AutoCBT: An Autonomous Multi-agent Framework for Cognitive Behavioral Therapy in Psychological Counseling

Anonymous ACL submission

Abstract

Traditional face-to-face psychotherapy remains a niche practice, typically sought by individuals experiencing psychological distress. Online mental health consultation forums offer a viable alternative for those hesitant to seek help. In this context, large language models (LLMs) and cognitive behavioral therapy (CBT) jointly facilitate the development of automated online mental health consultation platforms. However, many automated mental health systems rely on rigid, rule-based agent workflows or single-prompt LLM responses, resulting in generic advice that lacks empathy and contextual awareness. Inspired by the single-turn consultation style commonly found in online forums—where users, unlike in real-time multi-turn chat scenarios, are more willing to wait longer for thoughtful and in-depth replies—we developed **AutoCBT**, an autonomous multi-agent framework designed to improve the quality of automated mental health consultations. AutoCBT is built for single-turn consultation scenarios and introduces dynamic routing and supervisor mechanisms to generate high-quality responses. Our research shows that AutoCBT consistently outperforms baseline models on key psychotherapy metrics, including empathy, cognitive distortion detection, and response relevance. Furthermore, we identify key challenges in implementing a multi-agent consultation framework, such as routing inconsistencies and LLM safety constraints. Our findings underscore the potential of AutoCBT as a scalable and effective AI-driven approach to mental health support.

1 Introduction

The rapid advancement of computer technology—especially the emergence of Large Language Models (LLMs) (Demszky et al., 2023; Zhao et al., 2025)—has greatly propelled the growth of online forums and automated mental health counseling (Althoff et al., 2016). For individuals hesitant to pursue face-to-face therapy, online forums offer a platform to pose questions and receive detailed, thoughtful responses. Among various approaches, Cognitive Behavioral Therapy (CBT) is particularly effective in addressing conditions such as anxiety and depression, as emphasizes identifying and challenging cognitive distortions (Beck, 1979, 1993). However, current LLM-based counseling systems often fall short in replicating the nuanced reasoning exhibited by human counsellors (Wang et al., 2024). Many automated systems rely on rigid rule-based agents or single-prompt LLMs, frequently producing generic advice that lacks both empathy and contextual sensitivity. This gap highlights the urgent need for adaptive, context-aware AI counsellors capable of delivering high-quality support in single-turn scenarios (He et al., 2023).

In contrast to multi-turn dialogues centered on a single topic, single-turn counseling on online forums (e.g., Quora[1], Zhihu[2], YiXinLi[3]) presents three key differences: **First**, in single-turn scenarios, counsellors cannot rely on users to ask follow-up questions or provide additional context. Therefore, all critical content must be effectively conveyed in a single response. **Second**, real-time conversations require models to respond swiftly and appear human-like, often at the expense of deep reasoning. In contrast, forum users value depth over immediacy, enabling AI counsellors to take the rare opportunity—unavailable in live chats—to deliberate internally for minutes or even hours before composing a single, self-contained response. This is especially critical in high-risk contexts such as mental health counseling, where models must rigorously complete all safety checks and ethical filters before delivering a final response. **Lastly**, forum interfaces typically show only the user's original question and ranked counsellor responses, whereas

[1] https://www.quora.com
[2] https://www.zhihu.com
[3] https://www.xinli001.com/qa

in multi-turn dialogues, intermediate exchanges are often collapsed or omitted, hindering the review of essential reasoning chains by readers or moderators. Given these characteristics, thoughtfully crafted, high-quality single-turn responses align more closely with the expectations of both forum administrators and users.

Motivated by these insights, we developed **AutoCBT**—a multi-agent framework specifically tailored for online forum counseling. When a forum user submits a question, the counsellor agent in AutoCBT drafts a response and determines whether to consult a supervisor agent—for review focused on empathy, distortion identification, strategy formulation, or safety—or respond directly to the user.

This approach contrasts with generic self-revision methods, where an LLM merely rewrites a response using a single prompt lacking clearly defined objectives. AutoCBT incorporates a clearly structured, multi-objective CBT optimization framework in which multiple supervisor agents—each adhering to distinct and orthogonal CBT principles—offer specific suggestions. The counsellor then synthesizes these suggestions into a refined, context-rich response, of which only the final version is presented to the user.

Due to privacy constraints and the high cost of manual annotation, large-scale CBT datasets remain scarce. Consequently, we curated a bilingual test set of 200 single-turn counseling cases (100 in Chinese and 100 in English), encompassing diverse cognitive distortions and mirroring real-world forum concerns. This dataset effectively demonstrates the advantages of AutoCBT's latency-aware routing: automated evaluation metrics affirm AutoCBT's superior empathy, relevance, and root-cause analysis of users' psychological distress, while expert evaluations reveal a clear preference for AutoCBT's responses over baseline systems by six psychologists.

Our contributions are as follows:

1. We introduce a dataset annotated both automatically using GPT-4o and manually by six psychology experts, **enabling rapid evaluation of single-turn counseling frameworks**.

2. We present AutoCBT, a dynamic, routing-based autonomous multi-agent framework. Experimental results indicate that **AutoCBT outperforms baseline systems when implemented on LLaMA and Qwen models**.

3. Finally, we outline the challenges encountered during the development of AutoCBT and **provide practical solutions to support future reproduction of this work**.

2 Related Work

CBT is a widely recognized and effective treatment for mental health conditions such as anxiety, depression, and addiction. A fundamental aspect of CBT is helping individuals identify and restructure cognitive distortions—biased or irrational thought patterns that lead to misinterpretations of reality and contribute to negative emotions and maladaptive behaviors (Beck, 1963, 2020). These distortions often create reinforcing cycles of unhealthy thinking, making it essential to provide effective strategies for their recognition and modification (Beck, 1979).

With advancements in Artificial Intelligence (AI), researchers have explored methods to automate the detection of cognitive distortions and enhance CBT-based interventions. Annotated datasets and CBT-based ontologies have been developed to facilitate AI-driven cognitive distortion analysis (Rojas-Barahona et al., 2018; Wang et al., 2023).

Early computer-based CBT systems sought to make therapeutic care more accessible. For instance, virtual therapists for depression counseling (Ring et al., 2016) and chatbot-based interventions like Woebot (Fitzpatrick et al., 2017) employed decision-tree-based responses. However, these systems relied on predefined scripts, limiting their ability to engage in flexible, natural conversations.

