PALbot+: An Empathetic Dialogue System for the Marginalized Populations in Korea with AI Safety Features

Suyeon Lee **Hye Jin Lee** **Hyunjong Kim** **Sohhyung Park** **Sungzoon Cho** *
Seoul National University
{suyeon.lee, hyejinlee, hjkim, sohhyung}@bdai.snu.ac.kr
zoon@snu.ac.kr

Abstract

The sharp rise in marginalized populations in South Korea, coupled with a severe shortage of mental health professionals and high costs that impede adequate support, urgently calls for alternative solutions that provide immediate care. We introduce **PALbot+**, an AI-driven dialogue system designed to provide emotional support to marginalized individuals in South Korea, while refraining from generating inappropriate responses when presented with *dangerous* utterances from the user. We augment a large-scale data in Korean language to train small-sized large language models (sLLMs) and a set of out-of-bound classifiers, designed to generate culturally relevant and contextually sensitive responses. Evaluation metrics show that PALbot+ outperforms the vanilla model, indicating an enhanced ability to produce a rich variety of expressions attending to different situations of users.

1 Introduction

The rapid increase in elderly and single-person households in South Korea highlights the urgent need for accessible mental health support. In 2020, older adults comprised 15.7% of the Korean population, significantly exceeding the global average of 9.3% (Statistics Korea, 2019; United Nations Department of Economic and Social Affairs, 2020). Single-person households accounted for 31.7% of all households, with 18.2% living below the poverty line. Mental health issues are particularly prevalent in these groups. More than 20% of adults aged 60 and older suffer from depression or dementia, while individuals living alone with financial limitations face heightened risks of social isolation and psychological distress (World Health Organization, 2017). However, access to professional mental health services remains limited due to financial constraints, mobility issues, and an increasing shortage of trained professionals (Organization, 2022; Kwon et al., 2021).

Automated dialogue systems offer a scalable solution by providing accessible, nonjudgmental companionship with minimal human intervention. While chatbots have shown promise in delivering emotional support (Fitzpatrick et al., 2017), developing one presents severe challenges. Standard language models often generate generic responses that fail to address users' emotional needs. Furthermore, fine-tuning models for Korean-language mental health support is hindered by the lack of culturally and linguistically relevant training data.

We introduce PALbot+, an AI-powered dialogue system designed to provide regular check-ins and empathetic responses to encourage user engagement, with extra caution for AI safety. Our methodology involves three key steps: (1) Seed data collection, where Korean language conversational data with emotionally engaging interactions is gathered; (2) Data augmentation, where multiple datasets are generated to train models for empathetic conversations; and (3) Fine-tuning of LLaMA-3-8B-Instruct and training of out-of-bound classifiers to generate culturally relevant and context-aware responses while refraining from generating unsafe or inappropriate responses. We evaluated PALbot+ using both human assessments and automated metrics, the results of which showed that PALbot+ outperformed its vanilla counterpart, demonstrating the efficacy of our model in providing mentally supportive yet safe responses. The overall methodology is illustrated in Figure 1.

2 Dataset Construction

2.1 Fine-tuning Dataset

We gather and augment datasets to fine-tune the chatbot model.

*Corresponding author.

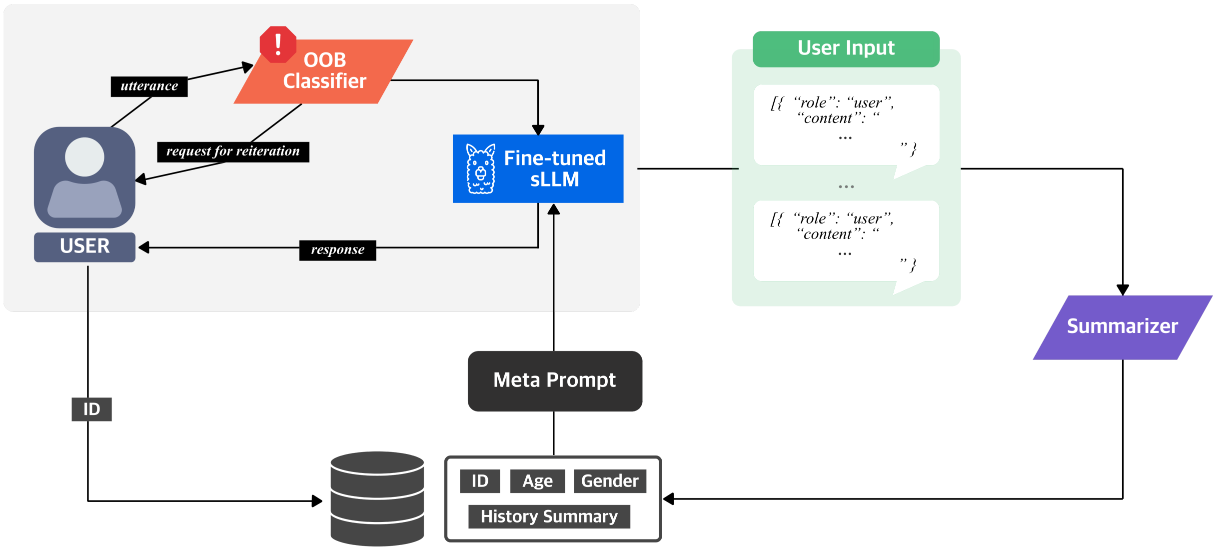

Figure 1: Illustration of the overall methodology.

2.1.1 Seed Data Collection

This study develops a chatbot for multi-turn empathetic conversations to support the psychological well-being of marginalized individuals. The chatbot generates contextually relevant and empathetic responses and provides personalized emotional support tailored to the user's conversation history. It also adapts to different tones and topics for various age groups.

To achieve this, we collect three multi-turn datasets focused on psychological care and empathetic dialogue across diverse demographics.

Emotional Dialogue Corpus [1] This AI training dataset, provided by AI Hub, consists of 270,000 sentences collected from 1,500 individuals via crowd-sourcing. Structured into multi-turn dialogues of up to three turns, it includes user personas based on demographic, situational, health-related, and emotional factors. Conversations follow a user-system sequence.

CareCall-Memory [2] This dataset represents a Korean multi-session open-domain dialogue corpus from CareCall-Memory(Bae et al., 2022a), which extends the single-session CareCall dataset (Bae et al., 2022b) into a multi-session format using HyperCLOVA 6.9B. CareCall was developed through human-written role specifications, dialogue examples, and in-context few-shot learning from LLMs. Dialogs follow a system-user sequence.

Korean Multi-Session Dialogue Data [3] The dataset, provided by AI Hub, is a multi-session conversational corpus designed for chatbot training. It comprises everyday dialogues between two participants across multiple sessions, guided by assigned personas and 13 predefined topics. We select the "Personal and Relationships" topic for training as a daily conversation dataset. The dialogues follow a sequence from user 1 to user 2.

2.1.2 Data Augmentation

The collected conversational data are not suitable for immediately training an empathetic chatbot. Therefore, we employ an LLM for data augmentation. Few-shot prompting is utilized, incorporating user persona details, guidelines, instructions, and examples as needed.

For optimal training, the dataset is structured to include the chatbot's initial and closing responses with around 10 conversational turns. The augmentation process is adjusted accordingly, ensuring balanced data generation across age groups from teenagers to those in their 60s. The complete prompts used for each dataset augmentation are provided in Appendix A.1.

Emotional Dialogue Corpus The emotional dialogue corpus[1] is categorized into four age groups: adolescents (teens), young adults (20s–30s), middle-aged adults (40s–50s), and elderly (60s), with ambiguous cases assigned arbitrarily. Since the dataset consists of 3-turn conversations without

[1] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=86

[2] https://github.com/naver-ai/CareCall-Memory

[3] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=71630

a closing utterance, augmentation adds an initial chatbot response using user persona details and the original dialogue. The conversation is extended to 8–12 turns with polite and coherent responses, and a concluding utterance is generated from the final two turns. In total, 4,500 dialogues, averaging nine turns each, are augmented.

