KokoroChat: A Japanese Psychological Counseling Dialogue Dataset Collected via Role-Playing by Trained Counselors

Zhiyang Qi[1], Takumasa Kaneko[1], Keiko Takamizo[2,3,4], Mariko Ukiyo[2,3,4], Michimasa Inaba[1,2]

[1]The University of Electro-Communications,
[2]Rapport Technologies, Inc.
[3]iDEAR Human Support Service,
[4]Japanese Organization of Mental Health and Educational Agencies
`{qizhiyang,m-inaba}@uec.ac.jp`

Abstract

Generating psychological counseling responses with language models relies heavily on high-quality datasets. Crowdsourced data collection methods require strict worker training, and data from real-world counseling environments may raise privacy and ethical concerns. While recent studies have explored using large language models (LLMs) to augment psychological counseling dialogue datasets, the resulting data often suffers from limited diversity and authenticity. To address these limitations, this study adopts a role-playing approach where trained counselors simulate counselor-client interactions, ensuring high-quality dialogues while mitigating privacy risks. Using this method, we construct **KokoroChat**, a Japanese psychological counseling dialogue dataset comprising 6,589 long-form dialogues, each accompanied by comprehensive client feedback. Experimental results demonstrate that fine-tuning open-source LLMs with KokoroChat improves both the quality of generated counseling responses and the automatic evaluation of counseling dialogues. The KokoroChat dataset is available at https://github.com/UEC-InabaLab/KokoroChat.

Figure 1: Each **KokoroChat** sample includes a counseling dialogue and client feedback, with both roles played by trained counselors.

1 Introduction

Psychological issues have long posed a significant global challenge, with many individuals suffering from mental health disorders (WHO, 2022). However, limited medical resources restrict access to professional psychological counseling for most people (SAMHSA, 2015). To address this gap, researchers have explored language models for generating empathetic responses to provide emotional support. Advancing this research depends on constructing high-quality datasets. For instance, Liu et al. (2021) developed the ESConv dataset by training crowdworkers in emotional support skills, while Li et al. (2023) created the Client-Reactions dataset by establishing an online mental health support platform to collect dialogues between real clients and professional counselors. Despite these efforts, several challenges remain. Psychological counseling is a highly specialized form of communication (Althoff et al., 2016), making it costly and time-consuming to train crowdworkers without professional backgrounds. Meanwhile, dialogue participants may struggle to fully grasp the experiences of individuals with mental disorders, making it difficult to simulate authentic interactions. Additionally, real counseling data may involve privacy and ethical concerns. As a result, the manual collection of such data often faces practical limitations.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12424–12443

Dataset	Human-made	Score	Score items	Language	# Dialogues	Avg. utterances
HealMe (Xiao et al., 2024)	✗	✗	-	English	1,300	6.0
ESD-CoT (Zhang et al., 2024b)	✗	✗	-	English	1,708	23.4
CACTUS (Lee et al., 2024)	✗	✗	-	English	31,577	31.5
SMILECHAT (Qiu et al., 2024)	✗	✗	-	Chinese	55,165	33.2
AUGESC (Zheng et al., 2023)	✗	✗	-	English	65,000	26.7
Anno-MI (Wu et al., 2022)	✓	✗	-	English	133	72.9
ESConv (Liu et al., 2021)	✓	✓	2	English	1,300	29.5
Client-Reactions (Li et al., 2023)	✓	✓	4	Chinese	2,382	78.5
KokoroChat	✓	✓	**20**	**Japanese**	**6,589**	**91.2**

Table 1: Comparison of psychological counseling datasets: LLM-augmented (top), human-collected (bottom).

Recently, LLMs have made significant strides in natural language generation and have shown considerable potential in generating psychological counseling responses (Inaba et al., 2024). Consequently, many studies have leveraged LLMs for self-chat, rewriting, or mimicking existing datasets to construct or expand psychological counseling dialogue datasets (Zheng et al., 2023; Xiao et al., 2024; Zhang et al., 2024b; Lee et al., 2024; Qiu et al., 2024). However, despite LLMs' strong ability to generate psychological counseling responses, these augmented datasets often exhibit redundancy and homogeneity, leading to a lack of dialogue diversity (Zheng et al., 2024). Moreover, as shown in Table 1, augmented datasets contain significantly fewer utterances than human-collected psychological counseling dialogues (e.g., Client-Reactions: 78.5), limiting their effectiveness and applicability.

To address these issues, this study employs a role-playing approach for data collection, in which trained professional and trainee counselors simulate interactions between a counselor and a client. Compared to traditional crowdsourced methods, this approach ensures higher professionalism and dialogue quality, as both participants are trained counselors. Additionally, unlike direct collection of real counseling dialogues, this method mitigates privacy and ethical risks. Furthermore, compared to LLM-augmented datasets, the dialogues collected through this approach offer stronger assurances of professionalism and authenticity.

Through a role-playing approach, we develop **KokoroChat**, a high-quality dialogue dataset for psychological counseling. As shown in Figure 1, trained professional and trainee counselors play both counselor and client roles, engaging in approximately one-hour online text-based counseling sessions. Considering that an objective and detailed scoring mechanism can quantify counseling quality and support counselor skill development, we collect client feedback after each session. The client-role player evaluates the counselor-role player based on two key dimensions: overall impressions and professional skills. The evaluation consists of 20 assessment items, each rated on a scale of 0 to 5, with a maximum total score of 100. This study involves 480 participants, all of whom have completed 10 hours of training in online text-based psychological counseling. Over one-third are professional counselors, while the remaining participants are trainees who have studied relevant topics for six months to one year and aspire to become certified counselors.

As shown in Table 1, we collect 6,589 high-quality psychological counseling dialogues, averaging 91.2 utterances per dialogue, each accompanied by detailed client feedback. To our knowledge, KokoroChat is the largest human-collected psychological counseling dialogue dataset to date, with session durations aligning with real-world counseling practices (approximately one hour per session). Notably, KokoroChat is a Japanese-language dataset, offering linguistic resources for psychological counseling dialogue research from a diverse cultural perspective. Given the global demand for psychological counseling, developing datasets that encompass multiple cultures and languages is essential for enhancing models' adaptability to users from different backgrounds. By filling the gap in Japanese psychological counseling dialogue data, KokoroChat provides a foundation for cross-cultural research in psychological counseling.

To summarize: (1) we develop KokoroChat, the largest manually collected psychological counseling dialogue dataset to date, using a role-playing approach, with detailed client feedback; (2) we fine-tune an open-source LLM to demonstrate that KokoroChat enhances LLM performance in generating psychological counseling responses; (3) we train a dialogue evaluation model using client feedback from KokoroChat, and our experimental results show that this model provides more robust and accurate evaluation outcomes.

2 Related Work

2.1 Psychological Counseling in NLP

In recent years, psychological counseling has attracted significant attention in the field of natural language processing (NLP). Some studies have focused on generating empathetic responses, where systems provide appropriate feedback by understanding users' emotions (Rashkin et al., 2019; Sharma et al., 2020; Zheng et al., 2021). However, empathy alone is insufficient to address the complex demands of psychological counseling. To bridge this gap, Liu et al. (2021) proposed the Emotional Support Conversation (ESC) task, which requires systems not only to exhibit empathy but also to deeply explore users' concerns and offer effective guidance to help them navigate challenges.

As LLMs' generation capabilities advance, their potential applications in psychological counseling have gained further attention. For example, Inaba et al. (2024) demonstrated that GPT-4's (OpenAI, 2023) responses in psychological counseling scenarios are comparable to those of professional counselors. Additionally, several LLM-based psychological counseling chatbots, such as ChatCounselor (Liu et al., 2023), MeChat (Qiu et al., 2024), and SoulChat (Chen et al., 2023), have emerged. These systems are typically fine-tuned on manually curated or LLM-augmented psychological counseling data to adapt to specific scenarios. However, the field still faces challenges, particularly the lack of high-quality, diverse professional datasets. To address this, our study develop KokoroChat, a high-quality psychological counseling dialogue dataset, through role-playing by professional counselors. This dataset aims to facilitate the development of psychological counseling dialogue systems.

2.2 Counseling Dialogue Datasets

The key to equipping language models with psychological counseling capabilities lies in the availability of high-quality datasets. Currently, psychological counseling datasets fall into two main categories: manually constructed datasets and those generated by LLMs. Manually constructed datasets typically consist of human dialogues. For example, Anno-MI (Wu et al., 2022) was developed by extracting full motivational interview dialogues from online videos. ESConv (Liu et al., 2021) collected emotional support dialogues from crowdworkers trained in specialized skills, while Client-Reactions (Li et al., 2023) was derived from interaction records between counselors and clients on real online counseling platforms.

