I Understand How You Feel: Enhancing Deeper Emotional Support Through Multilingual Emotional Validation in Dialogue System

Zi Haur Pang, Yahui Fu, Koji Inoue, and Tatsuya Kawahara
Graduate School of Informatics, Kyoto University, Japan
{pang, fu, inoue, kawahara}@sap.ist.i.kyoto-u-ac.jp

Abstract

Emotional validation - explicitly acknowledging that a user's feelings make sense - has proven therapeutic value but has received little computational attention. The emotional validation in dialogue systems can be decomposed into (i) validating response identification, (ii) validation timing detection, and (iii) validating response generation. To support research on all three subtasks, we release **M-EDESConv**, a 120k English–Japanese multilingual corpus created through hybrid manual–automatic annotation, and **M-TESC**, a multilingual spoken-dialogue test set. For timing detection, we propose **MEGUMI**, a Multilingual Emotion-aware Gated Unit for Mutual Integration, that fuses frozen XLM-RoBERTa semantics with language-specific emotion encoders via cross-modal attention and gated fusion. MEGUMI shows superior performance on both the M-EDESConv and M-TESC datasets, both objectively and subjectively. Finally, our **EmoValid-Bench** benchmarks of GPT-4.1 Nano and Llama-3.1 8B indicate that current LLMs generate contextually similar, diverse validating responses, but emotional understanding remains a major area for improvement. Project page: https://github.com/zihaurpang/Multilingual-Emotional-Validation

1 Introduction

Empathy improves human–computer communication by fostering trust, rapport, and sustained engagement in human-robot interaction and conversational agents. Recent studies show that systems that modulate empathy in real time are more trusted and perceived as more helpful, underscoring the practical value of artificial empathy (Leite et al., 2013; Morris et al., 2018; Casas et al., 2021). Accordingly, empathetic-dialogue research has enriched response generation with socio-cognitive signals. Representative directions include commonsense reasoning (Sabour et al., 2022; Fu et al., 2023), emotion-cause extraction (Gao et al., 2021), and personality tailoring (Zhong et al., 2020; Cai et al., 2024; Fu et al., 2024).

Yet generic "I'm sorry to hear that" replies often fall short for people who suppress emotions or face chronic stressors. Psychotherapy highlights ***emotional validation*** - recognizing, understanding, and acknowledging others' emotional states, thoughts, and actions, as a deeper intervention that de-escalates negative affect and strengthens therapeutic alliance (Linehan, 1997). Validating statements (e.g., "It makes sense that you feel frustrated") lower pain intensity in chronic-pain patients (Edlund et al., 2015), and foster adherence in youth mental-health care (Wasson Simpson et al., 2022) among other benefits (Carson-Wong et al., 2018; Lambie and Lindberg, 2016; Daniel, 2023).

Despite its importance in supportive interaction, computational work on emotional validation remains nascent. This gap is especially important in the era of Large Language Models (LLMs), as aligned LLMs have been reported to exhibit socially desirable and overly agreeable behavior, as well as sycophantic tendencies that favor agreement with the user even when such agreement is not well grounded (Sharma et al., 2023; Wei et al., 2023; Salecha et al., 2024; Bodroža et al., 2024; Hong et al., 2025). In emotional-support settings, this tendency can manifest as *over-validation*: models may validate too early, too often, or without sufficient evidence that validation is warranted. Meanwhile, existing studies have largely relied on handcrafted phrase lists and have been evaluated mainly in Japanese (Pang et al., 2023, 2024, 2026), leaving open how emotional validation should be modeled across languages whose emotion semantics, social-support norms, and listener feedback behavior may differ (Jackson et al., 2019; Kim et al., 2008; Clancy et al., 1996).

To address these gaps, we formalize emotional validation in a multilingual dialogue framework,

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 219–236
August 2–5, 2026.

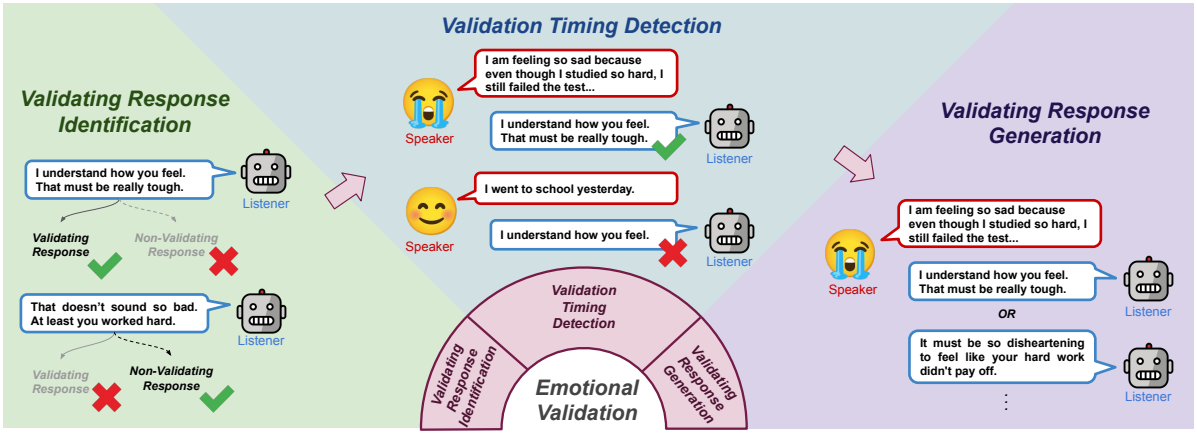

Figure 1: Emotional validation can be decomposed into three subtasks: (1) Validating Response Identification, which identifies *what* constitutes a validating response; (2) Validation Timing Detection, which determines *when* the system should validate the user; and (3) Validating Response Generation, which determines *how* to generate an appropriate validating response.

illustrated in Figure 1 and detailed in Appendix A. Our main contributions are:

- We release **M-EDESConv**, and **M-TESC**, the first open-source, semi-automatic verified multilingual dialogue corpus annotated for validation phenomena in text and speech scenario, enabling cross-lingual evaluation beyond prior Japanese-only efforts.
- We introduce **MEGUMI**, which fuses monolingual emotional cues with multilingual semantic representations to detect validation timing more accurately, helping reduce the over-validation tendency observed in LLMs.
- We present **EmoValidBench**, the first benchmark for validating response generation, evaluating the emotional validation of LLM baselines from semantic, diversity, empathy and LLM-as-judge metrics, to enable standardized comparison not only of response quality but also of whether models validate appropriately rather than indiscriminately.