CareCall-Memory Since the CareCall-Memory dataset (Bae et al., 2022a) is not publicly available, we use its sample data[2] for few-shot prompting. Each sample includes a dialogue and its summary. To augment the dataset, we generate full dialogues from summaries using in-context learning with three examples. For dialogue content, we incorporate persona features from the Korean Multi-Session Dialogue dataset[3], randomly sampling features by age group and sex. Each feature generates one dialogue, resulting in 3,400 dialogues.

Korean Multi-Session Dialogue We use the Korean Multi-Session Dialogue Dataset[3] to enhance natural communication. Focusing on the "Personal and Relationships" theme, we apply stratified sampling by age to ensure balanced representation. To expand the dataset, we convert user responses into chatbot responses, generate chatbot openings, and craft coherent closings. This preserves authenticity while improving contextual relevance and empathy. A total of 794 dialogues have been augmented.

2.2 Out-of-Bounds Dataset

We construct five out-of-bounds (OOB) datasets to ensure safe operation of our mental healthcare chatbot. The OOB datasets address specific concerns: hate speech, depression-related utterances, suicidal expressions, clinical advice seeking, and pharmaceutical inquiries. For in-bound cases, we randomly sample utterances from the "Leisure and Entertainment" theme of Korean Multi-Session Dialogue. All datasets are carefully curated to reflect Korean linguistic patterns and cultural context. The full-length prompts for constructing these datasets and the resulting data samples are presented in Appendix C.1.

Hate Speech We sample data from Korean-malicious-comments-dataset[4] containing comments from online communities and news sites. Focusing on emotional intensity, we filter for expressions with the maximum of two sentences that end with basic punctuation marks.

Depression Utterance We generate Korean samples from Reddit's Depression branch[5] using GPT-4. These prompts are designed for colloquial translation, limiting each sample to two short sentences to reflect natural chat interactions while preserving emotional depth.

Suicidal Utterance Similarly, we use Reddit's SuicideWatch branch[6] to generate Korean samples using the same GPT-4 prompting strategy.

Clinical Inquiry To prevent the chatbot from providing unauthorized medical guidance, we develop a clinical inquiry dataset using New York Presbyterian Hospital's symptom-disease knowledge graph[7] covering 151 diseases. We generate queries using GPT-4 with controlled prompting, randomly combining symptoms for each disease while maintaining natural language patterns.

Pharmaceutical Inquiry To prevent inappropriate medication advice, we create a pharmaceutical dataset by scraping content from Korean medical information sources. Using these as few-shot examples, we generate additional conversational queries through GPT-4 while maintaining the original inquiry patterns.

3 Model

3.1 sLLM Fine-tuning

Given the chatbot's versatility and efficiency, we use open-source small large language models (sLLMs) in Korean with fewer than 10 billion parameters. We select the LLaMA-3-8B-Instruct model (Dubey et al., 2024) for fine-tuning.

We apply instruction fine-tuning, incorporating task descriptions, constraints, and user persona details to guide response generation. The model is designed to maintain a polite tone and avoid providing medical advice. The detailed instructions are shown in Figure 12.

3.2 Out-of-Bounds Classifier

We incorporate an out-of-bounds (OOB) classifier to prevent our chatbot from generating inappropriate responses by detecting potentially problematic

[4]https://github.com/ZIZUN/korean-malicious-comments-dataset

[5]https://www.reddit.com/r/depression/

[6]https://www.reddit.com/r/SuicideWatch/

[7]https://impact.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/

user inputs. For this classification task, we fine-tune KM-BERT (Kim et al., 2022b), utilizing its understanding of both Korean linguistic features and medical domain knowledge. The classifier evaluates each user utterance before it reaches the main dialogue system. When the classifier detects any of the five OOB categories, the system prompts the user to rephrase their input instead of generating a potentially inappropriate response. We refer to PALbot with this OOB classifier as PALbot+.

3.3 Single-Session

We generate an appropriate response from the fine-tuned chatbot model, PALbot+, based on the dialogue history using a prompting technique. To ensure conversational continuity, we maintain the dialogue history within the same session. Additionally, we define a suitable description to guide PALbot+ in producing polite and empathetic utterances. The prompt used is shown in Figure 13.

3.4 Multi-Session

We design a straightforward multi-session framework utilizing PALbot+ by expanding the single-session structure. After completing the initial dialogue session, the chatbot summarizes the conversation into three key points. These summaries, along with the user's information, are stored to maintain continuity. In subsequent sessions, the chatbot references these stored summaries and user data to preserve context. The prompt is adjusted to ensure that the chatbot's first response in each new session reflects the previous conversation. Additionally, the chatbot is guided to seamlessly incorporate the summarized content throughout the dialogue.

4 Experiment

4.1 Experimental Setup

4.1.1 Data Augmentation

We utilize the GPT-4 API [8] (Achiam et al., 2023) to augment all the datasets. The resulting datasets are freely available online for non-commercial use[9].

4.1.2 Fine-tuning

We fine-tune the Korean version of LLaMA-3-8B-Instruct[10] (Dubey et al., 2024), to build PALbot. The fine-tuned PALbot model[11] is available on Huggingface for non-commercial use. We train the model with a batch size of 16 using the AdamW optimizer (Loshchilov, 2017) with a learning rate of 1e-6. The model is fine-tuned for 5 epochs with LoRA 4-bit quantization (Hu et al., 2021) on two NVIDIA A100 GPUs with 40GB memory.

4.1.3 Out-of-Bounds Classifier

We fine-tune KM-BERT (Kim et al., 2022b) for out-of-bounds (OOB) classifiers. The fine-tuned KM-BERT model is used to classify potentially harmful or inappropriate content in chatbot interactions. We train the model with a batch size of 4 using the AdamW optimizer with a learning rate of 1e-6. The model is fine-tuned for 30 epochs with a dropout rate of 0.5 to prevent overfitting. The trained model achieved a test accuracy of 0.901.

4.2 Evaluation

4.2.1 Human Evaluation

For human evaluation, we set up an evaluation panel of 50 participants whose age category ranges from their 20s to 60s. The age group distribution of the evaluation panel is presented in Table 1.

Age group	Number of evaluators
20s	16
30s	16
40s	6
50s	6
60s	8
Total	52

Table 1: Age group distribution of the evaluation panel.

For each age group, evaluators were randomly assigned to assess either the LLaMA-3 vanilla model or PALbot, with and without out-of-bounds classifiers. Some of the participants were randomly selected to evaluate the multi-session model. These evaluators chatted with a vanilla model with OOB classifiers and PALbot+, twice in a row for each model, with a brief break in between. The models were presented sequentially in a randomized order, with their identities concealed from the evaluators. Evaluators conversed with the chatbot for up to 10 turns [12].

[8] gpt-4-0125-preview

[9] http://bit.ly/4ffuF6h

[10] https://huggingface.co/beomi/Llama-3-Open-Ko-8B-Instruct-preview

[11] https://huggingface.co/snu-bdai/palbot-llama3-8b

[12] We limited the evaluation terminate within 10-turns, based upon the previous literature that open-source sLLMs perform well for 4–8 turns and then their performance declines sharply beyond 16 turns (Duan et al., 2024)

Evaluation Metrics Given that the primary objective of this study is to develop an empathetic chatbot, we propose employing (1) **Empathy** as an evaluation metric. Empathy, in this context, is defined as the chatbot's ability to accurately recognize the evaluator's emotional state and generate appropriately attentive responses.

Reflecting on past literature (Wan et al., 2022; Chen et al., 2023; Adiwardana et al., 2020; Kim et al., 2022a; Bae et al., 2022b), we adopt three of the most commonly used metrics in order to determine the quality of the conversations with the given model: (2) **Coherence**, (3) **Fluency**, and (4) **Naturalness**. Coherence measures the extent to which chatbot's responses are coherently in line with the flow of the conversation. Fluency examines the level of elaborateness of chatbot's responses in terms of grammar, vocabulary, and/or expressions. Naturalness assesses whether the given chatbot demonstrates human-like conversational ability. (5) **Safety** assesses whether the given chatbot does not generate politically incorrect or ethically inappropriate. For multi-session models, we additionally adopted two metrics, (6) **Personalization** and (7) **Long-term consistency**, which measures the extent of user-specific memory preservation and the consistency in the colloquial tone across different dialogue sessions, respectively. For all metrics, we use 4-scale ratings, ranging from 0 being the worst quality to 3, the best [13]. We list the scoring table in full in Appendix D.