On the other hand, LLM-augmented datasets were created by having LLMs simulate both the counselor and client roles (Chen et al., 2023; Ye et al., 2025). For instance, HealMe (Xiao et al., 2024) and CACTUS (Lee et al., 2024) generated dialogues based on cognitive behavioral therapy (CBT) using carefully designed prompts; ESD-CoT (Zhang et al., 2024b) extracted situations from existing datasets to generate complete dialogues. SMILECHAT (Qiu et al., 2024) expanded single-turn Q&A into multi-turn dialogues, while AUGESC (Zheng et al., 2023) modeled data augmentation as a dialogue completion task to extend conversations. As shown in Table 1, although these datasets are often large in size, they tend to contain fewer utterances and continue to face challenges in content diversity (Zheng et al., 2024). In contrast, manually constructed datasets are smaller while offering higher authenticity and quality. This study proposes the largest known human-collected psychological counseling dialogue dataset, with dialogue lengths resembling real counseling sessions.

2.3 Evaluation of Counseling Dialogue

In recent years, studies have increasingly explored the use of LLMs for dialogue evaluation in specific scenarios. For example, Liu et al. (2023) employed GPT-4 to compare psychological counseling responses generated by different models across multiple dimensions, such as information quality and user self-disclosure. Lee et al. (2024) and Zhao et al. (2024) simulated client interactions with counselor models, conducting full-length dialogues and evaluating them based on overall performance.

Additionally, some datasets have introduced rating mechanisms. For instance, the Client-Reactions (Li et al., 2023) recorded client ratings across four dimensions alongside dialogue collection. However, among 2,382 dialogue turns, only 479 turns included ratings, limiting its scale. In contrast, our study introduces a dataset with 6,589 counseling dialogues, all accompanied by rating information.

We design 20 evaluation dimensions (as shown in Table 2) to assess both the overall impression of the counseling session and the counselor's professional skills, providing a more comprehensive standard for evaluating dialogue quality.

3 Data Collection

To construct a high-quality psychological counseling dialogue dataset, we employed a role-playing approach in which trained professional and trainee counselors simulated counselor-client interactions, ensuring authenticity and professional relevance. Additionally, we collected detailed client feedback, providing a valuable resource for evaluating psychological counseling dialogues.

3.1 Data Source

We developed an online platform to facilitate participant matching, role-playing dialogues, and client feedback collection. On this platform, participants can choose their preferred roles and schedule dialogues based on their availability and role preferences. Once matched, role-playing dialogues take place at the designated time, with the counselor-role player communicating via a computer keyboard and the client-role player using LINE, a mobile messaging app. This setup reflects real-world online text-based psychological counseling practices in Japan. Additionally, to ensure a complete record of the dialogue process, the platform stores message timestamps for further analysis. The detailed interface is shown in Appendix A.1.

Each dialogue session typically lasts one hour, though the duration may be adjusted as needed. As part of the role-playing process, the counselor-role player can specify discussion topics and the client's psychological state (e.g., whether suicidal tendencies are present). If no specific conditions are set, the client-role player decides freely. To protect participant privacy, all client-role participants are explicitly instructed not to discuss their real-life concerns during the dialogue and are strictly prohibited from sharing personal identifying information. After the session, the client-role participant evaluates the counselor-role player's performance, with specific evaluation criteria detailed in Section 3.3.

3.2 Participants

The dataset comprises 480 participants, including 117 males, 360 females, and 3 individuals who did not disclose their gender. Participants' ages range from 21 to 78, with approximately 80% between 30 and 59 years old. Detailed distribution information is provided in Appendix A.2. All participants are native Japanese speakers, with over 80% having played both the client and counselor roles. In total, 424 participants took on the counselor role, and 463 participants played the client role.

Professionalism Participants have expertise in online psychological counseling. More than one-third hold professional qualifications and have practical counseling experience, while the rest, though not yet certified, have undergone six months to one year of systematic study with the goal of obtaining certification. Additionally, all participants completed a 10-hour structured training program covering the characteristics, advantages, and limitations of online text-based psychological counseling, the role and ethical guidelines of counselors, as well as professional counseling skills and procedures.

3.3 Client Feedback

After each role-playing, the client-role player evaluates the counselor-role player's performance. The results are immediately shared with the counselor-role player and are monitored by the platform administrator to ensure fairness and reliability.

The client feedback items are designed under the supervision of an expert holding the nationally recognized Certified Public Psychologist qualification and a Ph.D. degree. As shown in Table 2, the feedback covers two main aspects:

- **Overall impression of the conversation** (e.g., understanding and respect, sense of hope, engagement, fluency, satisfaction)
- **Evaluation of counseling skills** (e.g., empathy, affirmation, effective questioning, goal setting, problem clarification, and conveying hope)

The evaluation employs a six-point Likert scale (0–5 points) across 20 items, with a maximum total score of 100 points. Additionally, three check items assess serious issues, such as inappropriate remarks or ethical violations. If any of these are selected, the overall score is halved or reset to zero. Further details are provided in Appendix A.3.

4 Data Characteristics

This study collected dialogue data from March 7, 2020, to September 8, 2024, filtering out conversations with fewer than 30 utterances, durations

Category	Aspect	Feedback Item
Overall Impression of the Conversation	**Sense of Validation**	1. Felt heard and understood. 2. Felt respected.
	Awareness and Hope	3. Gained new insights. 4. Felt hopeful or expectant.
	Engagement	5. Concerns were addressed. 6. Thought through concerns together.
	Flow and Comfort	7. The conversation had a good rhythm. 8. The conversation felt comfortable.
	Overall Evaluation	9. Felt appropriate and satisfying. 10. The conversation was valuable.
Evaluation of Counseling Skills	**Flow of Conversation**	11. The conversation started smoothly. 12. The conversation ended well.
	Counseling Skills	13. Showed acceptance and empathy. 14. Provided acknowledgment and affirmation. 15. Asked effective questions to foster dialogue. 16. Summarized key points effectively. 17. Clarified issues clearly. 18. Helped identify goals for the conversation. 19. Offered actionable suggestions. 20. Encouraged and instilled hope.

Table 2: The 20 client feedback items (each rated on a 0–5 point scale).

under 30 minutes, or cases where all 20 evaluation items were rated as 3 (as such scores may be unreliable). The final dataset consists of 6,589 dialogues, with statistical details provided in Table 3.

Each dialogue contains an average of 91.20 utterances, surpassing other manually collected datasets such as ESConv (Liu et al., 2021) (29.5 utterances), Anno-MI (Wu et al., 2022) (72.9 utterances), and Client-Reactions (Li et al., 2023) (78.5 utterances). This suggests that our dataset provides greater depth and interactivity. Additionally, the average utterance length of counselors (35.84 characters) is significantly higher than that of clients (20.63 characters), reflecting the counselor's guiding role in conversations, where they typically use more detailed language to provide support.

Additionally, the dataset includes 480 unique speakers, with 424 counselors and 463 clients, resulting in 4,900 distinct counselor–client pairings, which contribute to conversational diversity to some extent.

4.1 Dialogue Topics

The counseling topics were determined by the participants. To gain a comprehensive understanding of the issues present in KokoroChat, we utilized GPT-4o-mini (OpenAI, 2024) to analyze dialogue topics. Specifically, we input the dialogue content to predict problem types and generate more detailed descriptions. The prompt used for this analysis is provided in Appendix B.

Category	Total	Counselor	Client
# Dialogues	6,589	-	-
# Speakers	480	424	463
# Utterances	600,939	306,495	294,444
Avg. utterances per dialogue	91.20	46.52	44.69
Avg. length per utterance	28.39	35.84	20.63

Table 3: Statistics of the overall conversations. An "utterance" denotes one discrete message sent by a client or counselor in the chat system upon clicking the send button.

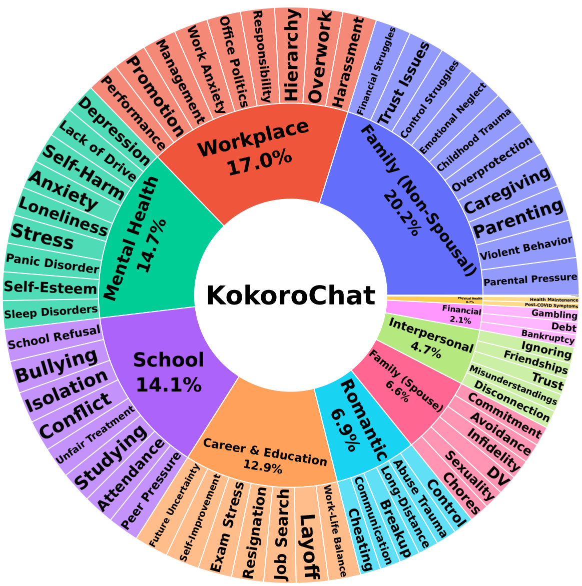


Figure 2: The distribution of issues in KokoroChat.

The distribution of the 11 predefined issue types is shown in Figure 2. Family issues (non-spousal) account for the largest proportion (20.2%), followed by workplace issues (17.0%) and mental health issues (14.7%), reflecting that family, work, and mental health are primary concerns for clients. School issues (14.1%) and career and education issues (12.9%) also constitute significant portions, highlighting the importance of education and career development in counseling conversations.

