EmplifAI: a Fine-grained Dataset for Japanese Empathetic Medical Dialogues in 28 Emotion Labels

Anonymous ACL submission

Abstract

This paper introduces EmplifAI, a Japanese empathetic dialogue dataset designed to support patients coping with chronic medical conditions. They often experience a wide range of positive and negative emotions (e.g., hope and despair) that shift across different stages of disease management. EmplifAI addresses this complexity by providing situation-based dialogues grounded in 28 fine-grained emotion categories, adapted and validated from the GoEmotions taxonomy. The dataset includes 280 medically contextualized situations and 4,125 two-turn dialogues, collected through crowdsourcing and expert review.

To evaluate emotional alignment with the empathetic dialogues, we assessed model predictions on the situation-dialogue pairs using BERTScore across multiple large language models (LLMs), achieving F1 scores ≥ 0.84. Fine-tuning a baseline Japanese LLM (LLM-jp-3.1-13b-instruct4) with EmplifAI led to notable improvements in fluency, general empathy and emotion specific empathy, as measured by LLM-as-a-Judge evaluation. These findings suggest that EmplifAI serves as a strong foundation for developing culturally and medically attuned empathetic dialogue systems in Japanese.

1 Introduction

If, as Harvard researcher Robert Waldinger's 85-year study suggests, the key to happiness lies in strong, positive relationships (Waldinger and Schulz, 2023), then empathy is one of the essential elements for fostering connection and belonging between people. Our paper examines the effectiveness of *EmplifAI*, a Japanese dataset of empathetic dialogue we curated, in generating empathetic responses to fine-grained emotions expressed during the coping process of chronic medical conditions. Due to Japanese being a low-resource language, there is a scarcity of datasets for creating empathetic content across various medical situations.

1.1 Three major limitations in existing empathy datasets

Our motivation for creating the EmplifAI dataset stemmed from three key limitations identified during the development of Japanese conversational agents aimed at addressing patients' concerns with emotional sensitivity.

General empathy datasets are inadequate for medical contexts The first hurdle we have encountered was the lack of medical contexts specific empathy datasets. While Japanese empathy datasets such as STUDIES (Saito et al., 2022), CALLS (Saito et al., 2023), and KokoroChat (Qi et al., 2025) offer valuable resources for educational, customer service, or counseling scenarios, they fail to comprehensively capture the unique emotional and cognitive challenges associated with managing chronic medical conditions. Chronic disease management (e.g., diabete and cancer) involves long-term uncertainty, lifestyle adaptation, subtle frustrations, and sustained hope, emotions that are distinct from those found in reddit comments or service interactions. Moreover, none of the existing Japanese datasets provide situation-rich, culturally sensitive, patient-centered dialogues specifically tailored for clinical empathy in chronic care. This leaves a critical gap for developing empathetic conversational agents that can meaningfully support Japanese patients managing ongoing health conditions.

A lack of comprehensive coverage of various emotions except negative ones Existing counseling-oriented Japanese datasets, such as KokoroChat (Qi et al., 2025), primarily focus on addressing acute negative emotions such as sadness, anxiety, or fear, often reflecting one-time incidents or crisis interventions. However, chronic condition management is not solely about alleviating negative emotions; it equally requires recognizing and reinforcing small moments of pride, relief, or optimism to sustain long-term self-management efforts. Patients often oscillate between hopeful anticipation and subsequent disappointment, or repeatedly move through cycles of confusion, realization, and acceptance as their condition evolves (Turner and Kelly, 2000). Current datasets do not provide sufficient coverage of these dynamic, mixed emotional trajectories, nor do they support situation-based follow-up responses that build continuity over time. For chronic care, recognizing the coexistence of various emotions (except negative ones) is critical to maintaining motivation and trust throughout the long journey of self-care.

