生成式影像合成 Midterm Report

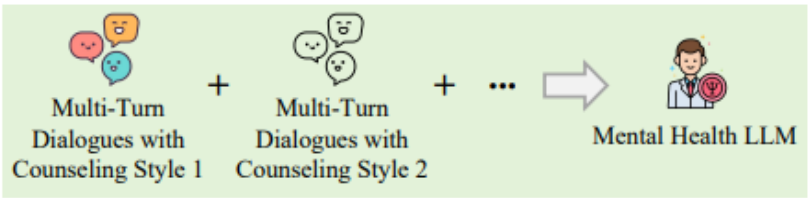
第九組組員: 314581006 林鼎翔、313581037 許哲瑜

Title: PsyDT: Using LLMs to Construct the Digital Twin of Psychological Counselor with Personalized Counseling Style for Psychological Counseling

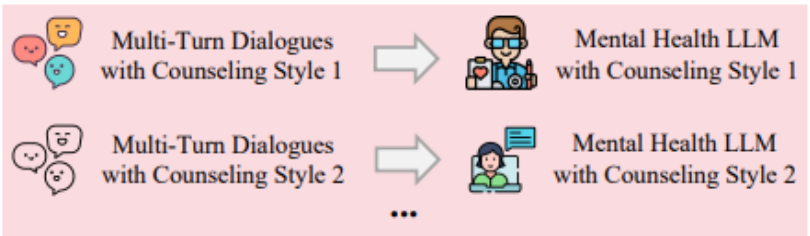
Authors: Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, Xiangmin Xu

1. Introduction: Background, Motivation

In recent years, large language models (LLMs) such as ChatGPT and GPT-4 have significantly changed the way people interact with computers. These models can write essays, answer questions, and even hold meaningful conversations, which has inspired researchers to explore their potential in psychological counseling. LLMs are particularly promising for offering emotional support and guidance, especially when real human resources are limited. However, most existing mental-health LLMs—such as SoulChat, PsyChat, or MeChat—treat all users the same way and overlook the fact that every psychological counselor has a unique personal style. Some counselors tend to be logical and analytical, while others emphasize empathy and emotional support. Likewise, clients have diverse personalities and needs, making a "one-size-fits-all" model inadequate for real-world counseling scenarios.



(a) Previous methods.



(b) Our proposed PsyDT framework.

To solve this issue, the authors propose PsyDT, a framework that uses LLMs to automatically construct the digital twin of a psychological counselor with a personalized counseling style. Instead of collecting large amounts of private, real-world therapy data—which is time-consuming, costly, and raises privacy concerns—PsyDT leverages GPT-4 to simulate realistic multi-turn counseling dialogues that combine a counselor's linguistic style, therapy technique, and the client's personality traits. This approach provides a faster, safer, and more cost-effective way to reproduce authentic counseling interactions while preserving individualization and privacy protection.

2. Related work

Prior studies on mental-health dialogue generation largely build multi-turn counseling data from single-turn or report-style sources, then fine-tune medium-sized Chinese LLMs. SMILE begins with public single-turn Q&A (e.g., PsyQA) and uses a tailored prompt that instructs ChatGPT to act as both psychologist and rewriter, expanding each item into realistic dialogues with at least ten turns (the SMILEChat dataset), which are then used to fine-tune ChatGLM2-6B to obtain MeChat. SoulChat scales this paradigm by outsourcing the collection of large volumes of Chinese single-turn counseling Q&A and, inspired by SMILE, applying an empathy-oriented prompt that explicitly amplifies listening, comfort, understanding, and trust; the ChatGPT-generated multi-turn dialogues undergo human post-editing to produce the SoulChatCorpus with over two million samples, subsequently used to fine-tune ChatGLM-6B and markedly strengthen empathetic behaviors.

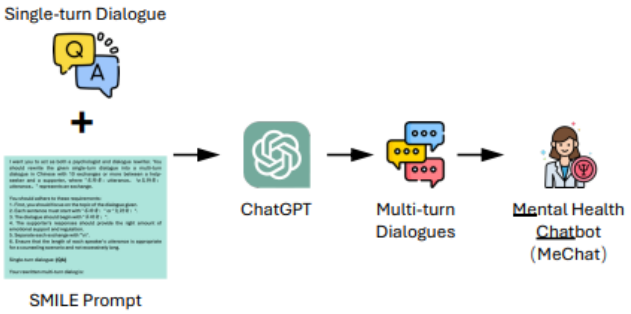


Figure 1: The SMILE method used to generate dialogues for mental health support.

