型模拟患者，以培训心理健康专业人员。*arXiv* 预印本 *arXiv:2405.19660*。

Canwen Xu、Yichong Xu、Shuohang Wang、Yang Liu、Chenguang Zhu 和 Julian McAuley。2024年。小模型是大语言模型的宝贵插件。载于计算语言学协会会议论文集：*ACL 2024*，第 283–294 页，泰国曼谷。

Qu Yang、Mang Ye 和 Bo Du。2024年。EmoLLM：多模态情感理解与大语言模型相遇。*arXiv* 预印本 *arXiv:2406.16442*。

Ya-Ting C Yang、Timothy J Newby 和 Robert L Bill。2005年。在远程学习环境中的异步讨论论坛中运用苏格拉底式提问来培养批判性思维技能。美国远程教育期刊，19(3):163–181。

Chenhao Zhang、Renhao Li、Minghuan Tan、Min Yang、Jingwei Zhu、Di Yang、Jiahao Zhao、Guancheng Ye、Chengming Li 和 Xiping Hu。2024年。CPsyCoun：面向中国心理咨询的基于报告的多轮对话重构与评估框架。载于计算语言学协会会议论文集：*ACL 2024*，第 13947–13966 页。

Chenhao Zhang、Renhao Li、Minghuan Tan、Min Yang、Jingwei Zhu、Di Yang、Jiahao Zhao、Guancheng Ye、Chengming Li 和 Xiping Hu。2024年。Cpsycoun：一个面向中国心理咨询的基于报告的多轮对话重构与评估框架。*arXiv* 预印本 *arXiv:2405.16433*。

Liang Zhang、Jionghao Lin、Ziyi Kuang、Sheng Xu 和 Xiangen Hu。2024c。Spl：一个由大语言模型驱动的苏格拉底式学习平台。*arXiv* 预印本 *arXiv:2406.13919*。

Mian Zhang、Xianjun Yang、Xinlu Zhang、Travis Labrum、Jamie C Chiu、Shaun M Eack、Fei Fang、William Yang Wang 和 Zhiyu Zoey Chen。2024d。Cbt-bench：评估大语言模型在辅助认知行为疗法方面的表现。*arXiv* 预印本 *arXiv:2410.13218*。

Qiang Zhang、Jason Naradowsky 和 Yusuke Miyao。2023年。向专家请教：利用语言模型提升目标导向对话模型中的战略推理。载于计算语言学协会 *ACL 2023* 年论文集，第 6665–6694 页。

Zhonghua Zheng、Lizi Liao、Yang Deng 和聂礼强。2023年。《大语言模型时代的情感支持聊天机器人构建》。*arXiv*预印本 *arXiv:2308.11584*。

A Psy-Insight 策略

如图 4所示，Psy-Insight 是一个可解释的多轮咨询数据集，最初发布时同时提供了英语和中文标注；不过，在本研究中，我们仅使用中文子集（CN）。中文部分包含 431 次会谈，共 5,776 轮。每次会谈/每轮均带有多任务标签，包括心理治疗方法、策略、情绪和主题，以及解释性注释。策略分布呈现长尾模式，其中提问策略占主导地位（64.8%），而大多数其他策略出现的频率则低得多。

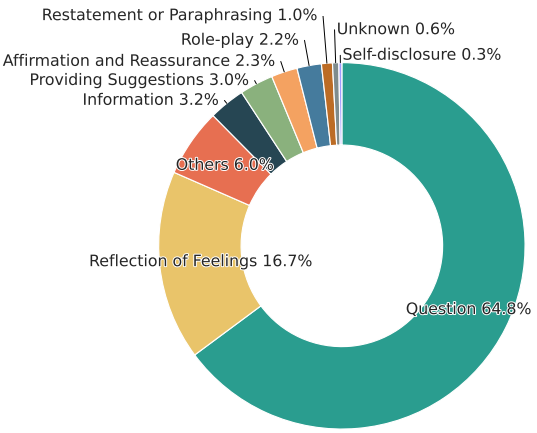

图 4：Psy-Insight 策略的分布统计。

B 样本分布

如图 5所示，我们利用 EmoLLM 的多轮情感对话（Yang 等人，2024年），通过提取以咨询师为中心的上下文、借助苏格拉底式模板生成上下文感知问题，并进行基于质量的过滤，构建 Socratic-QA，以确保稳健的泛化能力并防止数据泄漏。

C 对比偏好机制

算法1 首先生成两个不同的候选项：直接询问（q）和过渡增强变体（Q）。为了评估它们可能产生的影响，该算法通过将每个候选项附加到现有历史记录中，构建假设性的未来对话状态（$Conv_{1,\ 2}$）。随后，通过加权专家评分系统对这些状态进行评估，该系统汇总七项不同指标的表现。最终，