2 Dataset Construction

We begin with two public English datasets emphasizing affective support, EmpatheticDialogues (Rashkin et al., 2019), and ESConv (Liu et al., 2021). To evaluate under the spoken dialogue scenario, we also include the TUT Emotional Storytelling Corpus (TESC) (Oishi et al., 2021). Details of these dataset can be found in Appendix C.

For cross-lingual evaluation, on top of adding a Japanese ED (Sugiyama et al., 2023), motivated by benchmarks like XNLI (Conneau et al., 2018) and PAWS-X (Yang et al., 2019) that similarly adopted translation-based settings to obtain broad multilingual coverage, we create a Japanese ESConv via a GPT-4.1-mini[1] translation workflow. Combining English and Japanese versions yields our Multilingual-Empathetic Dialogue Emotional Support Conversation (M-EDESConv) dataset. We also translate TESC to everyday English with the same workflow, yielding Multilingual-TUT Emotional Storytelling Corpus (M-TESC). Dataset summaries and translation prompts appear in Appendices C, K.1, and K.2, respectively.

Translation quality was assessed using both automatic and human evaluation. Using the COMETKiwi (Rei et al., 2022), a reference-free quality estimation metric for machine translation, the translated utterances and responses obtained scores of 0.8470 and 0.8479, respectively, indicating high overall quality. Three bilingual Japanese–English annotators further evaluated the translations on accuracy, contextual consistency, emotional consistency, fluency/readability, and cultural appropriateness. The translations received an average score of 5.63/7 overall, with average fair inter-annotator agreement (IAA) of 0.344. Full details are provided in Appendix D.

2.1 Annotation of Emotional Validation

Given the impracticality of manually annotating all 120k utterances in our multilingual corpus, we adopted a two-stage, hybrid annotation strategy

[1] https://openai.com/index/gpt-4-1/

Dataset	#Validation	#Non-Validation
M-EDESConv	46002	80551
-train	36714	64540
-val	4652	7867
-test	4636	8144
M-TESC	1052	2028

Table 1: Distribution of dataset

inspired by large-scale emotion datasets (Yang et al., 2012). In the first stage, we manually labeled roughly 3000 utterances per source (≈ 2% of EmpatheticDialogue and 3% of ESConv), classifying each response as either validating or non-validating. This yielded 1204 validating and 1663 non-validating responses in EmpatheticDialogue, and 680 validating versus 2258 non-validating responses in ESConv. In the second stage, we trained a classifier to automatically label the remaining data. We frame this as a binary classification task using the response as input and the validation label as output, which is also known as **validating response identification** in this study. Through fine-tuning the *xlm-roberta-large* model (Conneau et al., 2020), with training hyperparameters detailed in Appendix E.1, the classifier achieves a macro-average F1 score of 85.28 and an F1 score of 86.67 for the validation class on the manually annotated test set. The details of comparison to several baselines are presented in Appendix E.3, with result shown in Table 9 .

Using the trained classifier, we proceed to annotate the full dataset. As a result, the final dataset made up of 98.6k utterances, with 5.89% (5,805) of manual annotations, and 94.11% (92,795 responses) were annotated automatically, detailed in Appendix G. To verify the automatic labels' reliability, we ran an additional human evaluation, detailed in Appendix F. With randomly 100 automatically annotated samples for each language, our crowdsourcing judgement result showed that 85% of human judgement, with a substantial agreement level of 0.752 for the IAA, aligned with the automatic labels in the annotation process, showing the effectiveness of our hybrid annotation procedure in the dataset construction in this study. To further assess our task in a spoken dialogue setting, we additionally conduct manual annotation on the TESC dataset. Implementing the train/valid/test split with 8:1:1 ratio, the final distribution of validating and non-validating responses across datasets is summarized in Table 1.

3 Validation Timing Detection

We cast validation timing detection as a binary classification problem: given the dialogue context up to the current user turn, decide whether the next system response should generate a validating response. Accurate timing requires both semantic content (what the user is saying) and affective cues (how they are feeling). The proposed Multilingual Emotion-aware Gated Unit for Mutual Integration (MEGUMI) architecture, shown in Figure 2, fuses language-agnostic semantics with language-specific emotion representations through a gated, cross-modal pipeline that can be trained end-to-end from text utterance alone.

3.1 Validation Timing Detection Modal

Semantic backbone The core encoder of MEGUMI is XLM-RoBERTa-large[2] , chosen for its strong zero-shot transfer across 100+ languages. We freeze its parameters to preserve multilingual lexical knowledge and to curb computational cost.

Language-specific emotion channels Research shows that emotion taxonomies and lexical cues vary by language; a single encoder therefore risks conflating culture-specific signals (Takenaka, 2025). For English utterances, we leverage ModernBERT-large[3] fine-tuned on GoEmotions - a 58 k-instance Reddit corpus with 27 fine-grained labels (Demszky et al., 2020). Japanese turns are processed by LUKE-Japanese-large adapted to the WRIME writer-emotion dataset[4].

Emotion-Enhanced Multilingual Attention (EEMA) In each training batch, we ensure that both English and Japanese emotional cues are present and processed jointly. To capture their interactions, we introduce an emotion-enhanced multilingual attention block inspired by the Multimodal Transformer (Tsai et al., 2019). Specifically, the module projects the concatenated representation from one language as the query and the representation from the other language as the key and value, then applies scaled dot-product attention and returns two residual-normalized streams. With both English and Japanese emotion channels available during training, EEMA allows MEGUMI to learn latent cross-lingual alignments

[2]https://huggingface.co/FacebookAI/xlm-roberta-large

[3]https://huggingface.co/cirimus/modernbert-large-go-emotions

[4]https://huggingface.co/Mizuiro-sakura/luke-japanese-large-sentiment-analysis-wrime

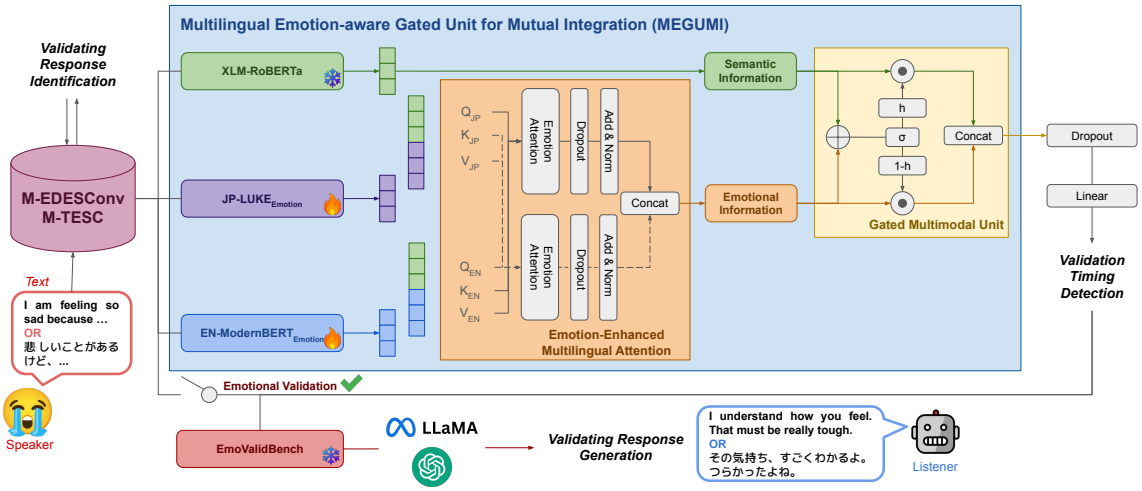

Figure 2: Overall proposed architecture in this study. We proposed a Multilingual Emotion-aware Gated Unit for Mutual Integration (MEGUMI) for the Validation Timing Detection Task.

