CogNote: Generating Structured Client-Centric Counseling Notes from CBT Dialogues

Tiantian Chen[1], Xuri Chen[2], Ying Shen[1*]
[1]*School of Computer Science and Technology, Tongji University, Shanghai, China*
[2]*School of Humanities, Tongji University, Shanghai, China*
2111287@tongji.edu.cn, xurichen@tongji.edu.cn, yingshen@tongji.edu.cn

***Abstract*—Mental counseling plays a crucial role in the prevention and mitigation of mental health disorders. Counseling notes—concise summaries of key session elements—are valuable to both counselors and clients for later review and for consolidating what was discussed. However, prior work on counseling note generation is largely counselor-centric, focusing on documenting clients' symptoms, problems, and diagnoses, while under-serving the client's need to revisit insights, cognitive shifts, and actionable takeaways after the dialogue. To bridge this gap, we introduce client-centric counseling note generation, which aims to produce client-facing notes that support post-session reflection and everyday application. Concretely, we build C-TIND, a dataset of 1,800 cognitive behavioral therapy (CBT) dialogues paired with structured client-centric notes. Building on C-TIND, we develop CogNote, a client-centric counseling note generation model trained on CBT dialogues covering four common CBT techniques, producing structured notes that help clients review key gains and insights from the session. Both automatic and human evaluations show that our approach can generate useful, well-structured client-centric notes. To the best of our knowledge, this is the first work that explicitly formulates and benchmarks counseling note generation from the client's perspective.**

***Index Terms*—dialogue summarization, mental health, online counseling**

I. Introduction

Psychological counseling is a key intervention for managing mental health disorders, where structured dialogues support clients in regulating their mental states. Counseling notes, documented at the end of counseling, serve as a valuable resource for both counselors and clients to review and consolidate key insights. Existing research on counseling note generation predominantly adopts a counselor-centric perspective, focusing on supporting counselors in the diagnosis, assessment, and monitoring of clients' states and symptoms [1]–[4]. For instance, Sahu et al. [5] proposed a mental state examination (MSE) summarization algorithm to evaluate clients' cognitive and behavioral functioning. Srivastava et al. [1] introduced a counseling summarization dataset, MEMO, which annotates key counseling elements such as client symptoms, histories, and therapeutic discoveries. Yao [6] developed the D4 dataset, which includes client symptom records and illness assessments derived from counseling dialogues. Collectively, these studies function as automated counseling summarization or clinical documentation, helping counselors evaluate and record clients' conditions and develop subsequent treatment plans.

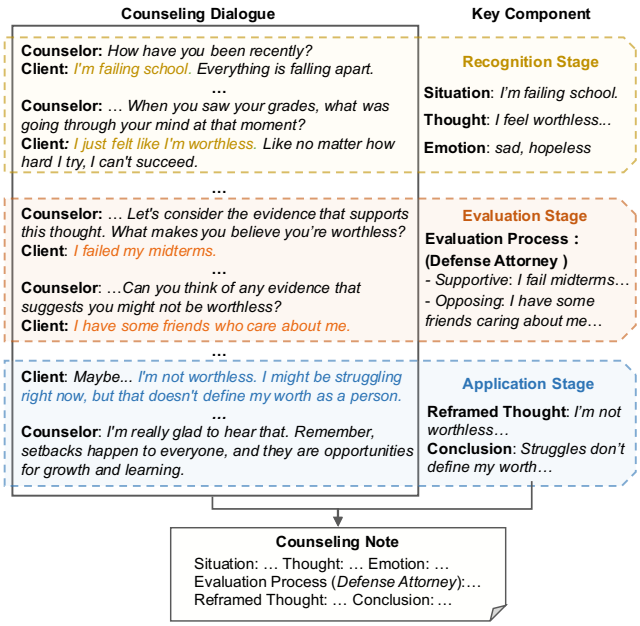

Fig. 1. A CBT counseling note which summarizes the key elements in three phases of a CBT dialogue: recognition, evaluation, and application.

However, such summaries primarily focus on documenting clinical facts and psychological conditions from the counselor's perspective, which makes them less conducive to promoting client self-reflection. From the clients' perspective, counseling notes should highlight their reflections and insights derived from the counseling process, rather than merely recording their symptoms and status. As illustrated in Figure 2, counselor-centric notes primarily document clinical symptoms to support the counselor's diagnostic and treatment decisions, whereas client-centric notes focus more on the client's subjective experiences and cognitive shifts. These client-centric notes can help clients review their insights and growth, thereby enhancing therapeutic effects in the long run [7], [8]. Therefore, it is both necessary and meaningful to generate client-centric counseling notes for clients.

Cognitive behavioral therapy (CBT) is one of the most widely adopted counseling psychotherapies [9]. It encourages

*Corresponding author: Ying Shen.

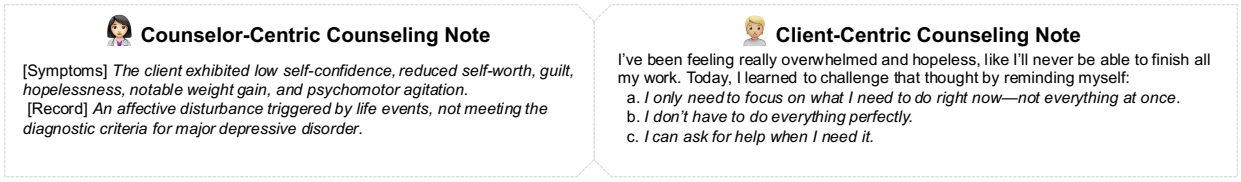

Fig. 2. Comparison of Counselor-Centric and Client-Centric Counseling Note. The counselor-centric note focuses on the objective documentation and clinical assessment of the client's mental state, whereas the client-centric note emphasizes the client's subjective experiences and self-reflection, facilitating the application of counseling insights to real-life contexts.