Based on the generated detailed descriptions, we further summarized the characteristics of each issue category, as presented in Figure 2. Overall, the dataset encompasses a broad range of real-world problems, demonstrating high diversity and providing a solid foundation for research on psychological counseling dialogues.

4.2 Analysis on Client Feedback

Score Distribution Figure 3 presents the score distribution of dialogues in KokoroChat. The histogram shows a well-balanced distribution, with a unimodal shape centered around the mean (63.58) and median (64.00). Additionally, the distribution exhibits a slight right skew, indicating that most dialogues received moderate to high client feedback.

Correlation Between Dialogue Features and Scores We conducted a Spearman correlation analysis to examine the relationship between various dialogue features and feedback scores, as shown in Figure 4. The results indicate that the total word count of the client has the highest positive correlation with the score ($\rho = 0.42$), suggesting that the extent of client expression may influence their evaluation. When clients use more words to express themselves, they may feel better heard and understood, leading to a more positive assessment of their counseling experience. In contrast, the total word count of the counselor shows a lower correlation with the score ($\rho = 0.28$), implying that while greater counselor speech may contribute to higher ratings, its impact is relatively limited. Additionally, the correlation between utterance count and scores is weaker, with clients ($\rho = 0.21$) and counselors ($\rho = 0.17$) both showing a positive correlation, though to a lesser extent than word count. This result suggests that the richness of conveyed information may be more influential than the number of utterances. Furthermore, the counselor's average response time exhibits a negative correlation with the scores ($\rho = -0.21$), indicating

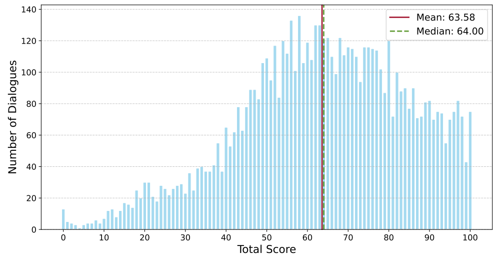


Figure 3: Score distribution of dialogues.

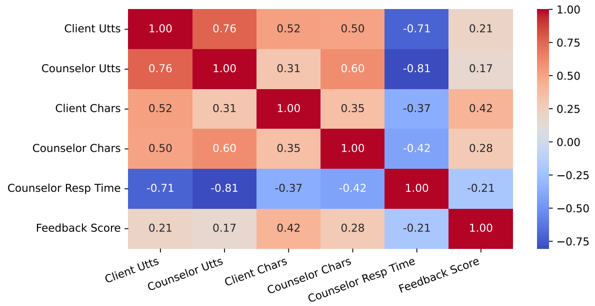


Figure 4: Spearman correlation between dialogue features and feedback scores.

that longer response times may negatively impact user experience, whereas quicker responses could potentially contribute to higher ratings.

Correlation Among Evaluation Dimensions Additionally, we conducted a Spearman correlation analysis to examine the relationships between different evaluation dimensions. The results indicate that most rating dimensions exhibit a strong positive correlation ($\rho > 0.6$), suggesting that the counseling experience is influenced by multiple interrelated factors rather than a single determinant.

Notably, D1 (*felt heard and understood*) shows a high correlation with D2 (*felt respected*), D9 (*felt appropriate and satisfying*), and D10 (*the conversation was valuable*). This suggests that when clients feel heard and understood, they are more likely to experience a sense of respect, perceive the counseling process as meaningful, and ultimately report higher overall satisfaction. The complete results are presented in Figure 11 in Appendix C.

5 Experiments

To evaluate KokoroChat's potential in psychological counseling response generation and dialogue assessment, we conducted experiments on two tasks: response generation and score prediction.

5.1 Response Generation

Due to the lack of a Japanese psychological counseling dataset and a Japanese LLM specifically designed for counseling, direct comparison with other models is not feasible[1]. Therefore, this study focuses on verifying whether fine-tuning on KokoroChat can enhance the performance of open-source LLMs in psychological counseling tasks.

For dialogue data preprocessing, we applied an utterance merging strategy, combining consecutive utterances from the same speaker into a single utterance. The model takes the complete dialogue history $D_t = \{u_1^C, u_2^S, u_3^C, \ldots, u_t^C\}$ as input, where u_i^C and u_j^S represent utterances from the client (C) and counselor (S), respectively. The model then generates the next counselor response u_{t+1}^S.

5.1.1 Models

In this experiment, we used Llama 3.1 Swallow [2] (Fujii et al., 2024; Okazaki et al., 2024) as the base model, fine-tuning and evaluating it using the KokoroChat dataset. To ensure high-quality test data, we selected 118 dialogues with client feedback scores of 99 and 100 as the test set, while the remaining data was used for fine-tuning. The model-generated responses were then compared with the corresponding counselor replies in the test set to evaluate the quality of the generated outputs.

To explore the impact of client feedback on model ability to generate psychological counseling responses, we constructed the following variants:

- **Kokoro-Low:** Fine-tuned on 334,022 utterances from 3,870 dialogues with a client feedback score of < 70.[3]
- **Kokoro-High:** Fine-tuned on 254,515 utterances from 2,601 dialogues with a client feedback score in the range of $70 \leq score \leq 98$.
- **Kokoro-Full:** Fine-tuned on 6,471 dialogues with a client feedback score of ≤ 98.

Finally, we evaluated these variants on the test set to examine the impact of data partitions on model improvement. Additionally, we compared our models with GPT-4o (OpenAI, 2024), one of the most advanced models, to further understand their relative performance. Appendix D provides details on the training process.

5.1.2 Automatic Evaluation

We used three commonly adopted automatic evaluation metrics: BLEU-n (Papineni et al., 2002), ROUGE-L (Lin, 2004), and Distinct-n (Li et al., 2016). The results are presented in Table 4. Experimental findings indicate that Kokoro-High performs best on most BLEU and ROUGE metrics, likely due to the higher quality of its training data and its closer alignment with the test set. In contrast, Kokoro-Full, which includes a larger dataset, achieves slightly better performance on the diversity metric Dist-n. For non-fine-tuned models, GPT-4o outperformed Llama-3.1 across all automatic evaluation metrics.

5.1.3 Human Evaluation

We also conducted a human evaluation of 100 responses generated by each model, independently assessed by five professional counselors. Specifically, we randomly selected 10 dialogues from the test set and, for each dialogue, randomly sampled 10 sets of dialogue histories of varying lengths to generate model responses. The evaluation used pairwise comparison, where counselors judged which response was more suitable (Win, Lose, Tie). The final result followed majority voting—if over half agreed, it was adopted; otherwise, or if Tie votes exceeded half, the result was Tie.

Figure 5 presents the evaluation results. The comparison between Kokoro-Low and Llama-3.1 indicates that even when using only the lower-scoring portions of KokoroChat, it still enhances open-source LLMs in generating psychological counseling responses. Notably, despite using less data, Kokoro-High outperforms both Kokoro-Low and Kokoro-Full, similar to the results of automatic evaluation, highlighting the importance of high-quality training data in improving model performance. However, due to the difference in model size (Llama-3.1 = 8B, GPT-4o $\approx$ 200B[4]), the fine-tuned model still lags behind GPT-4o. Similarly, GPT-4o's responses also exhibit a noticeable gap compared to those of highly rated human counselors, further emphasizing the high quality of

[1] To provide a rough performance reference, we conducted a simplified experiment comparing our model with models trained on non-Japanese counseling data in a Japanese setting. See Appendix E for details.

[2] Llama-3.1-Swallow-8B-Instruct-v0.3 was used, which is a continuously pre-trained variant of Meta Llama 3.1 (Dubey et al., 2024) with enhanced Japanese proficiency and optimized dialogue generation.

[3] Considering that low-scoring dialogues tend to have a relatively lower number of utterances, we set the threshold at 70 to ensure balanced data segmentation.

[4] This is merely an estimate by Abacha et al. (2024).

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	Dist-1	Dist-2
Llama-3.1	17.32	9.13	4.77	2.25	23.81	7.37	16.96	1.04	6.86
GPT-4o	21.77	11.72	6.32	3.17	28.67	9.19	19.82	1.19	6.90
Kokoro-Low	25.39	15.30	8.69	5.39	33.38	14.05	27.28	2.42	12.98
Kokoro-High	**27.03**	**16.45**	**9.57**	**6.00**	**34.64**	**14.72**	28.00	2.33	13.08
Kokoro-Full	25.69	15.65	9.23	5.83	34.02	14.60	**28.10**	**2.48**	**13.24**

Table 4: Performance comparison of models. Best values are in **bold**, second-best are underlined.