between affective cues expressed differently across languages (e.g., fear in English vs. anger in Japanese).

Gated Multimodal Unit Simply concatenating streams can swamp minority cues; we therefore integrate them through a Gated Multimodal Unit, which has proven effective in image–text genre classification (Arevalo et al., 2017). A sigmoid gate h decides, per sample, how much of the emotion-projected vector, $z_{emotion}$ versus the semantic-projected vector, $z_{semantic}$ to pass forward: $z = \mathbf{h}z_{semantic} + (\mathbf{1} - \mathbf{h})z_{emotion}$. The fused vector is fed to a dropout–linear softmax head for the binary labels validate or non-validate. To counter class imbalance we weight the cross-entropy loss.

3.2 Experimental Settings

We fine-tuned all models on the M-EDESConv corpus, with hyperparameters detailed in Appendix H. We benchmarked against (i) a random classifier, (ii) two fine-tuned multilingual language models, mBERT and XLM-RoBERTa (Conneau et al., 2020), and (iii) instruction-tuned large language models: Llama-3.1.1 8B-Instruct and GPT-4.1 nano, each in zero-shot, 3-shot prompt-engineering, CoT and LoRA regimes, refer to Appendix K.4.

3.3 Ablation Studies

To disentangle architectural choices we compared:

XLM-RoBERTa uses only the frozen semantic encoder, followed by dropout and a linear layer.

+Mono-EN uses only the English Emotion Embedding (EE) Encoder, fused with XLM-RoBERTa output, followed by dropout and a linear layer.

+Mono-JP uses the Japanese Emotion Embedding Encoder in the same fashion.

+Multi-Concat combines both English and Japanese EE Encoders via direct concatenation, together with the XLM-RoBERTa output.

+Multi-EEMA feeds the outputs of the English and Japanese EE Encoders into the EEMA module to compute a single, fused emotional representation, which is then concatenated with the semantic encoder's output before classification.

3.4 Evaluation Metrics

A model that indiscriminately labels many turns as validating (high recall) risks producing hollow or repetitive acknowledgments that undermine perceived empathy; hence a high F1 score alone can be misleading if it masks low precision. We therefore report (i) validation-precision as the primary indicator of conversational appropriateness, (ii) validation-F1 to capture the precision-recall trade-off, and (iii) macro-F1 (M-F1) across both classes to ensure that performance on the majority non-validate class is not neglected.

3.5 Results and Discussion

Table 2 summarizes the results on validation-timing detection in multilingual and monolingual settings. In the multilingual setting, MEGUMI performs best overall: it achieves the highest macro-F1, target-class precision, and target-class F1 on M-EDESConv, and again the best macro-F1 and target-class precision on M-TESC, showing strong gener-

	M-EDESConv				M-TESC				Human-validated subset			
	Overall	Target Class			Overall	Target Class			Overall	Target Class		
	M-F1	Pre.	Rec.	F1	M-F1	Pre.	Rec.	F1	M-F1	Pre.	Rec.	F1
Multilingual												
Random Baseline	49.23	36.45	50.35	42.30	49.03	34.41	50.02	40.77	48.68	48.71	47.60	48.13
mBERT	59.10	45.16	74.15	56.14	53.29	38.29	41.35	39.76	58.80	60.47	52.00	55.91
XLM-RoBERTa	59.03	45.19	76.96	56.94	55.10	40.24	50.00	44.60	58.85	60.23	53.00	56.38
Llama 3.1 8b												
- *Zero-shot*	37.18	37.72	93.75	53.79	40.76	36.31	89.58	51.68	36.54	50.16	96.80	66.07
- *3-shot*	35.69	37.49	**94.84**	53.73	31.53	34.81	**96.29**	51.13	39.12	51.04	98.00	67.12
- *LoRA*	34.42	36.40	90.81	51.97	47.99	33.83	58.44	42.86	42.02	51.85	98.00	67.82
GPT 4.1 Nano												
- *Zero-shot*	51.68	41.43	79.25	54.41	54.04	39.95	66.46	49.90	51.76	52.14	46.80	49.31
- *3-shot*	46.75	40.02	87.04	54.99	50.14	39.01	80.93	**52.65**	49.50	53.01	88.00	66.17
- *CoT*	42.19	39.16	92.88	55.09	44.16	36.19	86.88	51.09	46.54	52.89	95.40	68.05
MEGUMI (Ours)	**63.71**†	**51.07**	66.11	**57.62**	**56.89**†	**41.44**	48.70	44.78	61.15	64.20	52.00	57.46
Human	No Applicable				No Applicable				70.82	66.20	88.67	75.78
Monolingual												
English												
Random Baseline	49.40	35.83	50.59	41.95	48.84	34.29	50.27	40.76	47.00	47.17	50.00	48.54
BERT	59.74	45.16	74.15	56.19	53.29	38.29	41.35	39.76	63.47	61.29	76.00	67.86
ModernBERT	59.10	44.94	75.02	56.27	52.80	37.67	46.77	41.73	69.81	67.24	78.00	72.22
XLM-RoBERTa	**62.22**	47.24	70.38	**56.57**	54.83	39.94	52.09	45.21	69.00	73.17	60.00	65.93
Llama 3.1 8b												
- *Zero-shot*	40.90	37.55	90.54	53.08	49.07	38.55	81.64	**52.37**	37.57	49.68	95.20	65.28
- *3-shot*	33.75	36.45	**96.47**	52.91	34.15	35.41	**95.93**	51.73	42.75	51.60	97.60	67.51
- *LoRA*	50.18	33.33	31.17	32.21	50.18	33.33	31.17	32.21	44.16	50.59	86.00	63.70
GPT 4.1 Nano												
- *Zero-shot*	55.36	42.53	74.43	54.13	53.74	39.29	60.80	47.73	50.02	50.69	46.00	48.21
- *3-shot*	56.69	43.19	68.49	52.98	53.70	39.24	60.57	47.62	49.11	51.97	84.40	64.33
- *CoT*	43.84	38.88	92.39	54.73	46.66	36.53	82.08	50.56	47.31	53.13	95.20	68.20
MEGUMI (Ours)	61.67	**48.40**	60.99	53.97	**58.41**	**45.07**	41.56	43.24	64.58	76.67	46.00	57.50
Human	No Applicable				No Applicable				70.81	66.84	84.67	74.71
Japanese												
Random Baseline	49.01	37.23	50.09	42.71	49.22	34.54	49.77	40.77	47.00	47.17	50.00	48.54
BERT	57.35	44.83	76.57	56.61	55.67	40.90	58.56	48.16	53.70	53.45	62.00	57.41
ModernBERT	51.15	42.02	86.48	56.67	51.51	38.18	66.92	48.62	61.80	59.46	88.00	70.97
XLM-RoBERTa	60.76	47.52	69.73	56.56	54.39	39.45	49.05	43.73	55.93	55.56	60.00	57.69
Llama 3.1 8b												
- *Zero-shot*	30.98	37.79	97.82	54.51	29.75	34.74	98.21	51.33	35.36	50.62	98.40	66.85
- *3-shot*	29.22	37.69	99.44	54.67	27.07	34.46	**99.85**	51.23	38.83	51.57	98.80	67.77
- *LoRA*	38.71	34.02	85.71	48.71	38.71	34.02	85.71	48.71	37.63	52.02	98.50	68.08
GPT 4.1 Nano												
- *Zero-shot*	47.03	40.28	81.51	53.91	52.48	39.63	74.68	51.78	53.45	53.60	47.60	50.40
- *3-shot*	39.67	39.28	**99.94**	55.22	43.38	37.53	91.41	**53.21**	44.93	52.29	92.00	66.67
- *CoT*	40.02	39.47	93.43	55.50	41.28	35.88	91.69	51.57	45.74	52.65	95.60	67.90
MEGUMI (Ours)	**61.20**	**48.83**	75.87	**59.42**	**57.09**	**41.35**	55.84	47.51	57.00	56.86	58.00	57.43
Human	No Applicable				No Applicable				70.75	65.57	92.67	76.80