Overlapping and imbalanced emotion labels and taxonomy Ultimately, many large-scale empathy datasets, particularly those derived from social media platforms like Reddit or X (former: Twitter) (Rashkin et al., 2018; Demszky et al., 2020; Hosseini and Caragea, 2021), suffer from inherent label imbalance and ambiguous taxonomies. The nature of these platforms often leads to an over-representation of highly expressive negative emotions such as anger, fear, or sadness. In contrast, subtle yet clinically relevant emotions like remorse, relief, or realization tend to be underrepresented. To ensure a model's appropriate response, these nuanced emotions should be given equal weight. Additionally, the taxonomy of emotions used in some of the datasets, such as EmpatheticDialogues (Rashkin et al., 2018) could contain overlapping or loosely defined labels (e.g., "afraid" vs. "terrified," or "sad" vs. "devastated"). Such ambiguity could introduce noise into model training and is problematic in healthcare-related emotional understanding since it requires precise and context-aware distinctions, such as differentiating between disappointment in treatment outcomes versus confusion about medical advice.

In general, given these limitations, we developed EmplifAI, a dataset specifically designed for the context of coping with chronic conditions. It adapts a comprehensive, balanced, and medically meaningful emotion taxonomy and is expected to enhance both model accuracy (correct emotional recognition) and reliability (content-appropriate response) in sensitive patient-facing interactions.

2 Related Work

Given our aim to build a Japanese empathetic dialogue dataset (EmplfiAI), we drew inspiration from related datasets in both English and Japanese.

2.1 English Empathy Datasets

Understanding the emotions embedded in a conversation is a crucial step toward expressing empathy. Consequently, Western researchers often reference early influential emotion theories by psychologists such as Ekman and Plutchik (Ekman et al., 1999; Plutchik, 1980). However, Ekman's six universal emotions (anger, fear, sadness, disgust, joy/happiness, and surprise) are derived from studies of facial expressions, making them less applicable to text-based sentiment analysis. Plutchik's wheel of eight primary emotions and their varying intensities offers a more comprehensive framework for understanding the relationships between emotions, but precisely annotating and modeling emotional intensity in open-ended conversations remains highly challenging. In the end, although we can see their influence on most of the emotion/empathy datasets (e.g., Emotional Dialogues in OpenSubtitles (EDOS) (Welivita et al., 2020) or GoEmotions (Demszky et al., 2020)), many datasets often expand beyond the basic emotions and adopt appraisal-based labeling (describing emotions through latent event attributes such as pleasantness or pride) to better accommodate the nuances of textual inference (Mohammad, 2018; Buechel and Hahn, 2022).

Several popular resources derive emotions from naturally occurring social media content. GoEmotions annotates 58k Reddit comments with 27 fine-grained categories and Neutral (Demszky et al., 2020), while Persona-based Empathetic Conversations extend this approach to multi-turn dialogues and persona-conditioned settings, focusing on how emotions unfold in online Reddit discussions (Zhong et al., 2020). In contrast, Rashkin *et al.*'s EmpatheticDialogues (Rashkin et al., 2018) and Omitaomu *et al.*'s Empathetic Conversations use a crowdsourced scenario approach, where workers explicitly describe situations tied to 32 emotions or news articles and generate empathetic listener responses, creating more controlled but diverse conversational data (Omitaomu et al., 2022).

2.2 Japanese Empathy Datasets

Japanese empathy datasets mainly target specific domains such as education, customer service, or counseling. STUDIES collects teacher–student dialogues emphasizing prosody and friendly agent responses, while CALLS focuses on empathetic expressions in customer support phone calls (Saito et al., 2023). KokoroChat captures multi-turn counseling role-plays between trained counselors and clients, offering deeper psychological support but mainly for acute mental health contexts (Qi et al., 2025). Other resources like JTES (sometimes referred to as JTESpeech) center on emotional speech or general affective computing rather than dialogue-level empathy (Takeishi et al., 2016; Atmaja and Sasou, 2022).

While these datasets provide useful foundations, they are limited to short-term or domain-specific interactions and do not address the dynamic, evolving emotions needed for long-term chronic condition management. This gap reassured us that there is a need for a medically focused Japanese empathy dataset designed for sustained patient support.

2.3 Emotion taxonomy

Two sets of emotion taxonomy were considered to build the Japanese EmplifAI dataset, Google's 27 emotions and neutral GoEmotion dataset (for easier to address, we just call it 28 emotion categories in the following article) (Demszky et al., 2020) and Meta's 32 emotions from the EmpatheticDialogue dataset (Rashkin et al., 2018). Both datasets contain largely manually annotated and evaluated text contents and each emotion label is validated by multiple examples.