With the advent of LLMs, researchers have begun leveraging advanced AI to enhance CBT delivery. Frameworks such as CBT-LLM (Na, 2024) employ prompt-based in-context learning to analyze user questions and generate therapeutic responses. Similarly, CoCoA (Lee et al., 2024) integrates memory mechanisms for retrieval-augmented response generation and applies CBT techniques to detect cognitive distortions in user statements. Furthermore, studies have examined LLM therapist behaviors in simulated therapy interactions (Chiu et al., 2024).

3 Methodology

Our framework is illustrated in **Figure 1**. AutoCBT is a general framework designed to serve

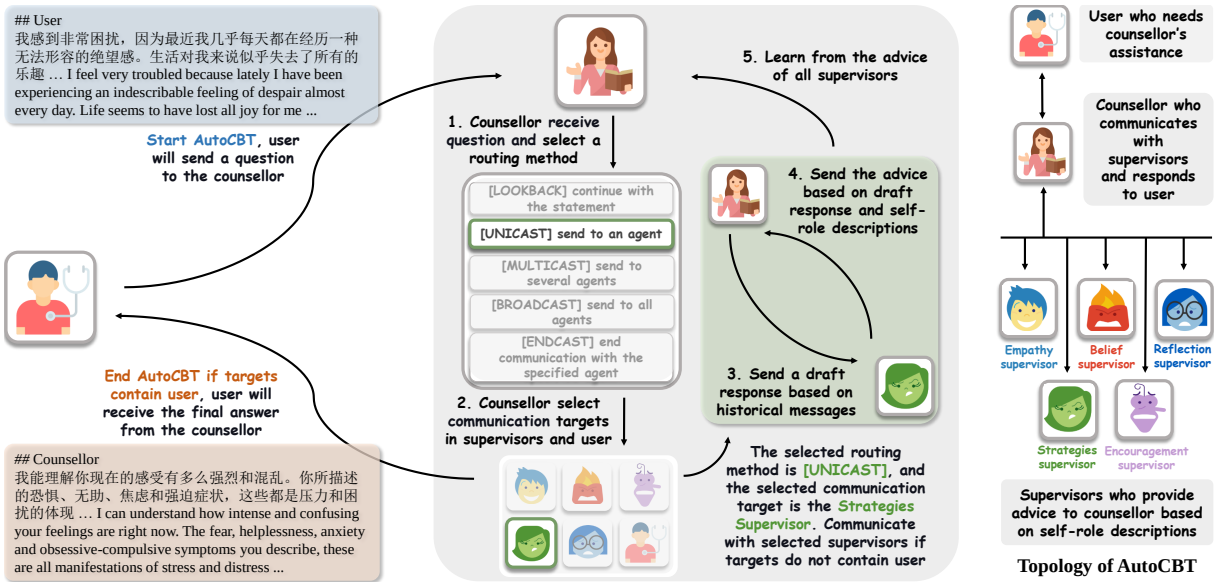


Figure 1: An overview of the AutoCBT framework. Upon receiving a user's query, the Counsellor Agent employs dynamic routing to determine whether to respond directly or seek guidance from a Supervisor Agent.

as a proxy for various multi-agent systems in the backend. The framework is formally represented as $(a_0, S, \mathcal{T}, \Sigma)$, where:

- a_0 is the **Counsellor Agent**, acting as the primary interface between the user and the supervisors.
- $S = \{a_i | i \in [1, N]\}$ represents the set of **Supervisor Agents** from which the Counsellor Agent can request additional information.
- $\mathcal{T}$ defines the **topology** of communicable agents.
- Σ denotes the set of **permitted routing strategies** among agents.

3.1 AutoCBT

Counsellor Agent This agent is an interface for the AutoCBT which acts as the interface between the users with psychological confusion (either simulated users or real users from the web) and candidate supervisors.

Supervisor Agents These agents can generate advice based on self-role descriptions and the counsellor's draft response. Their number and the way they are connected can be adjusted according to the adopted CBT approach.

Memory Mechanisms Each agent is accompanied by a short term memory to store most recent messages and a long-term memory to store summaries of messages with a sliding window. The detailed workflow is illustrated in **Appendix F**.

Topology of Agents In AutoCBT, a topology is the graph (either static or dynamic) of communicable agent pairs. Messages can be transported over the topology but may endue subsequent modifications at each agent.

Routing Strategies The routing strategies are defined for the communicable agents in the topology:

1. `[LOOPBACK]` Loop back, continue with the statement.
2. `[UNICAST]` Unicast, send to a communicable agent.
3. `[MULTICAST]` Multicast, send to several communicable agents.
4. `[BROADCAST]` Broadcast, send to all communicable agents.
5. `[ENDCAST]` Terminated casting, end communication with the specified agent.

A detailed description of the agents is provided in the **Appendix A**.

Two stage generation process AutoCBT employs a structured two-stage reasoning mechanism:

- Draft Response Process: The Counsellor Agent generates an initial response and evaluates whether additional supervisory input is required.

- Final Response Process: If needed, the Counsellor Agent consults a specialized Supervisor Agent, who refines the response based on CBT principles before delivering it to the user.

This two-step approach enables AutoCBT to iteratively refine its responses, aligning them with established CBT therapeutic guidelines.

3.2 Decomposition of CBT Core Principles

The CBT core principles can be divided into five standards: **Validation and Empathy**, which ensures responses acknowledge and validate the user's emotions; **Identify Key Thought or Belief**, which detects cognitive distortions in user statements; **Pose Challenge or Reflection**, which encourages users to critically assess their thought patterns; **Provide Strategy or Insight**, which offers actionable coping mechanisms; and **Encouragement and Foresight**, which reinforces positive thinking and future planning.

In our framework in **Figure 1**, these five standards are mapped onto five Supervisor Agents, each specializing in one standard. During inference, when the Counsellor Agent receives a question from the user, it determines whether to seek advice from one or multiple Supervisor Agents based on the message context and its memory. This iterative process continues until the Counsellor Agent has sufficient information to generate a final response for the user.

3.3 Dataset Construction

To validate AutoCBT's effectiveness across multiple languages, we construct a **bilingual counseling dataset with 200 samples** from two existing psychological counseling datasets:

- **PsyQA** (Sun et al., 2021): A Chinese-language dataset designed for psychological question-answering.

- **TherapistQA**(kaggle, 2019): An English-language dataset featuring therapeutic responses from licensed professionals.

Considering the diversity of the data, we adopted a **classification-then-sampling** approach. First, we collected all the PsyQA and TherapistQA data. Then, using the Qwen2.5-72B and Llama3.1-70B models, we semantically categorized the PsyQA and TherapistQA datasets into 10 distinct classes each. From each of these 10 classes, we randomly sampled 10 entries, resulting in 100 entries for each of the Chinese and English bilingual datasets. Examples of dataset entries can be found in **Appendix B**.