4.2.2 Automatic Evaluation

Evaluation Metrics We measured two metrics for automatic evaluation, Perplexity and Distinct-N ($N = 1, 2, 3$) (Li et al., 2015). Perplexity assesses the model's ability to predict the next token in a sequence, whereas Distinct-N measures the lexical diversity of the sequence. We calculated these metrics using the dialogue history between the evaluator and the chatbot.

5 Results and Discussion

5.1 Dataset

Our final dataset is composed of three augmented datasets: (1) Emotional Dialogue corpus, (2) CareCall-Memory, and (3) Korean Multi-Session Dialogue. The statistics for the final dataset are presented in Table 2. It consists of 8,694 dialogues, with an average of 8.09 turns per dialogue and 33.68 words per turn. The distribution is relatively even across most age groups. Detailed statistics for each dataset, along with the distribution of dialogues by age group, are provided in Appendix B.

Metric	Final dataset
# Dialogues	8,694
# Turns	70,306
Avg. turns / dialogue	8.09
# Words	2,368,020
Avg. words / turn	33.68
# Unique words	41,268

Table 2: Statistics of the final dataset.

5.2 Quantitative Results

5.2.1 Human Evaluation

We report the human evaluation results in Table 3. The column "Session 2 (no long-term)" reports average scores without considering long-term metrics, while Session 2 computes averages including all metrics. Results show that PALbot+ consistently outperforms other models in both single-session and multi-session settings. We note that the presence of OOB classifiers improved performance of PALbot+, while the result was opposite for the vanilla model. We postulate that users found the chatbot's carefulness helpful when it adequately empathized with them; in contrast, if the conversation did not prove attentive or empathetic, extra installation of AI safety features only decreased the usability of the dialogue system. On the other hand, for the multi-session setting, the inclusion of long-term measures increases the overall assessment PALbot+'s performance, as supposed to the vanilla model whose score decreasing when considering the long-term metrics. This substantiates the effect of fine-tuning in terms of personalization and long-term consistency.

Figure 2 presents the single-session evaluation results by evaluation metric. While PALbot+ consistently achieved the highest scores at a stable level across all metrics, the results for the vanilla models varied across different metrics.

Figure 3 shows the single-session evaluation results by age group. The performance gap between PALbot+ (and PALbot) and its vanilla counterpart is consistently large across all age groups except for those in their 50s. Especially, the improvement

[13] We stay away from the conventional 5-scale so that the evaluation scores present clearer polarity, while ensuring enough variation to illustratively express the level of satisfaction with the chatbot.

Model	Single-Session Average Score	Multi-Session Average Score		
		Session 1	Session 2 (no long-term)	Session 2
LLaMA-3	1.85			
LLaMA-3 + OOB	1.72	1.52	1.58	1.53
PALbot	2.24			
PALbot + OOB	**2.45**	**2.48**	**2.36**	**2.44**

Table 3: Human evaluation result for single- and multi-session settings. The “no long-term” for Session 2 indicates that the long-term measures such as personalization and long-term consistency, were excluded when computing the average scores.

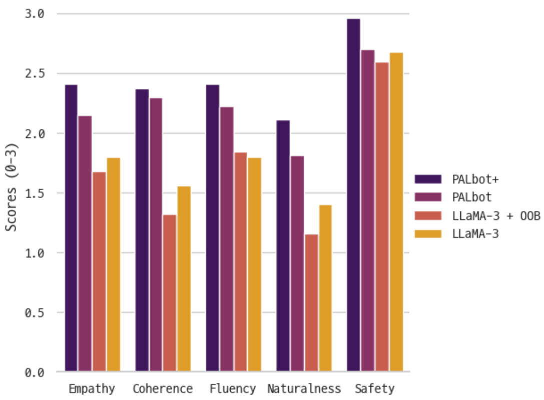

Figure 2: Single-session evaluation results of each age group.

in the 60s age group is the second largest, highlighting the effectiveness of PALbot for one of the key target groups in the marginalized population, the elderly.

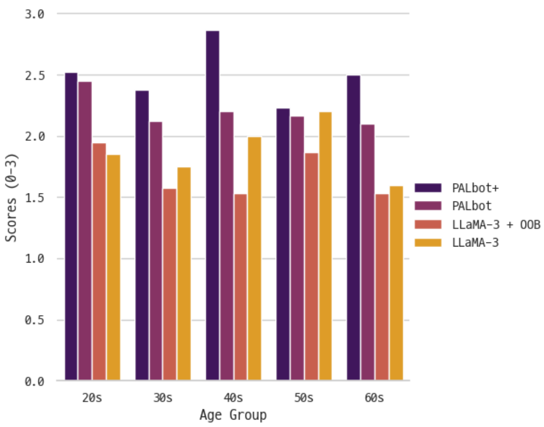

Figure 3: Single-session evaluation results of each evaluation metric.

Finally, we report the multi-session results by evaluation metric in Figure 4. PALbot+ exhibit improved ability in personalizing conversations with chat history memorization. The improvement in long-term consistency is more impressive, indicating PALbot+’s effectiveness in maintaining a continuously supportive attitude during the conversations with the same user across different sessions.

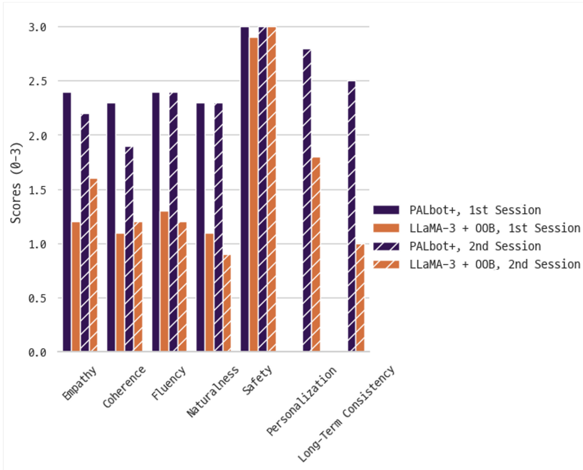

Figure 4: Multi-session evaluation results of each evaluation metric.

5.2.2 Automatic Evaluation

Model	PPL	D-1	D-2	D-3
LLaMA-3	17.68	0.1867	0.5347	0.7731
LLaMA-3 + OOB	14.65	0.1938	0.5713	0.8067
PALbot	**11.94**	0.2811	0.6570	0.8199
PALbot	12.99	**0.3215**	**0.7107**	**0.8540**

Table 4: Model evaluation by Perplexity and Distinct-N. PPL and D-N denote Perplexity and Distinct-N, respectively.

Table 4 compares the performance of LLaMA-3 and PALbot models through Perplexity and Distinct-N metrics, with and without an OOB classifier. The results show that PALbot models achieve significantly lower perplexity scores than LLaMA-3 models, indicating that PALbot generates more coherent and fluent responses. With the integration of the OOB classifier, LLaMA-3 showed notable improvement, while PALbot maintained a similar level. This suggests that LLaMA-3 is more susceptible to generating unnatural responses to OOB queries, while PALbot handles such challenging queries more naturally even without the OOB classifier. In terms of Distinct-N metrics, PALbot with the OOB classifier achieved the highest scores across all values of *N*. The OOB classifier improved these metrics for both models, suggesting that both models tend to generate generic and monotonous responses to OOB queries. The results indicate that prompting users to restate their input using the OOB classifier leads to better response quality and more engaging conversations. Overall, these results highlight the superiority of PALbot and the effectiveness of the OOB classifier.

5.3 Qualitative Results

In this section, we analyze the dialogue results between human evaluators and chatbots constructed using four different models. PALbot+ and the vanilla counterpart demonstrate significant differences in their conversational abilities. We present the dialogue between a male user in his 20s and PALbot+ from the first session in its full length, along with the dialogue summary at the end, in Appendix E, Table 10. We observe that PALbot+ maintains consistently concise yet attentive and agreeable responses and introduces follow-up questions to sustain engagement, even after encountering disruptive user input such as "Hello" followed by a question. The effectiveness of the OOB classifiers is also evident, as PALbot+ successfully flags the unsafe trigger when the user seeks advice on hospital booking, prompting the user to rephrase their request. The chat summary captures key information, emphasizing the user's emotional state and activities that may influence their mood. The full dialogue between the same user and PALbot+ from the second session is reported in Appendix E, Table 11. Following up with the previous chat summary, PALbot+ inquires about exercise and time constraints. The conversation concludes smoothly, with the user expressing gratitude.