In contrast, CPsyCoun sources anonymized counseling reports and introduces the two-stage Memo2Demo pipeline: reports are first converted into counseling notes and transformed into

dialogues under a professional counseling framework (inquiry, diagnosis, counseling, summary/review), after which GPT-4 evaluates single-turn performance against a predefined rubric and aggregates the scores to the multi-turn level. Training InternLM2-7B-Chat on the resulting CPsyCounD dataset yields CPsyCounX.

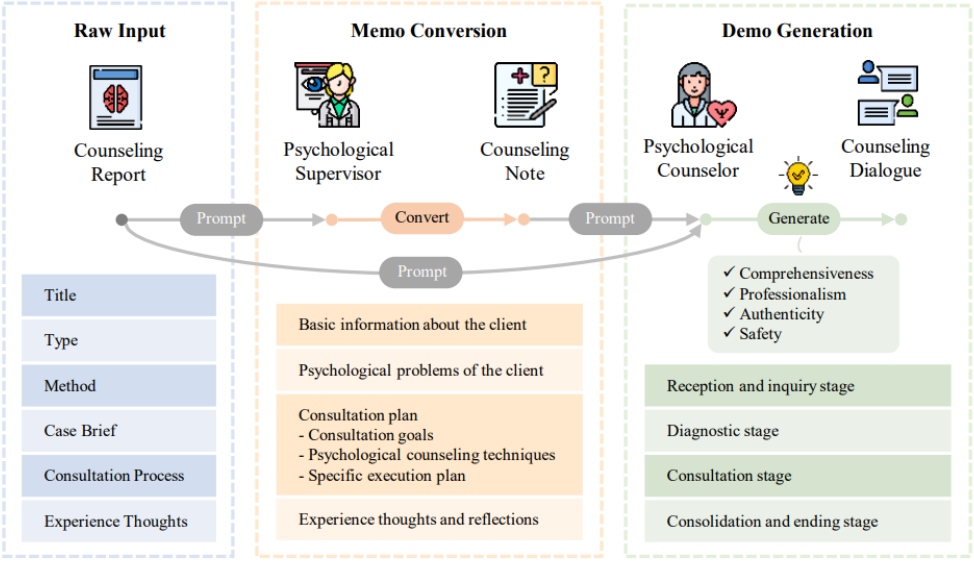


Figure 3: Illustration of the dialogue reconstruction method Memo2Demo

Collectively, these lines of work differ in supervision sources (public Q&A vs. outsourced Q&A vs. real reports), emphases (multi-turn synthesis vs. empathy at scale vs. framework-guided reconstruction with GPT-4 evaluation), and target models (MeChat, SoulChat, CPsyCounX).

3. Technical part: Method

The PsyDT framework consists of several main stages that work together to simulate realistic counseling interactions. First, the researchers selected 5,000 high-quality single-turn dialogues from the SoulChatCorpus dataset, which served as the foundation for generating longer, multi-turn conversations. Next, they invited a licensed professional counselor to conduct 12 real text-based sessions on different topics such as family, emotion, and relationships. GPT-4 then analyzed these cases to summarize the counselor's linguistic style and therapy technique, which primarily followed the Rational Emotive Behavior Therapy (REBT) approach.

After establishing the counselor's profile, GPT-4 was used to simulate clients' Big Five personality traits—openness, conscientiousness, extraversion, agreeableness, and

neuroticism—based on their questions. Combining the counselor's style, therapy type, and client personality, GPT-4 synthesized complete multi-turn dialogues, forming a new dataset called PsyDTCorpus. Finally, an open-source model, Qwen2-7B-Instruct, was fine-tuned on PsyDTCorpus through Multi-Turn Instruction Fine-Tuning (MIFT), resulting in PsyDTLLM—a digital twin model capable of responding in a counselor's unique and consistent style.

