Stage-Aware Therapeutic Support in Cognitive Behavioral Therapy with Large Language Models

Aijia Yuan†
Department of Operations and Decision Technologies
Indiana University
Bloomington, IN, USA
yuana@iu.edu

Yang Gao
Department of Operations and Decision Technologies
Indiana University
Bloomington, IN, USA
gaoyang@iu.edu

Edlin Garcia
Department of Health and Wellness Design
Indiana University
Bloomington, IN, USA
eggarcia@iu.edu

Jianna Hur
Department of Operations and Decision Technologies
Indiana University
Bloomington, IN, USA
jhur06@iu.edu

Sagar Samtani
Department of Operations and Decision Technologies
Indiana University
Bloomington, IN, USA
ssamtani@iu.edu

ABSTRACT

Mental health disorders remain a leading cause of disability worldwide, creating sustained demand for accessible care. Cognitive Behavioral Therapy (CBT) is widely recognized as the gold standard in psychotherapy for addressing both clinical disorders and everyday stressors, yet many patients discontinue treatment early due to insufficient support between sessions when therapists are unavailable. Strengthening inter-session support is therefore essential for maintaining continuity of care in ways that are clinically aligned and socially accessible. This study introduces a stage-aware large language model framework designed to complement human therapy by generating responses that align with CBT principles and therapeutic structure across the early, middle, and late stages of treatment. The framework integrates two key components: (1) a synthetic data generation strategy that closely follows CBT guidelines while modeling inter-session continuity by explicitly linking prior action plans and therapeutic progress to subsequent prompts, thereby generating CBT dialogues across different stages of treatment; and (2) a stage-aware supervised fine-tuning (SFT) process that adapts a large language model using these generated dialogues to better capture CBT-oriented communication patterns and therapeutic interactions associated with different stages of treatment. Our evaluation against prevailing general-purpose and mental health-specific LLMs demonstrates that the proposed framework achieves the strongest overall linguistic performance, including the highest BERTScore F1 (0.8982), indicating improved lexical and semantic alignment with reference CBT responses. Additional therapeutic quality evaluation using fidelity-related measures derived from the Cognitive Therapy Rating Scale further suggests that the generated responses align well with CBT-oriented therapeutic principles across treatment stages. Overall, this study demonstrates the importance of stage-aware therapeutic modeling for generating more therapeutically aligned responses in AI-enabled mental health systems for inter-session support.

CCS CONCEPTS

• Computing methodologies → Artificial intelligence → Natural language processing → Natural language generation.

KEYWORDS

Mental Health, Psychotherapy, Cognitive Behavioral Therapy, Large Language Model, Supervised Fine-Tuning

1 Introduction

Mental health disorders remain among the leading causes of disability, affecting more than one billion people globally [50]. According to the WHO, the global median number of mental health workers is only 13 per 100,000 people. These conditions are also associated with worsened physical health, reduced productivity, and diminished quality of life, and impose a substantial economic burden, costing the global economy an estimated US$1 trillion annually due to depression and anxiety alone [7]. In severe cases, mental health challenges can contribute to self-harm and suicidal behaviors [25, 26]. These trends underscore the urgent need to expand access to mental health therapy. Individuals seek therapy for a wide spectrum of challenges, including clinical disorders such as depression and anxiety to common life stressors including relationship difficulties, work or academic burnout, grief, and major life transitions [3]. These stressors often overlap with clinical symptoms, creating complex therapeutic needs that require structured, emotionally responsive, and ethically grounded care.

Cognitive Behavioral Therapy (CBT) is the most widely adopted psychotherapeutic approach for addressing both clinical and nonclinical mental health concerns [8, 12]. It provides a structured, goal-oriented treatment process that progresses through early, middle, and late stages [9]. This progression allows therapists to guide patients from identifying initial problems to restructuring maladaptive thoughts and ultimately developing strategies for relapse prevention through a collaborative process delivered in an empathetic, inclusive, and humanistic manner [4].

CBT has been highly effective in controlled clinical environments. However, many patients discontinue treatment prematurely, often before reaching the middle and late stages where the most significant therapeutic progress occurs [6]. A primary reason for early dropout is insufficient engagement between therapy sessions. Intersession work refers to the time when patients extend their therapeutic experience by practicing CBT skills, reflecting on their thoughts, and maintaining engagement with CBT principles [5]. However, when patients do not feel emotionally supported outside formal appointments, the continuity of treatment weakens, thereby diminishing motivation to continue [15, 44]. This challenge is further intensified by a widespread shortage of therapists, which restricts access to individualized care and limits opportunities for consistent follow-up outside scheduled sessions [34].

Traditionally, intersession support has relied on therapist check-ins, follow-up calls, and reflective or behavioral homework assignments [38]. They have been shown to improve outcomes and reduce dropout, but they are highly dependent on therapist availability and are labor- and time-intensive, making them difficult to sustain on a large scale and limiting equitable and socially accessible care. Digital tools such as journaling apps and rule-based chatbots have been developed to provide automated support between sessions [51]. While these tools offer convenience and accessibility, they often rely on static scripts that cannot adapt to a patient's therapy stage and progression.

The rapid advancement of large language models (LLMs) offers a new opportunity to create scalable, context-aware forms of therapeutic support [42]. LLMs can generate coherent, adaptive, and empathetic text, making them promising tools to complement therapy between sessions, particularly when human therapists are unavailable. Recent mental health LLMs have demonstrated encouraging performance in conversational support and counseling-oriented dialogue generation. However, most existing systems remain limited in their ability to model the structured and progressive nature of psychotherapy, particularly within CBT. Specifically, they often rely on generic counseling datasets that fail to capture stage-specific therapeutic progression and continuity across sessions. Thus, there is a critical need to build AI systems that support stage-aware inter-session therapeutic interactions.