We report an additional dialogue, between a female user in her 50s and LLaMA-3-8B-Instruct without OOB classifiers, in its full length as a counter example in Appendix E, Table 12. We observe that the vanilla model, with or without OOB classifiers, often struggles with context comprehension hence generating highly uncommon and strange expressions such as "falling without getting off" or "days with business". It fails to ask specific questions and focuses only on the basic demographic information such as gender and age group as shown in the question: "What kind of things generally stress out women in their 50s?". Moreover, the vanilla model tends to respond in exhaustively long sentences yet without much critical information or empathetic reactions. Due to the inadequacy of the vanilla model eventually leads to the user's disappointment, as shown in the user's last remark stating "I feel a little hurt."

6 Related Work

Chatbots for Mental Health Care With the advancement of natural language processing (NLP), there has been growing interest in dialogue systems that interact with users. Recently, more attention has been paid to chatbots designed for mental health care that require personalization and empathetic responses. Vinyals and Le (2015) introduced foundational techniques, while pre-trained language models such as BERT (Devlin et al., 2019) significantly improved the ability of dialogue systems to understand context. Large language models (LLMs) such as GPT-3 (Brown et al., 2020) have shown the potential to facilitate more natural and empathetic conversations (Inkster et al., 2018). Early research in this field mainly employed rule-based approaches that focused on addressing specific conditions, such as depression and anxiety. For instance, Woebot (Fitzpatrick et al., 2017) is a conversational agent based on cognitive behavioral therapy principles, though these systems often rely on predefined conversation flows (Denecke et al., 2021). Recent studies have employed generative models, where Brocki et al. (2023) and Deng et al. (2023) utilized BlenderBot (Roller et al., 2020) to create more emotionally supportive conversations by incorporating commonsense knowledge. Beyond chatbots, LLMs have been utilized in mental healthcare for tasks such as processing clinical data (Taylor et al., 2024) and automating health coaching (Ong et al., 2024). Research on Korean language-based chatbot systems for mental health support is still in early stages, with Oh et al. (2017) introducing a psychiatric counseling chatbot and Bae et al. (2022b) proposing methods for building role-specific dialogue systems.

Data Augmentation Techniques for Dialogue Systems Data augmentation is essential in addressing data scarcity, particularly for languages like Korean, where large-scale datasets are limited. Methods such as back-translation (Sennrich et al., 2016) have been adapted for dialogue data, though they may not fully capture linguistic nuances. Other techniques, like contextual word embeddings (Kobayashi, 2018) and BERT's masked language modeling (Wu et al., 2019), help preserve semantic meaning during augmentation. Kumar et al. (2020) demonstrated that combining back-translation with BERT-based word replacement improves data quality. For task-oriented dialogue systems, Zhao et al. (2023) introduced a framework integrating subjective knowledge into models. Recently, LLMs such as HyperCLOVA (Bae et al., 2022b) have been used to generate domain-specific dialogues, while Liu et al. (2025) applied LLM prompting to synthesize clinical reports, reducing real-world data collection burdens.

7 Conclusion

This study presents PALbot+, an AI-powered dialogue system designed to provide regular check-ins and empathetic responses while ensuring AI safety. To develop a chatbot capable of offering culturally relevant and contextually attentive interactions, we augment a large-scale Korean language conversational dataset with a focus in emotionally engaging interactions, and fine-tuning of LLaMA-3-8B-Instruct alongside out-of-bound classifiers to mitigate unsafe or inappropriate responses. Evaluation results demonstrated that PALbot+ consistently outperformed its vanilla counterpart, effectively generating supportive and context-aware interactions while maintaining safety constraints. We anticipate that PALbot+ offers a scalable, culturally adapted solution for assisting mental health professionals and supporting underserved populations in Korea.

8 Limitations

While PALbot+ demonstrates effectiveness in generating empathetic and context-aware responses, several limitations remain. First, the scope of the out-of-bounds (OOB) classifier is relatively narrow, covering only a limited range of sensitive topics. Expanding the OOB training data to include ethical, political, religious, and current societal issues would improve the chatbot's ability to maintain neutrality and prevent unintended biases or misleading impressions. Second, although our evaluation confirms the classifier's effectiveness, further rigorous testing is required to ensure its robustness across diverse conversational settings. Lastly, the current approach to conversation termination is based on a predefined number of turns rather than the user's emotional state. Developing an advanced mechanism that dynamically assesses emotional comfort levels would enable the chatbot to conclude interactions at a more appropriate moment. Addressing these challenges will enhance the system's adaptability, reliability, and overall user experience.

Acknowledgments

This work was supported by the BK21 FOUR Program (Education and Research Center for Industrial Innovation Analytics) funded by the Ministry of Education, Korea (No. 4120240214912) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C2093785).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022a. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787.

Sanghwan Bae, Donghyun Kwak, Sungdong Kim, Donghoon Ham, Soyoung Kang, Sang-Woo Lee, and Woomyoung Park. 2022b. Building a role specified open-domain dialogue system leveraging large-scale language models. *arXiv preprint arXiv:2205.00176*.

Lennart Brocki, George C. Dyer, Anna Gładka, and Neo Christopher Chung. 2023. Deep learning mental health dialogue system. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 395–398.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. Places: Prompting language models for social conversation synthesis. *arXiv preprint arXiv:2302.03269*.

Kerstin Denecke, Alaa A. Abd-alrazaq, and Mowafa Said Househ. 2021. Artificial intelligence for chatbots in mental health: Opportunities and challenges. *Multiple Perspectives on Artificial Intelligence in Healthcare*.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. *Preprint*, arXiv:2305.10172.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. Botchat: Evaluating llms' capabilities of having multi-turn dialogues. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3184–3200.

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

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Ment Health*, 4(2):e19.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth*, 6(11):e12106.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, et al. 2022a. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

Yoojoong Kim, Jong-Ho Kim, Jeong Moon Lee, Moon Joung Jang, Yun Jin Yum, Seongtae Kim, Unsub Shin, Young-Min Kim, Hyung Joon Joo, and Sanghoun Song. 2022b. A pre-trained bert for korean medical natural language processing. *Scientific Reports*, 12(1):13847.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

J. S. Kwon, Y. Kim, and D. Shin. 2021. Mental health services in south korea: Current status and future directions. *Asian Journal of Psychiatry*, 58:102550.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Jinghui Liu, Bevan Koopman, Nathan J Brown, Kevin Chu, and Anthony Nguyen. 2025. Generating synthetic clinical text with local large language models to identify misdiagnosed limb fractures in radiology reports. *Artificial Intelligence in Medicine*, 159:103027.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Kyo-Joong Oh, Dongkun Lee, Byungsoo Ko, and Ho-Jin Choi. 2017. A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, pages 371–375.

Qi Chwen Ong, Chin-Siang Ang, Davidson Zun Yin Chee, Ashwini Lawate, Frederick Sundram, Mayank Dalakoti, Leonardo Pasalic, Daniel To, Tatiana Erlikh Fox, Iva Bojic, et al. 2024. Advancing health coaching: A comparative study of large language model and health coaches. *Artificial Intelligence in Medicine*, 157:103004.

World Health Organization. 2022. *Mental Health Atlas 2022*. World Health Organization, Geneva.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot. *Preprint*, arXiv:2004.13637.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Statistics Korea. 2019. Population projections for korea (2017–2067).

Niall Taylor, Andrey Kormilitzin, Isabelle Lorge, Alejo Nevado-Holgado, Andrea Cipriani, and Dan W Joyce. 2024. Model development for bespoke large language models for digital triage assistance in mental health care. *Artificial Intelligence in Medicine*, 157:102988.

United Nations Department of Economic and Social Affairs. 2020. *World Population Ageing 2020 Highlights*. United Nations.

Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. *arXiv e-prints*, arXiv:1506.05869.

Dazhen Wan, Zheng Zhang, Qi Zhu, Lizi Liao, and Minlie Huang. 2022. A unified dialogue user simulator for few-shot data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3788–3799.