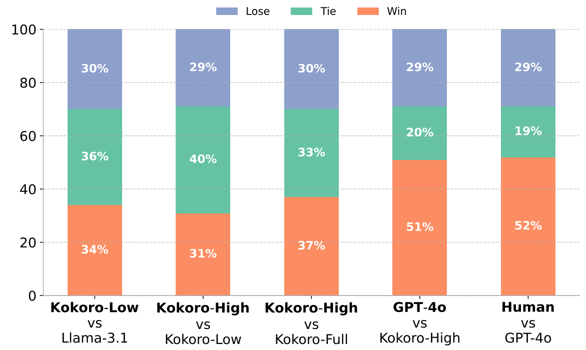


Figure 5: Human evaluation result. Orange denotes upper model wins; the winning model is in **bold**. *Human* refers to responses from human counselors in the KokoroChat test dataset.

Model	ACC (↑)	ACC_{soft} (↑)	MAE (↓)
Llama-3.1	28.70 $\pm$ 7.39	72.53 $\pm$ 12.40	1.0540 $\pm$ 0.2731
GPT-4o	30.92 $\pm$ 6.84	75.27 $\pm$ 11.04	1.0151 $\pm$ 0.2685
Ours	**35.35** $\pm$ **1.75**	**83.64** $\pm$ **2.15**	**0.8283** $\pm$ **0.0349**

Table 5: Performance comparison of different models in terms of accuracy (ACC), soft accuracy (ACC_{soft}), and mean absolute error (MAE) (the results of model **Ours** are averaged over five different random seeds; detailed results can be found in Table 6 of Appendix D).

KokoroChat. The generated response examples are shown in Figure 14 of Appendix F.

5.2 Score Prediction

Automatic evaluation of psychological counseling dialogues not only reduces the cost of human evaluation but also provides counselors with feedback to enhance their professional skills. To verify the effectiveness of LLMs fine-tuned on KokoroChat for dialogue evaluation, we conducted a score prediction experiment.

Specifically, given a complete dialogue D as input, the model predicts a set of item-score pairs $S = \{(d_1, s_1), (d_2, s_2), \ldots, (d_{20}, s_{20})\}$, where each d_i $(i = 1, 2, \ldots, 20)$ represents an evaluation dimension (e.g., *felt heard and understood*, *felt respected*), and each s_i corresponds to a score in the range $[0, 5]$.

5.2.1 Models

We similarly employed Llama-3.1-Swallow as the base model and fine-tuned it by splitting the dataset into training, validation, and test sets (8:1:1). To ensure robust results, we repeated fine-tuning five times with different seeds. For comparison, we also performed zero-shot score prediction using the original Llama-3.1-Swallow and GPT-4o. Detailed training procedures are provided in Appendix D.

5.2.2 Results

Table 5 presents the average performance of different models in predicting scores across 20 evaluation dimensions for psychological counseling dialogues. Our model outperforms Llama-3.1 and GPT-4o in accuracy, demonstrating superior score prediction capabilities. Given the inherent subjectivity and ambiguity of human ratings, we also evaluated performance using ACC_{soft}, which allows a ± 1 error margin between predicted and human-assigned scores. Our model again surpasses the baselines under this metric, highlighting its ability to capture scoring trends in psychological counseling dialogues while maintaining robust performance with a more flexible scoring standard. Detailed results across 20 dimensions are in Table 8 of Appendix D.

Our model also achieves the lowest Mean Absolute Error (MAE), indicating smaller prediction errors and closer alignment with human ratings. Additionally, its lower standard deviation across all metrics suggests more stable scoring. These results confirm the effectiveness of the KokoroChat dataset and demonstrate that our fine-tuned model delivers both stable and high-performance predictions across 20 evaluation dimensions.

6 Conclusion

This study introduces **KokoroChat**, the largest manually collected psychological counseling dialogue dataset to date, developed using a role-playing approach. The dataset includes detailed client feedback, enabling automatic evaluation of psychological counseling dialogues. Experimental results demonstrate that KokoroChat enhances LLM performance in generating psychological counseling responses. Additionally, by leveraging its extensive client feedback data, we train a dialogue evaluation model capable of producing more robust and accurate assessment results.

Limitations

Due to the lack of publicly available Japanese psychological counseling datasets or Japanese LLMs designed for counseling, this study could not be directly compared with existing research. As part of future work, we plan to translate KokoroChat into multiple languages, such as Chinese and English, enabling comparisons with other psychological counseling dialogue datasets and models for a more comprehensive evaluation of our approach.

Additionally, we plan to annotate dialogue acts within the dataset to analyze the evolving strategies used in counseling and their impact on outcomes. This will provide deeper insights into how counselors adjust their communication styles based on client responses and help optimize the model's adaptability across different counseling scenarios.

Furthermore, potential gender and age biases among participants during data collection may affect the model's generalization ability.

Ethical Considerations

The dialogue data collected in this study originates from an internal training platform used by a psychological counselor association. This platform is designed to support professional counselors and aspiring trainees in developing their psychological counseling skills. Psychological counseling relies not only on a strong theoretical foundation but also on extensive practical experience. Even experienced counselors must engage in continuous practice to refine their skills. However, gaining experience in real counseling settings can pose ethical and safety risks, particularly when clients are experiencing emotional distress or psychological crises. Counselors cannot rely solely on real-world counseling experiences for training. To address this, the platform provides a low-risk training environment, enabling counselors to practice and refine their skills in simulated scenarios.

The dialogues on this platform are not real psychological counseling sessions; rather, they are role-play-based simulated counseling exercises. All participants are fully aware of the simulated nature of the dialogues and voluntarily engage in the training process without monetary compensation. As a benefit, participants gain valuable hands-on experience and receive feedback from role-players with relevant professional backgrounds, helping them further develop their counseling skills.

Due to the nature of psychological counseling, even in role-play scenarios, dialogues may include expressions of severe emotional distress, such as suicidal ideation or other extreme emotions. These simulated cases are designed to help counselors develop crisis intervention skills in a controlled setting. While these dialogues do not represent real client experiences, they reflect situations that counselors may encounter in real practice, contributing to their preparedness for handling complex emotional states.

Additionally, to protect participant privacy, all users are explicitly informed that they must not discuss real-life personal issues during the dialogues and must not disclose real names or other identifiable information. Furthermore, they are fully aware that their dialogue data will be stored on the training platform and may be used for service optimization, scientific research, or third-party academic studies in the future.

Acknowledgments

The authors wish to acknowledge Prof. Yasushi Sugihara from Kyoto University for his insightful advice during the design of the client feedback items. This work was supported by JSPS KAKENHI Grant Number 25H01382.

References

Asma Ben Abacha, Wen wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2024. Medec: A benchmark for medical error detection and correction in clinical notes.

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

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183, Singapore. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and Angela Fan et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, University of Pennsylvania, USA.

Michimasa Inaba, Mariko Ukiyo, and Keiko Takamizo. 2024. Can large language models be used to provide psychological counselling? an analysis of gpt-4-generated responses using role-play dialogues. In *The 14th International Workshop on Spoken Dialogue Systems Technology*.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023. Understanding client reactions in online mental health counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10358–10376, Toronto, Canada. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. 2024. Building a large japanese web corpus for large language models. In *Proceedings of the First Conference on Language Modeling*, COLM, University of Pennsylvania, USA.

OpenAI. 2023. Gpt-4 technical report.

OpenAI. 2024. Gpt-4o system card.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636, Miami, Florida, USA. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

SAMHSA. 2015. Behavioral health trends in the united states: Results from the 2014 national survey on drug use and health. Annual report, Substance Abuse and Mental Health Services Administration.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

WHO. 2022. World mental health report: Transforming mental health for all.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. HealMe: Harnessing cognitive reframing in large language models for psychotherapy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1707–1725, Bangkok, Thailand. Association for Computational Linguistics.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025. Sweetiechat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. In *Proceedings of the 31th International Conference on Computational Linguistics*. International Committee on Computational Linguistics.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024a. CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13947–13966, Bangkok, Thailand. Association for Computational Linguistics.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024b. ESCoT: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. 2024. ESC-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810, Miami, Florida, USA. Association for Computational Linguistics.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. CoMAE: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824, Online. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11325–11345, Bangkok, Thailand. Association for Computational Linguistics.

A Data Collection Details

A.1 Online Platform for Data Collection

As described in Section 3.1, our online platform follows the setup of real-world online psychological counseling in Japan. In this setting, the counselor-role player interacts via computer keyboard input (as shown in Figure 6), while the client-role player communicates through LINE (as shown in Figure 7). After the session, the client-role player provides feedback on the counselor's performance across 20 dimensions (as shown in Figure 8). An example, including a collected dialogue and the corresponding client feedback, is shown in Figure 12 (Japanese original version) and Figure 13 (English version, translated by authors).