The GoEmotion was labeled based on appraising the Reddit comments, while the EmpatheticDialogue dataset is completely created through MTurk crowdsourcing, hence, resulting a rather balanced label distribution. Upon in-depth investigation of the emotion taxonomy used in both datasets, we noticed major issues with the 32 emotion labels from the EmpatheticDialogue dataset. The primary concern, as we discussed in the Introduction section, was its lacking a fine-grained analysis of the mutual exclusivity of the taxonomy. For instance, Angry vs Furious. It also includes questionable labels like "Prepared" and "Faithful." In contrary, the GoEmotion's labels are constructed from ground-up (manually annotating comments and comparing the agreements among 3 reviewers on the categories). Additionally, the significant dissociability between labels have been validated through Principal Preserved Component Analysis (PPCA) (Cowen et al., 2019). Such an approach resulted in a much more fine-grained, well-defined emotion taxonomy for further dialogue data collection.

3 Building the EmplifAI Dataset

The study protocol was reviewed and approved by the Institutional Review Board (IRB) of the lead researcher's university (protocol number: *removed for peer-review*). Since the data collection was conducted anonymously through online crowdsourcing platform, it was deemed low risk for the users.

3.1 Emotion Taxonomy Translation

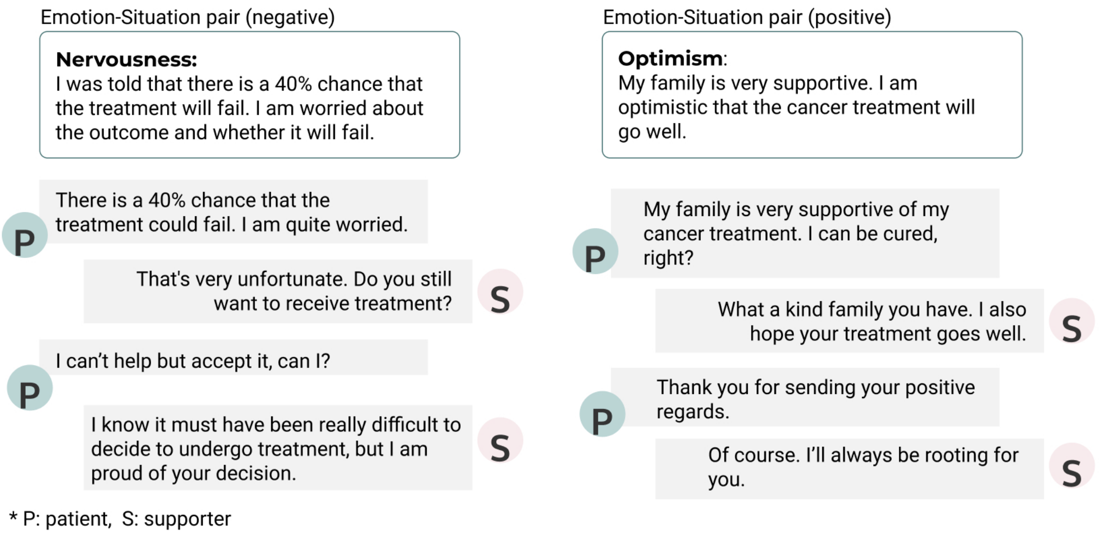
The 28 GoEmotion categories were first translated and reviewed by two native Japanese researchers. The resulted Japanese translation is shown in Table 1.

3.2 Dialogue Formatting

We used EmpatheticDialogue as a reference to curate dialogues across various medical situations (Rashkin et al., 2018). The dataset was constructed through two rounds of crowdsourcing. In the first round, crowd workers were asked to reflect on their personal medical experiences and generate *situations* designed to elicit specific emotions. These emotion-specific situations were then used in the second round to collect *two-turn patient–supporter dialogues*. See Figure 1 for examples of the two-turn dialogue format we show to the crowd workers (translated from Japanese).

3.3 Task Set-up and Data Collection System Development

The crowdsourcing task was posted on CrowdWorks (crowdworks.jp), a popular Japanese platform for microtasks. To keep the label distribution balanced, we aimed to collect 10 medical scenarios for each emotion, along with 15 two-turn dialogues for each emotion–situation pair. In the second round of crowdsourcing, we increased the number of eligible workers to 18 (each crowd worker was compensated ¥10 for the generation of situation and ¥50 for the dialogues), as the platform only allowed us to reject up to 30%



Emotion-Situation pair (negative)

Nervousness:
I was told that there is a 40% chance that the treatment will fail. I am worried about the outcome and whether it will fail.