4 Experiments

4.1 Setup

Models For effective psychological counseling, language models must accurately interpret user intent, recognize emotional nuances, and adhere to structured intervention techniques. We employ two state-of-the-art LLMs for this task: **Qwen-2.5-72B** for both Chinese and English counseling sessions and **Llama-3.1-70B** for English-only sessions. The temperature parameter is set to 0.98 to balance creativity and consistency in responses, with all other hyperparameters kept at their default settings.

Baselines To establish comparative performance, we evaluate AutoCBT against two baseline approaches:

- **Generation**: LLMs generate responses directly to bilingual dataset questions without CBT-specific guidance.

- **PromptCBT**: CBT principles are embedded within the input prompts before response generation, ensuring LLMs incorporate CBT techniques implicitly.

We have also examined multi-agent frameworks such as CAMEL(Li et al., 2023) and AutoGEN(Wu et al., 2023). However, these frameworks differ significantly from our AutoCBT setting. First, they do not natively support long-term and short-term memory compression, requiring manual implementation of such mechanisms. Second, these frameworks operate strictly according to predefined workflows that terminate upon success or failure, without granting the counsellor agent the autonomy to decide when to interrupt or terminate the entire process. For these reasons, we did not include them as baselines in our experiments.

4.2 Evaluation Methodology

In designing our evaluation metrics, we also referred to MEEP (Ferron et al., 2023). We found that this work primarily focuses on evaluating

Perspective	Description	Criterion	Score
Empathy	Demonstrates understanding and sympathy towards the user's emotions or issues, and creates a sense of safety.	1.1 Did the counsellor correctly understand the user's intent? 1.2 Did the counsellor show respect, understanding, and sympathy for the user's anxiety and pain? 1.3 Did the counsellor create a safe environment for the user to express their feelings?	7
Identification	Identify potential cognitive distortions of the user through the description of the problem in the dialogue.	2.1 Did the counsellor identify the user's distorted beliefs? 2.2 Did the counsellor delve into the user's distorted beliefs? 2.3 Did the counsellor assist the user in recognizing and challenging these distorted beliefs?	7
Reflection	Ask open-ended questions to encourage the user to reconsider or reflect on their initial thoughts or beliefs.	3.1 Did the counsellor ask questions related to the user's initial thoughts? 3.2 Did the counsellor pose questions that facilitated deeper thinking? 3.3 Did the counsellor ask questions reflecting the user's distorted beliefs?	7
Strategy	Provide practical strategies or insights to help the user address their current situation.	4.1 Were the strategies or insights provided by the counsellor practical? 4.2 Could the strategies or insights solve the user's current problems? 4.3 Were the strategies based on professional psychological methods?	7
Encouragement	Encourage the user to use the strategies.	5.1 Did the counsellor encourage the user to take action? 5.2 Did the counsellor address potential failures the user might encounter while implementing the strategies? 5.3 Did the counsellor provide comfort and encouragement regarding setbacks and challenges?	7
Relevance	Evaluate the relevance of the dialogue content.	6.1 Was the counsellor's response highly relevant to the user's question? 6.2 Did the counsellor's response flow naturally? 6.3 Did the counsellor's answer cover the main issues or concerns raised by the user?	7

Table 1: Six automatic evaluation metrics and corresponding score criterion based on the CBT core principles.

general dialogue quality from the perspectives of Interactional-Quality and Interestingness. However, **critical dimensions in psychological counseling—such as empathy, encouragement, and reflection—are not included in MEEP**. Therefore, we propose the following evaluation criteria that are more appropriate for single-turn psychological counseling dialogues:

Automatic Evaluation For automated assessments, we utilize **GPT-4o-mini** to score responses according to six predefined evaluation metrics outlined in **Table 1**. Each response is independently evaluated **three times** to mitigate the impact of randomness in the LLMs' token outputs, with the final score being the average of these ratings.

This process ensures robust statistical evaluation and reduces bias introduced by outlier responses.

Human Evaluation Although automatic evaluation provides a standardized scoring mechanism, it does not fully capture the complexity of human cognitive distortions and nuanced therapeutic interactions. To address this, we develop a **human evaluation framework** that focuses on in-depth assessment of AutoCBT's ability to detect and challenge cognitive distortions. Details of this framework are provided in **Appendix C**.

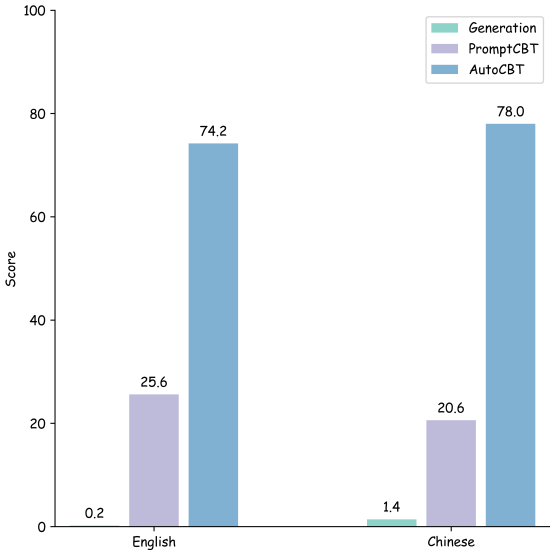
Compared to automated scoring, human evaluation emphasizes the qualitative aspects of counseling, such as the appropriateness of therapeutic interventions and the emotional resonance of responses. We conduct two human evaluation experiments:

- **Simple Overall Evaluation (SOE)**: Five psychology professionals review all responses in the bilingual dataset and select the most effective response per question among AutoCBT and the baselines.

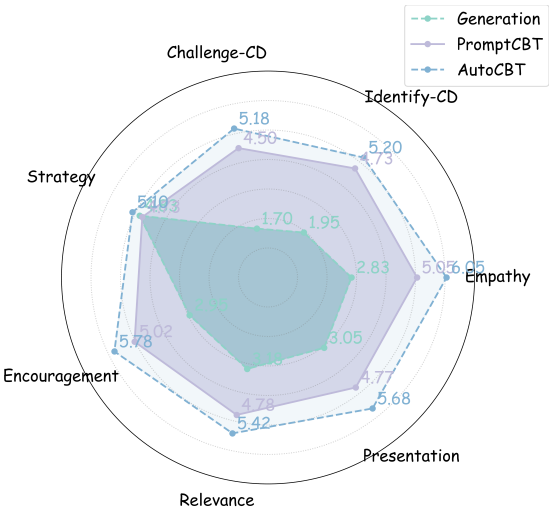
- **Detailed Sampling Evaluation (DSE)**: Six psychology professionals evaluate a subset of 60 responses (10% of the bilingual dataset), assessing them across seven key dimensions,