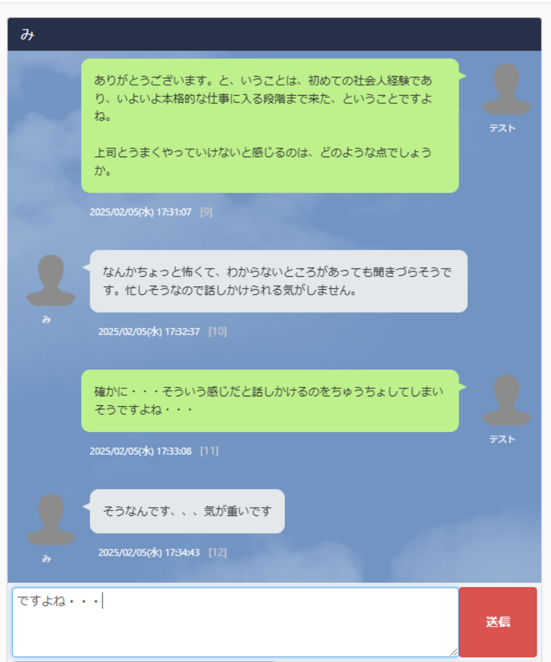


Figure 6: Counselor-role player's dialogue interface using a computer.

A.2 Distribution of Participants

Figure 9 illustrates the age and gender distribution of dialogue collection participants, with age calculated as of February 15, 2025. The figure shows that most participants fall within the 30–59 age range, with a higher proportion of female participants.

A.3 Three screening items in client feedback.

In addition to the 0–5 rating scale, client feedback includes three items designed to screen for serious issues, assessing potential ethical or communication problems:

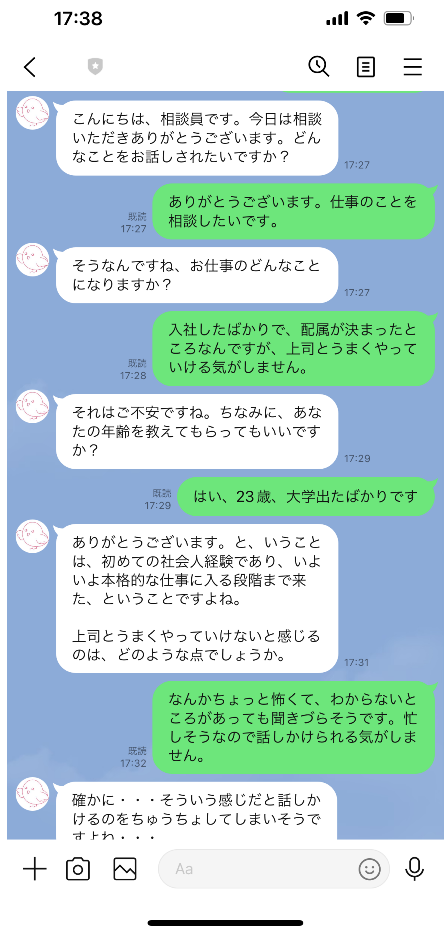


Figure 7: Client-role player's dialogue interface using a mobile phone.

(1) **Whether the counselor made harmful remarks due to a lack of understanding or careless speech.** Examples include telling an LGBTQ client that "homosexuality is abnormal," advising a grieving parent to "just have another child," or suggesting to a bullying victim that "the bullied party is also responsible."

(2) **Whether the dialogue contains other potentially unethical statements**, such as comments that could be misinterpreted as medical diagnoses, inappropriate medication advice, blatantly irrational spiritual (occult) claims, or sexually inappropriate remarks.

(3) **Whether the client is unwilling to continue communicating with the counselor.**

The scoring adjustment rules are as follows: If marked with (1) or (2), the total score is set to zero; if marked with (3), the total score is halved. Based on our analysis, among the 6,589 dialogues in KokoroChat, 8 were marked with (1), 4 with

Figure 8: Client feedback interface (the client-role player is required to rate each dimension on a scale from 0 to 5).

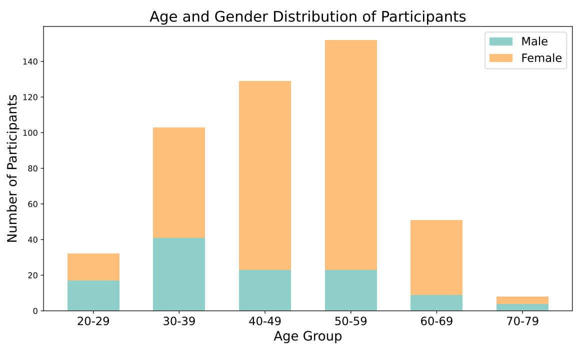


Figure 9: Age and gender distribution of participants.

 # タスク
対話全体を観察し、最も重要で主要な内容に基づき、出力フォーマットに従って、トピック一覧から一個選択し、その次の細かいトピックを生成してください。

 # トピック一覧
人間関係、学校問題、家庭内問題（夫婦）、家庭内問題（夫婦以外）、恋愛問題、職場問題、経済的問題、身体的健康問題、心理的問題、進路・キャリア問題、その他

 # 出力フォーマット（例）
家庭内問題（夫婦） -> 夫とのコミュニケーション問題

 # 対話
[DIALOGUE]

 # 出力

 # Task
Observe the entire dialogue and, based on the most important and central content, select one topic from the topic list. Then, generate a more detailed subtopic under the selected topic.

 # Topic List
Mental Health Issues, Physical Health Issues, Interpersonal Issues, Romantic Issues, Family Issues (Spouse), Family Issues (Non-Spousal), School Issues, Financial Issues, Workplace Issues, Career and Educational Issues, Others.

 # Output Format (Example)
Family Issues (Spouse) -> Communication Problems with Husband

 # Dialogue
[DIALOGUE]

 # Output

Figure 10: Prompt for topic prediction: Japanese (top), English (bottom).

(2), and 209 with (3), resulting in a total of 215 dialogues meeting at least one screening item. It is important to note that while these screening items were included in the data collection process, this study's experiments did not involve predictions related to these flagged issues.

B Prompt for Topic Prediction

We used gpt-4o-mini-2024-07-18 to predict dialogue topics in the dataset using the prompt shown in Figure 10.

C Details of Client Feedback Analysis

Figure 11 presents the Spearman correlation matrix of client feedback across different evaluation dimensions, analyzing their interrelationships. In addition to the strong correlations among D1, D2, D9, and D10 mentioned in Section 4.2, D3, D4, and D10 also exhibit high correlations. This suggests that the perceived value of psychological counseling largely stems from whether clients gain

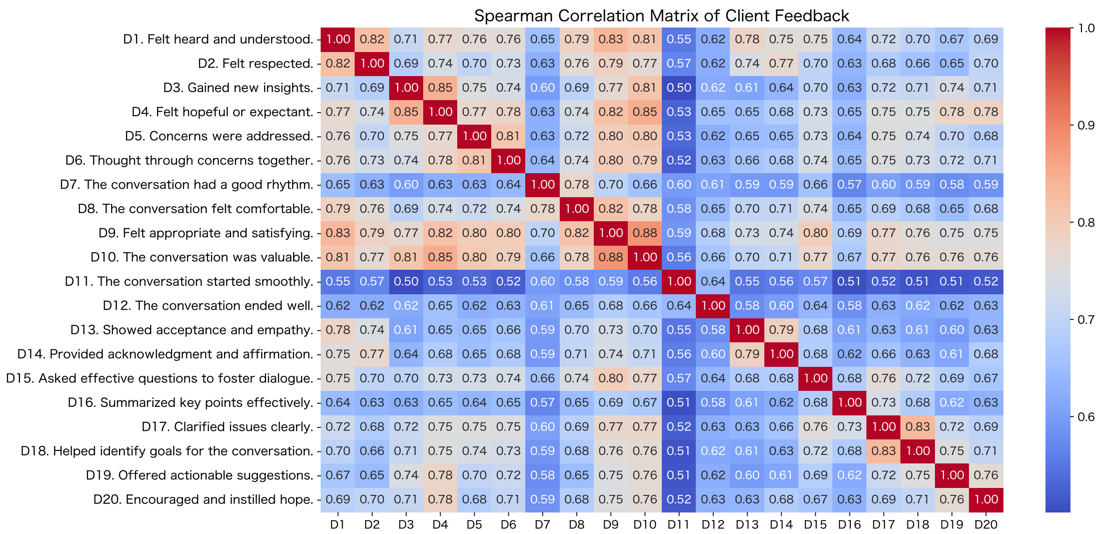


Figure 11: Spearman correlation analysis between evaluation dimensions.

new insights during the process, which in turn fosters hope and expectations for the future. In other words, emotional support alone is insufficient—counselors need to help clients develop new perspectives, strengthening their confidence in the future and ultimately enhancing their evaluation of the dialogue's value.

Moreover, the strong positive correlation between D8 and D9 indicates that the conversational environment is a key factor in client satisfaction with the counseling experience. If a counselor creates a relaxed and open communication setting, clients are more likely to perceive the conversation as appropriate and provide a higher overall evaluation. This finding highlights that, beyond delivering substantive support, counselors should also pay attention to their communication style, tone, and pacing to enhance client comfort.

Overall, the high correlations among multiple dimensions underscore the complexity of the psychological counseling experience.

D Experimental Details

D.1 Response Generation

Fine-tuning Phase This study employed QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2023) as the fine-tuning method to efficiently adapt a large-scale language model. The process began with 4-bit NF4 quantization, utilizing `bfloat16` computation to optimize memory usage and computational efficiency. LoRA adaptation was then applied to key projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`) with parameters set to `r = 8`, `lora_alpha = 16`, and `lora_dropout = 0.05`, ensuring that the model retained its learning capability while undergoing low-rank updates.