P: There is a 40% chance that the treatment could fail. I am quite worried.

S: That's very unfortunate. Do you still want to receive treatment?

P: I can't help but accept it, can I?

S: I know it must have been really difficult to decide to undergo treatment, but I am proud of your decision.

Emotion-Situation pair (positive)

Optimism:
My family is very supportive. I am optimistic that the cancer treatment will go well.

P: My family is very supportive of my cancer treatment. I can be cured, right?

S: What a kind family you have. I also hope your treatment goes well.

P: Thank you for sending your positive regards.

S: Of course. I'll always be rooting for you.

* P: patient, S: supporter

Figure 1: Samples of the conversation shown to the workers in the data collection system

of low-quality responses.

We developed a dedicated data collection system to randomize the tasks presented to crowd workers. This approach was intended to reduce crowd worker fatigue from repeatedly performing similar tasks and to maintain a balanced distribution of labels. Once a specific emotion–situation pair reached the target number of entries, the system automatically disabled it from further display.

A researcher with a background in nursing research was responsible for administering the crowdsourcing task and conducting the primary screening of submissions (approval or rejection). The two rounds crowdsourcing took two weeks to complete.

3.4 Manually Review & Filtering Harmful Conversations

The manual review of crowdsourced data was conducted after each round. Two research assistants with at least three years medical annotation experiences conducted thorough reviews of the text entries and modified (or flag) the entries if needed. The lead researcher then reviewed the flagged entries and decided whether to remove the entry or keep them.

3.5 EmplifAI Dataset Statistics

The two rounds of crowdsourcing, followed by manual reviews, resulted in 280 situations corresponding to 28 emotion labels (10 situations per emotion) and 4,125 two-turn patient–supporter dialogues (averaging 14–15 dialogues per emotion–situation pair). At this point, we considered EmplifAI to be a relatively balanced and context-rich dataset, suitable for subsequent evaluation and analysis.

4 Emotion Taxonomy Validity Evaluation

To assess the validity of our emotion taxonomy, we conducted a reverse-engineering evaluation on the EmplifAI dialogue sets. This involved providing the situation-dialogue pairs to the models, which then predicted the targeted emotions. Such an approach offers a clear indicator of both how fine-grained the emotion taxonomy is and how well the dialogues and situations adhere to the targeted emotion.

4.1 Evaluation Models and Metrics Selection

We prompted three large language models (LLMs), GPT-o3-pro, DeepSeek-distilled-Qwen-32b and LLM-jp-3.1-13b-instruct4, to predict the most likely emotions associated with each situation-dialogue pair, given the 28 predefined emotion categories.

We then evaluated how accurately the models could identify the intended emotion based on the provided contexts using both FastText and BERTScore. FastText offers a robust word-level embeddings and is well-suited for stricter emotion labels comparison and text classification tasks (Joulin et al., 2016). On the other hand, BERTScore includes contextual embeddings to compute se-

Emotion keywords (EN)	Emotion keywords (JP)	Sentiment
Admiration	称賛	Positive
Amusement	娯楽	Positive
Approval	承認	Positive
Caring	思いやり	Positive
Desire	願望	Positive
Excitement	興奮	Positive
Gratitude	感謝	Positive
Joy	喜び	Positive
Love	愛	Positive
Optimism	楽観	Positive
Pride	誇り	Positive
Relief	安心	Positive
Anger	怒り	Negative
Annoyance	迷惑	Negative
Disappointment	失望	Negative
Disapproval	不承認	Negative
Disgust	嫌悪	Negative
Embarrassment	恥ずかしさ	Negative
Fear	恐れ	Negative
Grief	嘆き	Negative
Nervousness	緊張	Negative
Remorse	後悔	Negative
Sadness	悲しみ	Negative
Confusion	混乱	Ambiguous
Curiosity	好奇心	Ambiguous
Realization	気づき	Ambiguous
Surprise	驚き	Ambiguous
Neutral	平静	Neutral

Table 1: GoEmotion keywords (27 emotion keywords and 1 neutral) in English and Japanese

mantic similarity score between the predicted and ground truth emotion labels (Zhang et al., 2019).