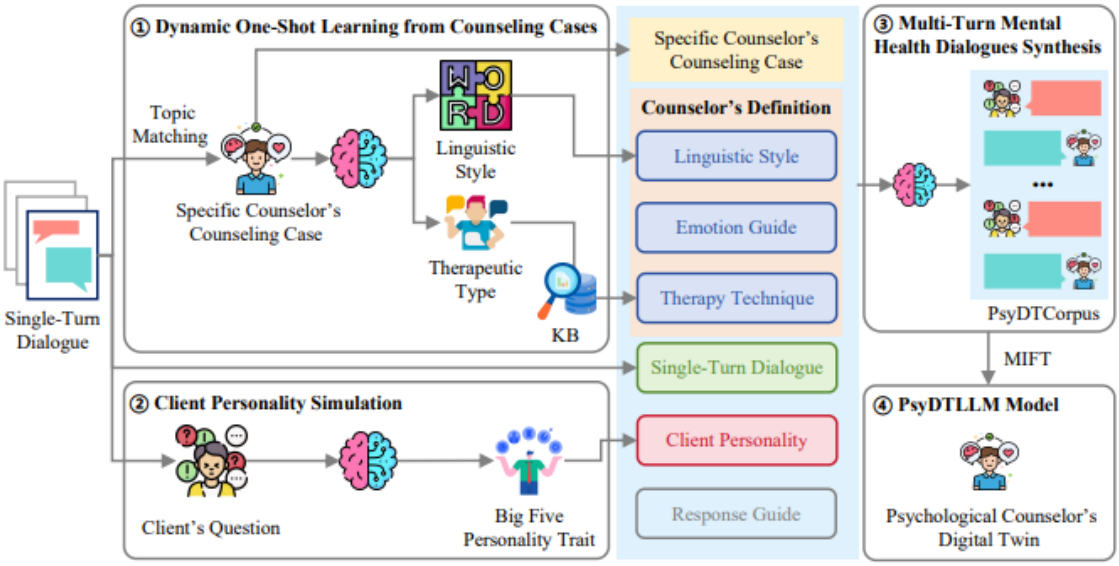
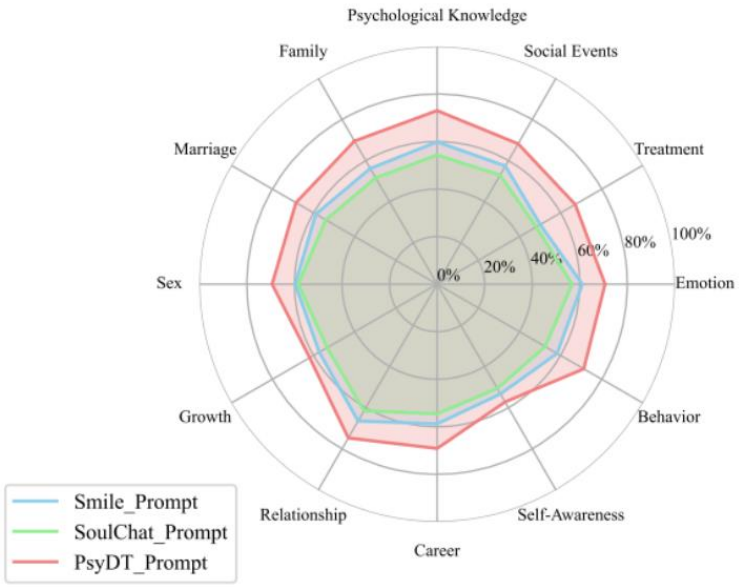


Figure 2: Illustration of multi-turn dialogues synthesis method of PsyDT framework and PsyDTLLM model.

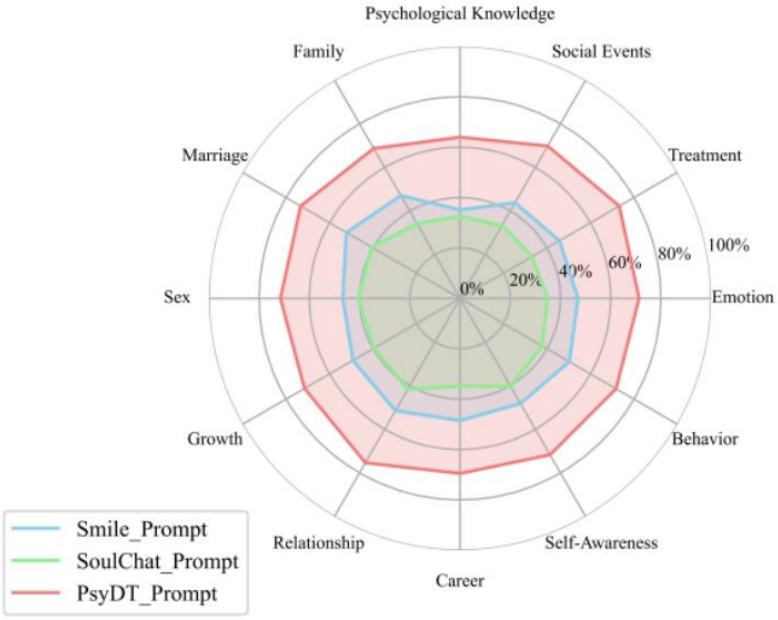
4. Result: Ablation Study, Experimental Result