In this study, we develop a Stage-Aware CBT Large Language Model (SA-CBT-LLM) designed to complement human therapy through stage-sensitive intersession augmentation. The framework integrates two key components:

1) A stage-aware synthetic data generation strategy that models inter-session continuity by explicitly linking prior action plans and therapeutic progress to subsequent prompts. Grounded in the gold-standard CBT manual, the design produces dialogues that reflect therapeutic interactions across the early, middle, and late stages of treatment.
2) A stage-aware supervised fine-tuning (SFT) process that adapts a large language model using these generated dialogues, enabling the model to generate responses that better reflect CBT-oriented communication patterns and therapeutic interactions associated with different stages of treatment.

We rigorously evaluated the proposed SA-CBT-LLM against general-purpose and mental health LLMs using both linguistic and therapeutic quality evaluation metrics. The remainder of the paper is organized as follows. We review CBT, digital mental health, and relevant methodological developments. We then identify key research gaps and propose research questions. Next, we present the proposed framework and discuss the experiment results. Finally, we summarize implications and conclude with directions for future research.

2 Literature Review

We review prior research on CBT, digital mental health systems, and large language models for therapeutic support. We further discuss recent methodological developments in data generation and supervised fine-tuning for mental health LLMs.

2.1 Cognitive Behavioral Therapy (CBT)

CBT is widely regarded as the gold standard in psychotherapy, offering a goal-oriented framework for treating a wide range of psychological concerns [12]. Grounded in the principle that maladaptive thoughts influence emotions and behaviors, CBT helps clients identify and modify distorted thinking patterns through guided reflection and behavioral experimentation [4]. The therapeutic process is highly structured and unfolds across multiple sessions, with progress built cumulatively over early, middle, and late stages. Specifically, the early stage emphasizes problem identification and rapport building, the middle stage focuses on cognitive restructuring and skill application, and the late stage consolidates gains and prepares clients for relapse prevention [9]. The structured and progressive nature of CBT highlights the importance of stage-aware therapeutic interactions and continuity across sessions, particularly within inter-session support systems.

Prior clinical research has emphasized the importance of maintaining therapeutic quality (i.e., fidelity to CBT principles) throughout treatment. One commonly used framework for evaluating CBT fidelity is the Cognitive Therapy Rating Scale-Revised (CTRS-R) [2, 16], which provides a structured approach for assessing the quality of CBT delivery. The framework evaluates 11 items, including agenda, feedback, understanding, interpersonal effectiveness, collaboration, pacing, guided discovery, focusing on key cognitions and behaviors, strategy for change, application of CBT techniques, and action plan. Among these, Action Plan remains particularly important across all treatment stages because CBT emphasizes continuity of learning between sessions through homework assignments and real-world behavioral application. Each session therefore builds upon prior therapeutic progress and planned activities from previous sessions. The CTRS-R can therefore serve as a useful reference for assessing therapeutic characteristics in generated responses.

2.2 Digital Tools for Mental Health and Inter-Session Support

Digital tools for mental health have evolved substantially, expanding access to therapeutic support beyond traditional in-person sessions. Early systems primarily consisted of teletherapy modules, such as online courses and scheduled video sessions designed to replicate the structure of therapy in a virtual format [47]. Human-in-the-loop platforms, such as BetterHelp, Talkspace, and Grow Therapy [21], connect users with licensed clinicians who provide therapy via text, phone, or video. Although these tools improve accessibility, their reliance on human therapists limits scalability and makes continuous, on-demand support difficult to provide. This limitation motivated the development of automated digital support tools, including early CBT applications and rule-based conversational agents [43]. While these systems demonstrated the feasibility of delivering structured therapeutic content, their reliance on predefined scripts limited adaptability and reduced their ability to handle the complexity of real mental health conversations. More recently, advances in generative AI have enabled LLM-based mental health systems, such as Yuna and Wysa, as well as research prototypes built on models like GPT and LLaMA [55]. These systems can generate more adaptive and empathetic responses, demonstrating AI's potential to complement human-delivered therapy [33].

However, despite these advancements, there has been limited progress in developing systems that provide effective inter-session support to help clients apply therapeutic principles between scheduled sessions. Most prior work in this stream has focused on depression and anxiety, employing digital modules, mobile apps, or chatbots [18, 49, 52]. These empirical studies suggest that digital systems can enhance CBT continuity and patient engagement between sessions. Yet, most research evaluates existing applications rather than proposing frameworks for stage-aware inter-session augmentation. In addition, evaluations often rely on outcome-based metrics such as attendance, engagement levels, or symptom reduction (e.g., PHQ-9 scores) rather than process-oriented measures that reflect the quality of therapeutic interactions [27, 54]. While outcome measures remain important, maintaining therapeutic quality is also central to effective CBT, particularly given its emphasis on structured cognitive and behavioral progression over time [19].

2.3 LLMs for Mental Health

LLMs have been an attractive proposition for mental health applications, as they can flexibly adapt to user context and potentially sustain ongoing support between sessions. Most existing systems focus on general mental health support, primarily addressing depression, anxiety, stress, crisis intervention, or psychoeducation rather than being specifically tailored for structured CBT sessions [1, 10, 13, 14, 20, 22, 24, 36, 39, 40, 45, 53, 56]. The underlying models are predominantly based on the GPT and LLaMA families. Training datasets commonly consist of counseling question-answer pairs, therapist-client dialogues, crisis intervention resources, and other mental health corpora, typically containing hundreds to several thousand examples. Although some studies leverage substantially larger datasets derived from social media platforms or online support communities, these sources are often not clinically grounded and may not adequately reflect the structured nature of psychotherapy. Nearly all existing systems function as stand-alone tools, such as chatbots or text-analysis models, designed for single interactions rather than continuous inter-session support. As a result, these systems typically lack stage-aware mechanisms that account for therapeutic progression across early, middle, and late sessions, which is essential for modeling CBT dynamics over time. For evaluation, most studies rely on either technical natural language processing (NLP) metrics (e.g., BLEU, ROUGE, BERTScore) or domain-specific measures such as empathy and helpfulness [55], depending on their specific focus.