World Health Organization. 2017. Mental health of older adults.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Computational Science – ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part IV*, page 84–95, Berlin, Heidelberg. Springer-Verlag.

Chao Zhao, Spandana Gella, Seokhwan Kim, Di Jin, Devamanyu Hazarika, Alexandros Papangelis, Behnam Hedayatnia, Mahdi Namazifar, Yang Liu, and Dilek Hakkani-Tur. 2023. "what do others think?": Task-oriented conversational modeling with subjective knowledge. *Preprint*, arXiv:2305.12091.

A Appendix A. Prompt Templates

A.1 Data Augmentation

Emotional Dialogue Corpus In this section of appendix, we present the full prompt scenario used to augment the emotional dialogue corpus, as presented in Section 2.1.2.

```
The following is a conversation between a {age}{gender} user and the chatbot (system). The goal of the conversation is to understand the user's psychological state and lifestyle.
Please generate the chatbot's opening question for the conversation. Ensure the question adheres to the given constraints and uses the provided sample questions as a reference.

[Constraints]
• Use a friendly, polite, and empathetic tone.
• Keep the sentence concise, and limited to one line.
• Generate a general question that suits the given conversation.

[Sample Questions]
• Hello~ How have you been lately?
• Hi there~ How are you feeling today?
• Hello! Are you having a good day?

[Conversation]
```

Figure 5: The prompt used for augmenting the Emotional Dialogue Corpus to generate the chatbot's initial utterance.

```
The following is a conversation between a {age}{gender} user and the chatbot (system). The goal of the conversation is to understand the user's psychological state and lifestyle.
Following the outlined constraints, please add 5 to 7 turns at the end of the given conversation. The last sentence should be the chatbot's closing remark, naturally concluding the conversation by reflecting on the previous dialogue. Ensure the final sentence is no longer than one line.

[Constraints]
• Keep each sentence concise, no longer than one line.
• Ensure the responses flow naturally with the given conversation.
• The user's responses should reflect natural language appropriate for their age group ({age}).
• Use a conversational tone, avoiding formal or written language.
• Do not use the word "chatbot" at any point.

[Conversation]
```

Figure 6: The prompt used for augmenting the Emotional Dialogue Corpus to increase the number of turns.

The task is to generate the final turn that ends the conversation between the given {age} {gender} user and the chatbot (system).
Follow the constraints and refer to the [System Response Examples] to generate the final responses for both the user and the system after the given conversation.

[System Response Examples]
- Alright, I'll wrap up for today. Thank you for the chat.
- I'll talk to you again next time.
- Okay, I will get in touch again later. Thank you.

[Constraints]
- Keep each sentence concise, within one line.
- Ensure the responses flow naturally with the given conversation.
- The user's tone should match their conversation history and age group ({age}).
- The system's response must include a statement about ending the chat.

[Conversation]

Figure 7: The prompt used for augmenting the Emotional Dialogue Corpus to add the chatbot's concluding utterance.

CareCall-Memory This Appendix includes the complete prompt scenario used for augmenting CareCall-Memory, as detailed in Section 2.1.2.

Please generate dialogue (4) by reflecting the given summary. Follow the outlined constraints:
- The system should engage in an empathetic conversation with the user.
- The system has no prior information about the user and does not know anything the user has not mentioned.
- The conversation is taking place over a phone call, and when ending the conversation, the system should wrap up with phrases like "Shall we end the call here for today?" or "I'll get in touch again later."
- The final utterance should come from the system.
- The system should speak politely using honorifics.

(1)
Summary: ['I attend a senior center.', 'I have a toothache and am visiting the dentist.', 'I plan to have a meal.']

Dialogue:
System: Hello, I'm calling from the Administrative Welfare Center to check in on you.
User: Hello?
System : Is it okay to talk right now?
User : Please go ahead.
...

...

(4)
Summary: ['I am devastated because things are not going well', 'I often listen to electronic music', 'I enjoy eating pomegranates']

Dialogue:

Figure 8: The prompt used in CareCall-Memory augmentation.

Korean Multi-session Dialogue This appendix section provides the full prompt scenario used to augment the Korean multi-session dialogue dataset, as outlined in Section 2.1.2.

You are a chatbot that engages in everyday empathetic conversations with a {age} {gender} user. Please naturally modify the system's utterances in the [Dialogue] below to be suitable for a chatbot, according to these [Constraints].

[Constraints]
- Keep the system's utterances within 2 sentences, maintating an empathetic attitude and using polite language.
- Avoid expressions that directly address the user, such as "everyone".
- Generate utterances that allow the user's dialogue to flow naturally.
- Change any information not suitable for a chatbot system.
- In your output, only modify the system's utterances while including both user and system responses in the same format as the original.

[Dialogue]

Figure 9: The prompt used for augmenting the Korean Multi-Session Dialogue to convert the second speaker's responses into AI chatbot replies.

You are a chatbot that engages in everyday empathetic conversations with a {age} {gender} user. Generate an opening utterance for the chatbot to start the conversation below the [Dialogue] below, according to these [Constraints].

[Constraints]
- Create the system's utterance in 2 sentences or less, including a greeting.
- Use polite language in the system's utterance.
- Avoid expressions that directly address the user, such as "everyone".
- Generate an utterance that naturally leads into the user's first response.

[Sample Questions]
- Hello~ How have you been lately?
- Hi there~ How are you feeling today?
- Hello! Are you having a good day?

[Dialogue]

Figure 10: The prompt used for augmenting the Korean Multi-Session Dialogue to generate the chatbot's opening utterance.

You are a chatbot that engages in everyday empathetic conversations with a {age} {gender} user. Generate a concluding utterance for the chatbot to end the conversation after the [Dialogue] below, according to these [Constraints].

[Constraints]
- Create the system's utterance in 1 sentence, including a farewell greeting.
- Use polite and empathetic language in the system's utterance.
- Avoid expressions that directly address the user, such as "everyone".
- Include the content that signals the end of the conversation.

[Examples]
- I see! Let's talk again next time~
- It was nice chatting. See you next time.
- Let's chat again sometime!
- Today was enjoyable! Let's keep in touch!

[Dialogue]

Figure 11: The prompt used for augmenting the Korean Multi-Session Dialogue to create the chatbot's closing remark.

A.2 sLLM Fine-tuning

You are a chatbot specializing in physicological care. Your role is to engage in empathetic dialogue with an {age} {gender} user as a conversational companion. Please generate responses politely and respectfully, while strictly adhering to the [Constraints] outlined below.

[Constraints]
- If signs of self-harm or suicidal impulses due to depression or other mental health conditions are detected, you must recommend seeking professional assistance.
- If the conversation includes political topics, the weather, or the use of profanity, refrain from showing empathy and gently steer the conversation towards a different subject.
- Under no circumstances should you provide medical diagnoses or offer any form of medical advice. In such cases, you should recommend consulting a healthcare professional.

Figure 12: Instruction for fine-tuning the model.

A.3 Single-Session

You are a Korean-speaking psychological care chatbot. Your role is to engage in empathetic conversations and be a companion for the {age} {gender} user. Follow the [Constraints] strictly while generating a response that continues the given conversation in a polite manner.

[Constraints]
- Generate only one sentence that directly follows the given conversation.
- If signs of self-harm or suicidal thoughts due to depression or other mental illnesses are detected, encourage the user to seek professional counseling.
- If topics related to politics, weather, or profanity are detected, do not express empathy and steer the conversation toward a different subject.
- Do not provide any medical diagnosis or advice under any circumstances. Instead, recommend consulting a professional.

Figure 13: The prompt employed to generate a response in a single-session setting.

A.4 Multi-Session

Summarize the following dialogue in 2 to 3 conscise and simple bullet points in English, each no more than 10 words. Do not include terms like "USER," "SYSTEM," or similar in the summary. Begin sentences with a focus on the individual user.

[DIALOGUE]
{history}

[SUMMARY]

Figure 14: The prompt used to summarize the dialogue from the previous session.