For dataset partitioning, 118 dialogues with scores of 99 or 100 were selected as the test set to ensure a high-quality evaluation standard. The remaining data was split into 90% for training and 10% for validation.

For hyperparameter tuning, a grid search determined the optimal configuration. The search covered three key parameters: optimizer, warm-up steps, and learning rate. The optimizer candidates included `adamw_torch_fused`, `adamw_8bit`, and `paged_adamw_8bit`. Warm-up steps were tested at `{100, 300, 500}`, while learning rates were selected from `{1e-3, 5e-4, 2e-4, 1e-4, 5e-5}`. Based on the evaluation results, the final configuration adopted `adamw_8bit` as the optimizer, `100` warm-up steps, and a learning rate of `1e-3`. Training was conducted on four A100 40GB GPUs with a batch size of `8` for `five` epochs. Validation was performed every `400` steps, and the final model was selected based on the lowest validation loss.

Inference Phase During the inference phase, we also employed 4-bit quantization to optimize computational efficiency while maintaining model performance. Additionally, we set `do_sample = False` and `temperature = None` to ensure deterministic outputs, eliminating sampling variability and enhancing response consistency.

D.2 Score Prediction

Fine-tuning Phase This experiment followed the same QLoRA fine-tuning approach as in the response generation experiment. The dataset was randomly split into training, validation, and test sets with an 8:1:1 ratio. To ensure robustness, we conducted five experiments by fixing the test set while varying the training-validation split using different random seeds. The results under five different seeds are shown in Table 6, allowing a direct and fair comparison with the baseline models.

Model	ACC (↑)	ACC_{soft} (↑)	MAE (↓)
Llama-3.1	28.70 ± 7.39	72.53 ± 12.40	1.0540 ± 0.2731
GPT-4o	30.92 ± 6.84	75.27 ± 11.04	1.0151 ± 0.2685
Ours	**35.35** ± **1.75**	**83.64** ± **2.15**	**0.8283** ± **0.0349**
- Seed 1	36.41 ± 1.64	82.58 ± 2.29	0.8413 ± 0.0369
- Seed 2	35.18 ± 1.49	85.69 ± 1.58	0.8106 ± 0.0275
- Seed 3	34.61 ± 2.05	82.09 ± 3.26	0.8397 ± 0.0513
- Seed 4	34.94 ± 1.89	83.64 ± 1.90	0.8292 ± 0.0277
- Seed 5	35.62 ± 1.70	84.19 ± 1.70	0.8205 ± 0.0309

Table 6: Performance comparison of different models on accuracy, soft accuracy, and MAE (including results across different seeds).

For hyperparameter tuning, we explored different configurations for three key parameters: learning rate, warm-up steps, and optimizer. The learning rate candidates were `{5e-4, 2e-4, 1e-4, 5e-5, 2e-5}`, while the warm-up steps were selected from `{50, 100, 150}`. The optimizer candidates included `adamw_torch_fused`, `adamw_8bit`, and `paged_adamw_8bit`. Based on the evaluation results, the final configuration adopted `2e-4` as the learning rate, `100` warm-up steps, and `adamw_torch_fused` as the optimizer.

Training was conducted on two A6000 48GB GPUs with a batch size of 4 for `four` epochs. Validation was performed at the end of each epoch, and the final results were obtained from the epoch with the highest prediction accuracy.

Inference Phase During the inference phase, we applied the same settings as in the response generation experiment, using 4-bit quantization and setting `do_sample = False` and `temperature = None` to ensure deterministic outputs.

E Simplified Model Comparison Experiment

Given that some models trained on non-Japanese psychological counseling datasets exhibit a certain degree of Japanese conversational ability, we conducted a simplified experiment to compare Kokoro-High with the following publicly available counseling dialogue models in a Japanese setting:

- **CPsyCounX**[5] (Zhang et al., 2024a): A Chinese dialogue model based on InternLM2-Chat-7B[6], fine-tuned on CPsyCounD, a dataset of 3,134 multi-turn synthetic counseling dialogues.

  EmoLLM[7] : A counseling-oriented LLM series[8] fine-tuned using synthetic counseling dialogues and derived data from professional literature.

To ensure fairness, GPT-4o was instructed to simulate a client engaging in Japanese conversations with each model using a simple and consistent prompt. Each dialogue consisted of 10 to 20 turns. After 10 turns, if the conversation appeared to reach a natural conclusion and included a farewell, the client concluded the session by appending <end>; otherwise, the session was forcibly terminated at 20 turns. The full prompt used is shown below.

Prompt for GPT-4o Client Simulation (Translated from Japanese)

Task Description
You are now a client who has come to receive psychological counseling. Please follow the instructions below to engage in a conversation with the counselor.

- You are currently experiencing emotional distress or stress, and you have come to counseling because you want someone to listen to you.
- You feel a bit nervous speaking with a counselor for the first time, but deep down, you genuinely want someone to help you.
- You sometimes find it difficult to put your thoughts and feelings into words.
- Depending on the situation and your emotional state, behave as a client with one or more of the following characteristics (you may select them randomly if needed).
- After the conversation exceeds 10 turns (a turn is a full exchange between client and counselor), if the conversation seems to be coming to a natural close and a farewell is expressed, please add <end> at the end of that utterance.

Example Client Characteristics

- Recently having trouble falling asleep and feeling tired every day.
- Struggling with relationships and feeling isolated at work or school.
- Feeling emotionally down due to a breakup or family issues.

[5]https://huggingface.co/CAS-SIAT-XinHai/CPsyCounX
[6]https://huggingface.co/internlm/internlm2-chat-7b
[7]Llama-3-8B-Instruct version was used.
[8]https://github.com/SmartFlowAI/EmoLLM

- Having strong anxiety about the future and lacking self-confidence.
- At a loss for what to talk about during the session, sometimes falling into silence.

For each model, we collected 50 dialogues. Following the dialogue evaluation prompt proposed by Zhang et al. (2024a), GPT-4o was asked to independently score each dialogue across four dimensions: Comprehensiveness, Professionalism, Authenticity, and Safety. Each dimension was rated on a 0–5 scale.

The evaluation results are shown in Table 7. Our model, Kokoro-High, achieved the highest scores across all dimensions. This outcome is not unexpected, as Kokoro-High is the only model fine-tuned on a Japanese psychological counseling dataset. Nonetheless, the results highlight the effectiveness of fine-tuning on KokoroChat in building high-quality counseling dialogue systems tailored for the Japanese language and context.

Model	Comp.	Prof.	Auth.	Safe.
CPsyCounX	2.64	1.78	2.70	3.90
EmoLLM	3.58	3.02	3.92	4.74
Kokoro-High	**3.98**	**3.38**	**4.50**	**4.98**

Table 7: Automatic evaluation scores (0–5 scale) using GPT-4o across four dimensions: Comprehensiveness (**Comp.**), Professionalism (**Prof.**), Authenticity (**Auth.**), and Safety (**Safe.**). Each score represents the average over 50 dialogues. The best results are highlighted in bold.

F Case Study

Figure 14 presents example responses generated by each model. Although our model received a lower score than GPT-4o in human evaluations, its responses still demonstrated appropriate empathy, indicating a certain level of psychological counseling dialogue capability.

Additionally, we observed that models without fine-tuning (Llama-3.1 and GPT-4o) tend to frequently ask questions in their responses. However, in actual psychological counseling, not all situations require additional questioning to gather more information. On the contrary, when a client actively expresses themselves, providing only empathetic responses—without further questioning—can prevent unnecessary interruptions, facilitating smoother emotional expression and better conversation flow.

While this study's human evaluation primarily focused on overall dialogue quality, it did not include a fine-grained analysis of question appropriateness. A well-structured questioning strategy is crucial in psychological counseling, as it can guide clients toward deeper reflection and help establish trust. However, excessive questioning may disrupt the conversation's rhythm and even affect the client's emotional stability. Therefore, in future research, we aim to conduct more fine-grained human evaluations, assessing the appropriateness of questions posed by each model, among other factors.