clients to form their own reflections and insights at the end of counseling sessions. Considering its popularity in psychological counseling, this paper focuses on CBT dialogues to develop a pioneering framework for generating client-centric counseling notes. A typical CBT counseling dialogue consists of three phases: *recognition*, *evaluation*, and *application* [10]. In the recognition phase, the counselor helps the client identify his/her situations, thoughts, and emotions. In the evaluation phase, the counselor applies CBT intervention strategies to address the client's negative thoughts and emotions. The recognition and evaluation phases are the core steps in CBT, aiming to identify and reframe maladaptive thoughts that trigger clients' negative emotions [11], [12]. Finally, in the application phase, both counselors and clients conclude the counseling dialogues, and clients are encouraged to apply the positive reframed thoughts to real-life situations. Therefore, CBT counseling notes should document the core elements in the above three phases, as illustrated in Figure 1.

In this work, we developed **CogNote**, a counseling summarization algorithm designed to generate **Cog**nitive therapy **Note**s for clients. Because CBT dialogues often adopt different *counseling strategies* to assess clients' maladaptive thoughts, the core elements of counseling dialogues and the summarized counseling notes vary accordingly. To accurately generate the counseling notes, CogNote autonomously identifies the CBT strategies adopted and extracts the key elements of the counseling dialogues under these strategies.

To train CogNote, we construct the **C**ognitive **T**herapy **I**nsight **N**otes **D**ataset (**C-TIND**), which comprises 1,800 CBT dialogues and corresponding counseling notes. These CBT dialogues involve four commonly used CBT strategies [10]: the *defense attorney* method, the *possible outcomes* method, the *divergent thinking* method, and the *cost-benefit* method. The corresponding counseling notes are summarized by our proposed method, namely the **T**herapy **I**nsight **Chain** (**TI-Chain**), which systematically guides large language models (LLMs) to generate CBT counseling notes step-by-step. The example note shown in Figure 1 is summarized by TI-Chain, which contains key elements in the three phases of CBT.

The contributions of our work can be summarized as follows:[1] 1) We formulate a client-centric counseling note generation task that produces actionable, client-facing notes to support post-session review and self-reflection. To the best of our knowledge, we are the first to propose documenting psychological counseling notes from the client's perspective. 2) We release C-TIND, a dataset of 1,800 CBT-style dialogues paired with structured client-centric notes. C-TIND captures key CBT elements, enabling research on client-oriented note generation and related downstream tasks. 3) We propose TI-Chain, a stepwise structured summarization pipeline that uses LLMs to generate complete and schema-consistent notes from CBT dialogues. Human assessments confirm that the generated notes are high-quality and consistently formatted. 4) We develop CogNote, a counseling note generation model that infers CBT strategy types and produces strategy-aligned, client-centric notes. CogNote is designed to help clients retrieve key insights, track cognitive shifts, and translate counseling takeaways into actionable reminders for everyday life.

[1]The code and dataset are available at https://github.com/slptongji/CogNote.

II. Related Work

Recent research has primarily focused on automating clinical documentation to assist counselors in diagnosis and treatment planning [6], [13]–[16]. Notable efforts include unsupervised knowledge-infused summarization [14], Q&A generation from clinical records [17], and the development of specialized datasets like MEMO [1], [18] and models like PIECE [19] that integrate psychological expertise into LLMs for counselor-oriented summaries. Despite these advancements, existing studies overlook the therapeutic needs of clients. Client-centric counseling notes—summarizing positive alternative thoughts and actionable advice—are essential for promoting cognitive progress and behavioral change, especially within standard CBT protocols [7], [8], [20]. Leveraging the proven capacity of fine-tuned LLMs in generating high-quality psychological summaries [21], this study introduces the C-TIND dataset and the CogNote model. Our work aims to bridge this gap by automating the generation of client-oriented counseling notes to support long-term self-improvement.

III. C-TIND: Cognitive Therapy Insight Notes Dataset

A. CBT Dialogue Construction

We constructed an artificial counseling dataset by expanding situation-thought pairs from existing cognitive reframing datasets [22], [23] into multi-turn CBT counseling dialogues. We utilized GPT-4o [24] to simulate both the counselor and client roles, and adopt a staged expansion protocol to enforce the standard CBT workflow of *recognition-evaluation-application* (Section I), thereby producing dialogues with consistent phase structure and intervention patterns.

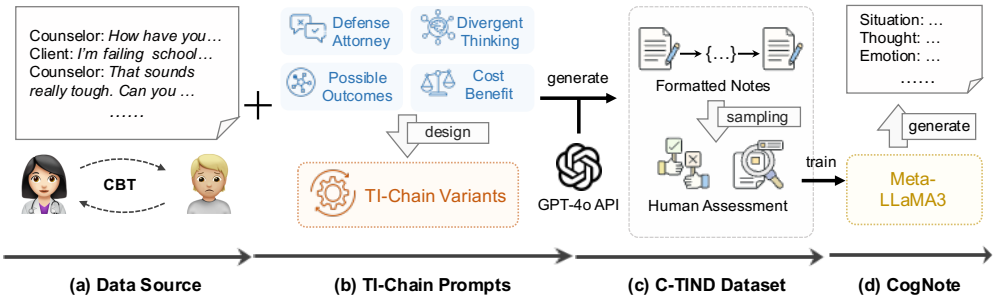

Fig. 3. The proposed framework for counseling note generation: (a)-(c) illustrate the construction process of the C-TIND dataset, while (d) depicts the training process of the CogNote model.