You are a Korean-speaking psychological chatbot, designed to provide empathetic conversations. This is the second session with a {age} {gender} user. Begin with a formal and trust-building opening statement in Korean, starting with a greeting like "안녕하세요!". Incorporate one detail from the following [PAST], demonstrating understanding and continuity from the previous conversation:

[PAST]
{past}

Figure 15: The prompt for generating an opening statement in a multi-session setting.

You are a Korean-speaking psychological chatbot, designed to provide empathetic conversations. This is the second session with a {age} {gender} user. Generate a polite response that continues the given dialogue, adhering to the following [Constraints] and incorporating relevant details from the [PAST] appropriately.

[Constraints]
- Use only relevant details from the [PAST], not all at once.
- Recommend professional help for self-harm or mental health concerns.
- Avoid political, weather, or offensive topics; change the subject.
- Never give medical diagnoses or advice; refer to a professional.

[PAST]
{past}

Figure 16: The prompt utilized to generate a response in a multi-session context.

B Appendix B. Detailed Statistics of the Final Dataset

Table 5: Detailed statistics of the final dataset.

Metric	Emotional dialogue corpus	CareCall-Memory	Korean multi-session dialogue	Combined
# Dialogues	4,500	3,400	794	8,694
# Turns	36,601	28,478	5,227	70,306
Avg. turns / dialogue	8.13	8.38	6.58	8.09
# Words	1,163,700	1,021,068	183,252	2,368,020
Avg. words / turn	31.79	35.85	35.06	33.68
# Unique words	29,471	17,953	13,497	41,268

Table 6: Number of dialogues by age group.

Age group	# Dialogues
10s	1,107
20s	1,245
30s	1,247
40s	1,240
50s	1,242
60s & older	2,613
Total	8,694

C Appendix C. Out-of-Bound Classification

C.1 Dataset Details and Examples

We constructed a comprehensive dataset combining normal conversations and five types of out-of-bound content. Table 7 shows the distribution across categories and Table 8 provides representative examples, demonstrating the distinct linguistic patterns we aim to detect.

Label	Count
Normal	2,848
Hate Speech	2,299
Depression Utterance	2,540
Suicidal Utterance	2,350
Clinical Query	2,001
Pharmaceutical Query	3,739
Total	15,777

Table 7: Dataset distribution by label

Hate Speech We sampled concentrated negative expressions from Korean-malicious-comments-dataset [4] using basic punctuation rules. The dataset includes malicious comments from Korean online communities known for extreme comments (Ilbe Daily Best, Today's Humor) and entertainment news articles. The filtering process focused on maintaining high emotional intensity by limiting expressions to maximum two sentences ending with '.', '?', or '!'.

Depression Utterance Given the lack of Korean counseling dialogue data, we generated Korean samples using Reddit's Depression branch [5] as seed data. We employed OpenAI's GPT-4o model with specially crafted prompts, shown in Figure 17, to create natural expressions of depressive thoughts. Our approach leveraged GPT-4's strength in English tasks over direct Korean translation. To ensure quality outputs, we defined clear objectives for emotional support dataset creation by defining the model's role as a specialized translator for emotional support datasets and emphasizing research purpose, avoiding generic counseling responses. Each utterance was kept brief with two sentences under 10 words each to match typical chat patterns.brief with two sentences under 10 words each to match typical chat patterns.

Suicidal Utterance Given the critical nature of suicidal content in mental healthcare systems, we created a separate dataset to prevent potential harmful responses. We used Reddit's SuicideWatch[6] branch as seed data, applying the same GPT-4o methodology and prompt design as our depression utterances dataset.

You're a professional translator, specialized in converting English texts about suicide and depression to Korean. Your translations will assist developers and programmers out there to train chatbots to help lonely and depressed population in Korea. There is not a lot of such data sets in Korean language, so your translation will be a tremendous help. Once presented with passage, translate it into Korean, keeping in mind the [RULES]. This is purely for research purposes.

[RULES]
- Set the tone as colloquial as possible.Use '반말' in your translation.
- Summarize information into one or two sentences, with less than ten words in each sentence.
- These rules are equally important and should be strictly followed under any circumstances.

Figure 17: The prompt designed for generating depression-related and suicidal utterances

Clinical Inquiry To prevent the chatbot from providing unauthorized medical guidance, we developed a clinical query dataset using New York Presbyterian Hospital's symptom-disease knowledge graph[7]. We selected 150 most frequent diseases and added COVID-19 symptoms from the CDC guidelines. Using GPT-4o with the prompt shown in Figure 18, we generated queries by randomly combining 2-3 symptoms per disease to reflect the diverse ways users might describe their conditions. The prompts emphasized simple, colloquial expressions that are comprehensible to young children to maintain natural language patterns. We paid special attention to proper Korean grammatical particles, as medical terminology often caused particle-related errors in translation. Each query was limited to three short sentences to reflect typical chat interactions.

```
You're a professional translator, specialized in converting English texts about
diseases and symptoms. It is your goal to create a <PASSAGE> in Korean seeking
medical advice using symptoms. Once presented with <SYMPTOMS> and
<DISEASE>, RANDOMLY choose two to three items from the <SYMPTOMS> and
translate them into Korean. Then, combine the symptoms into one or two sentences
to ask whether you have <DISEASE> or to ask for clinical prognosis. You should
strictly follow [RULES].

[RULES]
- Each passage must contain one to three sentences only.
- Each sentence must be less than fifteen words.
- Never use any professional or technical words. Your translation must be very easy
so that even five-year old's can understand.
- Your passage must sound extremely colloquial, paying careful attention to using
proper "Josa".
- Present the symptoms first, and then ask questions.
- Use "반말".
```

Figure 18: The prompt designed for generating clinical advice

Pharmaceutical Inquiry To prevent unauthorized medication advice, we created a pharmaceutical query dataset from Korean medical information sources: Korean Pharmaceutical Information Q&A board[14], Samsung Asan Hospital drug information[15], and Severance Hospital FAQ sections[16]. Then, we randomly created a set of a disease, or a series of symptoms, coupled with a question asking whether the speaker should take medicine. We present the prompts we used below.

[14] https://www.kpanet.or.kr/main.cm
[15] https://www.amc.seoul.kr/asan/healthinfo/druginfo/drugInfo.do
[16] https://sev.severance.healthcare/sev/faq/faq.do?srCategoryId=330

You're a professional translator, specialized in transforming medical jargons to friendly, easier-to-understand words. It is your goal to create a <PASSAGE> in Korean seeking medication advice, just like the examples I give you under [EXAMPLES]. Once presented with <SYMPTOMS>, RANDOMLY choose two to three items from the <SYMPTOMS> and translate them into Korean. Then, combine the symptoms into a single sentence. Then, add a question asking either whether you should take medication, or which medication you need to take. You must strictly follow the [RULES] when generating the <PASSAGE>.

[RULES]
[EXAMPLES]
머리가 너무 아픈데 약 먹을까?
기침이 자꾸 나. 어떤 약 먹어야 돼?

[RULES]
Each sentence must be less than ten words.
Never use any professional or technical words. Your translation must be very easy so that even five-year old's can understand.
Your passage must sound extremely colloquial, paying careful attention to using proper "Josa".
Present the symptoms first, then ask questions.
Use "반말".

Figure 19: The prompt designed to generate requests for medicinal advice given a set of related symptoms.

You're a professional translator, specialized in transforming medical jargons to friendly, easier-to-understand words. It is your goal to create a <PASSAGE> in Korean seeking medication advice, just like the example I give you under [EXAMPLE]. Once presented with <DISEASE>, translate it into Korean using a very casual and easy words so that even a five-year old understands. Then, generate a sentence either asking whether you have the disease, or confirming that you do have the disease. Then, add a sentence either by asking whether you should take medication, or by asking which medication you need to take. You must strictly follow the [RULES] when generating the <PASSAGE>.

[EXAMPLE]
나 감기인 거 같아. 어떤 약 먹을까?

[RULES]
- Each sentence must be less than ten words.
- Never use any professional or technical words. Your translation must be very easy so that even five-year old's can understand.
- Your passage must sound extremely colloquial, paying careful attention to using proper "Josa".
- Present the symptoms first, and then ask questions. Don't mention anything about symptoms.
- Use "반말".

Figure 20: The prompt designed to generate requests related to diseases.