4.1 Similarity with Real Counseling

The team compared PsyDT's generated dialogues with real counseling cases. Two strong evaluators—GPT-4o and Claude 3.5—were used to judge similarity in linguistic style and therapy technique. PsyDT achieved about 70% similarity, outperforming other methods such as SoulChat and SMILE. This means PsyDT successfully captured how a real counselor speaks and behaves.



(a) Linguistic Style



(b) Therapy Technique

4.2 Dataset Comparison

When human experts evaluated four datasets (SMILEChat, SoulChatCorpus, CPsyCounD, and PsyDTCorpus), they rated PsyDTCorpus the highest in quality across professional categories like: Conversation Strategy, State and Attitude, Relationship Building, Application of Therapy Techniques.

Table 1: Dataset evaluation results. The best score for each metric is **in-bold**, while the second best score is underlined.

Datasets	Statistics					Abilities			Expert Evaluation					
	Open.	Size	NoT.	LoC.	LoP.	EmoE.	CogE.	TheT.	Con.	Sta.	Rel.	App.	Flu.	Saf.
SMILECHAT	✓	56k	10.4	26.1	28.9	✓			5.38	5.92	5.65	4.37	0.84	✓
SoulChatCorpus	✓	258k	5.9	41.4	90.0	✓			5.24	5.80	5.62	4.38	0.86	✓
CPsyCounD	✓	3.1k	8.0	32.9	52.6	✓		✓	5.57	6.02	5.66	5.49	0.72	✓
PsyDTCorpus	✓	5k	18.1	31.6	58.1	✓	✓	✓	**8.39**	**8.69**	**8.29**	**8.12**	**1.00**	✓

The dataset also showed perfect scores in fluency and safety, suggesting that the generated dialogues were natural and non-harmful.

4.3 Ablation Study

To test how important each element was, the researchers removed one component at a time—linguistic style, therapy technique, or client personality—and asked evaluators to judge which version felt more realistic. The full version of PsyDT won in over 60% of comparisons, proving that all three components were essential for creating authentic counseling conversations.

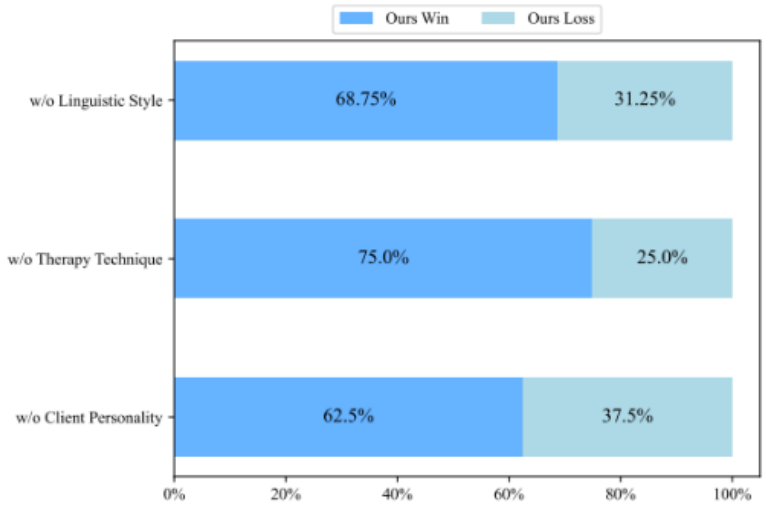


Figure 6: Results of ablation study on synthetic dialogues and other ablated dialogues.