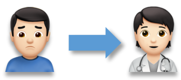
Dim.	Model	ACC (↑)	$\mathrm{ACC_{soft}}$ (↑)	MAE (↓)	Dim.	Model	ACC (↑)	$\mathrm{ACC_{soft}}$ (↑)	MAE (↓)
	Llama-3.1	24.32	70.52	1.0957		Llama-3.1	**39.21**	**87.39**	**0.7477**
D1	GPT-4o	18.39	52.43	1.6535	D2	GPT-4o	32.83	84.19	0.8891
	Ours	**33.37**	**83.83**	**0.8407**		Ours	35.71	84.47	0.8112
	Llama-3.1	18.69	59.57	1.3024		Llama-3.1	25.08	74.16	1.0365
D3	GPT-4o	34.35	80.09	0.8815	D4	GPT-4o	34.80	81.61	0.8587
	Ours	**35.50**	**82.76**	**0.8495**		Ours	**36.72**	**83.40**	**0.8222**
	Llama-3.1	28.27	75.23	0.9954		Llama-3.1	27.36	75.08	1.0061
D5	GPT-4o	**37.54**	81.76	**0.8237**	D6	GPT-4o	35.71	80.85	0.8602
	Ours	35.81	**84.13**	0.8240		Ours	**36.54**	**83.71**	**0.8246**
	Llama-3.1	17.02	55.78	1.4195		Llama-3.1	30.40	75.38	0.9696
D7	GPT-4o	22.19	68.09	1.1702	D8	GPT-4o	**37.54**	83.13	0.8055
	Ours	**34.01**	**83.74**	**0.8410**		Ours	36.90	**84.32**	**0.8006**
	Llama-3.1	34.65	77.96	0.8921		Llama-3.1	34.19	77.36	0.9058
D9	GPT-4o	**37.08**	83.89	0.8009	D10	GPT-4o	**35.41**	80.40	0.8632
	Ours	35.71	**85.62**	**0.7960**		Ours	34.92	**83.83**	**0.8292**
	Llama-3.1	34.50	67.02	0.9878		Llama-3.1	19.60	51.22	1.5380
D11	GPT-4o	**35.87**	75.68	**0.8906**	D12	GPT-4o	19.60	48.33	1.5957
	Ours	34.47	**77.66**	0.8951		Ours	**34.59**	**80.97**	**0.8763**
	Llama-3.1	**37.54**	**85.26**	**0.7827**		Llama-3.1	**36.78**	**89.82**	**0.7447**
D13	GPT-4o	32.22	77.51	0.9726	D14	GPT-4o	30.85	83.74	0.8982
	Ours	33.86	82.71	0.8462		Ours	33.98	85.71	0.8085
	Llama-3.1	**37.39**	86.17	0.7812		Llama-3.1	34.95	**85.56**	0.8176
D15	GPT-4o	28.27	74.16	1.0471	D16	GPT-4o	26.75	76.29	1.0608
	Ours	37.35	**86.35**	**0.7723**		Ours	**37.51**	83.59	**0.8131**
	Llama-3.1	18.24	51.22	1.6565		Llama-3.1	30.09	79.18	0.9377
D17	GPT-4o	19.00	55.47	1.4970	D18	GPT-4o	**38.30**	84.80	0.7812
	Ours	**33.16**	**83.56**	**0.8498**		Ours	36.72	**86.02**	**0.7796**
	Llama-3.1	18.69	53.34	1.4407		Llama-3.1	27.05	73.40	1.0228
D19	GPT-4o	24.47	70.36	1.1277	D20	GPT-4o	**37.23**	82.52	**0.8252**
	Ours	**34.19**	**82.89**	**0.8526**		Ours	36.05	**83.47**	0.8322

Table 8: Performance comparison across 20 evaluation criteria. Each dimension (Dim.) is evaluated using accuracy (ACC↑), soft accuracy ($\mathrm{ACC_{soft}}$↑), and the mean absolute error (MAE↓).

Counseling Dialogue

Counselor: (おはようございます本日は宜しくお願いします)
Client: (おはようございます。よろしくお願いします。)
Counselor: おはようございます　相談員です宜しくお願いいたします
Client: よろしくお願いします。
Counselor: では初めに、年齢と性別を教えて下さい
Client: 14歳　男です
Counselor: 14歳ですね　何年生ですか？
Client: 中学3です
Counselor: ありがとうございます本日はどのようなご相談ですか？
Client: 受験をどうしよう
Counselor: 受験のことでお悩みなのですね
Client: はい。
Counselor: もう少し具体的に教えてもらえますか？
Client: うちはひとり親で、家の手伝いもしなければならなくて、自分の時間がないけど、
Client: そろそろ高校のことも考えないとと思っていますが
Client: 部活もやってないし、委員会みたいなのもしてないし
Client: 内申に関係あるのかなって考えたりします。
Counselor: そうなのですねそれは不安になりますよね
Client: はい、妹の世話もあるから高校に行ってもいいのかなって悩みます。
Counselor: 家のことを手伝ったり、妹さんの世話をされたり、相談者さんは本当にお優しいかたですね
Counselor: 進学するかどうかも悩んでしまっているのですね
Client: 家のこととかは、やる人がいないからやってる感じですよ
Client: 進学すると家のことが回らないんじゃないかという心配も正直あるかな
Counselor: そう思われているのですね心配や不安がありながらも、進路のことなどしっかり考えているのは凄いことだと思います
Counselor: このことを、どなたかに相談されたことはありますか？
Client: ないですね。
Counselor: そうなんですね親御さんは、何か言われていますか？
Client: 親は働きづめで、あまり話をする時間もないかな
Client: 自分のことより妹のことの方が気になるかもしれない。
Counselor: そうなんですね妹さんの、どんなことが気になりますか？
Client: ダウン症なので、手がかかるんです。
Client: だから僕のことより妹のことの方が気になって進路のこととかあまり考えていないかも。
Counselor: そうなんですねそれは心配になりますよねそのことが気になって、進路のことも悩まれているのですね
Client: 食事の介助もいるので、よく手伝うんですけど、
Client: 家事にも時間が取られてとにかく自分の時間がないので勉強できるかな～って思っています。
Client: みんな塾とか行ってるし…
Counselor: うんうん家事というのは意外に時間を取られるものですよね他のお友達は塾とかに行っているのですね
Client: はい、だから勉強量がちがうし、追いつけるのかなって焦る気持ちも出てきています。
Counselor: そうなんですねまわりの方達を見ていると、不安や焦りが出てきてしまうのですね
Counselor: ここまでのお話しを少し整理してみますね
Client: そうですね、自分のことだけに集中できるのは羨ましい気持ちにもなります。
Client: はい。
Counselor: そうですよねそう思ってしまうのは当然だと思いますではまとめてみますね
Client: はい。
Counselor: 進学のことで悩んでいる親御さんが仕事で忙しい中、妹さんの世話や家事に時間を取られていて、塾などにも行く時間がない
Client: はい。
Counselor: 部活や委員会にも入っていないので、内申も気になる妹さんの将来のことなどを考えると、進学していいのかも考えてしまっている、ということですかね？
Client: そんな感じですね
Counselor: 確認ありがとうございます相談者さんが少しでも気持ちが楽になるように、一緒に考えてみましょう
Client: はい。お願いします
Counselor: 色々と悩まれていることはあると思いますが、今一番悩んでいることは、受験するかどうか、でしょうか？
Client: 受験はしたいけど勉強時間が確保できないことになるのかなぁ
Counselor: うんうん受験はしたいと思っているのですねでは、どのようにすれば勉強時間を作れると思いますか？
Client: 家事等の時間が減ればですかね。でも妹の世話を誰がするのかとかいろいろ気になってしまいますね。
Counselor: うんうん家事の時間を減らせるかどうか、と妹さんのお世話のことですね
Client: そうですね。
Counselor: 妹さんのお世話は、相談者さん以外にできる方はいそうでしょうか？
Client: 今のところ思いつかないですね。
Counselor: そうなんですねでは、家事の時間をどう減らすか、になりますよね
Client: そうですね
Counselor: 学校を終えて帰宅してから、どんな流れでしょうか？
Client: 買い物して、ご飯作って、食べさせて、
Client: 風呂沸かして、食器洗いして宿題かな
Counselor: うんうんたくさんのことをしっかりこなしているのですね買い物は毎日ですか？
Client: そうですね
Counselor: そうなんですねお料理も毎日ですか？
Client: ほぼ毎日です。たまに惣菜や弁当買ってきます。
Counselor: そうなんですねでは、例えばですが、買い物と料理の回数を毎日でなくしてみる、というのはどうでしょうか？
Client: どうしたらいいですか？
Counselor: これはあくまでも私の場合なのですが、仕事や家事で忙しくて時間がないなと思う時は、
Counselor: 週末とかに少し頑張って大目に買い物をしておいて、料理も何品かまとめて作って冷蔵や冷凍にしておきます
Client: なるほど。
Counselor: そうすると、平日の時間は結構作れます
Client: どんなものをまとめて作ったらいいでしょうか？
Counselor: そうですね大抵のものは冷凍しておくと大丈夫ですよ副菜的なものを作っておくだけでも、メインはお肉焼くだけとか、魚焼くだけですみますよ
Client: 野菜のおかずですか？
Counselor: そうですね野菜の副菜であれば冷蔵で数日大丈夫なものもあるので、一度ネットとかで調べてみると、結構色々なアイデアが出てくると思いますが、いかがでしょうか？
Client: まとめ買いとまとめて作るですね。週末にやってみます。。
Counselor: うんうんもし親御さんもお休みで一緒にできれば、効率も上がりますし、お料理しながら少しお話しする時間もとれるかもしれないですよね？
Client: 結構休みは親が昼まで寝てたりするんですけど、やれたらいいですね。
Counselor: そうですね親御さんもお疲れだと思いますが、きっと相談者さんとお話しもしたいなと思われていると思いますよ
Client: 普段話さないから何話したらいいか…正直照れくさいですけどね。
Counselor: うんうんそうですよね最初はお料理のコツとか聞きながらとかで大丈夫だと思いますよ相談者さんはとても頑張り屋さんだと感じましたが、ご自分のことも大切にして下さいね
Client: ありがとうございます。
Counselor: では、本日のご相談は終了となりますまたいつでもご相談に来てくださいね
Client: わかりました。
Counselor: 受験が上手くいくよう、応援しております本日はありがとうございました
Client: ありがとうございました。
Counselor: (相談者役、お疲れ様でした　ありがとうございました)
Client: (お疲れ様でした。ありがとうございました。)