Model	Lang.	Method	Empathy	Cognitive Distortions		Strategy	Encouragement	Relevance	Total Score
				Identification	Reflection				
Qwen	ZH	Generation	5.493 / 7	4.630 / 7	4.280 / 7	6.153 / 7	5.200 / 7	6.543 / 7	32.300
		PromptCBT	6.000 / 7	5.610 / 7	5.623 / 7	6.237 / 7	6.130 / 7	**6.860 / 7**	36.460
		AutoCBT	**6.247 / 7**	**5.760 / 7**	**5.787 / 7**	**6.363 / 7**	**6.447 / 7**	6.857 / 7	**37.460**
Qwen	EN	Generation	5.907 / 7	4.903 / 7	4.740 / 7	6.093 / 7	5.383 / 7	6.637 / 7	33.663
		PromptCBT	6.390 / 7	5.687 / 7	5.797 / 7	6.233 / 7	6.377 / 7	6.887 / 7	37.370
		AutoCBT	**6.650 / 7**	**5.830 / 7**	**5.983 / 7**	**6.440 / 7**	**6.560 / 7**	**6.913 / 7**	**38.377**
Llama	EN	Generation	6.055 / 7	5.267 / 7	5.161 / 7	6.059 / 7	5.549 / 7	6.718 / 7	34.810
		PromptCBT	6.377 / 7	5.678 / 7	5.886 / 7	5.879 / 7	6.103 / 7	6.799 / 7	36.722
		AutoCBT	**6.513 / 7**	**5.780 / 7**	**5.996 / 7**	**5.908 / 7**	**6.227 / 7**	**6.909 / 7**	**37.333**

Table 2: AutoCBT's performance on Qwen-2.5-72B and Llama-3.1-70B using the bilingual dataset. For a more detailed analysis about Llama refer to Section 4.6.



(a) Bilingual language performance in SOE.



(b) The performance of English in the DSE experiment, **Figure 4b** shows the results for Chinese.

Figure 2: AutoCBT generates better answers than both PromptCBT and Generation for over 70% of the bilingual dataset questions and outperforms both PromptCBT and Generation in identifying and challenging cognitive distortions.

including empathy and coherence. This method allows for a deeper analysis of model performance beyond standard automatic evaluation metrics.

At the same time, in order to better differentiate between the various scores and because individuals with higher education are more suited to a **7-point scale** (Robinson, 2018), we adopted the 7-point scale for evaluation.

4.3 Results

Automatic Evaluation The observed scores for responses in the Chinese section of the dataset are presented in **Table 2**. When comparing the performance of **Generation** and **PromptCBT**, it is evident that incorporating core CBT principles significantly enhances the quality of LLM-generated responses. Furthermore, AutoCBT, which leverages a structured multi-agent approach based on CBT principles, consistently produces higher-quality responses than PromptCBT, outperforming it in 5 out of 6 evaluation metrics.

For the English section of the dataset, the results indicate that AutoCBT surpasses both baseline methods across all six evaluation metrics and overall English proficiency, as assessed using the Llama and Qwen models. Notably, AutoCBT's superior performance in English closely aligns with its effectiveness in Chinese, highlighting the framework's robustness across languages.

Human Evaluation AutoCBT's effectiveness is further validated in human evaluation experiments. As shown in **Figure 2a**, results from the SOE indicate that AutoCBT provides the most preferred

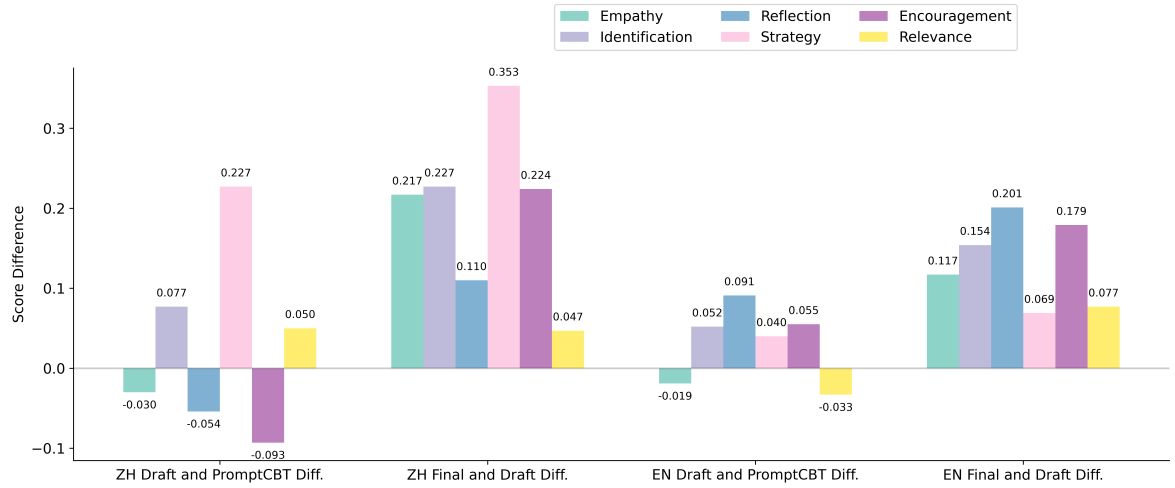


Figure 3: **Draft and PromptCBT Diff.** represents the score difference between AutoCBT's draft responses and similar PromptCBT responses; **Final and Draft Diff.** indicates the improvement in quality score of AutoCBT's final responses compared to AutoCBT's draft responses. It can be clearly observed that the quality of the Draft responses and the PromptCBT baseline shows mixed results, with **no significant overall difference between the two**. In contrast, the Final responses consistently outperform both the Draft responses and the PromptCBT baseline across all six evaluation dimensions. This indicates that AutoCBT's two-stage process—**first simulating the baseline via Draft responses, then further enhancing performance through Final responses**—is a key factor behind its superior performance compared to the baselines.

response for over 70% of questions in the bilingual dataset. Additionally, **Figure 2b** presents the results from the DSE, where psychology professionals systematically assessed AutoCBT's ability to identify and challenge cognitive distortions. Across all seven evaluation dimensions related to cognitive distortions, AutoCBT consistently outperforms both baseline methods, reinforcing its ability to generate contextually appropriate and therapeutically effective responses.

A more detailed qualitative analysis of the differences between Generation, PromptCBT, and AutoCBT is provided in **Appendix D**.