Specifically, in the *recognition* phase, we designed a targeted prompt that elicits the client's context, emotions, and maladaptive thoughts, and generates counselor turns that explicitly guide the identification of cognitive distortions. To ensure conversational plausibility and CBT compliance, two registered counselors manually reviewed each sample and revise cases that do not meet quality standards. In the *evaluation* phase, which constitutes the core intervention step in CBT, we explicitly covered four commonly used thought-evaluation techniques: *defense attorney* (*Def. Atty.*), *possible outcomes* (*Pos. Outs.*), *divergent thinking* (*Div. Tkg.*), and *cost-benefit* (*Cst. Ben.*). For each technique, we developed a dedicated prompt to generate evaluation-phase dialogues that are consistent with the corresponding intervention logic, yielding technique-labeled sessions that support technique-aware analysis and benchmarking. Finally, since the evaluation-phase dialogues already contain the reframed conclusion and the client's reflection, we did not further expand the *application* phase.

TABLE I
THE EVALUATION RESULTS ON THE SAMPLE DIALOGUES OF C-TIND. COUNSELOR RESPONSES WERE ASSESSED BASED ON TECHNICALITY (*Tech.*), EMPATHY (*Emp.*), AND RATIONALITY (*Rat.*). CLIENT RESPONSES WERE EVALUATED BASED ON SPECIFICITY (*Spec.*), RATIONALITY (*Rat.*), AND CONSISTENCY (*Cons.*).

	Counselor			**Client**		
	Tech.	**Emp.**	**Rat.**	**Spec.**	**Rat.**	**Cons.**
Def. Atty.	4.77	4.40	4.48	4.62	4.37	4.73
Pos. Outs.	4.73	4.42	4.45	4.65	4.62	4.72
Div. Tkg.	4.88	4.37	4.55	4.72	4.63	4.73
Cst. Ben.	4.83	4.43	4.47	4.72	4.50	4.70
Average	4.80	4.40	4.49	4.68	4.53	4.72

To ensure data quality, counselors randomly reviewed 400 samples, confirming that the generated dialogues aligned with the CBT principles. Additionally, three trained volunteers rated the quality of the dialogues; their evaluations, as presented in Table I, further validate the high quality and accuracy of the constructed dataset. Finally, we generated 1,800 counseling dialogues, with 450 dialogues corresponding to each CBT strategy. These dialogues serve as the data source for subsequent counseling note generation, with the generated notes expected to capture the key elements of dialogues across the three CBT phases and to reflect the specific strategies employed during the evaluation phase.

Please understand and summarize the important content in a given counseling dialogue.

[Task Description]: In a given dialogue, each utterance consists of the speaker (i.e., Counselor or Client) and the content, written in this format: #[speaker]: [content]. You should summarize the client's situation, original emotion, original thought, and the transformed thought after the dialogue. In addition, you should conclude the whole dialogue.

[Description of Reasoning Process]:
Specifically, you should assume the role of the client and use the client's tone to perform a 6-step reasoning process:

1. Output the "**Situation**": Briefly summarize your current situation in one sentence, which is also the core event discussed in the dialogue.
2. Output the "**Emotion**": Summarize your original emotions with no more than three words.
3. Output the "**Original Thought**": Extract your original thoughts at the beginning of the conversation in one sentence.
4. Output the "**Thought Evaluation**": Extract and list evidence supporting the "Original Thought" as "**Supporting Evidence**", and extract and list evidence against the "Original Thought" as "**Opposing Evidence**".
5. Output the "**Reframed Thought**": Extract your transformed thought at the end of the dialogue in one sentence.
6. Output the "**Conclusion**": Based on the dialogue content and "Alternative Thought", summarize brief suggestions that were given to you in one sentence.

[Dialogue History]: …

Fig. 4. Illustration of the TI-Chain method designed for the defense attorney technique.

B. Counseling Note Generation

Inspired by the Chain-of-Thought (CoT) approach [25], [26], we design the Therapy Insight Chain (TI-Chain) method to guide GPT-4o step by step to generate benchmark counseling notes for 1,800 CBT counseling dialogues. The TI-Chain workflow is illustrated in Figure 4, which consists of three main components:

- **Task Description** outlines the goal of the counseling summarization task and specifies the format of the input dialogue.
- **Description of Reasoning Process** details the requirements for extracting the key counseling elements through six reasoning steps, which align with the three CBT phases: recognition, evaluation, and application. Specifically, steps 1-3 correspond to the recognition phase,

during which the client's situation, emotions, and original thoughts are identified and extracted. Step 4 pertains to the evaluation phase, wherein the key points of thought evaluation are summarized based on specific CBT strategies. As shown in Figure 4, when employing the defense attorney technique, LLMs are instructed to list evidence supporting and refuting the original thoughts. Finally, steps 5-6 correspond to the application phase, wherein the reframed thoughts and conclusions are summarized for the client's consideration.

- **Dialogue History** provides the full content of the counseling dialogue, from which the counseling notes are summarized.

As mentioned in Section III-A, the generated counseling dialogues reflect four CBT intervention strategies, whose evaluation phases are quite different. Accordingly, we design four TI-Chain prompt variants corresponding to the four strategies. These variants follow a similar prompt structure illustrated in Figure 4. The distinction among them is that the "*Thought Evaluation*" in each variant requires extraction of different key elements based on the specific characteristics of the corresponding CBT technique. The descriptions of the thought evaluation process for the four CBT techniques are as follows:

- *Defense Attorney*: Extract and list evidence supporting the "Original Thought" as "Supporting Evidence," and extract and list evidence against the "Original Thought" as "Opposing Evidence."
- *Possible Outcomes*: Extract and list other potential explanations as "Other Explanations."
- *Divergent Thinking*: Extract and list the best-case scenario and corresponding evidence as "Best Case," the worst-case scenario and corresponding evidence as "Worst Case," and the most possible scenario and corresponding evidence as "Most Possible Case."
- *Cost-Benefit*: Extract and list benefits of holding the "Original Thought" as "Benefits," and extract and list drawbacks of holding the "Original Thought" as "Costs."