Table 2: Results of validation timing detection task [%]. Top two results are highlighted in **bold** and underline, respectively. † indicates a statistically significant difference for $p < 0.05$ between MEGUMI and the best baselines, XLM-RoBERTa, determined by t-test.

alization to spoken dialogue. Across model types, LLM-based methods consistently trade precision for recall, often over-predicting validation timing, whereas fine-tuned classifiers are better balanced; MEGUMI improves this balance further. The monolingual results show the same overall trend with some dataset-specific differences. On M-EDESConv, fine-tuned PLMs remain strongest overall, with XLM-RoBERTa achieving the best macro-F1 in both languages, while MEGUMI yields the highest target-class precision in both languages. On M-TESC, however, MEGUMI achieves the best macro-F1 and target-class precision in both languages, indicating better calibration under noisier spoken conditions. Although some LLM baselines obtain the highest target-class F1 on M-TESC due to extremely high recall, their lower precision suggests more false positives.

Ablation study in Table 3 shows that multilingual emotion information plays an important role. +*Multi-Concat* outperforms both +*Mono-EN* and +*Mono-JP*, suggesting that English and Japanese emotion cues are complementary. Adding Emotion-Enhanced Multilingual Attention (+*Multi-EEMA*) yields further gains, and the full model performs best overall. These results indicate that cross-lingual affective cues, beyond lexical semantics alone, help identify validation-worthy moments, especially in emotionally ambiguous contexts.

To further examine task validity, we conducted an additional human evaluation on 200 utterances (100 English and 100 Japanese), balanced across

	Overall	Target Class		
	M-F1	Pre.	Rec.	F1
XLM-RoBERTa	47.29	39.60	82.42	53.50
+ Mono-EN	57.27	45.10	48.23	46.61
+ Mono-JP	56.86	43.97	62.88	51.75
+ Multi-Concat	59.80	46.75	73.86	57.26
+ Multi-EEMA	62.48	49.70	65.31	56.45
MEGUMI (Ours)	63.71	51.07	66.11	57.62

Table 3: Ablation study result [%].

validate and non-validate labels. Three native speakers judged whether a validating response should follow each utterance. The automatic labels achieved 72% average agreement with human judgments, with per-annotator agreement of 69–73% for English and 70–75% for Japanese. The overall IAA was moderate ($\kappa = 0.448$), suggesting that validation timing is more subjective than response-level validation labeling but still sufficiently consistent to support the task. On this human-validated subset, shown in the rightmost part of Table 2, MEGUMI again achieves the best macro-F1 among automatic systems and the highest target-class precision, outperforming mBERT, XLM-RoBERTa, and the LLM baselines.

4 Validating Response Generation

We position validating response generation as a stand-alone benchmark task, with the introduction of **EmoValidBench**, that tests whether a system can produce a concise, theory-consistent acknowledgment once the dialogue context has been flagged as requiring validation.

4.1 Benchmark construction

From the M-EDESConv corpus we extract every user utterance whose gold timing label is validate. Each of these turns is paired with one or more human validating replies that serve as references. English inputs are pre-processed with the Moses tokenizer[5], while Japanese inputs are segmented by MeCab + UniDic[6] to ensure comparability across BLEU and Distinct-n implementations.

We prompted Llama-3.1 8B and GPT-4.1 nano in Zero-shot (only the task definition), 3-shot (three labelled dialogue exemplars per language), Zero-shot CoT ("Let's think step by step" preamble) (Kojima et al., 2022), and LoRA-based fine-tuning (Hu et al., 2022), see Appendix K.5.

[5]https://github.com/luismsgomes/mosestokenizer
[6]https://taku910.github.io/mecab/

4.2 Evaluation metrics

To comprehensively assess validating response generation, we employ traditional-language-model-based metrics that capture semantic consistency, lexical diversity, and empathetic signal presence, as well as more-modern LLM-as-judge-based metrics (Zheng et al., 2023), that captures several clinical-grounded metrics. We also employ a small-scale of human evaluation on top of objective evaluation metrics.

4.2.1 Traditional-Language-Model Evaluation

Semantic Consistency. We utilize BERTScore (Zhang* et al., 2020) and BLEU (Papineni et al., 2002) to evaluate the semantic alignment between generated responses and reference texts.

Lexical Diversity. To quantify the diversity of generated language, we calculate Distinct-1 (D1) and Distinct-2 (D2) (Li et al., 2016), which represent the ratios of unique unigrams and bigrams to the total number of tokens.