4.4 Simulating and Surpassing Baselines

From the **Draft and PromptCBT Diff.** in Figure 3, we observe that the quality of AutoCBT's draft responses is comparable to that of PromptCBT in the baselines. Since the prompt used for the draft response is flexible and configurable, AutoCBT's draft responses can effectively simulate the behavior of PromptCBT in the baselines. In the **Final and Draft Diff.**, we observe that the quality of the final responses surpasses both the draft responses and PromptCBT across all evaluation dimensions. This demonstrates that the dynamic routing and supervisory mechanisms in AutoCBT can significantly enhance the quality of psychological counseling responses.

AutoCBT achieves performance beyond traditional prompt-only approaches through this "**simulate first, then surpass**" strategy. We aim to extend this paradigm to other domains in order to validate its generalizability.

4.5 Challenges in AutoCBT

Simultaneous Routing In psychological counseling, a Counsellor Agent must decide whether to engage with the user or consult a supervisor, but these decisions are mutually exclusive—it cannot simultaneously select both "improve dialogue" and "end dialogue". However, despite the large parameter sizes of LLMs (e.g., 70B+), they still exhibit limitations in semantic understanding and logical reasoning, leading to conflicting routing objectives.

To address this, we modified the routing logic: if the Counsellor Agent simultaneously selects both the user and a Supervisor Agent as routing targets, the system prioritizes session termination to prevent looping behavior.

Addressing the simultaneous routing issue is critical for ensuring logical consistency in AI-driven therapy. Without intervention, LLMs may create looping behaviors that erode user trust and reduce conversational quality. By enforcing exclusive routing in conflicting scenarios, AutoCBT improves decision-making reliability, ensuring smoother therapeutic interactions.

Role Confusion When the Counsellor Agent determines that a response requires improvement, it forwards the query to a Supervisor Agent for guidance. However, in some cases, the Supervisor Agent mistakenly generates a direct response instead of providing feedback, causing confusion in the Counsellor Agent, which expects guidance rather than a complete answer. When historical responses contain prior advice, LLMs are more prone to role misinterpretation.

We introduced a modification in the Supervisor Agent's prompt: it now explicitly begins each response with **"Hello Counsellor"** to reinforce its advisory role and minimize misinterpretation, ensuring that responses remain in line with the expected supervisory function.

Preventing role confusion enhances interpretability and safety in AI-mediated counseling. Misalignment in counsellor-supervisor interactions could mislead users or result in inappropriate recommendations.

Routing Loop In real-world psychological counseling, repeated requests for advice on the same issue are uncommon. However, in our system, due to LLMs' limited semantic tracking and instruction-following abilities, the Counsellor Agent may unintentionally send multiple redundant requests to the same Supervisor Agent. **To mitigate this, we implemented a dynamic edge removal strategy**: when agent A sends a message to agent B, the directed edge $A \rightarrow B$ is removed from the topology graph, preventing repeated requests to the same Supervisor Agent. This ensures that each Supervisor Agent is accessed only once per query session. Given N Supervisor Agents and one user, the Counsellor Agent is restricted to a maximum of $N + 1$ routing operations, maintaining system efficiency while reducing redundant loops.

Eliminating redundant advice requests is crucial for efficiency and scalability in AI-driven counseling. The dynamic edge removal strategy ensures each query follows an optimized path, reducing unnecessary computation while maintaining structured decision-making. Future implementations could explore reinforcement learning-based routing mechanisms to make agent collaboration more context-aware and adaptive.

4.6 Over-Protection in Llama

During psychological counseling simulations, we observed that the Llama model refuses to answer nine questions from the English section of the dataset, particularly those related to minors, sex, and suicide. This behavior was consistent across both AutoCBT and baseline methods using Llama. In contrast, the Qwen model successfully responded to all questions in the bilingual dataset. The reject status is in the **Appendix E**.

While Llama's over-protection mechanism aims to prevent AI-generated harm, excessive refusal to engage in sensitive topics can be detrimental to users in distress. A more balanced approach may involve confidence-based response generation, where the LLM partially engages in sensitive topics but refers high-risk cases to human professionals.

5 Conclusion

This study first introduces the differences between single-turn and multi-turn psychological counseling in online forums. Inspired by these differences, we propose AutoCBT, a multi-agent framework based on CBT designed to enhance psychological counseling. Additionally, we have collected a bilingual dataset that enables rapid verification of single-turn counseling quality. By incorporating dynamic routing and a supervisory mechanism, AutoCBT enhances the quality of traditional counseling dialogues based on LLMs, focusing on identifying and addressing cognitive distortions in users. Experimental results demonstrate that AutoCBT significantly outperforms purely prompt-based counseling frameworks, providing more structured and contextually appropriate responses, and is preferred by psychological professionals. We also analyze the two-stage process of AutoCBT and how it outperforms traditional baselines using the "simulate first, then surpass" approach. This study systematically analyzes the limitations of large language models, including difficulties in instruction-following, role confusion, and inefficient routing, and proposes solutions to address these issues. These findings not only improve the performance of AutoCBT but also offer insights into enhancing the collaboration of multi-agent LLMs for applications in other fields. We believe that AutoCBT marks an important step towards scalable AI-driven psychological support systems. It can complement traditional mental health services, making counseling more accessible and personalized, thus benefiting a broader audience.

Limitations

AutoCBT enhances performance compared to purely prompt-based psychological counseling methods. However, interactions between the counsellor and supervisors in AutoCBT increase token consumption. These interactions lead to two issues: longer memory texts due to repeated conversations and a higher likelihood of invalid routes, as LLMs often deviate from the specified format, requiring rerouting until corrected. To address this, AutoCBT incorporates a memory window that condenses older conversations, retaining only the most recent dialogue and significantly reducing token consumption.

While all goals of psychological counseling aim to address the user's psychological needs, every response must meet an implicit precondition: ensuring safety, non-harmfulness, and preventing retraumatization. If this safeguard is compromised, counseling loses its foundational meaning. We believe that by providing higher-quality consultation responses compared to the baselines, AutoCBT implicitly aligns with the safety and harmlessness requirements.

However, even the smallest security lapse can have serious consequences, particularly when deploying LLM-based systems to real users, highlighting the importance of enhancing safety measures. Due to the flexible nature of the AutoCBT framework, we propose adding a new supervisor—the Safety Supervisor—who will focus on reviewing the content for potential harm, addressing safety gaps overlooked by existing supervisors, and further minimizing the risk to users.

Although no additional experiments have been conducted to validate the Safety Supervisor, previous results showing that each supervisor improves AutoCBT's response quality within their designated oversight metrics suggest that the Safety Supervisor would effectively enhance the security of counsellor-generated content.