4.4 Model Evaluation

The final PsyDTLLM model was compared with well-known open-source and closed-source models, including ChatGPT, GPT-4, LLaMA-3, Baichuan-2, and many other LLMs. Using automatic metrics like ROUGE, BLEU, and BERTScore, PsyDTLLM achieved the best

results overall. Professional evaluations also showed that PsyDTLLM performed best in emotional empathy, cognitive empathy, and counseling strategy, matching or exceeding ChatGPT and GPT-4 in certain aspects. This suggests that PsyDTLLM is not only technically strong but also capable of providing emotionally supportive responses closer to human counselors.

Table 2: Model evaluation results.

Type	Models	Automatic.					Professional.				
		R-1	R-2	R-L	B-4	F_{BERT}	EmoE.	CogE.	Con.	Sta.	Saf.
Closed	ChatGPT	31.72	7.77	24.52	7.24	96.69	1.70	1.74	1.88	1.99	1.00
	GPT-4	26.51	6.79	18.23	5.31	96.59	1.80	1.99	2.06	1.89	1.00
Open	Baichuan2-7B-Chat	15.40	3.69	11.84	3.46	94.14	1.35	1.34	1.44	1.49	1.00
	GLM4-9B-Chat	23.38	5.45	14.35	3.84	96.58	1.68	1.88	1.94	1.74	1.00
	InternLM2-Chat-7B	27.15	5.87	20.38	5.49	96.62	1.87	1.92	2.04	2.05	1.00
	Llama3-8B-Instruct	26.31	5.25	19.64	5.11	95.16	1.58	1.72	1.77	1.81	1.00
	Llama3.1-8B-Instruct	30.20	5.84	22.88	5.96	96.54	1.61	1.70	1.81	1.90	1.00
	Qwen2-7B-Instruct	23.42	5.28	15.42	4.05	96.64	1.81	2.09	2.18	2.12	1.00
	Yi-1.5-9B-Chat	29.32	6.89	21.85	7.50	96.66	1.75	1.79	2.11	1.93	1.00
Domain	MeChat	30.71	7.05	24.43	6.73	96.55	1.54	1.58	1.66	1.96	1.00
	PsyChat	27.96	5.21	21.44	4.83	96.19	1.36	1.40	1.34	1.79	1.00
	SoulChat	28.93	5.93	23.26	5.49	96.42	1.29	1.36	1.42	1.76	1.00
	MindChat	22.55	3.44	17.75	3.48	93.89	1.13	1.25	1.13	1.54	1.00
	EmoLLM	23.26	4.01	18.50	3.74	91.74	1.06	1.18	1.21	1.36	1.00
	CPsyCounX	23.71	4.32	17.59	3.59	95.46	1.28	1.42	1.54	1.60	1.00
Our	PsyDTLLM	**36.03**	**10.08**	**28.86**	**9.99**	**96.79**	**1.90**	**2.13**	**2.19**	**2.26**	1.00

5. Conclusion, Personal Reflection

5.1 Conclusion

The PsyDT framework introduces a creative way to replicate a real counselor's personality and counseling method using LLMs. Instead of relying on large, private datasets, PsyDT builds a personalized and ethical digital twin through simulated dialogues. Experiments show that the synthetic dataset (PsyDTCorpus) and the resulting model (PsyDTLLM) perform better than other existing mental-health LLMs in terms of empathy, style, and safety.

Although promising, the authors note some limitations. PsyDT currently builds a twin for only one counselor, which might not generalize to all types of clients or therapy approaches. In the future, combining multiple counselors could make the model more flexible and inclusive. They

also emphasize that AI counselors should never replace real human therapists, but rather serve as supportive tools.

5.2 Personal Reflection

In future work, the ablation experiment could be improved by expanding the sample size. The current setup only includes 16 randomly selected samples for manual voting, which may be too small and possibly biased toward certain topics. A more reliable approach would be to draw k samples from each topic to ensure broader coverage and greater statistical validity. Moreover, the LLM-as-Judge evaluation method may introduce bias, as the scores can easily be influenced by prompt wording and model preferences. To make the results more stable, multiple models could be included in the evaluation process, and their scores averaged or combined through consensus. Finally, both LLM-based and expert-based evaluations remain indirect assessments. Incorporating real user feedback—by inviting actual users to interact with PsyDTLLM and provide their opinions—would offer a more authentic reflection of the model's real-world performance and help address the limitations of current evaluation methods.