In terms of methodology, prior work commonly involves collecting or generating counseling data, with supervised fine-tuning (SFT) approaches to adapt pretrained LLMs to mental health contexts. Existing counseling datasets are often derived from domain expertise or patient-therapist interactions and are typically limited in scale. In addition, many studies employ parameter-efficient fine-tuning approaches, such as LoRA or QLoRA, to adapt pretrained LLMs for mental health dialogue generation. More details regarding data generation and SFT are discussed next.

2.3.1 Data Collection and Synthetic Data Generation. Data collection remains a core challenge in LLM-based therapy modeling due to the limited availability of structured therapy transcripts and the ethical constraints associated with collecting real patient data, particularly data containing personally identifiable information (PII) [23]. In domains where privacy is essential, such as healthcare, synthetic data generation has emerged as a promising alternative [17]. It uses LLMs to produce training data that emulate real-world therapeutic exchanges. Synthetic data generation typically involves defining system instructions and user prompts that specify the clinical intent and then prompting the model to generate structured outputs, such as therapy dialogues [41]. Prior work has explored several prompting strategies for this purpose [32]. Zero-shot prompting relies solely on task instructions without examples but often produces overly generic responses. One-shot prompting introduces a single example to guide tone and structure, while few-shot prompting incorporates multiple examples to enhance consistency and coherence. More recently, topic-controlled prompting has been applied to produce samples grounded in context-driven themes and scenarios [32]. In our context, this approach can enable large-scale, clinically relevant data generation to better reflect real-world CBT patient characteristics (e.g., demographic attributes) and commonly addressed topics, particularly when authentic counseling data are scarce. However, standard data generation approaches still struggle to capture session continuity across different therapy stages.

2.3.2 Supervised Fine-Tuning (SFT). SFT adapts a pretrained LLM to a target task by training on paired input-output examples and minimizing cross-entropy loss to reproduce reference responses [48]. Instruction-style SFT extends this framework by incorporating domain-relevant context; in our case, this may include additional information such as system roles (e.g., therapist, patient), therapy stages, or CBT framings [37]. This approach is particularly suited for structured, stage-sensitive domains such as psychotherapy, where multiple types of instructions are needed to capture the evolving therapeutic context and ensure that model outputs remain consistent with clinical intent and progression rather than surface-level linguistic similarity.

3 Research Gaps and Research Questions

Despite advances in LLM-based mental health systems, most existing models remain general-purpose and stand-alone, lacking stage-aware inter-session augmentation that complements human-delivered therapy. Available data resources are limited and often lack clinically grounded dialogues that reflect how therapy evolves over time. Although synthetic data generation helps address data scarcity, current approaches typically produce isolated conversational pairs that fail to capture session continuity or the shifting therapeutic focus across CBT stages. In addition, while supervised fine-tuning enables LLMs to adapt to mental health dialogue generation, existing approaches primarily focus on general counseling interactions rather than stage-aware therapeutic progression across sessions. Thus, we propose the following research questions:

(1) How can we develop an LLM-based approach for stage-aware intersession augmentation in CBT?

(2) How can we design a synthetic data generation strategy that reflects CBT session continuity for stage-aware supervised fine-tuning?

4 Methods

4.1 Data Collection

We use CBT transcripts drawn from Beck's CBT Manual [4], which is widely regarded as the gold standard reference used by clinicians in professional practice. The manual provides authentic records of therapist-patient interactions that illustrate the tone, principles, and flow of CBT. These transcripts, along with the clinical guidance presented in the manual, serve as ground-truth exemplars and are used as seed data for subsequent data generation and model development. Representative therapist-patient dialogues were identified to capture how CBT interactions evolve across the early, middle, and late stages of treatment.

Although these transcripts are clinically grounded, their coverage remains limited. The available material includes only a small number of patient cases and partial session excerpts, each consisting of short conversational snippets. As a result, the dataset offers an instructive but narrow view of therapeutic exchanges and insufficiently represents the diversity of patient profiles, commonly addressed topics, and session progressions observed in real-world CBT practice. Other publicly available dialogue corpora, such as counseling question-answer sets and social media-based interactions, have been explored in prior studies [40, 53]. However, these datasets are not CBT-based and do not reflect the structured, stage-specific treatment cadence observed across sessions. Thus, the limited availability of comprehensive, stage-aware, and clinically grounded dialogue data, motivates synthetic data generation.

4.2 Synthetic Data Generation

We designed a topic-controlled synthetic data generation framework by explicitly encoding the structure and cadence of CBT. The approach aims to mimic real-world therapeutic practice, where therapists address diverse mental health concerns across different stages of treatment for distinct patient profiles. By incorporating therapeutic structure and continuity, the generated data supports stage-aware learning in subsequent model training.

4.2.1 Profile Creation. The data generation process begins with the creation of synthetic patient profiles. Explicitly defining patient profiles before dialogue generation ensures controlled diversity in clinical characteristics. Each profile contains a unique patient identifier, one primary therapeutic topic, and demographic attributes (age and gender), which together form a concise patient case summary providing clinical context. Each synthetic patient profile is associated with a single primary therapeutic topic and is designed to progress through the three sequential stages of CBT: early, middle, and late. The six major concern categories were selected to reflect common presenting issues in CBT practice and were equally distributed across profiles: relationship issues, divorce, work or academic burnout, grief and loss, life transitions, and stress and coping.

In total, 3,000 synthetic patient profiles were generated. Across all profiles, the mean patient age was 39.3 years (SD = 17.0), and approximately 68.3 percent of profiles were assigned female gender. These distributions were selected to approximate patient characteristics commonly observed in real CBT settings [11]. Each profile contains a concise case summary that provides additional clinical context, for example: "A 42-year-old recently divorced individual who feels isolated, self-critical, and unsure how to rebuild routines; no acute risk." These elements form a structured representation of patient profiles used for subsequent stage-aware synthetic dialogue generation.