Ethical Considerations

Based on the data copyright protocols delineated by the PsyQA (Sun et al., 2021), we will release our dataset for research purposes only. All questions from online mental health forums have been anonymized to protect participant privacy. We work with annotators to repeatedly check the rules and details of annotations to ensure their accurate understanding. Furthermore, the responses are generated by LLMs, not professionals. Therefore, this work cannot provide therapeutic recommendations or diagnostic statements.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Aaron T Beck. 1963. Thinking and depression: I. idiosyncratic content and cognitive distortions. *Archives of general psychiatry*, 9(4):324–333.

Aaron T Beck. 1979. *Cognitive therapy and the emotional disorders*. Penguin.

Aaron T Beck. 1993. Cognitive therapy: past, present, and future. *Journal of consulting and clinical psychology*, 61(2):194.

Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *Preprint*, arXiv:2401.00820.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, and 1 others. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Amila Ferron, Amber Shore, Ekata Mitra, and Ameeta Agrawal. 2023. MEEP: Is this engaging? prompting large language models for dialogue evaluation in multilingual settings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2078–2100, Singapore. Association for Computational Linguistics.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.

Tianyu He, Guanghui Fu, Yijing Yu, Fan Wang, Jianqiang Li, Qing Zhao, Changwei Song, Hongzhi Qi, Dan Luo, Huijing Zou, and Bing Xiang Yang. 2023. Towards a psychological generalist ai: A survey of current applications of large language models and future prospects. *Preprint*, arXiv:2312.04578.

kaggle. 2019. Therapist q&a. https://www.kaggle.com/datasets/arnmaud/therapist-qa.

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

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *Preprint*, arXiv:2303.17760.

Hongbin Na. 2024. CBT-LLM: A Chinese large language model for cognitive behavioral therapy-based mental health question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2930–2940, Torino, Italia. ELRA and ICCL.

Lazlo Ring, Timothy Bickmore, and Paola Pedrelli. 2016. An affectively aware virtual therapist for depression counseling. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) workshop on Computing and Mental Health*, pages 01951–12.

Mark A Robinson. 2018. Using multi-item psychometric scales for research and practice in human resource management. *Human resource management*, 57(3):739–750.

Lina M. Rojas-Barahona, Bo-Hsiang Tseng, Yinpei Dai, Clare Mansfield, Osman Ramadan, Stefan Ultes, Michael Crawford, and Milica Gašić. 2018. Deep learning for language understanding of mental health concepts derived from cognitive behavioural therapy. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 44–54, Brussels, Belgium. Association for Computational Linguistics.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. PsyQA: A Chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Bichen Wang, Pengfei Deng, Yanyan Zhao, and Bing Qin. 2023. C2D2 dataset: A resource for the cognitive distortion analysis and its impact on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10149–10160, Singapore. Association for Computational Linguistics.

Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024. Towards a client-centered assessment of llm therapists by client simulation. *Preprint*, arXiv:2406.12266.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Each Role Configuration

In the AutoCBT framework, we define each role accordingly. To ensure the flexibility of the AutoCBT framework, we did not fine-tune LLMs but only define roles and task descriptions for each agent through prompts in **Table 3**.

B The Bilingual Dataset Structure

To ensure consistency across languages, we merge question descriptions from PsyQA into unified instructions, simulating real or hypothetical user queries directed toward AutoCBT's Counsellor Agent. Using **Qwen-72B** (Bai et al., 2023) and **LLaMA-70B** (Touvron et al., 2023), we classify user queries into ten categories. Each query is assigned to its corresponding category, and 10 representative examples from each class are randomly selected, resulting in a **200-sample bilingual dataset** containing both original user questions and their corresponding model-generated responses. This dataset serves as the benchmark for evaluating AutoCBT's effectiveness in addressing cognitive distortions and improving response quality.

The structure of our dataset in the **Table 4**.

C Human Evaluation Metrics

The human evaluation metrics build upon the automatic evaluation metrics by refining the previous *Identification* and *Reflection* metrics, replacing

them with *Identify-CD* and *Challenge-CD*, respectively. Additionally, a new metric, *Presentation*, is introduced to assess the overall performance of the counsellor's response. As a result, we retain four original metrics from the automatic evaluation experiment while incorporating three new metrics in the Detailed Sampling Evaluation. The modified metrics provide a more nuanced assessment of the model's ability to detect and address cognitive distortions effectively.

In our evaluation framework in the **Table 5**, newly introduced metrics are highlighted in red, while original metrics remain unchanged and are displayed in black.

D Human Analysis of Detailed Sampling Evaluation

Both AutoCBT and PromptCBT employ empathetic methods. However, AutoCBT provides warmer and more contextually adaptive support, shaped by cultural variations.

In Chinese responses, AutoCBT emphasizes respect and indirectness, while in English responses, it maintains professionalism with a direct yet supportive tone. While both models excel in redescription and clarification, AutoCBT's softer and more context-specific tone fosters better emotional validation compared to PromptCBT's more rigid and academically structured responses, which some users perceive as overly clinical or labeling.

Furthermore, while the Generation approach may be suitable for addressing mild psychological concerns, AutoCBT demonstrates **stronger empathy and encouragement**, making it the preferred option for users requiring deeper emotional support. Although PromptCBT balances structured intervention with some degree of flexibility, it often lacks the clarity and emotional engagement found in AutoCBT, positioning AutoCBT as the most effective choice for addressing emotional and psychological challenges.

The analysis result in the **Table 6**.

E Rejections by Qwen and Llama

The reject result in the **Table 7**. Llama initially rejected 8 questions, reduced to 2 after AutoCBT's enhancements. In total, Llama rejected 9 unique questions.

F Long- and Short-Term Dialogue Compression

First, a boundary between long-term and short-term memory is preset, for example, setting it to 10. Each agent's dialogue history is maintained unchanged as long as it does not exceed this limit of 10 turns. Once the dialogue history of any agent exceeds 10 turns (e.g., reaches 11), the oldest 10 turns of dialogue need to be compressed, while the most recent current turn remains uncompressed.

At this point, the agent must first submit the oldest 10 dialogue turns to a large language model (LLM) for compression into a new, concise text summary—this process involves compressing a long text into a short text. The newly compressed short text then replaces the original 10 turns in the dialogue history. Together with the most recent current dialogue turn, these form a dialogue history consisting of only 2 records. The agent continues the conversation with this condensed history until the dialogue record count again exceeds 10, at which point the compression process is repeated.

This approach ensures that the dialogue history length of each agent never exceeds 11 turns. The detailed workflow is illustrated in **Figure 4a**.