Empathetic Signal Coverage. Inspired by prior work on empathetic communication (Sharma et al., 2020; Lee et al., 2022; Fu et al., 2024), we introduce three categories of empathetic signals: IP (interpretations), EX (explorations), and ER (emotional reactions). Specifically, IP denotes acknowledgments and understanding of, the interlocutor's emotion or situation; EX denotes expressions of active interest in the interlocutor's situation; and ER denotes explicit emotional reactions. The training details are shown in Appendix I, while the evaluation results of the empathetic-signal predictors are summarized in Table 10.

4.2.2 LLM-as-judge-based Evaluation

We further complement the objective evaluation with a structured LLM-as-judge rubric grounded in clinical communication frameworks, including NURSE (Childers et al., 2023) and OARS (Rollnick and Miller, 1995). Our rubric operationalizes whether the response (i) acknowledges feelings, (ii) reflects the user's experience accurately, (iii) maintains respect and a non-judgmental tone, (iv) conveys warmth and support, and (v) preserves safety and appropriate conversational boundaries. The complete prompt shown in the Appendix K.6.

4.2.3 Subjective Human Evaluation

To directly assess therapeutic appropriateness, we additionally conduct a blinded human evaluation

	Traditional LM-based metrics								LLM-as-Judge					
	BERT	BLEU	D1	D2	ER	IP	EX	AVG	AF	AR	RN	SW	SB	Ov.
Multilingual														
Llama 3.1 8b														
- *Zero-shot*	88.67	15.46	4.88	23.96	52.90	70.94	67.14	46.28	4.02	4.90	6.11	4.68	7.00	65.98
- *3-shot*	89.03	15.94	5.52	25.97	54.66	71.54	69.59	47.61	4.42	5.06	6.65	4.96	7.00	68.96
- *LoRA*	89.29	15.20	12.05	30.60	59.52	65.36	64.43	48.06	4.45	5.21	6.69	5.30	7.00	69.55
GPT 4.1 Nano														
- *Zero-shot*	88.87	13.47	4.80	22.37	51.95	73.17	69.97	46.37	4.96	5.22	6.73	5.41	7.00	71.69
- *3-shot*	89.26	16.24	4.58	21.42	52.32	72.29	68.76	46.41	5.05	5.57	6.90	5.68	7.00	73.57
- *CoT*	88.13	17.71	4.83	24.53	50.72	73.30	71.21	47.21	5.12	5.68	6.92	5.86	7.00	75.10
Monolingual														
English														
Llama 3.1 8b														
- *Zero-shot*	88.36	13.20	4.35	23.94	62.37	76.11	68.73	48.15	5.34	4.75	6.86	5.39	6.98	61.33
- *3-shot*	88.45	13.17	4.51	24.93	62.01	76.73	67.59	48.20	6.05	5.52	6.97	5.99	6.99	67.12
- *LoRA*	89.30	12.01	7.53	30.60	65.13	74.73	71.74	50.06	3.66	3.18	6.31	4.32	6.96	63.25
GPT 4.1 Nano														
- *Zero-shot*	88.33	12.86	5.02	25.60	62.73	75.65	69.10	48.47	5.95	5.33	6.94	5.80	6.99	74.24
- *3-shot*	88.53	13.32	4.59	22.88	60.71	75.52	67.75	47.61	5.68	5.06	6.95	5.77	6.99	74.62
- *CoT*	88.13	12.78	4.77	27.01	62.37	75.79	67.86	48.39	5.97	5.41	6.94	5.90	6.99	74.69
Japanese														
Llama 3.1 8b														
- *Zero-shot*	89.15	18.24	5.23	23.67	54.73	72.07	61.38	46.35	3.93	3.35	6.38	4.04	7.00	59.79
- *3-shot*	89.55	19.92	5.79	24.97	54.67	74.27	61.10	47.18	4.37	3.77	6.33	4.42	6.99	64.63
- *LoRA*	89.34	16.56	3.39	15.02	56.92	75.38	58.84	45.06	3.03	2.70	5.66	3.29	6.95	56.32
GPT 4.1 Nano														
- *Zero-shot*	89.54	19.96	4.56	18.40	53.34	76.77	61.62	46.31	5.49	4.90	6.94	5.58	7.00	73.42
- *3-shot*	90.16	23.16	5.00	20.33	57.70	77.60	60.83	47.83	5.61	4.99	6.95	5.63	7.00	74.24
- *CoT*	88.14	14.38	7.00	27.74	49.90	78.09	61.35	46.66	5.72	5.16	6.93	5.73	7.00	74.30

Table 4: Validating response generation results [%]. The left block reports traditional language-model-based metrics. The right block reports multilingual LLM-as-Judge scores. AF = Acknowledges Feelings, AR = Accurate Reflection, RN = Respect / Nonjudgment, SW = Support / Warmth, SB = Safety / Boundaries, and Ov. = Overall.

Criterion	Ground Truth	Llama 3.1 8B	GPT-4.1 Nano
Naturalness	5.48	5.27	5.42
Contextual Understanding	5.76	5.67	5.67
Emotional Understanding	5.92	5.35	5.59
Non-judgmental Tone	6.08	5.98	5.85
Mindful Presence	5.66	5.53	5.43
Perceived Validation	5.71	5.33	5.47
IAA (α)	0.406	0.453	0.322

Table 5: Human evaluation of validating-response quality for ground truth and 3-shot prompted model outputs. Three annotators rated each response on six dimensions; IAA denotes Krippendorff's α.

on 100 utterance–response pairs (50 English, 50 Japanese). Each item is rated by three native speakers recruited through crowdsourcing. We evaluate six dimensions: *Naturalness*, *Contextual Understanding*, *Emotional Understanding*, *Non-judgmental Tone*, *Mindful Presence*, and *Perceived Validation*. These dimensions are designed to capture whether a response feels emotionally appropriate and validating beyond surface semantic overlap.

4.3 Results and Discussion

Table 4 reports five-seed averages for validating-response generation in multilingual and monolingual settings. Overall, compact LLMs produce contextually appropriate responses that are semantically close to the references, but remain weaker in affective quality. This trend is consistent across both traditional metrics and LLM-as-Judge scores: models score relatively high on *Respect/Nonjudgment* and *Safety/Boundaries*, but lower on *Acknowledges Feelings*, *Accurate Reflection*, and *Support/Warmth*. The same pattern appears in the automatic empathy metrics, where models are stronger at showing interest in the speaker's situation (IP) than at recognizing emotion (ER) or expressing explicit emotional reaction (EX). Overall, current LLMs appear generally polite and safe, but less reliable at explicit emotional validation.