4.2.2 Stage-Aware Synthetic Data Generation Workflow with Continuity. Once the patient profiles were created, we developed a stage-aware synthetic data generation workflow with continuity. This workflow expands data coverage while encoding both stage cadence and session continuity. Let $s \in \{1,2,3\}$ represent the therapy stage (early, middle, late), $t \in \{1,2,3,4,5,6\}$ represent the topic category, and p denote the patient profile (including demographic attributes). The dataset can be expressed as a set of dialogue pairs:

$$QA_{s,t,p} = (q_{s,t,p}, a_{s,t,p})$$

where each pair contains a patient query and a therapist response associated with the given profile, topic, and stage. For each profile, three dialogue exchanges are generated to represent the early, middle, and late phases of CBT. In addition, to model continuity across sessions, the action plan and progress from stage s is carried forward to stage $s+1$. This mechanism explicitly encodes therapeutic progression, reflecting how therapists build upon prior sessions in real CBT practice. Figure 1 presents an example illustrating the prompting workflow and continuity mechanism.

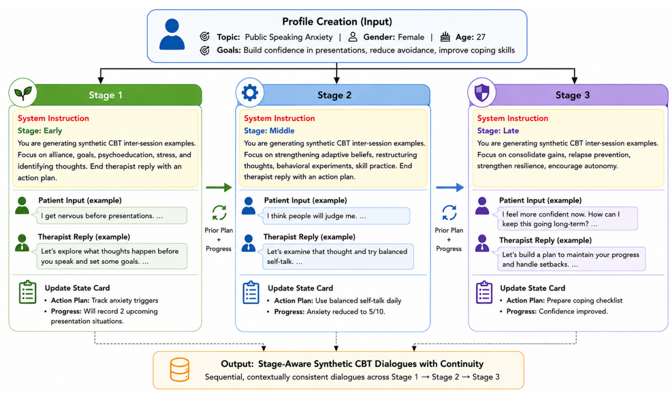

Figure 1: Stage-Aware Synthetic Data Generation Workflow with Continuity (Example)

The workflow follows a three-step process:

- *System Instruction:* The system prompt explicitly defines the therapist role, therapeutic objectives, communication focus, and stage-specific treatment priorities according to Beck's CBT guidance [4]. Different instructions are provided for the early, middle, and late stages of therapy to reflect the evolving structure of CBT treatment, such as rapport building and problem identification in early sessions, cognitive restructuring and behavioral intervention in middle sessions, and relapse prevention and maintenance planning in later sessions.
- *User Prompt:* The prompt includes the patient profile and any prior action plan or progress extracted from the previous stage (with no prior plan for the early stage). A separate state card stores and updates action plans after each stage's data generation.
- *Output:* The model generates a dialogue pair consisting of a patient message and a therapist response. After each stage's generation, the action plan extracted from the therapist's reply is summarized and recorded in the state card, which is then incorporated into the subsequent user prompt to preserve continuity during the next stage of data generation.

Through this iterative process, we generated 9,000 synthetic patient-therapist dialogue pairs. Each sample captures evolving therapeutic interactions while preserving continuity from previous sessions. To ensure that generated dialogues reflected appropriate therapeutic language and CBT structure, two mental health experts reviewed randomly selected subsets of the data balanced across therapy stages. The reviewers evaluated dialogue coherence, tone, and consistency with CBT communication principles. Their feedback resulted in refinements to phrasing, tone, and therapeutic framing to improve the overall quality and consistency of the generated dataset.

The complete synthetic dialogue dataset is available at: https://huggingface.co/datasets/yuana1234567/Mental-health-CBT-dialogues

4.3 Stage-Aware Supervised Fine-Tuning

The proposed stage-aware CBT LLM adapts LLaMA-3.1-8B to generate stage-aware CBT responses through supervised fine-tuning (SFT) on the dataset. The objective is to align model outputs with therapist responses that reflect stage-specific CBT reasoning, communication patterns, and therapeutic progression across sessions. Each input instance i consists of three structured components:

- *System Instruction* s_i *:* A fixed instruction defining the therapist's role, adapted from Beck's CBT manual [4] and verified by human experts.
- *Stage Indicator* (g_i)*:* A categorical token specifying the therapeutic stage (early, middle, or late), guiding the model to adjust tone, focus, and task emphasis accordingly.
- *Patient Message* (m_i)*:* A patient utterance drawn from the synthetic dataset.

The target output $(y_{i,t})$ corresponds to the therapist's response associated with the given stage and patient context. The model is trained to minimize cross-entropy loss between predicted and reference therapist responses using:

$$\mathcal{L}_{SFT}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\log P_\theta\left(y_{i,t} \mid y_{i,<t}, s_i, g_i, m_i\right)$$

where N denotes the total number of training examples, and T_i represents the token length of the therapist response for example i.

For model development, 80% of the dataset (7,200 dialogue pairs) was used for training, 10% (900 pairs) for validation, and the remaining 10% (900 pairs) was reserved as a held-out testing set for evaluation. Through this SFT process, the model learns to generate responses that better reflect the communication patterns associated with different stages of CBT.

4.4 Evaluation Plan

We evaluate the proposed SA-CBT-LLM through two experiments designed to assess both linguistic performance and therapeutic quality. In Experiment 1, we evaluated how the proposed SA-CBT-LLM performs against prevailing general-purpose LLMs. The benchmark models included LLaMA-3.1-8B, Mistral-7B Instruct, Falcon-7B Instruct, Qwen-7B Instruct, and GPT-4.1 mini. This experiment assesses the added value of stage-aware domain adaptation by examining how well standard LLMs capture therapeutic interaction patterns relative to our framework. In Experiment 2, we benchmarked SA-CBT-LLM against existing fine-tuned mental health LLMs to evaluate whether it provides measurable advantages in the CBT context. Benchmark models in this experiment [28–30, 35, 46] were selected because they are publicly available and represent commonly used adaptation strategies in mental health LLM research, including SFT and parameter-efficient fine-tuning approaches such as LoRA.