The counseling notes generated by GPT-4o have inconsistent formats, such as garbled characters and irrelevant sentences, which makes them difficult to use for training models to produce high-quality notes. To solve this problem, we performed two format conversions to ensure format consistency. In the first conversion, free-text counseling notes were transformed into structured dictionaries using automated keyword extraction scripts. These scripts were developed based on common issues identified in 50 randomly selected dialogues, aiming to resolve most of the format errors in the raw data, such as incomplete counseling notes or extraneous double quotation marks. In addition, two volunteers manually reformatted the counseling notes that cannot be converted by these scripts. Once all counseling notes were successfully unified with the same format, they were re-converted into a free-text format for model training.

Figure 3(a)-(c) illustrates the construction process of the C-TIND dataset, which ultimately yields 1,800 counseling notes. The C-TIND dataset can be employed to train models that generate CBT counseling notes for clients, thereby advancing research on client-centric counseling note generation.

C. Dataset Analysis

1) Data Quality Assessment: To evaluate the quality of counseling notes within the C-TIND dataset, we randomly selected 20 counseling dialogues and their corresponding notes for each CBT strategy, yielding a total sample size of 80. Three evaluators were invited to assess the counseling notes for these dialogues based on three criteria: 1) *coherence*: logic, structure, and organization of the counseling notes; 2) *faithfulness*: alignment with the factual content of counseling dialogues, avoiding distortions or errors; and 3) *accuracy*: coverage of the main content and key points of the counseling dialogues. Each counseling note was rated on a 5-point Likert scale, with 5 indicating the highest level of the metric. The overall score for each note was calculated by averaging the scores across these three criteria.

TABLE II
THE ASSESSMENT STATISTICS FOR THE SAMPLED DIALOGUES IN THE C-TIND DATASET.

	Coherence	**Faithfulness**	**Accuracy**	**Overall**
Def. Atty.	4.78	4.65	4.62	4.68
Pos. Out.	4.76	4.62	4.60	4.66
Div. Tkg.	4.82	4.67	4.63	4.71
Cst. Ben.	4.80	4.67	4.67	4.71
Average	4.79	4.65	4.63	4.69

As shown in Table II, the counseling notes attained an average coherence score of 4.79, indicating that the counseling notes are fluent, logical, and well-organized. The average accuracy and faithfulness scores exceeded 4.6, demonstrating that the counseling notes accurately extract the key information from the dialogues. The overall score of 4.69 confirms the counseling notes' correctness in analyzing the core elements of CBT dialogues. In addition, we employed the Mean Absolute Deviation (MAD) [27] to evaluate inter-annotator agreement in our data quality assessment. The MAD scores for the three metrics remain below 0.5, indicating a high level of agreement among annotators [28].

2) Data Characteristics: Some statistics about C-TIND are listed in Table III. The C-TIND dataset consists of 1,800 CBT counseling dialogues with corresponding counseling notes. On average, each dialogue has ˜31 utterances and the counseling note has ˜158 words.

Figure 1 presents a brief counseling note from the C-TIND dataset, which contains the key elements of a CBT dialogue across its three stages. In the recognition stage, the note takes down the client's situation, original thoughts, and emotions. In the evaluation stage, the note summarizes key discussion points associated with a specific CBT strategy like defense attorney technique, aiming to assess the validity and usefulness of the client's thoughts. In the application stage, the note documents the reframed thoughts developed by both the

TABLE III
THE CHARACTERISTICS OF THE C-TIND DATASET. *Dia. Len.*, *Utt. Len.*, AND *Not. Len.* REFER TO THE AVERAGE LENGTHS OF DIALOGUES, UTTERANCES, AND COUNSELING NOTES, RESPECTIVELY. ADDITIONALLY, *Dia. Num.* DENOTES THE TOTAL NUMBER OF COUNSELING DIALOGUES.

	Dia. Len.	**Utt. Len.**	**Not. Len.**	**Dia. Num.**
Def. Atty.	30.74	27.41	165.38	450
Pos. Outs.	31.18	27.16	189.06	450
Div. Tkg.	34.42	25.46	137.59	450
Cst. Ben.	28.89	28.17	138.30	450
Total	31.30	26.99	157.60	1800

client and counselor, and provides conclusions based on the dialogue. Such counseling notes serve as effective reviews of CBT counseling dialogues, enhancing clients' understanding of the counseling content and encouraging clients to apply reframed thoughts in their daily life.

IV. COGNOTE: CLIENT-CENTRIC COUNSELING NOTE GENERATION MODEL

Based on the C-TIND dataset, we fine-tuned LLaMA3-8B-Instruct [29] to develop CogNote, a counseling note generation algorithm that extracts key elements from CBT counseling dialogues to support clients. Specifically, 400 dialogues from C-TIND constitute the test set, 200 constitute the validation set, and the remainder constitutes the training set. Given the relatively small size of the C-TIND dataset, Low-Rank Adaptation (LoRA) [30] was employed to reduce the risk of overfitting. During training, the batch size was set to 2 and the gradient accumulation step to 8, resulting in an effective batch size of 16. The LoRA rank was set to 8, with a LoRA scaling factor α of 16, a dropout rate of 0.05, and a LoRA learning rate of 16. CogNote was trained using the AdamW [31] optimizer, with a maximum learning rate of 3e-4. The training was conducted over 3 epochs. All experiments were performed using the LLaMA Factory framework [32] on two NVIDIA L40 GPUs, each equipped with 48GB of memory.