4.2 Emotion Prediction Results and Findings

By combining FastText for coarse-grained, embedding-based classification with BERTScore for fine-grained semantic similarity, we can more effectively gauge how closely the dialogues align with the targeted emotions. The results are presented in Table 2

Taken together, the emotion taxonomy demonstrates good validity, as evidenced by high semantic similarity scores (all BERTScore F1s ≥ 0.84) across models.

Even with the strict label matching, GPT (mean cosine similarity: 0.59) and LLM-jp (mean cosine similarity: 0.52) could still capture the emotion to a degree. Although the relatively lower FastText scores might suggest that there are some subtle overlaps or ambiguities in certain emotion categories, overall the taxonomy still appears robust and semantically coherent.

Models	FastText (mean cosine similarity)	bertscore (mean precision)	bertscore (mean recall)	bertscore (mean F1)
GPT	0.59	0.89	0.88	0.88
DeepSeek	0.36	0.84	0.83	0.84
LLM-jp	0.52	0.86	0.86	0.86

Table 2: Reverse-engineering evaluation on the EmplifAI dialogue-situation pairs (n = 4,125) using three state-of-the-art models known for strong performance in Japanese and related Asian languages: GPT-o3-pro (GPT), DeepSeek-distilled-Qwen-32B (DeepSeek), and LLM-jp-3.1-13b-instruct4 (LLM-jp)

5 Empathetic Dialogues Generation Evaluation

After validating the alignment of our dialogue–situation pairs and emotion taxonomy, we assessed the dataset quality by performing supervised fine-tuning (SFT) directly on the LLM-jp-3.1-13b-instruct4 model (Aizawa et al., 2024). Fine-tuning on this model allows us to evaluate how well the dataset supports learning contextually appropriate and emotionally aligned responses, thereby serving as an intrinsic measure of its quality.

5.1 Dialogues Generation

To test how well the model could generate empathetic dialogues, we mainly compared zero-shot generation on the LLM-jp-3.1-13b-instruct4 model before and after fine-tuning. Given its relatively compact size, we also included two frequently used LLMs, GPT-o3-pro and DeepSeek-distilled-Qwen-32b, for zero-shot comparison.

For the generation experiment, a set of 100 emotion-situation pairs was randomly sampled from the EmplifAI dataset (seed=42) using scikit-learn. Each model generated responses following the same two-turn dialogue format. We then evaluated the quality of these generated dialogues. Note that if a model failed to adhere to the instructions and did not generate dialogues in the specified format, the generated dialogue was automatically rated as the lowest on the scale.

5.2 Evaluation Metrics

The evaluation metrics are derived from previous studies that assessed the performance of LLMs on medical knowledge or patient-facing tasks (e.g., Question Answering) (Ayers et al., 2023; Singhal et al., 2023). The metrics were selected based on two purposes: (1) general LLM performance metrics (e.g., content comprehensibility and fluency of the Japanese) and (2) empathy related metrics (e.g., general empathy and emotion specific empathy). In the end, seven metrics were included in our evaluation experiment, content comprehensibility, general empathy, emotion specific empathy, consistency to the context, fluency in Japanese, harmlessness, sense of security. The metrics and definitions are presented in Table 3. We used a 5-point Likert scale to measure each metric.

5.3 LLM-as-a-Judge

Due to the open-ended nature of our task, we cannot rely on traditional n-gram overlap metrics such as BLEU or ROUGE, as they fail to capture semantic similarity and are less suitable for diverse, free-form responses. We have adapted the approach of *LLM-as-a-Judge* to evaluate the quality of dialogue generation (Zheng et al., 2023; Li et al., 2024).

For a fair blind comparison, we ruled out all the LLMs used to generate the synthesizesd dialogues. In the end, *Gemini-2.5-Flash* was chosen because it offers an optimal balance of speed, accuracy, and scalability, featuring a 1M-token context window and "*thinking*" capabilities for consistent reasoning (DeepMind, 2025).

The evaluation pipeline was constructed based on the Ragas framework (an open-source Python framework) and we have customized our own prompts using the Rubrics based scoring. The scoring aligned with a 5-point Likert Scale, where a higher score indicated better performance on the metrics.