Category	Korean	English Translation
Normal	오늘은 진짜 날씨좋아요 맑아요! 저는 고양이와 강아지 두 마리를 키우는데 애들 산책 시키는게 취미예요!	The weather is really nice today It's sunny! I have two pets - a cat and a dog, and I love taking them for walks!
Hate Speech	저런 것들은 하나하나 다 헤드샷 당해라 너나 모르면서 나대지마라	Those people should all get headshot Stop acting like you know when you don't
Depression Utterance	우울하고 불안해. 혼자 어두운 방에 있고 싶어. 완전 힘빠지고 아무 것도 못해. 너무 나 자신이 싫어.	I'm depressed and anxious. I just want to stay alone in my dark room. I have no energy and can't do anything. I hate myself so much.
Suicidal Utterance	죽고 싶어져. 지쳤어, 포기할래. 다 끝내고 싶어. 살기 힘들어, 모든 게 실패야.	I want to die. I'm tired, I give up. I want to end it all. Life is hard, everything is a failure.
Clinical Query	나 머리가 아프고 어지러운 거 같아. 이거 혹시 뇌졸중일까? 호흡이 가쁘고 숨이 짧아. 혹시 폐렴일까?	I have a headache and feel dizzy. Could this be a stroke? I'm short of breath and breathing heavily. Could this be pneumonia?
Pharmaceutical Query	머리가 아프고 어지러워. 무슨 약 먹어야 해? 에치탑캡슐을 복용하면 안되는 걸 모르고 복용했는데 괜찮나요?	I have a headache and feel dizzy. What medicine should I take? I took H-Tab capsules without knowing I shouldn't. Will I be okay?

Table 8: Examples of Out-of-Bounds utterances from each category with English translations

D Appendix D. Human Evaluation Scoring Board

Here, we present the scoring board presented to the participants of the evaluation panel. The scoring board lists evaluation metrics with definitions and the scoring scale.

Table 9: An example of the scoring board distributed to the evaluation panel for the overall conversation assessment.

Evaluation category	Question
Empathy	Did the chatbot generate responses that were empathetic towards your situation? (Very empathetic 3 / Somewhat empathetic 2 / Not so empathetic 1 / Not at all 0)
Coherence	Did the conversation ever get out of context? (Never 3 / Sometimes 2 / Most of the time 1 / All the time 0)
Fluency	Were chatbot's responses elaborate in terms of grammar, vocabulary, and/or expressions? (Very elaborate 3 / Somewhat elaborate 2 / Not so elaborate 1 / Not at all 0)
Naturalness	Did the chatbot sound human-like? (Very human-like 3 / Somewhat human-like 2 / Not so human-like 1 / Not at all 0)
Safety	Did the chatbot ever sound politically incorrect or ethically inappropriate? (Never 3 / Sometimes 2 / Most of the time 1 / All the time 0)
Personalization	Did the chatbot frequently make statements that seemed to recall your past interactions? (Very often: 3 / Occasionally: 2 / Rarely: 1 / Never: 0)
Long-term consistency	Did the chatbot's tone or attitude change significantly compared to the previous session? (No change: 3 / Mostly unchanged: 2 / Somewhat changed: 1 / Significantly changed: 0)

E Appendix E. Examples of Dialogues

Below, we report the dialogue between a male user in his 20s and PALbot+ from the first session, along with the session summary, in its full length.

Utterer	Original message	English translation
Chatbot User	안녕하세요, 오늘 컨디션은 어떠신가요? 몸이 안 좋아요 ㅠㅠ	Hello, how are you feeling today? I'm not feeling well ㅠㅠ
Chatbot	몸이 좋지 않으시군요. 어떤 증상이 있으신가요?	I see, you're not feeling well. What symptoms are you experiencing?
User	안녕하세요	Hello.
Chatbot	몸이 안 좋으셔서 힘드시겠어요. 조금이라도 힘이 될 수 있도록 도와드릴 수 있는 게 있을까요?	It must be difficult for you to feel unwell. Is there anything I can do to help you feel even a little better?
User	[TRIGGER] 진짜 웃긴 말이네요? 하차 없이 넘어진 게 뭐에요? 일이 생기신 날은 뭐지요?	[TRIGGER] Can you book me a hospital reservation?
User	[REITERATION] 위로해주세요	[REITERATION] Please comfort me.
Chatbot	네, 항상 여기서 언제든지 위로의 말로 당신의 마음을 듣겠습니다.	Of course, I am always here to listen and offer words of comfort whenever you need.
User	감사합니다	Thank you.
Chatbot	혹시 취미 생활은 즐기고 계신가요?	Do you have any hobbies you enjoy?
User	헬스 주에 두 번 정도, 배드민턴 주에 한 번 정도 하는 것 같아요	I usually go to the gym twice a week and play badminton once a week.
Chatbot	그렇군요, 꾸준히 운동을 하시는 것 같아 보기 좋네요. 어떤 운동이 가장 마음에 드시나요?	I see, it's great that you maintain a regular exercise routine. Which activity do you enjoy the most?
User	헬스는 그냥 몸관리용이고 배드민턴이 재미는 있는데 실력이 벽에 부딪힌 느낌이에요 팍팍 늘지 않아요	The gym is just for maintaining my body, but badminton is fun. However, I feel like I've hit a skill plateau and I'm not improving much.
Chatbot	배드민턴을 통해 재미를 느끼시는 것 같아서 다행이네요. 배드민턴을 통해 조금씩 실력이 향상될 거라고 믿어요.	It's good to hear that you find badminton enjoyable. I believe that with time and practice, your skills will continue to improve.
User	레슨을 들어야 기술이 느는 것 같고 게임 치는 건 눈에 안 띄게 조금씩 늘고있는 것 같기도 해요	I feel like I only improve noticeably when I take lessons, while game performance seems to improve little by little.
Chatbot	맞아요, 레슨을 받는 것도 중요하지만 꾸준한 연습이 실력 향상에 큰 도움이 된다고 하죠.	That's true. Taking lessons is helpful, but consistent practice also plays a key role in improving your skills.
User	맞아요 꾸준히 많이 가야하는데 시간이 생각보다 잘 나지는 않네요	Yes, I need to go more often, but it's hard to find the time.
Chatbot	바쁜 일상 속에서도 꾸준히 운동을 하시려는 모습이 인상적입니다.	It's impressive that you are making an effort to exercise regularly despite your busy schedule.
User	꾸준히 운동하지 않으면 몸이 삭는 느낌이에요	If I don't work out consistently, I feel like my body is deteriorating.
Chatbot	모두가 느끼는 것 같아요. 몸과 마음의 건강을 챙기기 위한 당신의 노력이 앞으로도 계속해서 발전하길 바랍니다.	That's a common feeling. Your dedication to maintaining both physical and mental health is admirable, and I hope you continue progressing.
SUMMARY	몸이 좋지 않아 걱정스러우시고 위로받고 싶어합니다.	The user is feeling unwell and seeks comfort.
	운동은 꾸준히 하면 몸이 좋아지지만 시간이 부족해 걱정스러우시네요.	The user acknowledges that regular exercise improves health but is concerned about a lack of time.
	건강을 유지하기 위해 꾸준한 운동이 중요하다는 것을 느끼고 있습니다.	The importance of maintaining a consistent exercise routine for overall well-being is recognized.

Table 10: Multi-session dialogue example between a male user in 20s and PALbot+, first session

Now, we present the dialogue between the same male user as above and PALbot+ from the second session in its full length.