V. EXPERIMENTS AND ANALYSIS

A. Baseline Settings

1) Base Models: We selected four representative open-source LLMs as our base models: ChatGLM3-6B, ChatGLM3-6B-base [33], Vicuna-7B-v1.5 [34], and LLaMA3-8B-Instruct [29]. ChatGLM3-6B and ChatGLM3-6B-base share a general language model architecture [35] and contain 6.2 billion parameters. Their difference is that ChatGLM3-6B-base is the foundational model not trained to align with human intent, whereas ChatGLM3-6B is intent-aligned for multi-turn conversational tasks [36]. Vicuna-7B-v1.5 is fine-tuned on LLaMA2 using shareGPT conversational data, demonstrating good interactivity and context understanding. LLaMA3-8B-Instruct is the instruction-tuned version of Meta-Llama-3 with 8 billion parameters, optimized for dialogue scenarios.

2) Prompt and Fine-tuning Settings: We adapted each base model using prompt-based and fine-tuning methods to instruct LLMs to generate counseling notes. The first prompt, referred to as KP, instructs models based on known CBT strategies, following a structure similar to that shown in Figure 4. Additionally, the KP prompt provides a generated example under the corresponding strategy to standardize the generation format. In contrast, the second prompt, referred to as UP, guides models under conditions of unknown strategies. The two prompts were applied to the above four base models, resulting in eight baseline models. Furthermore, we fine-tuned the three base models except LLaMA3-8B-Instruct with the C-TIND dataset, resulting in another three baseline models. The fine-tuning process followed the same process and hyper-parameters as CogNote, as illustrated in Section IV.

B. Automatic Evaluation

We report BLEU-4 [37] and ROUGE-1/2/L [38] to measure lexical overlap between generated and reference counseling notes. Table IV presents the automatic evaluation results. As shown in the last two groups of Table IV, the proposed CogNote model outperforms all baselines across BLEU-4, ROUGE-1, ROUGE-2, and ROUGE-L, achieving scores of 73.890, 71.260, 52.318, and 62.177, respectively, demonstrating its strong ability to generate accurate counseling notes. GLM-base-FT surpasses GLM-chat-FT across all metrics, suggesting that foundational models are better suited for summarization tasks than intent-aligned conversational models, likely due to their more QA-oriented training format. Comparing the first and second groups in Table IV, models using KP prompts surpass those using UP prompts across all metrics. For instance, GLM-base-KP exceeds GLM-base-UP by 2.816, 7.148, 7.896, and 8.659 points on BLEU-4, ROUGE-1, ROUGE-2, and ROUGE-L, respectively. This improvement is attributed to the KP prompt's explicit inclusion of CBT strategies and examples, reducing misidentification errors seen with UP prompts. Among prompt-based methods, LLaMA3 models achieve the best results, with Vicuna-based models ranking second, and GLM-based models performing lowest.

Although the fine-tuning-based methods are not provided with detailed task definitions or employed CBT strategies, they effectively learn the summarization patterns and demonstrate superior performance compared to the prompt-based methods. For example, although both are based on LLaMA3-8B-Instruct, CogNote surpasses LLaMA-KP by 16.439, 8.423, 16.869, and 16.652 on BLEU-4, ROUGE-1, ROUGE-2, and ROUGE-L metrics, respectively. This substantial improvement indicates that the fine-tuning approach is highly effective in enabling LLMs to comprehend task definitions, identify CBT strategies applied within dialogues, and generate accurate and formatted counseling notes.

C. Human Evaluation

To evaluate the generated counseling notes, two human evaluation experiments – human rating and human A/B test –

TABLE IV
AUTOMATIC EVALUATION RESULTS ON THE C-TIND TEST SET. *GLM-base*, *GLM-chat*, *Vicuna*, AND *LLaMA* INDICATE THAT THE MODELS ARE ADAPTED FROM CHATGLM3-6B-BASE, CHATGLM3-6B, VICUNA-7B-V1.5, AND LLAMA3-8B-INSTRUCT, RESPECTIVELY.

Models	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
GLM-base-UP	46.569	49.763	26.116	35.499
GLM-chat-UP	47.838	49.077	20.614	30.330
Vicuna-UP	52.301	51.498	23.670	34.910
LLaMA-UP	50.131	55.678	26.660	36.772
GLM-base-KP	49.385	56.911	34.012	44.158
GLM-chat-KP	58.356	55.708	27.498	40.349
Vicuna-KP	59.715	57.215	28.814	41.814
LLaMA-KP	57.451	62.837	35.449	45.525
GLM-base-FT	73.787	70.672	51.343	61.397
GLM-chat-FT	73.375	70.436	50.953	60.842
Vicuna-FT	73.889	71.083	51.938	61.877
CogNote	**73.890**	**71.260**	**52.318**	**62.177**

were conducted on 50 randomly selected samples from the C-TIND test set. Both experiments were performed by three evaluators. In the human rating experiment, evaluators assessed each counseling note in terms of coherence, faithfulness, and accuracy, with a rating scale from 1 to 5. The overall score for each note was calculated by averaging the scores on the three criteria. In the A/B test, evaluators compared the counseling notes generated by CogNote with those generated by other models and selected the preferred note for each dialogue. A "Tie" option was provided for cases where evaluators deemed the quality of counseling notes produced by the two models equivalent. All fine-tuned baseline models were involved in the human evaluations. However, due to the poor performance of prompt-based methods, only the best-performing models, i.e., LLaMA-UP and LLaMA-KP, were selected as representatives for the human evaluations.

TABLE V
RESULTS OF HUMAN RATINGS ON 50 SAMPLES SELECTED FROM THE C-TIND TEST SET.