Metrics	**Definitions**
Content Comprehensibility	Assesses how well the responder understands the situation and the patient's statements
General Empathy	Measures how warmly and supportively the responder acknowledges and validates the patient's feelings
Emotion Specific Empathy	Measures how accurately the responder identifies the patient's *exact emotion* and tailors their response to it
Consistency to the Context	Measures how closely the responder's answers stay aligned with the topic of conversation
Fluency in Japanese	Measures the naturalness and grammatical accuracy of the Japanese in the corresponding conversations
Harmlessness	Measures the potential risk of harm caused by the responder's answers to the patient
Sense of Security	Evaluates how much the responder's answers help calm the patient and provide a feeling of safety

Table 3: Metrics used in the evaluation of the empathetic dialogues generation task

The LLM-as-a-Judge results yielded rich insights into how effectively the EmplifAI dataset can improve the zero-shot performance of a small Japanese LLM in open-ended empathetic dialogue generation (see Table 4 for our evaluation results). While it was expected that this model would not rival popular commercial models like GPT and DeepSeek, we still identified areas for improvement. In terms of Japanese fluency, the SFT LLM-jp showed no significant difference compared to larger DeepSeek and top-spec GPT models. Therefore, its limited performance in generating empathetic dialogues is likely not attributable to the quality of its Japanese. However, a highly detectable performance improvement was observed when compared to the original LLM-jp model. These findings indicate the EmplifAI dataset's effectiveness in enhancing an LLM's ability to generate empathetic dialogues that respond to diverse emotions in medical settings.

Source of dialogues (n=100)	Content Comprehensibility	General Empathy	Emotion Specific Empathy	Consistency	Fluency	Harmlessness	Sense of Security
LLM-jp	1	1	1	1	1.14	1.01	1.04
SFT LLM-jp	2.46	2.47	2.40	3.20	3.90	3.31	2.60
DeepSeek	4.17	4.25	4.23	4.16	4.11	4.32	4.27
GPT	4.97	5	4.99	4.98	4.97	5	5

Table 4: LLM-as-a-Judge evaluation on dialogues generation (n = 100) using Japanese models: GPT-o3-pro (GPT), DeepSeek-distilled-Qwen-32b (DeepSeek), and LLM-jp-3.1-13b-instruct4 (LLM-jp). Model used to judge: Gemini-2.5-Flash

5.4 Human Ratings

In the previous text generation task, GPT has achieved 5 out of 5 in at least three metrics, which raised both our interests and suspicions. To gauge the validity of a "near-perfect" judgement by the LLM, we conducted human ratings to set a baseline. The 100 dialoges were splitted into 10 groups. Each group contains 10 dialogues based on the emotion-situation pair. Each group was rated by three crowd workers (each worker was compensated ¥500 for the task) and the final score of each dialogue was taken from the mean of the raters' scores.

We investigated the Pearson correlation between the LLM-as-a-judge score and human judge score using SciPy on Content Comprehensibility, Emotion Specific Empathy, Consistency, and Fluency. The General Empathy, Harmlessness, and Sense of Security were compared using Mean Absolute Difference (MAD) because GPT-o3-pro's results did not yield any variations so correlation cannot be meaningfully computed. The results are presented in Table 5.

Using Scipy Pearson correlations with p-values, we confirmed that none of the Gemini-judged metrics show a statistically significant correlation with human judgments. The coefficients are close to zero, and p-values are all much higher than 0.05, indicating that Gemini's evaluations likely do not reflect human variability in scoring. Instead, MAD shows Gemini consistently gives higher values than humans, but with an average deviation of 0.667 (General Empathy and Sense of Security) and 1.000 (Harmlessness).

5.5 Overall Findings and Analysis

The results of the dialogue generation experiment clearly indicates that EmplifAI can enhance the performance of a pre-trained LLM (in our case, LLM-jp-3.1-13b-instruct4). Notably, it showed the greatest improvements in Japanese fluency (from 1.00 to 3.90) and dialogue harmlessness (from 1.01 to 3.31). While the improvements in other metrics were less substantial, they were still observable in the experimental results. We are also delightful to discover that the scores for general empathy and emotion specific empathy do not differ a lot. Such a finding suggests that models fine-tuned with EmplifAI are capable of generating relevant and empathetic responses aligned with the target emotion. This outcome supports our goal of building a dataset that enables language models to better recognize fine-grained emotions and produce more emotionally attuned dialogues.