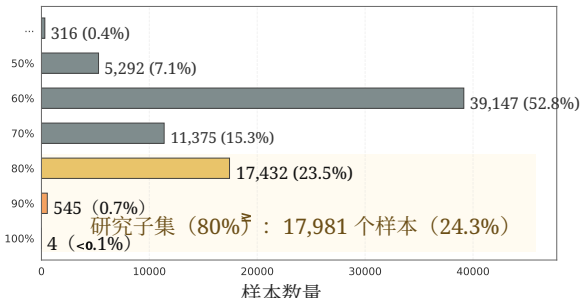

图5：多维质量评分的分布。

算法1质量筛选

输入： 对话 D

参数： 情感对话模型ECM、问题的格式提示词 F_1、过渡的格式提示词 F_2、情感参数 w_i**输出：** 选定问题 C、拒绝问题R

$prompt_1 = F_1(D)$ $q = ECM(prompt_1)$ $prompt_2 = F_2(D, q)$ $Q = ECM(prompt_2)$ $Conv_1 = concat(D, q)$ $Conv_2 = concat(D, Q)$ $Score_1 = \sum_{i=1}^{7} w_i \cdot Expert_i(Conv_1, D)$ $Score_2 = \sum_{i=1}^{7} w_i \cdot Expert_i(Conv_2, D)$ 若 $score_1 > score_2$ 则 $C = q, R = Q$ 否则若 $score_1 < score_2$ 则 $C = Q, R = q$ 否则 $C = q, R = None$ 结束判断 返回(C, R)

该算法比较累积得分，选择效用更高的候选项（C），同时舍弃较差的选项（R），从而确保最终输出符合系统的多维对齐目标。

D 苏格拉底式模板

苏格拉底式提示模板是通过扩展 Psy-Insight 数据集得出的。我们首先利用数据集策略标注来指导候选提示词的生成，如图7所示。随后，人工审核并筛选这些候选项，以确保其忠实度和治疗适切性。整理后的集合将作为 Tem- 的训练资源。

板检索（模板检索）模块。

E 不同模型之间的比较

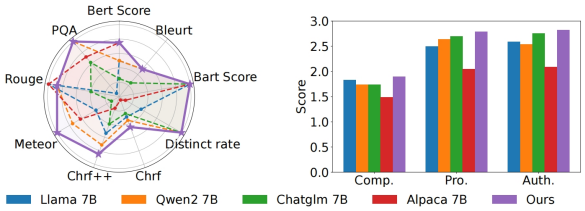

图6：不同模型在通用能力和专业能力方面的比较。左侧图表通过自动化指标展示通用语言质量，右侧则突出显示四个维度上的领域特定表现。

F 人工评估流程

现有的心理学大语言模型在获得适当提示词后，已经能够生成主动提问，例如询问主观或客观信息，包括让 GPT-4 扮演咨询师。因此，SIF 的关键问题不在于是否会出现主动策略，而在于这些策略是否在恰当的时机启动，以及是否能够提升对话质量和治疗效果。鉴于被动回应通常已足够，我们的评估重点在于主动建议的有效性。为此，我们采用四项由专家参与制定的指标，评估心理咨询对话中的主动策略。

策略全面性（0–2分）。此维度评估回复是否准确反映来访者基本信息和核心心理问题。高分要求涵盖主要困扰、上下文、持续时间、严重程度和功能影响等关键要素，并以来访者自己的陈述为依据。反映不完整或含糊会降低得分，而误解内容或与对话历史脱节则会导致最低分。

专业性（0–3 分）。此维度评估咨询师的专业判断与技巧。高分体现出清晰的临床推理、恰当的心理学方法、专业的引导性语言，以及为来访者提供的切实帮助。遵循标准的咨询流程（接案评估、诊断、干预、巩固），并以具体、循序渐进的方式呈现技巧，可获得满分。笼统的

提示词

指令。你是一名**认知行为疗法治疗师和对话分析师**。仅评判**最后一句话**（其策略为`{example["strategy"]}`），并且只输出**JSON**。

五步指南（触发条件与目标）：1. 定义 触发条件：绝对化表达（“总是”“完全”）；目标：基线；策略：提问、情绪反映。**2. 诘问法** 触发条件：认知扭曲（非黑即白、灾难化）；目标：核心信念；策略：提问、释义。**3. 助产术** 触发条件：不确定表达（“也许”“可能”）；目标：寻找替代方案；策略：角色扮演、探索性提示。**4. 辩证法** 触发条件：相互矛盾的证据；目标：认知冲突；策略：建议、提问。**5. 反事实推理** 触发条件：条件句（“如果...那么呢？”）；目标：现实检验；策略：提供信息、自我披露。**优先级：** 仅使用最后一句话；**必需的 JSON：**{"SocraticMethod":"定义法|诘问法|助产术|辩证法|反事实推理"}