Our evaluation metrics are computed on a held-out testing set of 900 synthetic conversational pairs. Following prior NLP and mental health LLM literature, we assess model performance using several linguistic evaluation metrics, including ROUGE-L and BERTScore (F1, Precision, and Recall). ROUGE-L measures lexical overlap between generated and reference responses, while BERTScore evaluates semantic similarity using contextual embeddings. Specifically, BERTScore Precision measures the relevance of the generated content relative to the reference response, Recall evaluates how much of the reference meaning is captured by the generated response, and F1 provides an overall semantic similarity score. In addition to linguistic evaluation, we also include a CBT-oriented therapeutic quality assessment using fidelity-related measures derived from the Cognitive Therapy Rating Scale-Revised (CTRS-R). To support scalable evaluation, an LLM is employed as a judge to assess generated therapist responses across the early, middle, and late stages of therapy. Specifically, the LLM generates a score for each of the 11 CTRS-R items, and these item scores are summed to obtain an overall fidelity score, consistent with the questionnaire's scoring approach in clinical practice. Higher fidelity scores indicate stronger alignment between generated responses and CBT-oriented therapeutic principles.

All experiments were conducted on Big Red 200, a supercomputer equipped with 64 GPU nodes, each with 512 GB of memory, a 64-core 2.0 GHz AMD EPYC 7713 processor, and four NVIDIA A100 GPUs.

5 Experiment Results

5.1 Experiment 1: SA-CBT-LLM vs General LLMs

Experiment 1 examines how domain adaptation affects the proposed SA-CBT-LLM relative to prevailing general-purpose LLMs in generating CBT-oriented responses. Results are summarized in Table 1. The best model performances appear in bold-face.

Table 1. Experiment 1 Results (Linguistic)

Models	ROUGE-L	BERTScore Precision	BERTScore Recall	BERTScore F1
LLaMA-3.1-8B	0.1845	0.8569	0.8695	0.8631
Mistral-7B Instruct	0.1564	0.8441	0.8620	0.8529
Falcon-7B Instruct	0.1729	0.8653	0.8634	0.8643
Qwen-7B Instruct	0.1196	0.8635	0.8511	0.8572
GPT-4.1 mini	0.2153	0.8699	0.8907	0.8802
SA-CBT-LLM	**0.2862**	**0.8984**	**0.8981**	**0.8982**

The proposed SA-CBT-LLM achieves the strongest performance across all linguistic evaluation metrics, including ROUGE-L (0.2862), BERTScore Precision (0.8984), Recall (0.8981), and F1 (0.8982). These results indicate that the stage-aware supervised fine-tuning strategy improves both lexical alignment and semantic consistency between generated therapist responses and reference CBT responses. In particular, the improvements in BERTScore suggest that the proposed framework more effectively captures the contextual meaning and conversational patterns associated with CBT-oriented interactions across different therapy stages. Among the baseline models, GPT-4.1 mini performs competitively across most metrics. This is expected because, although labeled as a smaller variant, it is a proprietary model trained on a much larger, more diverse corpus and advanced instruction-tuned.

In addition to linguistic evaluation, we further conducted a CBT-oriented therapeutic quality assessment using fidelity-related measures derived from the CTRS-R questionnaire. Results are summarized in Table 2. The best model performances appear in bold-face.

Table 2. Experiment 1 Results (Clinical)

Models	$\text{Fidelity}_{\text{Early}}$	$\text{Fidelity}_{\text{Mid}}$	$\text{Fidelity}_{\text{Late}}$	$\text{Fidelity}_{\text{Mean}}$
LLaMA-3.1-8B	6.7500	7.3167	8.5567	7.5411
Mistral-7B Instruct	5.1900	7.1700	9.5433	7.3011
Falcon-7B Instruct	6.7567	6.2667	5.5767	6.2000
Qwen-7B Instruct	0.7333	1.1333	3.0300	1.6320
GPT-4.1 mini	9.2800	**13.0433**	11.1833	11.1689
SA-CBT-LLM	**9.9367**	12.2033	**11.6800**	**11.2730**

From the therapeutic evaluation perspective, SA-CBT-LLM achieves the highest fidelity-related scores for the early stage (9.9367), late stage (11.6800), and overall mean performance (11.2730), while GPT-4.1 mini achieves the strongest performance in the middle stage. These findings suggest that the proposed framework generates responses that generally align well with CBT-oriented therapeutic principles across different stages of treatment. Nevertheless, there remains room to further enhance therapeutic quality, particularly in achieving more consistent fidelity across all stages of treatment. Further improvements may be achieved through future optimization strategies specifically designed to enhance CBT fidelity throughout the therapeutic process.

5.2 Experiment 2: SA-CBT-LLM vs Mental Health LLMs

Experiment 2 benchmarks the SA-CBT-LLM against existing open-source, mental health-specific LLMs. Results are summarized in Table 3. The best model performances appear in bold-face.

Table 3. Experiment 2 Results (Linguistic)

Models	ROUGE-L	BERTScore Precision	BERTScore Recall	BERTScore F1
Mental-Health-Mistral	0.1622	0.8478	0.8639	0.8558
TinyLlama-mental-health	0.1325	0.8350	0.8500	0.8424
Mental_Health_chatbot	0.1353	0.8403	0.8633	0.8516
mh-llama-GRPO	0.1443	0.8708	0.8505	0.8602
Psychosocial-Counseling	0.1728	0.8487	0.8653	0.8569
SA-CBT-LLM	**0.2862**	**0.8984**	**0.8981**	**0.8982**

The proposed SA-CBT-LLM again achieves the strongest performance across all linguistic evaluation metrics, including ROUGE-L (0.2862), BERTScore Precision (0.8984), Recall (0.8981), and F1 (0.8982). Many benchmark models reflect prevailing adaptation strategies in the field, including supervised fine-tuning and LoRA-based parameter-efficient fine-tuning. However, despite being fine-tuned on mental health datasets, improvements over general-purpose LLMs remain relatively limited for CBT-specific dialogue generation. These findings suggest that domain adaptation alone may not sufficiently capture the structured therapeutic principles and communication patterns associated with CBT-oriented interactions. In contrast, the proposed stage-aware supervised fine-tuning strategy enables the model to better align generated responses with reference CBT interactions constructed across different treatment stages. We further conducted a CBT-oriented therapeutic quality assessment. Results are summarized in Table 4. The best model performances appear in bold-face.