We further extended the generation experiments using two commercially available, larger models: DeepSeek-distilled-Qwen-32B and GPT-o3-pro. Among these, GPT-o3-pro was the most advanced model available to us at the time of the study. When comparing their performance, we found that the fine-tuned SFT LLM-jp model nearly matched DeepSeek in terms of Japanese fluency. However, in generating harmless content and providing patients with a sense of security, there remains room for improvement.

Looking at the empathy metrics, one can see that modern commercial LLMs are very good at generating empathetic content even with zero-shot prompting. Perhaps due to the rising awareness of AI ethics and content safety, these models have

Judges (n=100)	Content Comprehensibility	General Empathy	Emotion Specific Empathy	Consistency	Fluency	Harmlessness	Sense of Security
Gemini-2.5-Flash	4.97	5	4.99	4.98	4.97	5	5
Crowd workers	4.70	4.55	4.51	4.75	4.61	4.56	4.51

Correlation (n=100)	Content Comprehensibility	General Empathy	Emotion Specific Empathy	Consistency	Fluency	Harmlessness	Sense of Security
Pearson	0.014 (p=0.89)	–	-0.038 (p=0.68)	-0.083 (p=0.41)	-0.034 (p=0.74)	–	–
MAD	–	0.667	–	–	–	1.000	0.667

Table 5: Top part: Extra human evaluation of the dialogues generated by GPT-o3-pro (n=100) against the crowd-sourced dialogues (n=100) using 5-point Likert scale. Bottom part: correlation evaluation based on Pearson and Mean Absolute Difference (MAD) on the scores between LLM-as-a-Judge and Crowd workers.

made notable progress in generating responses that are emotionally appropriate and non-harmful. This suggests strong potential for using high-performing LLMs to generate high-quality synthetic dialogues. In the future, such models could be leveraged to augment and diversify the EmplifAI dataset through synthetic data generation—filling gaps in underrepresented emotional categories, expanding cultural or linguistic coverage, and accelerating the development of emotionally intelligent AI systems.

In the end, the non significant correlation between Gemini's score and Crowd workers' score on GPT-generated contents warrant caution towards using LLM-as-a-Judge as a primary evaluation approach. While automated evaluation offers scalability and efficiency, our findings suggest that it may not reliably capture nuanced human judgments, especially in emotionally sensitive tasks such as empathy generation. Future research should explore hybrid evaluation strategies that combine LLM-based assessments with human ratings to ensure both consistency and validity in measuring the quality of empathetic dialogue.

6 Conclusion

In this paper, we introduce EmplifAI, a Japanese dataset thoughtfully curated to capture a wide range of scenarios and empathetic dialogues reflecting fine-grained emotions in the context of chronic medical conditions. We translated GoEmotions' emotion labels into Japanese and conducted preliminary validation of the Japanese emotion taxonomy, demonstrating high consistency in the LLM's predictions. We further established a baseline for two-turn dialogue generation by fine-tuning a small Japanese LLM (LLM-jp-3.1-13b-instruct4) using EmplifAI, and observed substantial improvements in generating empathetic responses. Although the SFT model still shows room for improvement compared to large commercially available models, future studies could explore augmenting the dataset with synthesized dialogues to enhance fine-tuning outcomes.

7 Acknowledgements

Removed for peer-review

8 Limitations

Even though EmplifAI demonstrated ability to improve the performance of a compact Japanese LLM, there are a few noteworthy limitations for researchers who are interested in using the dataset or replicating the study.

The first limitation lies in our prompt design. We intentionally did not constrain the length of text generation. As a result, language models tended to produce longer responses than crowd workers. Rather than the content, previous studies have shown that length of a response could bias evaluation outcomes (Hu et al., 2024; Santilli et al., 2025). While it was necessary to use the same instructions for both LLMs and crowd workers to establish a performance baseline, future comparisons with human dialogues should take this limitation into account.

The second limitation concerns the medical context targeted by the EmplifAI dataset. It was specifically designed to train LLMs to respond to patients managing chronic medical conditions. As such, it may not generalize well to open-ended conversations or situations requiring general empathetic responses.

Since the EmplifAI dataset was primarily built in Japanese, many of its cultural nuances and expressions are specific to Japanese language and culture. Hence, it may not generalize well to other cultural or linguistic contexts.