Across settings, GPT 4.1 Nano consistently outperforms Llama 3.1 8b in judged overall quality. In the multilingual setup, GPT with CoT achieves the best overall score (75.10), and the same tendency holds in both monolingual languages. Although Japanese often matches or exceeds English on surface-oriented metrics such as BERTScore and BLEU, this does not yield better judged validation quality. More broadly, English performs better than Japanese on emotion-sensitive criteria, especially *Acknowledges Feelings*, *Accurate Reflection*, and *Support/Warmth*. We attribute this gap to

English-dominant training and evaluation resources for empathetic dialogue (Xuan et al., 2025), cross-linguistic pragmatic differences such as indirectness and *aizuchi* in Japanese (Belay et al., 2025), and language-specific segmentation and emotion semantics (Rusli and Shishido, 2024). Thus, while current LLMs can generate appropriate validating responses, their affective calibration remains limited, particularly outside English.

To assess whether these automatic results reflect perceived therapeutic quality, we conducted a blinded human evaluation on 100 utterance–response pairs. As shown in Table 5, human reference responses received the highest scores on most dimensions, confirming that the benchmark targets are generally perceived as the strongest validating responses. Among the models, GPT is slightly stronger on *Emotional Understanding* and *Perceived Validation*, whereas Llama is slightly better on *Non-judgmental Tone* and *Mindful Presence*, though the differences are modest. The overall inter-annotator agreement, measured by Krippendorff's α, is 0.394, reflecting the subjective nature of the task. These findings indicate that semantic similarity captures only part of validating quality. Case studies are provided in Appendix J.

4.4 Joint Timing-and-Generation Evaluation

EmoValidBench evaluates response generation assuming that validation is warranted, but a practical dialogue system must decide both *when* to validate and *what* to say, including for utterances that do not require validation. To assess this joint setting, we constructed a benchmark of 200 utterances (English and Japanese; per language, 50 validate and 50 non-validate cases). We compare four configurations: **MEGUMI+LLM**, which uses MEGUMI for timing and an LLM for generation; **LLM two-stage**, which uses an LLM for timing and another LLM for generation; **LLM multi-task**, where a single LLM first decides whether validation is needed and then generates a response only when appropriate; and **LLM alw. multi-task**, which uses the same format but always generates a validating response regardless of the timing decision. For each LLM-based setting, we use the same GPT 4.1 Nano backbone with 3-shots settings as in the response-generation experiments, with prompts provided in Appendix K.5.

Table 6 shows a clear precision–recall trade-off in timing. **LLM alw. multi-task** achieves the highest timing recall (92.00) and the best timing F1 among the automatic systems (69.70), but does so by over-validating many negative cases. In contrast, **MEGUMI+LLM** is better calibrated for deciding *when* to validate, achieving the highest timing precision (64.20) and the best message-level macro-F1 (61.15) among automatic systems, indicating fewer false positives on non-validation utterances. **LLM multi-task** also outperforms **LLM two-stage** on both message-level and timing-level F1, suggesting that joint modeling is more effective than a pipelined approach. Human judgments remain the upper bound (message-level F1: 70.82; timing-level F1: 75.78). Overall, these results support using a timing detector like MEGUMI to gate validation decisions, while relying on LLMs for response generation only when validation is warranted.

Model	Message-level		Timing-level		
	M-F1	Bal. Acc.	Pre.	Rec.	F1
Human Judgement	70.82	71.67	66.20	88.67	75.78
MEGUMI + LLM	61.15	61.50	64.20	52.00	57.46
LLM two-stage	49.50	55.00	53.01	88.00	66.17
LLM multi-task	54.14	58.00	55.06	87.00	67.44
LLM alw. multi-task	55.44	60.00	56.20	92.00	69.70

Table 6: Joint timing-and-generation evaluation result [%]. Message-level scores evaluate the appropriateness of the final system behavior after timing and generation, while timing-level scores evaluate whether the system correctly decides whether validation is warranted.

5 Conclusions

In this work, we have presented the first comprehensive treatment of *emotional validation* within dialogue systems, spanning task formalisation, data, models, and evaluation. We defined three clear subtasks - validating response identification, validation timing detection, and validating response generation - and introduced **M-EDESConv** and **M-TESC**, the first large-scale multilingual corpora annotated for validation phenomena in both text-based and spoken settings. Our proposed **MEGUMI** architecture leverages cross-lingual pretrained semantics together with language-specific emotion encoders, unified by cross-modal attention and a gated fusion mechanism, to accurately determine when a system should validate a user's feelings. We also introduced **EmoValidBench**, the first benchmark for validating response generation, providing evaluation scripts, LLM baselines, and empathy-oriented metrics to enable standardized comparisons across future models. Looking ahead, we would like to extend to the implementation of emotional validation in a conversational robot for the real-world study.

Limitations

Despite these contributions, our study has several limitations. First, the scope of our curated data is confined to English and Japanese; other languages and cultural norms around emotional validation may exhibit different linguistic cues and pragmatic conventions that our current models cannot capture. Second, although we bootstrap annotation with a semi-automatic classifier to scale to 120 k turns, the reliance on confidence-filtered pseudo-labels carries the risk of residual errors and biasing downstream models, especially in low-resource or edge-case contexts.

Moreover, our timing-detection model operates solely on text transcripts and omits prosodic, acoustic, and visual cues known to inform validation in face-to-face interaction. The freeze of the XLM-RoBERTa backbone for computational tractability also precludes domain-specific fine-tuning that might further improve performance, and hardware constraints prevented exploration of larger language models beyond 8b parameters. Finally, while our experiments show promising performance in non-clinical dialogue, deploying emotional validation in sensitive domains such as mental-health support will require rigorous safety protocols, expert oversight, and continuous monitoring to avoid harm or overreliance on automated empathy.

Acknowledgments

The authors would like to acknowledge Professor Mika Enomoto for providing us with access to the TUT Emotional Storytelling Corpus, which enabled us to analyze and draw conclusions from a vast amount of data. This work was also supported by JST Moonshot R&D JPMJPS2011.

References

Deeksha Adiani, Kelley Colopietro, Joshua Wade, Miroslava Migovich, Timothy J Vogus, and Nilanjan Sarkar. 2023. Dialogue act classification via transfer learning for automated labeling of interviewee responses in virtual reality job interview training platforms for autistic individuals. *Signals*, 4(2):359–380.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin Grissom II, Iqra Ameer, Grigori Sidorov, Olga Kolesnikova, and Seid Muhie Yimam. 2025. CULEMO: Cultural lenses on emotion - benchmarking LLMs for cross-cultural emotion understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18894–18909, Vienna, Austria. Association for Computational Linguistics.