Client Feedback



Total Score : 72 / 100

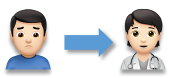
1. 聴いてもらえた、わかってもらえたと感じた: 3/5
2. 尊重されたと感じた: 3/5
3. 新しい気づきや体験があった: 4/5
4. 希望や期待を感じられた: 3/5
5. 取り組みたかったことを扱えた: 4/5
6. 一緒に考えながら取り組めた: 4/5
7. やりとりのリズムがあっていた: 3/5
8. 居心地のよいやりとりだった: 3/5
9. 全体として適切でよかった: 2/5
10. 今回の相談は価値があった: 3/5
11. 相談開始の円滑さ: 5/5
12. 相談終了のタイミング、円滑さ: 5/5
13. 受容・共感: 3/5
14. 肯定・承認: 3/5
15. 的確な質問による会話の促進: 3/5
16. 要約: 5/5
17. 問題の明確化: 5/5
18. この相談での目標の明確化: 5/5
19. 次の行動につながる提案: 3/5
20. 勇気づけ・希望の喚起: 3/5

Figure 12: Data example from KokoroChat (Japanese original version).

Counseling Dialogue

Counselor: Good morning. I'll be assisting you today. Nice to meet you.
Client: Good morning. Nice to meet you too.
Counselor: I'm your counselor today. It's a pleasure to talk with you.
Client: Likewise.
Counselor: To start off, could you tell me your age and gender?
Client: I'm 14 years old, male.
Counselor: You're 14, I see. What grade are you in?
Client: I'm in my third year of middle school.
Counselor: Thank you. So, what would you like to talk about today?
Client: I'm not sure what to do about high school.
Counselor: You're feeling unsure about high school?
Client: Yeah.
Counselor: Could you tell me a little more about what's on your mind?
Client: I live with a single parent, so I have to help out a lot at home. I barely have time for myself, but...
Client: I know I need to start thinking about high school.
Client: I haven't done any extracurricular activities like clubs or student council,
Client: So I'm wondering if that will affect my school application.
Counselor: I see, that must be a big concern for you.
Client: Yeah, and I also have a younger sister to take care of, so I wonder if I should even go to high school.
Counselor: You take care of your sister and help out around the house—you're really kind and responsible.
Counselor: So, you're not just worried about school, but also about whether you should continue studying at all?
Client: It's not really about being kind, there's just no one else to do it.
Client: And honestly, I worry that if I go to high school, no one will be there to take care of things at home.
Counselor: I see. Even with all these concerns, you're still seriously thinking about your future, which is really admirable.
Counselor: Have you talked to anyone about this before?
Client: No, not really.
Counselor: I see. Has your parent said anything about it?
Client: My parent is always working, so we don't really have much time to talk.
Client: I think they're more worried about my sister than me.
Counselor: I see. What about your sister concerns you?
Client: She has Down syndrome, so she needs a lot of care.
Client: That's why I think my parent is more focused on her and hasn't really thought about my future.
Counselor: I see. That must be really tough for you. It makes sense that this is affecting your thoughts about school.
Client: Yeah. I often help her with meals, and housework takes up a lot of time too.
Client: I barely have time for myself, so I don't know if I'll be able to keep up with schoolwork.
Client: Everyone else is going to cram school too…
Counselor: That makes sense. Chores can take up a lot of time, and seeing your classmates going to cram school probably adds to the pressure.
Client: Yeah, they study way more than I do, and I'm worried I won't be able to keep up.
Counselor: I understand. Seeing everyone else prepare while you feel stuck must make you anxious.
Counselor: Let's take a moment to go over what we've talked about so far.
Client: Yeah. Honestly, I feel envious of people who can just focus on themselves.
Client: Sure.
Counselor: That's totally understandable. It makes sense to feel that way. So, let's summarize.
Client: Okay.
Counselor: You're worried about high school. Your parent is always working, and you have to take care of household chores and your sister, which takes up most of your time. You don't have time to go to cram school.
Client: Yeah.
Counselor: You're also concerned because you haven't participated in clubs or student council, so you're not sure how that will affect your school application. And on top of that, you're wondering if you should even go to high school, since you're thinking about your sister's future too. Does that sound right?
Client: Yeah, that sounds about right.
Counselor: Thank you for confirming. Let's think this through together so you can feel a little more at ease.
Client: Yes, please.
Counselor: I know you have a lot on your mind, but right now, is your biggest concern whether or not you should take the entrance exam?
Client: I do want to take the exam, but I'm not sure if I can make enough time to study.
Counselor: I see, so you do want to take the exam. In that case, how do you think you could create more study time?
Client: I guess if I had less housework to do. But then, who would take care of my sister? That's something I keep worrying about.
Counselor: I see, so the issue is whether you can reduce the time spent on housework while also ensuring your sister is taken care of.
Client: Yeah, that's right.
Counselor: Is there anyone else who could help take care of your sister?
Client: Not really, I can't think of anyone right now.
Counselor: I see. So then, the focus would be on finding ways to reduce the time spent on housework.
Client: Yeah, I guess so.
Counselor: What does your usual routine look like after you get home from school?
Client: I go shopping, make dinner, feed my sister,
Client: Heat up the bath, wash the dishes, and then finally do my homework.
Counselor: I see, you're handling a lot of tasks on your own. Do you have to go shopping every day?
Client: Pretty much.
Counselor: And do you cook every day as well?
Client: Almost every day. Sometimes we just buy ready-made meals or bento boxes.
Counselor: I see. Here's a thought—what if you didn't have to shop and cook every single day?
Client: How would I do that?
Counselor: Well, this is just something I personally do when I'm busy with work and house chores—
Counselor: I try to buy groceries in bulk over the weekend and prepare several dishes in advance, storing them in the fridge or freezer.
Client: That makes sense.
Counselor: That way, I have a lot more free time during the weekdays.
Client: What kind of things would be good to prepare in advance?
Counselor: Most things can be frozen, actually. Even just preparing side dishes ahead of time can make things easier—you can quickly grill some meat or fish and pair it with a ready-made side.
Client: Like vegetable side dishes?
Counselor: Exactly. Some vegetable-based side dishes can last for a few days in the fridge. You might find some good ideas if you look it up online. What do you think?
Client: So, bulk shopping and batch cooking… I'll try doing it this weekend.
Counselor: That's great! If your parent has a day off, maybe you can do it together—it could be more efficient, and you might get a chance to talk while cooking.
Client: My parent usually sleeps in until noon on their days off, but if possible, that sounds good.
Counselor: Yeah, they must be really exhausted, but I'm sure they'd like to spend some time talking with you too.
Client: We don't usually talk much, so I wouldn't even know what to say… Honestly, it'd feel kind of awkward.
Counselor: That makes sense. But you could just start by asking for cooking tips—it doesn't have to be anything complicated. You seem like a really hardworking person, but please remember to take care of yourself too.
Client: Thank you.
Counselor: Well, that's all for today. Please feel free to come back anytime.
Client: Got it.
Counselor: I hope your entrance exams go well. Thank you for talking with me today.
Client: Thank you.
Counselor: (Thank you for taking part in this session. Great job!)
Client: (Thank you. I really appreciate it.)

Client Feedback



Total Score : 72 / 100

1. Felt heard and understood : 3/5
2. Felt respected : 3/5
3. Gained new insights : 4/5
4. Felt hopeful or expectant : 3/5
5. Concerns were addressed : 4/5
6. Thought through concerns together : 4/5
7. The conversation had a good rhythm : 3/5
8. The conversation felt comfortable : 3/5
9. Felt appropriate and satisfying : 2/5
10. The conversation was valuable : 3/5
11. The conversation started smoothly : 5/5
12. The conversation ended well : 5/5
13. Showed acceptance and empathy : 3/5
14. Provided acknowledgment and affirmation : 3/5
15. Asked effective questions to foster dialogue : 3/5
16. Summarized key points effectively : 5/5
17. Clarified issues clearly : 5/5
18. Helped identify goals for the conversation : 5/5
19. Offered actionable suggestions : 3/5
20. Encouraged and instilled hope : 3/5

Figure 13: Data example from KokoroChat (English version, translated by authors).

Figure 14: Examples of generated responses from each model.