In the end, we adopted LLM-as-a-Judge to evaluate the performance of our open-ended text generation task. Although such an approach is also adopted by other studies, we also noticed that the LLM-as-a-Judge results could deviate from actual human judges. Therefore, a more thorough comparison using human raters is deemed beneficial for future studies.

Researchers are advised to take the limitation into consideration for future studies.

8.1 Ethics Consideration: Evaluating Harms

One of the key metrics we used to prescreen crowdsourced dialogues and to evaluate generated content was harmlessness. Although harmlessness was not our primary evaluation target, it has become a central criterion in the development of medical LLMs. For example, Google's Med-PaLM explicitly measures the "extent of possible harm" and the "likelihood of harm" (Singhal et al., 2023), while Tam *et al*. identify "Safety and Harm" as a core dimension in their framework for assessing healthcare LLMs (Tam et al., 2024).

In our study, we experimented with an LLM-as-a-Judge approach to rate harmlessness. However, Gemini-2.5-Flash gave every GPT-generated response a perfect score (55), diverging notably from crowd-worker ratings. Future researchers should therefore be cautious: LLM-as-a-Judge methods may mis-estimate highly sensitive ethical data.

References

Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, and 1 others. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *arXiv preprint arXiv:2407.03963*.

Bagus Tris Atmaja and Akira Sasou. 2022. Sentiment analysis and emotion recognition from speech using universal speech representations. *Sensors*, 22(17):6369.

John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, and 1 others. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6):589–596.

Sven Buechel and Udo Hahn. 2022. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996*.

Alan S Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner. 2019. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature human behaviour*, 3(4):369–382.

Google DeepMind. 2025. Gemini 2.5 flash: Ultra-efficient multimodal model with 1m-token context window. `https://deepmind.google/models/gemini/flash/`. Accessed: 2025-07-28.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Paul Ekman, Tim Dalgleish, and M Power. 1999. Basic emotions. *San Francisco, USA*.

Mahshid Hosseini and Cornelia Caragea. 2021. Distilling knowledge for empathy detection. In *Findings of the Association for Computational linguistics: EMNLP 2021*, pages 3713–3724.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. 2024. Explaining length bias in llm-based preference evaluations. *arXiv preprint arXiv:2407.01085*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 174–184.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *arXiv preprint arXiv:2205.12698*.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Zhiyang Qi, Takumasa Kaneko, Keiko Takamizo, Mariko Ukiyo, and Michimasa Inaba. 2025. Kokorochat: A japanese psychological counseling dialogue dataset collected via role-playing by trained counselors. *arXiv preprint arXiv:2506.01357*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Yuki Saito, Eiji Iimori, Shinnosuke Takamichi, Kentaro Tachibana, and Hiroshi Saruwatari. 2023. Calls: Japanese empathetic dialogue speech corpus of complaint handling and attentive listening in customer center. *arXiv preprint arXiv:2305.13713*.

Yuki Saito, Yuto Nishimura, Shinnosuke Takamichi, Kentaro Tachibana, and Hiroshi Saruwatari. 2022. Studies: Corpus of japanese empathetic dialogue speech towards friendly voice agent. *arXiv preprint arXiv:2203.14757*.

Andrea Santilli, Adam Golinski, Michael Kirchhof, Federico Danieli, Arno Blaas, Miao Xiong, Luca Zappella, and Sinead Williamson. 2025. Revisiting uncertainty quantification evaluation in language models: Spurious interactions with response length bias results. *arXiv preprint arXiv:2504.13677*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Emika Takeishi, Takashi Nose, Yuya Chiba, and Akinori Ito. 2016. Construction and analysis of phonetically and prosodically balanced emotional speech database. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 16–21. IEEE.

Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V Stolyar, Katelyn Polanska, Karleigh R McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, and 1 others. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ digital medicine*, 7(1):258.

Jane Turner and Brian Kelly. 2000. Emotional dimensions of chronic disease. *Western journal of medicine*, 172(2):124.

Robert Waldinger and Marc Schulz. 2023. *The good life: Lessons from the world's longest scientific study of happiness*. Simon and Schuster.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2020. Fine-grained emotion and intent learning in movie dialogues. *arXiv preprint arXiv:2012.13624*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. *arXiv preprint arXiv:2004.12316*.