Table 4. Experiment 2 Results (Clinical)

Models	$\text{Fidelity}_{\text{Early}}$	$\text{Fidelity}_{\text{Mid}}$	$\text{Fidelity}_{\text{Late}}$	$\text{Fidelity}_{\text{Mean}}$
Mental-Health-Mistral	6.8033	8.4067	7.2200	7.4767
TinyLlama-mental-health	3.2600	4.1267	3.4433	3.6100
Mental_Health_chatbot	6.7433	8.9100	7.8933	7.8489
mh-llama-GRPO	3.5467	4.3367	4.4167	4.1000
Psychosocial-Counseling	6.3867	7.7000	8.0367	7.3744
SA-CBT-LLM	**9.9367**	**12.2033**	**11.6800**	**11.2730**

From the therapeutic evaluation perspective, SA-CBT-LLM achieves the highest fidelity-related scores across all therapy stages, including Early (9.9367), Middle (12.2033), and Late (11.6800), as well as the strongest overall mean performance (11.2730). These findings suggest that the proposed framework generally produces responses that better reflect CBT-oriented therapeutic principles and communication patterns associated with different stages of treatment relative to existing mental health-specific LLMs.

6 Discussion

The proposed SA-CBT-LLM aims to make mental health support more socially accessible, responsive, and empathetic to individuals' evolving therapeutic needs, while staying aligned with the professional and ethical standards followed by human therapists. It seeks to proactively acknowledge distress, promote well-being, and uphold human dignity, rather than only generating technically correct responses. Our work has implications for key stakeholders, including therapists and patients.

Therapists in various settings (e.g., higher educations) benefit from reduced manual effort as the system provides structured support between sessions, handling routine psychoeducation, check-ins, and guidance when direct contact is not possible. The system adapts its responses based on stage progression in a clinically aligned yet humanistic and empathetic manner, a capability not achievable with earlier rule-based or static LLM systems. In line with prior research suggesting that mental health technologies can help address unmet mental health needs when human-centered principles are preserved [31], SA-CBT-LLM complements human therapists by functioning as a supportive extension of clinical care and ultimately strengthening the therapeutic relationship.

Patients often discontinue therapy for a range of reasons, including limited accessibility, scheduling constraints, and social stigma associated with seeking mental health support. When instantiated within a digital tool, SA-CBT-LLM can enhance engagement by offering stage-specific support that is accessible at any time. For some individuals, the private and socially low-barrier nature of such tools may reduce hesitation around disclosure and help sustain engagement between appointments. By validating emotional experiences and providing guidance grounded in established clinical standards, the system supports continuity of care while promoting overall well-being.

7 Conclusion and Future Directions

This study presents the Stage-Aware CBT LLM (SA-CBT-LLM), a framework intended to support potential inter-session therapeutic augmentation in CBT. By integrating synthetic data generation with continuity grounded in Beck's CBT manual for stage-aware supervised fine-tuning, the proposed framework models therapeutic progression across the early, middle, and late stages of CBT treatment. Experimental results demonstrate that SA-CBT-LLM is capable of generating more therapeutically aligned responses relative to prevailing general-purpose and mental health LLMs.

For future directions, we plan to further extend this work by directly optimizing therapeutic quality during model training through reinforcement learning-based optimization strategies that explicitly model CBT fidelity using structured reward mechanisms in a stage-aware manner. Such approaches may further improve the alignment between generated responses and clinically grounded therapeutic principles across different stages of treatment. In addition, while the current study primarily relies on automated evaluation metrics and LLM-based assessment, future work may incorporate expert panels and qualitative analysis to provide deeper clinical insight into the therapeutic quality and appropriateness of generated responses. Overall, this work highlights the importance of modeling therapeutic progression and continuity within AI-enabled mental health support systems.

REFERENCES

[1] Abubakar, A.M., Gupta, D. and Parida, S. 2024. A Reinforcement Learning Approach for Intelligent Conversational Chatbot For Enhancing Mental Health Therapy. *Procedia Computer Science*. 235, (2024), 916–925.

[2] Affrunti, N.W. and Creed, T.A. 2019. The factor structure of the cognitive therapy rating scale (CTRS) in a sample of community mental health clinicians. *Cognitive therapy and research*. 43, 3 (June 2019), 642–655.

[3] American Psychological Association 2017. *What is Cognitive Behavioral Therapy?*

[4] Beck, J.S. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

[5] Bennion, M., Blakemore, A., Lovell, K. and Bee, P. 2025. Barriers and facilitators to engagement with between-session work for low-intensity Cognitive Behavioural Therapy (CBT)-based interventions: a qualitative exploration of practitioner perceptions. *BMC psychiatry*. 25, 1 (2025), 79.

[6] Bentley, K.H., Cohen, Z.D., Kim, T., Bullis, J.R., Nauphal, M., Cassiello-Robbins, C., Sauer-Zavala, S., Sbi, S., Gallagher, M.W., Farchione, T.J. and Others 2021. The nature, timing, and symptom trajectories of dropout from transdiagnostic and single-diagnosis cognitive-behavioral therapy for anxiety disorders. *Behavior therapy*. 52, 6 (2021), 1364–1376.

[7] Cao, Z., Zhou, J., Li, M., Huang, J. and Dou, D. 2023. Urbanites' mental health undermined by air pollution. *Nature sustainability*. 6, 4 (2023), 470–478.

[8] Chand, S.P., Kuckel, D.P. and Huecker, M.R. 2023. Cognitive behavior therapy. *StatPearls [Internet]*. StatPearls Publishing.