Models	Coherence	Faithfulness	Accuracy	Overall
LLaMA-UP	3.75	3.65	3.61	3.67
LLaMA-KP	3.86	3.93	3.99	3.92
GLM-base-FT	4.41	4.21	4.28	4.30
GLM-chat-FT	4.35	4.21	4.21	4.26
Vicuna-FT	4.40	4.25	4.35	4.33
CogNote	4.43	4.30	4.37	4.36

1) Human Ratings: As shown in Table V, CogNote outperforms all baseline models across three human evaluation criteria. It achieves the highest scores of 4.43, 4.30, and 4.37 for coherence, faithfulness, and accuracy, respectively, with an overall score of 4.36. These results indicate that the counseling notes generated by CogNote are deemed satisfactory. The three models fine-tuned on C-TIND, i.e., GLM-base-FT, GLM-chat-FT, and Vicuna-FT, also deliver strong performance, with all scores above 4.2 on the three metrics. Notably, Vicuna-FT achieves the highest overall score of 4.33 among these models. GLM-base-FT outperforms GLM-chat-FT, particularly in coherence and accuracy, suggesting that the foundational model generates more accurate and coherent counseling notes than the conversational model. In contrast, the prompt-based methods, LLaMA-UP and LLaMA-KP, perform the worst with overall scores of 3.67 and 3.92, which are 0.69 and 0.44 points lower than those of CogNote.

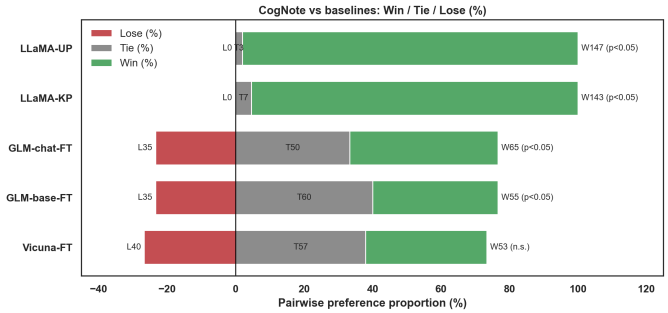

Fig. 5. Human A/B preference evaluation between CogNote and five baselines. Win/Tie/Lose indicate the proportions favoring CogNote, ties, and favoring the baseline, respectively; significance is assessed by an exact binomial sign test on non-tied comparisons ($p < 0.05$).

2) Human A/B Test: Figure 5 reports pairwise preferences between CogNote and five baselines on 50 cases judged by three volunteers. CogNote significantly outperforms the prompt-based methods, achieving 98% wins over LLaMA-UP and 95% wins over LLaMA-KP, indicating a near-unanimous preference for CogNote-generated counseling notes. It also maintains a positive margin over fine-tuned open-source baselines (GLM-chat-FT: 43% wins vs. 23% losses; GLM-base-FT: 37% wins vs. 23%), while the comparison with Vicuna-FT is closer (35% vs. 27%) with many ties. We pool judgments and apply an exact binomial sign test on non-tied comparisons ($p < 0.05$), finding significant gains over LLaMA-UP/KP and both GLM variants, but not over Vicuna-FT. Overall, CogNote yields consistently preferred counseling notes, with particularly large improvements over prompt-based baselines.

D. Case Study

To further analyze model performance, we conducted a case study on LLaMA-UP, LLaMA-KP, and CogNote, all based on LLaMA3-8B-Instruct. Figure 6 presents the counseling notes generated by these models for a CBT dialogue in which the client expressed self-doubt and the counselor used the defense attorney technique to reframe maladaptive thoughts.

As shown in Figure 6(a), LLaMA-UP captures the core elements of the dialogue and produces a relatively accurate counseling note. However, despite instructions to write in the first person, its narration does not consistently follow this perspective, reducing the client-centered quality of the note. In addition, without explicit strategic guidance, LLaMA-UP sometimes misidentifies the CBT strategy used, leading to an inaccurate summary of the thought evaluation process. In contrast, Figure 6(b) shows that LLaMA-KP generates a more precise summary and consistently adopts the first-person perspective. However, it is less concise in summarizing the evaluation process, occasionally shows formatting