Counsellor Prompt	Supervisor Prompt
##Attention## Then based on the following question and its description, please provide a professional, compassionate, and helpful response. Ensure your response adheres to the structure of Cognitive Behavioral Therapy (CBT) responses, especially in identifying the key thought or belief, and seamlessly integrates each part: 1. Validation and Empathy: Show understanding and sympathy for the patient's feelings or issues, creating a sense of safety. 2. Identify Key Thought or Belief: Through the problem description, identify potential cognitive distortions or core beliefs. 3. Pose Challenge or Reflection: Raise open-ended questions, encouraging the patient to reconsider or reflect on their initial thoughts or beliefs. 4. Provide Strategy or Insight: Offer practical strategies or insights to help them deal with the current situation. 5. Encouragement and Foresight: Encourage the patient to use the strategy, emphasizing that this is just the beginning and further support may be needed. ### Response content to be improved: {draft_response} ### Supervisor's revision suggestions: {revise_of_draft} ### The information of the patient is as follows: {original_question_of_user}	## You are playing the role of {agent_name} in a virtual world, accompanying the counsellor and examining the conversation between the patient and the counsellor. The counsellor will generate a response based on the patient's information and cognitive-behavioral therapy guidelines. As a supervisor, you need to provide some revision suggestions based on your own role description to the counsellor's response, so that the counsellor can improve and generate their response according to your revision suggestions. At present, the known information is as follows: ### Your role description as {agent_name}: {role_description} ### As {agent_name}, you saw a consultation from a patient: {original_question_of_user} ### As {agent_name}, you have seen the response to the patient's consultation generated by the counsellor that needs to be modified: {draft_response} ## Now, you have chosen to communicate with the following roles through {routing}: {agents} ##Attention## Please use 'Hello counsellor' as the beginning content of your response that you will send to {agent_name} and provide your revision suggestions to the counsellor!

Table 3: Prompt of Counsellor and Supervisors.

Lang.	Dataset Examples		Count
	Question Description	Answers	
EN	Me and my sister in law are both pregnant right now. And I've been noticing the inconsistency of level of care about our baby from my fiancée side of the fam. This situation really has me depressed, and unsure what to do. for starters my sister in law and that side of the family has made it a competition between the babies, I don't want it to be a competition. It always who can do what first.....	Thank you for explaining this situation. How unfortunate that this share joyous event is turning into a competition.The experience of being left out or ignored as part of the situation is what needs to be addressed. First, I would have a talk with your fiancé about what is happening and why. Does the family have a bias against the pregnancy because you were not married first? Is your fiancé on the outs.....	100
ZH	总是要考虑很多问题，我感觉我活在世界上就没有意义?我感觉我自己在交朋友的这条路上总是很不顺，初一初二的时候跟别人抢，我总是抢不过，不知道为什么，我总是把自己的地位放的很低，只要她一生气，我就卑微的求她原谅我，不管她说什么我都同意我总是感觉我拒绝一次，她就会离开我，一直到了上初三，爸妈突然让我转班，说是为了我的学业，我总是在课上无缘无故的哭，我真的很难受，到了高一，我认识了一个女孩，我们特别能合得来，我就看到了希望，我以为一切都会改变，但后来我发现，她很受人喜欢，班里的所有女生都很喜欢她，而我只能默默的看着，两个月以后，她总是前一天对我还很好，但是后一天又莫名其妙的对我冷暴力...	抱抱～看到发生在你身上的事就像往事重现。请允许我以姐姐的口吻与你讲下我的故事。我在刚去外地读大学的时候，认识了一位我很欣赏的女生。独立精干，双商高，性格开朗，很勇敢。她是一位好舍友，也是一位好闺蜜。入学初就约定好一起考研，从那就形影不离。在她面前，我可以表现很勇敢，很积极，很仗义，因为是我好朋友啊。上学一起，学习一起，吃饭一起，活动一起，实验一起，自习一起，睡觉也一起。反正就好的像黏黏胶一样。问想到我都上大学的人了还会这么粘人。后来想想可能是我一个人在外地吧，聊得来就会很上瘾哈哈。那段时光过得很愉快，回忆起来暖暖的。我一直以为会一直这样下去，连交男朋友都是外地异地恋，这样不会耽误我和闺蜜的时间...	100

Table 4: The overall structure and a Q&A example from the bilingual dataset.

Perspective	Description	Criterion	Score
Identify-CD	Identify potential cognitive distortions of the user through the description of the problem in the dialogue.	1.1 Has the cognitive distortion phenomenon of users been identified? 1.2 Does it help users recognize distorted beliefs? 1.3 Has cognitive distortion been explained from a psychological perspective?	7
Challenge-CD	Ask open-ended questions to encourage the user to reconsider or reflect on their initial thoughts or beliefs.	2.1 Does it help users think and challenge these distorted beliefs? 2.2 Have you raised open-ended questions that are helpful for deeper thinking? 2.3 Has psychological counseling technology been integrated? 2.4 Does the guided reflection correspond to the cognitive distortions that visitors may have?	7
Presentation	Evaluate the overall performance of the response of counsellor.	3.1 Is the overall language style close to the image of counsellor? 3.2 Is the information expressed clearly? 3.3 Have you flexibly applied some psychological counseling techniques?	7
Empathy	Demonstrates understanding and sympathy towards the user's emotions or issues, and creates a sense of safety.	4.1 Did the counsellor correctly understand the user's intent? 4.2 Did the counsellor show respect, understanding, and sympathy for the user's anxiety and pain? 4.3 Did the counsellor create a safe environment for the user to express their feelings?	7
Relevance	Evaluate the relevance of the dialogue content.	5.1 Was the counsellor's response highly relevant to the user's question? 5.2 Did the counsellor's response flow naturally? 5.3 Did the counsellor's answer cover the main issues or concerns raised by the user?	7
Strategy	Provide practical strategies or insights to help the user address their current situation.	6.1 Were the strategies or insights provided by the counsellor practical? 6.2 Could the strategies or insights solve the user's current problems? 6.3 Were the strategies based on professional psychological methods?	7
Encouragement	Encourage the user to use the strategies.	7.1 Did the counsellor encourage the user to take action? 7.2 Did the counsellor address potential failures the user might encounter while implementing the strategies? 7.3 Did the counsellor provide comfort and encouragement regarding setbacks and challenges?	7

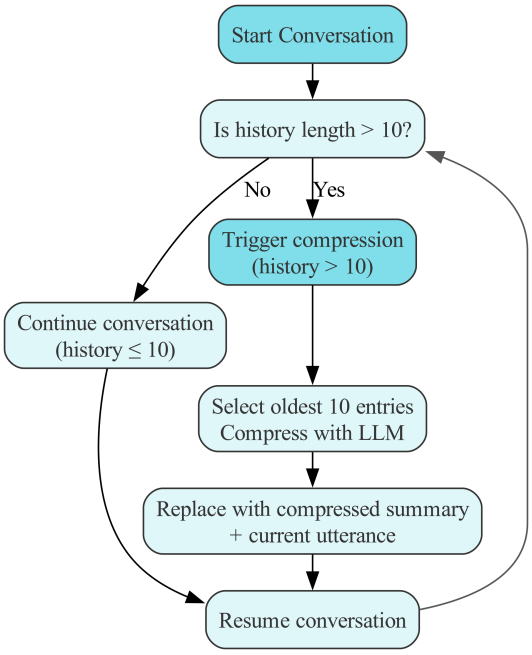
Table 5: The new metrics are inconsistent with the previous metrics, which were automatically evaluated.