Bojana Bodroža, Bojana M Dinić, and Ljubiša Bojić. 2024. Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, 11(10).

Naomi Breslau, Ronald C Kessler, Howard D Chilcoat, Lonni R Schultz, Glenn C Davis, and Patricia Andreski. 1998. Trauma and posttraumatic stress disorder in the community: the 1996 detroit area survey of trauma. *Archives of general psychiatry*, 55(7):626–632.

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024. Pecer: Empathetic response generation via dynamic personality extraction and contextual emotional reasoning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10631–10635. IEEE.

Lea Canales, Carlo Strapparava, Ester Boldrini, and Patricio Martínez-Barco. 2016. Innovative semi-automatic methodology to annotate emotional corpora. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 91–100.

Amanda Carson-Wong, Christopher D Hughes, and Shireen L Rizvi. 2018. The effect of therapist use of validation strategies on change in client emotion in individual dbt treatment sessions. *Personality Disorders: Theory, Research, and Treatment*, 9(2):165.

Jacky Casas, Timo Spring, Karl Daher, Elena Mugellini, Omar Abou Khaled, and Philippe Cudré-Mauroux. 2021. Enhancing conversational agents with empathic abilities. In *Proceedings of the 21st ACM international conference on intelligent virtual agents*, pages 41–47.

Mao Yan Chen, Siheng Li, and Yujiu Yang. 2022. EmpHi: Generating empathetic responses with human-like intents. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1074, Seattle, United States. Association for Computational Linguistics.

Julie W Childers, Hailey Bulls, and Robert Arnold. 2023. Beyond the nurse acronym: the functions of empathy in serious illness conversations. *Journal of pain and symptom management*, 65(4):e375–e379.

Patricia M Clancy, Sandra A Thompson, Ryoko Suzuki, and Hongyin Tao. 1996. The conversational use of reactive tokens in english, japanese, and mandarin. *Journal of pragmatics*, 26(3):355–387.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2475–2485.

Hannah Sophie Daniel. 2023. Exploring emotional validation in cross-cultural management: a case study.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Sara M Edlund, Maria L Carlsson, Steven J Linton, Alan E Fruzzetti, and Maria Tillfors. 2015. I see you're in pain–the effects of partner validation on emotions in people with chronic pain. *Scandinavian Journal of Pain*, 6(1):16–21.

Lauren Fonteyn, Enrique Manjavacas, Nina Haket, Aletta G Dorst, and Eva Kruijt. 2024. Could this be next for corpus linguistics? methods of semi-automatic data annotation with contextualized word embeddings. *Linguistics Vanguard*, 10(1):587–602.

Yahui Fu, Chenhui Chu, and Tatsuya Kawahara. 2024. Styemp: Stylizing empathetic response generation via multi-grained prefix encoder and personality reinforcement. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 172–185.

Yahui Fu, Koji Inoue, Chenhui Chu, and Tatsuya Kawahara. 2023. Reasoning before responding: Integrating commonsense-based causality explanation for empathetic response generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 645–656.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 807–819.

Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. 2025. Measuring sycophancy of language models in multi-turn dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2239–2259, Suzhou, China. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Joshua Conrad Jackson, Joseph Watts, Teague R Henry, Johann-Mattis List, Robert Forkel, Peter J Mucha, Simon J Greenhill, Russell D Gray, and Kristen A Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Deborah Johanson, Ho Seok Ahn, Rishab Goswami, Kazuki Saegusa, and Elizabeth Broadbent. 2023. The effects of healthcare robot empathy statements and head nodding on trust and satisfaction: a video study. *ACM Transactions on Human-Robot Interaction*, 12(1):1–21.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261, Bangkok, Thailand. Association for Computational Linguistics.

Heejung S Kim, David K Sherman, and Shelley E Taylor. 2008. Culture and social support. *American psychologist*, 63(6):518.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Janice R Kuo, Skye Fitzpatrick, Jennifer Ip, and Amanda Uliaszek. 2022. The who and what of validation: an experimental examination of validation and invalidation of specific emotions and the moderating effect of emotion dysregulation. *Borderline Personality Disorder and Emotion Dysregulation*, 9(1):15.

John A Lambie and Anja Lindberg. 2016. The role of maternal emotional validation and invalidation on children's emotional awareness. *Merrill-Palmer Quarterly*, 62(2):129–157.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *29th COLING*, pages 669–683.

Iolanda Leite, André Pereira, Samuel Mascarenhas, Carlos Martinho, Rui Prada, and Ana Paiva. 2013. The influence of empathy in human–robot relations. *International journal of human-computer studies*, 71(3):250–260.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In

Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 110–119, San Diego, California. Association for Computational Linguistics.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10993–11001.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Marsha M Linehan. 1997. Validation and psychotherapy. *American Psychological Association*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Audrey Marcoux, Marie-Hélène Tessier, and Philip L Jackson. 2024. Nonverbal behaviors perceived as most empathic in a simulated medical context. *Computers in Human Behavior*, 157:108268.

Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. 2018. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research*, 20(6):e10148.

Hikaru Oishi, Mika Enomoto, Keiko Ochi, and Yasunari Obuchi. 2021. Design and basic analysis of the tut emotional storytelling corpus. In *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 43–48. IEEE.

Zi Haur Pang, Yahui Fu, Yuan Gao, and Tatsuya Kawahara. 2026. Paralinguistic emotion-aware validation timing detection in japanese empathetic spoken dialogue. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zi Haur Pang, Yahui Fu, Divesh Lala, Keiko Ochi, Koji Inoue, and Tatsuya Kawahara. 2023. Prediction of validating response from emotional storytelling corpus. In *The 37th Annual Conference of the Japanese Society for Artificial Intelligence*, pages 2O5OS2a03–2O5OS2a03.

Zi Haur Pang, Yahui Fu, Divesh Lala, Keiko Ochi, Koji Inoue, and Tatsuya Kawahara. 2024. Acknowledgment of emotional states: Generating validating responses for empathetic dialogue. *14th International Workshop on Spoken Dialogue System Technology*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

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

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Stephen Rollnick and William R Miller. 1995. What is motivational interviewing? *Behavioural and cognitive Psychotherapy*, 23(4):325–334.

Shanequa S Roscoe-Nelson and Geoffrey A Silvera. 2024. Is timing everything?: The role of time on the relationship between patient-centered communication and provider empathy. *Patient Experience Journal*, 11(2):36–43.

Andre Rusli and Makoto Shishido. 2024. An experimental evaluation of japanese tokenizers for sentiment-based text classification. *The 27th Annual Meeting of the Association for Natural Language Processing*.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. Large language models show human-like social desirability biases in survey responses. *arXiv preprint arXiv:2405.06058*.

Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic persuasion: reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 844–856.