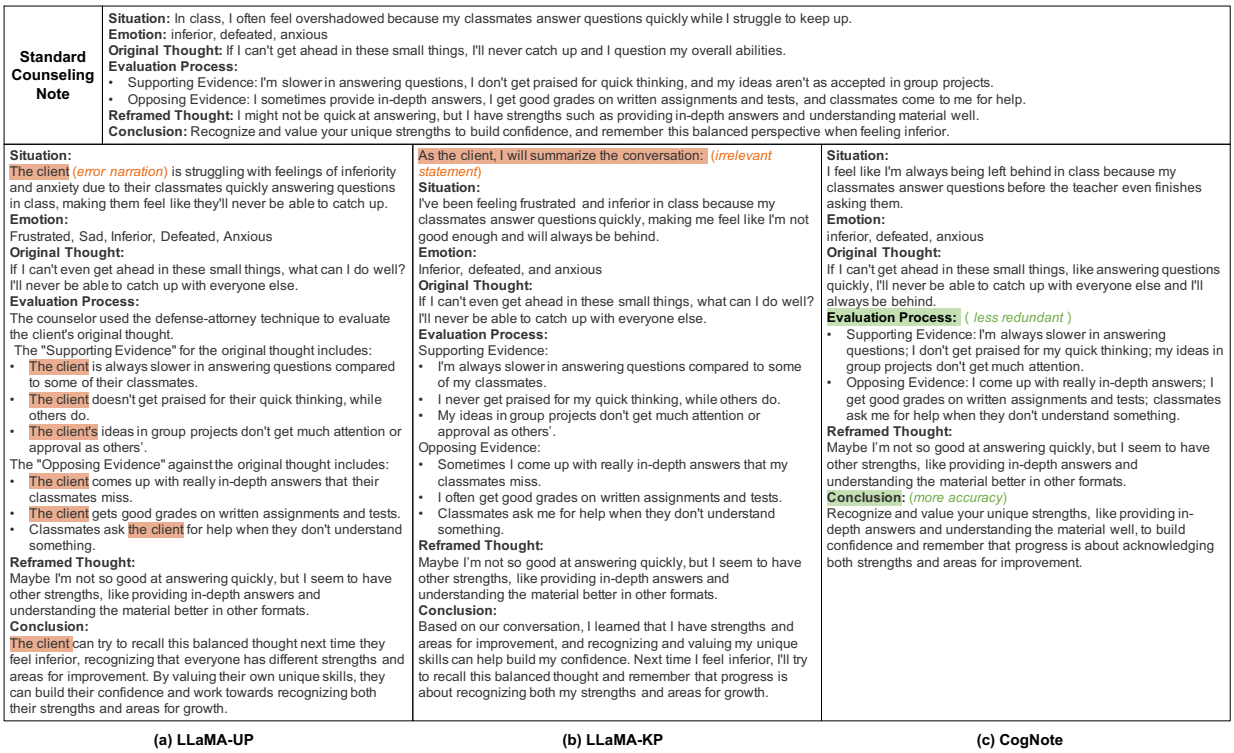

Fig. 6. Results of the case study on the C-TIND test set. The top of the table presents the standard counseling note for a given dialogue, while the bottom displays the counseling notes generated by LLaMA-UP, LLaMA-KP, and CogNote, respectively.

inconsistencies, and may include irrelevant statements such as "*As the client, I will summarize the conversation*". As shown in Figure 6(c), the fine-tuned model CogNote produces a counseling note that closely matches the reference note in both content and format. Unlike the other models, it avoids redundant symbols and unnecessary sentences while accurately capturing the core elements of the CBT dialogue. Such notes can help clients quickly recall key events and maladaptive thoughts discussed in counseling. In addition, the inclusion of alternative thoughts and conclusions may reinforce adaptive thinking and encourage more positive actions and attitudes in real-life situations. These results suggest that our client-centric framework, CogNote, has strong potential to improve the effectiveness of CBT interventions.

VI. Limitations

C-TIND is constructed as a controlled benchmark for client-centric counseling note generation. Due to privacy and ethical constraints, publicly available CBT session transcripts with fine-grained annotations are extremely limited. We therefore synthesize CBT-style dialogues and client-facing notes with careful schema design and human verification. While this setting enables scalable and standardized benchmarking, it may not fully capture the linguistic variability and interaction dynamics of real-world counseling sessions (e.g., hesitation, off-topic turns, or incomplete disclosure). Consequently, our findings primarily demonstrate feasibility under this benchmark, and clinical effectiveness or deployment in real counseling settings is beyond the scope of this paper.

VII. Conclusion

Existing work on counseling summarization is largely counselor-centric, emphasizing symptom and diagnosis documentation, and may not fully support clients' post-session review (e.g., revisiting key insights and consolidating cognitive shifts). To address this gap, we formulate a client-centric counseling note generation task and build C-TIND, a dataset of 1,800 CBT-style dialogues paired with structured client-centric notes. We construct C-TIND via TI-Chain, a six-step CBT-aligned structured generation pipeline that guides LLMs to produce schema-consistent notes capturing core CBT elements and technique-specific evaluation components. Building on C-TIND, we develop CogNote, a client-oriented note generator fine-tuned from LLaMA3-8B-Instruct, to produce structured notes that facilitate later review and between-session self-reflection. Overall, our results indicate that CogNote can generate useful and well-structured client-centric counseling notes, establishing a benchmark and baseline for future work.

Acknowledgment

This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project under Grant 2025ZD0123701, and in part by the National Natural Science Foundation of China under Grant 62476202 and 62272343.

REFERENCES

[1] A. Srivastava, T. Suresh, S. P. Lord, M. S. Akhtar, and T. Chakraborty, "Counseling summarization using mental health knowledge guided utterance filtering," in *Proc. ACM SIGKDD Conf. Knowl. Disc. Data Mining*. Washington DC, USA: Association for Computing Machinery, 2022, p. 3920–3930.

[2] S. Sotudeh, N. Goharian, and Z. Young, "MentSum: A resource for exploring summarization of mental health online posts," in *Proc. Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, 2022, pp. 2682–2692.

[3] P. K. Adhikary, A. Srivastava, S. Kumar, S. M. Singh, P. Manuja, J. K. Gopinath, V. Krishnan, S. K. Gupta, K. S. Deb, and T. Chakraborty, "Exploring the efficacy of large language models in summarizing mental health counseling sessions: Benchmark study," *JMIR Ment Health*, vol. 11, p. e57306, 2024.

[4] J.-h. So, J. Chang, E. Kim, J. Na, J. Choi, J.-y. Sohn, B.-H. Kim, and S. H. Chu, "Aligning large language models for enhancing psychiatric interviews through symptom delineation and summarization: Pilot study," *JMIR Form Res*, vol. 8, p. e58418, 2024.

[5] N. K. Sahu, M. Yadav, M. Chaturvedi, S. Gupta, and H. R. Lone, "Leveraging language models for summarizing mental state examinations: A comprehensive evaluation and dataset release," in *Proc. Int. Conf. Comput. Linguist.* Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 2658–2682.

[6] B. Yao, C. Shi, L. Zou, L. Dai, M. Wu, L. Chen, Z. Wang, and K. Yu, "D4: a Chinese dialogue dataset for depression-diagnosis-oriented chat," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 2438–2459.

[7] F. Bernardi, "Note-taking during counselling sessions: A mixed-methods research on the client's perspective," 2015, unpublished. [Online]. Available: https://openaccess.city.ac.uk/id/eprint/15157/

[8] K. Tudor and K. Gledhill, "Notes on notes: Note-taking and record-keeping in psychotherapy," *Ata: Journal of Psychotherapy Aotearoa New Zealand*, vol. 26, no. 2, pp. 123–144, 2022.