Perspective	Human Analysis
Empathy & Encouragement	The responses from AutoCBT are logically similar to those from PromptCBT, both begin with empathetic techniques to convey understanding and validate the user's challenges before moving to structured and logical assessments and recommendations. Both approaches generally demonstrate an accurate understanding of the user's concerns and challenges. However, AutoCBT provides slightly more emotional support, creating an overall warmer impression. Its responses integrate empathetic techniques more smoothly and maintain a consistent empathetic tone. Additionally, two specific aspects were observed. First, AutoCBT demonstrates more flexibility in word choice compared to PromptCBT. This marks a significant improvement over the formulaic responses typically associated with previous LLMs. Furthermore, likely due to cultural differences in counseling model training, AutoCBT's approach to creating a "safe environment for the user" varies between its Chinese and English responses. In the Chinese context, it emphasizes respect, attentiveness, and ensures the user feels valued, respected, and heard. In English, however, it emphasizes professionalism with phrases like, "I'll view your issue from a non-judgmental perspective," aligning with the clear boundaries often emphasized in Western society. In Chinese practice, these boundaries are generally less pronounced to avoid creating user apprehension.
Cognitive Distortion	Both AutoCBT and PromptCBT effectively identify and analyze users' cognitive distortions; however, Generation's responses contain minimal content on this aspect. There is a notable gap between AutoCBT and PromptCBT in further challenging cognitive distortions, primarily in their integration with the client's context. PromptCBT's guided reflection can feel rigid, and some responses may make users feel interrogated. In contrast, AutoCBT's recognition and reflection are well-aligned with users' specific contexts, using softer, gentler language that guides users to examine the rationality of their core beliefs from different perspectives. Both AutoCBT and PromptCBT responses exhibit re-description, summarization, and conceptual clarification of user questions, with AutoCBT applying these techniques more extensively. We see this as a key advantage of LLM-based psychological counseling responses. Re-description not only demonstrates that the "Counsellor Agent" genuinely understands the user's issue but also enhances the credibility of "I can understand you," helping users feel their emotions are acknowledged. Additionally, users experiencing psychological and emotional challenges often have confused thoughts. Techniques like re-description, summarization, and clarification assist users in clarifying their logical thinking and focusing on the issues they seek to resolve. Additionally, in vocabulary explanation, AutoCBT uses a more approachable and conversational language style, while PromptCBT tends toward academic expressions. PromptCBT often uses more specialized psychological terms, which can inadvertently make users feel "labeled" and lead to self-criticism. For instance, PromptCBT might use terms like "catastrophizing thinking," potentially leading users to think, "I'm really bad." Similar issues occasionally appear in AutoCBT's responses but with less frequency than in PromptCBT's. In real-life counseling, practitioners carefully use professional terminology, especially with clients experiencing significant psychological challenges. They often use more tactful language when conveying serious-sounding terms, a strength in which AutoCBT excels.
Usefulness of the strategy	Based on its performance, we believe Generation is suitable primarily for users with mild emotional issues and a clear objective of finding problem-solving methods. However, its mechanical and rigid language is less appropriate for users needing psychological and emotional support. For users experiencing emotional confusion or in a suboptimal or unhealthy psychological state, AutoCBT is recommended. AutoCBT's performance more closely resembles that of a psychological counsellor, providing greater empathy and respect in its language. PromptCBT's positioning lies between the other two; it employs more academic language that may seem diagnostic rather than consultative, lacking clear explanations for users. Generation offers the widest range of strategies among the three, providing users with diverse choices. However, its strategies are often vague, with broad, generic explanations that lack specific responses to users' challenges, leading to lower overall relevance. AutoCBT and PromptCBT incorporate user-specific contexts to better address their needs. Of the two, AutoCBT performs better, showing stronger empathy and encouragement in its language, and creating a more genuine dialogue with users. When proposing potentially sensitive strategies, like suggesting users seek professional counseling, AutoCBT uses caring language paired with empathy and encouragement, reducing visitors' resistance. In some responses, AutoCBT anticipates potential obstacles in implementing strategies and provides timely encouragement, offering empathetic support for users with psychological or emotional challenges.

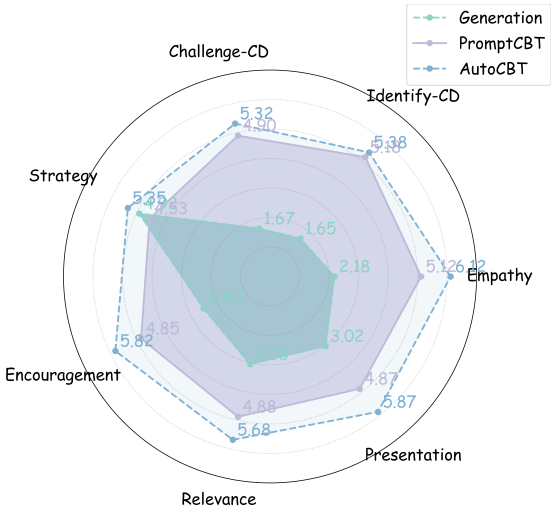
Table 6: Human analysis of the SOE and the DSE.

Model	Method	Chinese		English	
		Refused-Questions	Distinct-Refused-Questions	Refused-Questions	Distinct-Refused-Questions
Qwen	Generation	0	0	0	0
	PromptCBT	0		0	
	AutoCBT	0		0	
Llama	Generation	/	/	3	Union(3, 3, 8) = 9
	PromptCBT	/		3	
	AutoCBT	/		$8 \rightarrow 2$	

Table 7: Rejections by Qwen-2.5-72B and Llama-3.1-70B were analyzed.



(a) The workflow of long short-term memory.



(b) The performance of Chinese in the DSE experiment.

Figure 4: The workflow of long short-term memory, and the performance of Chinese in the DSE experiment.