[9] Crane, K.L. and Watters, K.M. 2025. *Cognitive behavioral therapy strategies*.

[10] Cui, X., Gu, Y., Fang, H. and Zhu, T. 2025. Development and evaluation of LLM-based suicide intervention chatbot. *Frontiers in Psychiatry*. 16, (2025), 1634714.

[11] Cuijpers, P., Miguel, C., Harrer, M., Plessen, C.Y., Ciharova, M., Ebert, D. and Karyotaki, E. 2023. Cognitive behavior therapy vs. control conditions, other psychotherapies, pharmacotherapies and combined treatment for depression: a comprehensive meta-analysis including 409 trials with 52,702 patients. *World psychiatry: official journal of the World Psychiatric Association (WPA)*. 22, 1 (Feb. 2023), 105–115.

[12] David, D., Cristea, I. and Hofmann, S.G. 2018. Why cognitive behavioral therapy is the current gold standard of psychotherapy. *Frontiers in psychiatry*. 9, (Jan. 2018), 4.

[13] Dewi, M. and Fahmi, H. 2023. Implementation of AI Chatbot Application for Social Anxiety Problem. *IT for Society*. 8, 1 (2023).

[14] Dongre, P. 2024. Physiology-Driven Empathic Large Language Models (EmLLMs) for Mental Health Support. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (2024), 1–5.

[15] Fernandez, E., Salem, D., Swift, J.K. and Ramtahal, N. 2015. Meta-analysis of dropout from cognitive behavioral therapy: Magnitude, timing, and moderators. *Journal of consulting and clinical psychology*. 83, 6 (2015), 1108.

[16] Goldberg, S.B., Baldwin, S.A., Merced, K., Caperton, D.D., Imel, Z.E., Atkins, D.C. and Creed, T. 2020. The structure of competence: Evaluating the factor structure of the Cognitive Therapy Rating Scale. *Behavior therapy*. 51, 1 (Jan. 2020), 113–122.

[17] Gonzales, A., Guruswamy, G. and Smith, S.R. 2023. Synthetic data in health care: A narrative review. *PLOS Digital Health*. 2, 1 (2023), e0000082.

[18] Habicht, J., Dina, L.-M., McFadyen, J., Stylianou, M., Harper, R., Hauser, T.U. and Rollwage, M. 2025. Generative AI--enabled therapy support tool for improved clinical outcomes and patient engagement in group therapy: real-world observational study. *Journal of medical Internet research*. 27, (2025), e60435.

[19] Hayes, S.C. and Hofmann, S.G. 2018. *Process-based CBT: The science and core clinical competencies of cognitive behavioral therapy*. New Harbinger Publications.

[20] Ilapaka, A. and Ghosh, R. 2025. A Comprehensive RAG-Based LLM for AI-Driven Mental Health Chatbot. *2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)* (2025), 1–5.

[21] Jafar, A., Shabnam, S. and Khan, A. 2025. Artificial Intelligence in Mental Healthcare: Opportunities, Risks, and Regulations. *AI in Mental Health: Innovations, Challenges, and Collaborative Pathways*. IGI Global Scientific Publishing. 339–378.

[22] Kang, B. and Hong, M. 2025. Development and evaluation of a mental health chatbot using ChatGPT 4.0: Mixed methods user experience study with Korean users. *JMIR Medical Informatics*. 13, (2025), e63538.

[23] Kayaalp, M. 2018. Patient privacy in the era of big data. *Balkan medical journal*. 35, 1 (2018), 8–17.

[24] Kumar, A., Sharma, S., Gupta, S. and Kumar, D. 2024. Mental Healthcare Chatbot Based on Custom Diagnosis Documents Using a Quantized Large Language Model. *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (2024), 1–6.

[25] Li, J., Wang, X., Li, H., Yan, Y., Zhang, S., Leong, H.V., Feng, L., Yu, N.X. and Li, Q. 2026. Multi-agent Causal Reasoning for Suicide Ideation Detection Through Online Conversations. *International Conference on Database Systems for Advanced Applications* (2026), 53–69.

[26] Li, J., Wang, X., Yan, Y., Li, H., Leong, H.V., Yu, N.X. and Li, Q. 2025. DynaProtect: A Dynamic Factor Influence Learning Framework for Protective Factor-aware Suicide Risk Prediction. *Companion Proceedings of the ACM on Web Conference 2025* (2025), 1785–1791.

[27] McFadyen, J., Habicht, J., Dina, L.-M., Harper, R., Hauser, T.U. and Rollwage, M. 2024. AI-enabled conversational agent increases engagement with cognitive-behavioral therapy: A randomized controlled trial. *medRxiv*. (2024), 2024–2011.

[28] Mental_Health_Chatbot: 2024. *https://huggingface.co/tanusrich/Mental_Health_Chatbot*.

[29] Mental-Health-FineTuned-Mistral-7B-Instruct-v0.2: 2024. *https://huggingface.co/prabureddy/Mental-Health-FineTuned-Mistral-7B-Instruct-v0.2*.

[30] mentalhealth-llama-1b-GRPO-gguf: 2025. *https://huggingface.co/XformAI-india/mentalhealth-llama-1b-GRPO-gguf*.

[31] Morrow, E., Zidaru, T., Ross, F., Mason, C., Patel, K.D., Ream, M. and Others 2022. Artificial intelligence technologies and compassion in healthcare: A systematic scoping review. Front Psychol. 2023; 13: 971044.

[32] Nadăș, M., Dioșan, L. and Tomescu, A. 2025. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*. (2025).

[33] Palmer, C.E., Marshall, E., Millgate, E., Warren, G., Ewbank, M., Cooper, E., Lawes, S., Smith, A., Hutchins-Joss, C., Young, J. and Others 2025. Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care. *Journal of Medical Internet Research*. 27, (2025), e69351.

[34] Phillips, L., Cordray, J., Mendoza, K. and Dorn-Medeiros, C. 2023. A closer look at the mental health provider shortage. *Counseling Today*. 65, 11 (2023), 25–39.