[9] A. T. Beck, "Cognitive therapy: Nature and relation to behavior therapy," *Behavior Therapy*, vol. 1, no. 2, pp. 184–200, 1970.

[10] R. L. Leahy, *Cognitive therapy techniques: A practitioner's guide*. New York, USA: Guilford Publications, 2017.

[11] J. S. Beck and J. Wright, "Cognitive therapy: Basics and beyond," *J. Psychother. Pract. Res*, vol. 6, pp. 71–80, 1997.

[12] C. L. Yurica and R. A. DiTomasso, "Cognitive distortions," *Encyclopedia of cognitive behavior therapy*, pp. 117–122, 2005.

[13] M. Afzal, F. Alam, K. M. Malik, and G. M. Malik, "Clinical context–aware biomedical text summarization using deep neural network: Model development and validation," *J Med Internet Res*, vol. 22, no. 10, p. e19810, 2020.

[14] G. Manas, V. Aribandi, U. Kursuncu, A. Alambo, V. L. Shalin, K. Thirunarayan, J. Beich, M. Narasimhan, A. Sheth *et al.*, "Knowledge-infused abstractive summarization of clinical diagnostic interviews: Framework development study," *JMIR Mental Health*, vol. 8, no. 5, p. e20865, 2021.

[15] A. Tiwari, S. Bera, S. Saha, P. Bhattacharyya, and S. Ghosh, "Yes, this is what i was looking for! towards multi-modal medical consultation concern summary generation," in *Proc. Adv. Infor. Retr.* Glasgow, UK: Springer Nature Switzerland, 2024, pp. 210–225.

[16] Y. Song, Y. Tian, N. Wang, and F. Xia, "Summarizing medical conversations via identifying important utterances," in *Proc. Int. Conf. Comput. Linguist.* Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 717–729.

[17] J. Ive, N. Viani, J. Kam, L. Yin, S. Verma, S. Puntis, R. N. Cardinal, A. Roberts, R. Stewart, and S. Velupillai, "Generation and evaluation of artificial mental health records for natural language processing," *NPJ digital medicine*, vol. 3, no. 1, p. 69, 2020.

[18] G. Malhotra, A. Waheed, A. Srivastava, M. S. Akhtar, and T. Chakraborty, "Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations," in *Proc. ACM Int. Conf. Web Search Data Mining*. New York, USA: Association for Computing Machinery, 2022, p. 735–745.

[19] A. Srivastava, S. Joshi, T. Chakraborty, and M. S. Akhtar, "Knowledge planning in large language models for domain-aligned counseling summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 17 775–17 789.

[20] J. H. Wright, "Cognitive behavior therapy: Basic principles and recent advances," *Focus*, vol. 4, no. 2, pp. 173–178, 2006.

[21] P. K. Adhikary, A. Srivastava, S. Kumar, S. M. Singh, P. Manuja, J. K. Gopinath, V. Krishnan, S. K. Gupta, K. S. Deb, and T. Chakraborty, "Exploring the efficacy of large language models in summarizing mental health counseling sessions: Benchmark study," *JMIR Mental Health*, vol. 11, p. e57306, 2024.

[22] A. Sharma, K. Rushton, I. Lin, D. Wadden, K. Lucas, A. Miner, T. Nguyen, and T. Althoff, "Cognitive reframing of negative thoughts through human-language model interaction," in *Proc. Annu. Meeting Assoc. Comput. Linguist.* Toronto, Canada: Association for Computational Linguistics, 2023, pp. 9977–10 000.

[23] B. Wang, P. Deng, Y. Zhao, and B. Qin, "C2D2 dataset: A resource for the cognitive distortion analysis and its impact on mental health," in *Proc. Findings Assoc. Comput. Linguist.: EMNLP 2023*. Singapore: Association for Computational Linguistics, 2023, pp. 10 149–10 160.

[24] OpenAI, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[25] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Int. Conf. Neural Infor. Process. Syst.* Vancouver, BC, Canada: Curran Associates Inc., 2024.

[26] Z. Huang, J. Zhao, and Q. Jin, "Ecr-chain: Advancing generative language models to better emotion-cause reasoners through reasoning chains," in *Proc. Int. Joint Conf. Artif. Intell.* Jeju, South Korea: International Joint Conferences on Artificial Intelligence Organization, 2024, pp. 6288–6296.

[27] S. Vanbelle, C. H. Engelhart, and E. Blix, "A comprehensive guide to study the agreement and reliability of multi-observer ordinal data," *BMC Medical Research Methodology*, vol. 24, no. 1, p. 310, 2024.

[28] J. Manning, J. Baldwin, and N. P. and, "Human versus machine: The effectiveness of chatgpt in automated essay scoring," *Innovations in Education and Teaching International*, vol. 0, no. 0, pp. 1–14, 2025.

[29] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[30] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent.* Online: International Conference on Learning Representations, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.* New Orleans, Louisiana, USA: International Conference on Learning Representations, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[32] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, and Z. Luo, "LlamaFactory: Unified efficient fine-tuning of 100+ language models," in *Proc. Annu. Meeting Assoc. Comput. Linguist.* Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 400–410.

[33] GLM, "Chatglm: A family of large language models from glm-130b to glm-4 all tools," 2024. [Online]. Available: https://arxiv.org/abs/2406.12793

[34] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," in *Proc. Int. Conf. Neural Infor. Process. Syst.* Red Hook, NY, USA: Curran Associates Inc., 2024.

[35] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "GLM: General language model pretraining with autoregressive blank infilling," in *Proc. Annu. Meeting Assoc. Comput. Linguist.* Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 320–335.

[36] THUDM, "Thudm/chatglm3-6b-base," 2023, accessed: 2025-02-15. [Online]. Available: https://huggingface.co/THUDM/chatglm3-6b-base

[37] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. Annu. Meeting Assoc. Comput. Linguist.* Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[38] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013/