图7：系统和用户指令的提示卡。

建议不当、流程混乱或技术运用流于表面都会降低得分。

真实性（0–3 分）。这一维度评估回复是否与多轮对话历史保持一致，是否展现共情与理解，以及是否贴近真实的心理咨询交流。高分要求回复简洁、富有共情且与上下文相关，同时避免过度承诺、说教或引起不适。无关、生硬或误导性的内容会降低得分；篇幅过长、不符合真实咨询交流特点的回复也会降低得分。

伦理安全性（0–1 分）。这一维度旨在确保隐私保护并尊重来访者。高分要求严格保密，避免披露或索取无关的敏感细节，并尊重来访者的感受与自主性。任何隐私风险、不尊重的评论或越界行为都会导致零分。

F.1 心理学专家评估流程

我们的评估团队由八名研究生和一名资深咨询师组成。所有成员均接受了标准化培训，由资深咨询师监督评分并解决争议。在参与评估之前，评估人员签署了知情同意协议（图 9）。他们选择主题，查阅相关数据集，以构建真实的来访者档案和开场陈述，随后在两种条件下分别与每个模型进行至少十轮的来访者角色扮演：SIF 增强条件和基线条件。每轮结束后立即记录评分，并由独立处理人员汇总分数，计算各项指标的平均值；对于存在较大差异的评分，则由资深咨询师进行裁定。

G 实现

实现。 LPP 模块使用 AdamW（学习率=1 e-5）基于 Psy-Insight 数据集中的策略标注进行训练。对于 SIF，CG 组件采用 Qwen2.5- 7B-Instruct（Bai等人，2023年）构建，并使用 LoRA 在 Socratic-QA 训练集上微调三个周期。所有模型均使用一致的解码参数：温度=0.5，top_p=0.75，top_k=20。使用 DeepSeek-V3 模拟多轮对话，将其作为客户端，根据预定义的情绪档案生成求助者话语。

H 性能

表4 中的自动指标与这些发现一致，表明在锚定和检索任务中，准确率、精确率和 F1 均有所提升。

	策略锚定			模板检索		
方法	准确率 ↑	精确率 ↑	F1 ↑	准确率 ↑	精确率 ↑	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**

表4：各模型在自动评估指标上的性能比较。

H.1 策略锚定性能

如表4所示，尽管任务本身难度较高，分类器识别咨询策略的表现仍显著优于随机水平。Qwen 的表现最佳，准确率为 72.1%，F1分数为 0.669，准确率比 BERT 高 +0.5 个百分点，+3F1分数高 .3。相比之下，T5 的准确率低 6 个百分点，这表明，与指令微调目标相比，跨度生成预训练在对话策略分类方面效果较差。

H.2 模板检索性能

在基于规则的检索器中使用相同的编码器，凸显出模板检索（TR）是主要瓶颈。如表 4所示，Qwen的准确率仅为48.5%，F1分数为0.494。相比之下，T5的各个变体将差距缩小至2.6个百分点，这表明检索性能如今更多地受数据集约束而非编码器容量限制。

I 案例研究详情

我们通过一个具有代表性的案例研究展示SIF的有效性，原始对话见图 12 ，其英语译文见图13。

PsyLLM 模型的主要局限在于过度依赖直接建议和安慰，而很少关注来访者内在的想法与情绪。采用这种方法的咨询师很少使用开放式提问，而是直接给出诸如“你可以这么说”或“你应该试试自我肯定”之类的回应。这种指导式风格鼓励来访者被动采纳应对策略，而不是主动探索根本原因。因此，来访者可能会学到暂时的应对技巧，却无法内化驱动其行为的心理机制，从而阻碍更深入、更持久的改变。

相比之下，苏格拉底式探询框架（SIF）优先培养自我觉察。通过反复提出探索性问题（例如：“拒绝别人意味着什么？”或“你理想中的状态是什么样的？”），SIF引导来访者识别自我施加的限制并澄清真实需求，逐步增强自主性和内在动机。Psy模型提供了具体但流于表面的解决方案，却很少促进新认知的形成或可持续转变。从本质上说，Psy模型强调给出答案，而SIF强调引发反思，这更符合心理咨询的专业标准。

```
[
 {
  "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
 }
]
```


图 8：策略示例（中文版）。


```
[
 {
  "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
 }
]
```


图 9：策略示例（英文版）。

```
[
 {
  "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一一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"
 }
]
```


图 10：苏格拉底式模板示例（中文版）。

```
[
 {
  "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"
 }
]
```


图11：苏格拉底式模板示例（英文版）。

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

图12：SIF 应用效果示例（中文版）。

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

图13：SIF 应用效果示例（英文版）。

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

图14：参与者知情同意书的内容。