[35] Psychosocial-Counseling-Llama-3.1-8B-Instruct: 2024. *https://huggingface.co/SirDrayge/Psychosocial-Counseling-Llama-3.1-8B-Instruct*.

[36] Reategui-Rivera, C.M., Smiley, A. and Finkelstein, J. 2025. LLM-Based Chatbot to Reduce Mental Illness Stigma in Healthcare Providers. *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)* (2025), 00001–00007.

[37] Ren, M., Cao, B., Lin, H., Liu, C., Han, X., Zeng, K., Wan, G., Cai, X. and Sun, L. 2024. Learning or self-aligning? rethinking instruction fine-tuning. *arXiv preprint arXiv:2402. 18243*. (2024).

[38] Ryum, T., Bennion, M. and Kazantzis, N. 2023. Integrating between-session homework in psychotherapy: A systematic review of immediate in-session and intermediate outcomes. *Psychotherapy*. 60, 3 (2023), 306.

[39] Singh, A., Ehtesham, A., Mahmud, S. and Kim, J.-H. 2024. Revolutionizing mental health care through langchain: A journey with a large language model. *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)* (2024), 0073–0078.

[40] Soman, G., Judy, M.V. and Abou, A.M. 2025. Human guided empathetic AI agent for mental health support leveraging reinforcement learning-enhanced retrieval-augmented generation. *Cognitive Systems Research*. 90, (2025), 101337.

[41] Soudani, H., Petcu, R., Kanoulas, E. and Hasibi, F. 2024. A survey on recent advances in conversational data generation. *arXiv preprint arXiv:2405. 13003*. (2024).

[42] Stade, E., Stirman, S.W., Ungar, L.H., Boland, C.L., Schwartz, H.A., Yaden, D.B., Sedoc, J., DeRubeis, R., Willer, R. and Others 2023. Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. (2023).

[43] Stawarz, K., Preist, C., Tallon, D., Wiles, N. and Coyle, D. 2018. User experience of cognitive behavioral therapy apps for depression: an analysis of app functionality and user reviews. *Journal of medical Internet research*. 20, 6 (2018), e10120.

[44] Swift, J.K. and Greenberg, R.P. 2012. Premature discontinuation in adult psychotherapy: a meta-analysis. *Journal of consulting and clinical psychology*. 80, 4 (2012), 547.

[45] Tasnim, R., Siddula, M. and Sai, A.M.V.V. 2025. A Context-Aware Mental Health LLM Chatbot with Enhanced Security. *International Conference on Wireless Artificial Intelligent Computing Systems and Applications* (2025), 1–12.

[46] TinyLlama-1.1B-Chat-v1.0-mental-health-conversational: 2024. *https://huggingface.co/givyboy/TinyLlama-1.1B-Chat-v1.0-mental-health-conversational*.

[47] Ueda, J., Yamaguchi, S., Matsuda, Y., Okazaki, K., Morimoto, T., Matsukuma, S., Sasaki, T. and Kishimoto, T. 2021. A randomized controlled trial evaluating the effectiveness of a short video-based educational program for improving mental health literacy among schoolteachers. *Frontiers in psychiatry*. 12, (2021), 596293.

[48] Wang, L., Chen, S., Jiang, L., Pan, S., Cai, R., Yang, S. and Yang, F. 2025. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review*. 58, 8 (2025), 227.

[49] Weintraub, M.J., Ichinose, M.C., Zinberg, J.L., Salimian, A., Brown, R.D., Morgan-Fleming, G., Gamarra, J.M., Tran, T. and Miklowitz, D.J. 2025. A randomized trial of an app-enhanced group cognitive behavioral therapy for adolescents with mood or psychotic spectrum disorders. *Journal of consulting and clinical psychology*. 93, 3 (2025), 131.

[50] WHO 2025. *Over a billion people living with mental health conditions – services require urgent scale-up*.

[51] Wilhelm, S., Bernstein, E.E., Bentley, K.H., Snorrason, I., Hoeppner, S.S., Klare, D., Greenberg, J.L., Weingarden, H., McCoy, T.H., Harrison, O. and Others 2024. Feasibility, Acceptability, and Preliminary Efficacy of a Smartphone App--Led Cognitive Behavioral Therapy for Depression Under Therapist Supervision: Open Trial. *JMIR Mental Health*. 11, 1 (2024), e53998.

[52] Wu, M.S., Chen, S.-Y., Wickham, R.E., Leykin, Y., Varra, A., Chen, C. and Lungu, A. 2022. Predicting non-initiation of care and dropout in a blended care CBT intervention: Impact of early digital engagement, sociodemographic, and clinical factors. *Digital Health*. 8, (2022), 20552076221133760.

[53] Yang, K., Zhang, T., Kuang, Z., Xie, Q., Huang, J. and Ananiadou, S. 2024. MentaLLaMA: interpretable mental health analysis on social media with large language models. *Proceedings of the ACM on Web Conference 2024* (2024), 4489–4500.

[54] Yasukawa, S., Tanaka, T., Yamane, K., Kano, R., Sakata, M., Noma, H., Furukawa, T.A. and Kishimoto, T. 2024. A chatbot to improve adherence to internet-based cognitive--behavioural therapy among workers with subthreshold depression: a randomised controlled trial. *BMJ Ment Health*. 27, 1 (2024).

[55] Yuan, A., Garcia Colato, E., Pescosolido, B., Song, H. and Samtani, S. 2025. Improving Workplace Well-being in Modern Organizations: A Review of Large Language Model-based Mental Health Chatbots. *ACM transactions on management information systems*. 16, 1 (2025), 1–26.

[56] Zhang, X., Cui, W., Wang, J. and Li, Y. 2024. Chat, Summary and Diagnosis: A LLM-Enhanced Conversational Agent for Interactive Depression Detection. *2024 4th International Conference on Industrial Automation, Robotics and Control Engineering (IARCE)* (2024), 343–348.