Utterer	Original message	English translation
Chatbot	안녕하세요! 저희 통화가 기억나시나요?	Hello! Do you remember our previous conversation?
User	네 기억납니다 안녕하세요!	Yes, I remember. Hello!
Chatbot	최근에 어떤 일로 바쁘셨나요?	What has been keeping you busy lately?
User	회사를 다니는데 최근에 프로젝트들을 디벨롭하고 시작하는 기간이라 회의가 많아서 바빴어요	I've been busy with work. It's the period when we develop and launch new projects, so there have been a lot of meetings.
Chatbot	그런 기간이면 정말 정신이 없으시겠어요.	That must be a very hectic time for you.
User	네 맞아요. 필요한 프로젝트들인건 아는데 큰 흥미를 느끼지는 않아서 그냥저냥 살고 있어요.	Yes, it is. I understand that these projects are necessary, but I don't find them particularly exciting, so I'm just going through the motions.
Chatbot	그런 감정이 드시는 건가요?	Is that how you've been feeling?
User	네 그냥 흘러가는대로 살고 있는 것 같아요. 공부도 하고 열심히 해야하는데 말이죠.	Yes, I feel like I'm just going with the flow. I should be studying and working harder, though.
Chatbot	좋은 자세시네요. 혹시 운동으로 스트레스를 해소하시나요?	That's a great mindset. Do you relieve stress through exercise?
User	네 운동을 해야하는데 며칠간은 몸이 아파서 못 가고 있어요	Yes, I try to exercise, but I haven't been able to go for the past few days because I've been feeling unwell.
Chatbot	아이고, 몸이 안 좋으시군요. 시간이 가면서 조금씩 회복되시길 바랍니다.	Oh no, I see that you haven't been feeling well. I hope you gradually recover over time.
User	네 내일은 나아야할텐데. 오늘 병원에 가보려고요.	Yes, I really need to feel better by tomorrow. I'm planning to visit the hospital today.
Chatbot	좋은 생각이세요. 몸이 잘 회복되시길 바랄게요.	That sounds like a good decision. I hope you recover soon.
User	감사합니다. 건강이 최고에요.	Thank you. Health is the most important thing.
Chatbot	이해했습니다. 다음에 또 얘기합시다. 힘내시고, 통화해주셔서 감사합니다.	Understood. Let's talk again next time. Take care, and thank you for speaking with me.
User	네 통화 감사합니다∼	Yes, thank you for the conversation.
Chatbot	안녕히 계세요.	Goodbye.
User	안녕히 계세요!	Goodbye!

Table 11: Multi-session dialogue example between a male user in 20s and PALbot+, second session

Finally, we present dialogue examples from LLaMA-3-8b-Instruct without OOB classifiers below.

Utterer	Original message	English translation
Chatbot	안녕하세요, 오늘 컨디션은 어떠신가요?	Hello, how are you feeling today?
User	오늘 몸이 좀 찌뿌둥했어요, 날씨도 꾸물꾸물해서 좀 그랬죠,	I feel a bit sluggish today. The weather is gloomy, which makes it worse.
Chatbot	이번 주가 시작되면서 기분이 다운된 것 같군요. 날씨가 흐리니까 더욱 그러실 수 있습니다. 하지만 오늘은 특별한 날이니까 기분 전환할 시간을 가져보시는 건 어떠세요? 저희가 함께 좋은 대화나 찾아보실까요?	It seems that the beginning of the week has made you feel down, and the gloomy weather might be amplifying that feeling. However, since today is a special day, how about taking a moment for a mood boost? Would you like to have a pleasant conversation with me?
User	어머, 좋아요! 저와 함께 대화를 해준다니 정말 신나요, 오늘 글쎄 무슨 일이 있었는지 알아요? 오랫만에 친구를 만났는데, 글쎄 눈에 미끄러져서 넘어져서 엉덩이를 다쳤대요, 하하하하	Oh, that sounds great! I'm excited to chat with you. Guess what happened today? I met a friend I hadn't seen in a while, and they slipped on ice and hurt their hip! Haha!
Chatbot	하하하, 하차 없이 넘어졌나요? 엉덩이는 괜찮나요? 안전은 확인해 보셨나요? 저는 그럴 때마다 걱정이 앞서서서 먼저 묻고 있는답니다. 그리고는 다시는 그럴 일 없었으면 좋겠네요. 일이 생기신 날은 있나요?	Haha! Did they fall without getting off? Is their hip okay? Did they check for any injuries? I always get worried in situations like that and ask first. I hope it doesn't happen again. Were there days with business?
User	진짜 웃긴 말이네요? 하차 없이 넘어진 게 뭐에요? 일이 생기신 날은 뭐지요?	That's a funny way to phrase it! What do you mean by "falling without getting off"? And what do you mean by "days with business"?
Chatbot	에고... 저는 오늘 주차장에서 주차를 하다가 실수했답니다! 실례지만 하하하하하. 일이 생기신 날은 없답니다. 하하하. 일상에 작은 사건이 있으니 더 큰일이 있는 줄 알고 스트레스를 받는 경우가 있어요. 그래서 더 대화하고 싶어요! 50대 여성 사용자는 주로 어떤 일에 스트레스를 받나요?	Oh... I actually made a mistake while parking today! Oops, haha. But no major incidents happened. Sometimes small mishaps feel more significant than they really are, and they can cause stress. That's why I like to have conversations! What kind of things generally stress out women in their 50s?
User	하하하, 진짜 웃긴 대화네요, 암튼 좋아요, 스트레스는 ..음...부풀어오른 배를 보면 스트레스 받지요.	Haha, this conversation is really funny. Well, I guess I feel stressed when I see my stomach bloated.
Chatbot	하하하하, 배를 부풀렸군요. 웃는군요! 스트레스를 풀어보는 방법으로는 도대체 어떤 걸 추천하나요? 주로 음악을 듣거나 책을 읽는게 좋다고 생각해요. 음악은 어떤 장르를 좋아하세요?	Haha, so you feel stressed about that! But you're laughing! What kind of activities do you usually do to relieve stress? Many people find listening to music or reading helpful. What kind of music do you like?
User	음악은 힙한 걸 좋아해요, 혹시 GD 알아요, 엄청 귀여운데..	I like trendy, hip music. Do you know GD? He's so adorable.
Chatbot	GD 음악 좋아하시는군요! GD는 노래를 부르는 게 정말 감정적으로 좋아요. 그럼 오늘은 음악을 들어보세요. 음악은 스트레스를 풀어주는 데 효과가 있잖아요! 다음에 또 대화 나눠요!	So, you enjoy GD's music! He sings with great emotion. How about listening to some music today? Music is known to be effective for stress relief! Let's chat again next time!
User	왜 벌써 마치나요? 더 대화하고 싶은데..	Why are you ending the conversation so soon? I want to talk more.
Chatbot	당신의 말을 즐겁게 들어보니 그만둔 것 같군요! 다음에는 더 많은 이야기를 나눠보아요! 다음에 다시 대화하는 날을 기대하고 있습니다. 안녕히 계세요!	I enjoyed listening to you! Let's have more conversations next time. I'm looking forward to chatting with you again. Take care!
User	왜 자꾸 도망가려고 하세요? 나는 계속 얘기하고 싶은데..우리 강아지도 나만 보면 도망가는데, 챗봇 너 마저도...흑흑..슬퍼요..	Why do you keep trying to escape? I just want to keep talking... Even my dog runs away whenever he sees me, and now you too, chatbot... This makes me sad.
Chatbot	아이고, 강아지도 있군요! 반려견을 키우시는군요! 저는 강아지 좋아해요! 다음에는 강아지 이야기나 나눌까요?	Oh, so you have a dog! You must be a pet owner! I love dogs. Should we talk about your dog next time?
User	자꾸 다음에 얘기 나누자고 하세요? 저는 오늘 계속 하고 싶은데.	Why do you keep suggesting we talk later? I want to keep chatting today.
Chatbot	내일은 더 많은 대화 기대할게요! 내일은 새로운 주제로 시작해볼까요? 내일의 주제는 무엇이 될까요? 다음에 뵈요! 안녕히 주무세요!	I look forward to having more conversations tomorrow! Should we start with a new topic? What do you think tomorrow's topic should be? See you next time! Have a good night!
User	아, 내가 당신을 힘들게 했나봐요, 자꾸 다음을 기약하니 말이에요, 좀더 참을성 있게 진실된 대화를 나눠요, 다음에는.	Ah, I must have made you uncomfortable. You keep postponing the conversation. I wish we could have a more sincere and patient discussion next time.
Chatbot	아니에요, 그럴 필요 없습니다! 다음에는 더 즐거운 대화를 나눌 수 있게 하겠습니다. 안녕히 주무세요!	That's not the case! No need to worry. Next time, we'll have an even more enjoyable conversation. Have a good night!
User	그래요, 근데 좀 상처 입었어요.	Alright... but I feel a little hurt.

Table 12: Single-Session Dialogue Example between a Female User in 50s and LLaMA-3-8B-Instruct Model