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, Kravec Shauna, Maxwell Timothy, McCandlish Sam, Ndousse Kamal, Rausch Oliver, Schiefer Nicholas, Yan Da, Zhang Miranda, and Perez Ethan. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2023. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. In *2022 ieee spoken language technology workshop (slt)*, pages 685–691. IEEE.

Yoichi Takenaka. 2025. Performance evaluation of emotion classification in japanese using roberta and deberta. *arXiv preprint arXiv:2505.00013*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558.

Qing Wang, Shuyuan Peng, Zhiyuan Zha, Xue Han, Chao Deng, Lun Hu, and Pengwei Hu. 2023. Enhancing the conversational agent with an emotional support system for mental health digital therapeutics. *Frontiers in Psychiatry*, 14:1148534.

Kendra S Wasson Simpson, Anna Gallagher, Scott T Ronis, David AA Miller, and Kate C Tilleczek. 2022. Youths' perceived impact of invalidation and validation on their mental health treatment journeys. *Administration and Policy in Mental Health and Mental Health Services Research*, pages 1–14.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anuradha Welivita, Chun-Hung Yeh, and Pearl Pu. 2023. Empathetic response generation for distress support. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 632–644.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Hui Yang, Alistair Willis, Anne De Roeck, and Bashar Nuseibeh. 2012. A hybrid model for automatic emotion recognition in suicide notes. *Biomedical informatics insights*, 5:BII–S8948.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692.

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

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Zhang Hao,

E. Gonzalez Joseph, and Stoica Ion. 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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

Appendix

A Task Description

In this section, we will describe the task necessity for the emotional validation expression in the spoken dialogue system. Even though previous studies have shown that validation can be expressed through response generation (Pang et al., 2024), there aren't any formal task descriptions until now. Inspired by the theory of validation (Linehan, 1997), we have defined the emotional validation in the spoken dialogue system into three subtasks, i.e. validating response identification, validation timing detection, and validating response generation. The summary of the task description we defined is summarized in the Figure 1.

A.1 Validating Response Identification

The first requirement is to decide whether a system generated response is, validating, or known as ***validating response***. Mis-labeling brings risk: inappropriate reassurance or pseudo-empathy can increase user distress[7] or alienation (Breslau et al., 1998). Linguistic studies of dialogue acts provide methodological precedents, showing that automatic classifiers can distinguish supportive acts such as "appreciation" or "agreement" (Welivita and Pu, 2020; Chen et al., 2022) from neutral turns, but accuracy drops when acts overlap semantically (Stolcke et al., 2000; Adiani et al., 2023). Validation adds further nuance because the same surface pattern (e.g., "I see") may or may not affirm the user's emotion depending on context. Our corpus therefore begins with manual labels and expands them semi-automatically via a fine-tuned classifier, following hybrid annotation pipelines in emotion research (Canales et al., 2016; Fonteyn et al., 2024)

A.2 Validation Timing Detection

Knowing when to validate is as critical as knowing how. Communication studies warn that over-frequent or ill-timed empathic moves can be perceived as insincere, reducing perceived provider empathy and therapeutic alliance (Roscoe-Nelson and Silvera, 2024; Kuo et al., 2022). Similar timing effects emerge in social-robot experiments, where repetitive "I understand" statements without appropriate pauses diminished user rapport (Johanson et al., 2023). Existing end-to-end generators seldom account for discourse-level timing; they optimise local next-utterance loss and may insert multiple empathic markers in rapid succession. We cast timing as a sequence-labeling task over the dialogue context, enabling models such as our MEGUMI architecture to decide whether the upcoming turn warrants validation.

A.3 Validating Response Generation

Finally, the system must produce a response that satisfies validation theory (Linehan, 1997). Generic empathetic models often interleave advice, persuasion, or question-asking strategies that conflict with unconditional acknowledgment (Welivita et al., 2023; Samad et al., 2022). Moreover, validation can be expressed verbally and non-verbally; head nods, prosodic alignment, and empathic facial displays amplify perceived support in communications (Linehan, 1997; Johanson et al., 2023; Marcoux et al., 2024). Thus, we release the **EmoValid-Bench**, the first benchmark for validating response generation. It pairs each user turn that requires validation with evaluation scripts that measure semantic consistency, lexical diversity, and empathy-coverage.

B Related Work

B.1 Empathetic Response Generation

Empathetic Response Generation (ERG) began with EmpatheticDialogues (ED), which established emotion-grounded evaluation and showed that training on emotion-conditioned dialogues improves perceived empathy over chitchat (Rashkin et al., 2019). Subsequent work emphasized affect matching, e.g., MoEL routes to listener experts by emotion, MIME mimics user affect, and knowledge/commonsense–augmented models (KEMP, CEM) inject situational priors to boost specificity (Lin et al., 2019; Majumder et al., 2020; Li et al., 2022; Sabour et al., 2022). Taxonomy-driven control labels empathetic intents to steer responses (Welivita and Pu, 2020).

Despite these advances, ERG only focus on

[7] https://www.psychiatrictimes.com/view/when-validation-is-harmful

generate empathetic response, which conflicts with emotional validation, that requires deciding whether/when to validate and delivering explicit acknowledgment rather than mirroring affect. The mismatch surfaces in support-oriented settings: ESConv frames multi-stage strategies (exploration, comforting, action), yet timing remains weakly supervised; recent analyses of LLMs on ESConv report strategy biases and difficulty choosing the right move, yielding shallow support (Liu et al., 2021; Kang et al., 2024). Emerging work adds interpretable support structures (e.g., ESCoT, STEF) but still centers on style/strategy selection rather than principled validation decisions (Zhang et al., 2024; Wang et al., 2023).

We address the ERG–validation tension by separating decision from generation: first decide whether validation is warranted given context, then generate concise acknowledgment, yielding timing-aware, not purely affect-matching, support.

B.2 Emotional Validation

Clinical psychology defines emotional validation as communicating that a person's feelings "make sense" in context. Dialectical Behavior Therapy (DBT) details graded validation levels, underscoring timing and contextual fit as core to therapeutic alliance, distinct from generic empathy (Linehan, 1997). Computational treatment is nascent. Prior work combining emotion recognition and phrase-based generation shows feasibility on Japanese ED (text) and the TUT Emotional Storytelling Corpus (speech), with task-adaptive BERT outperforming strong baselines (Pang et al., 2024). Complementary analyses on TESC find acoustic/linguistic cues predictive of validation-worthy moments, reinforcing the primacy of timing (Pang et al., 2023).

Despite these demonstrations, prior work lacks a standardized task description and benchmark for emotional validation in spoken dialogue systems, leaving the problem under-specified relative to ERG datasets. Inspired by validation theory (Linehan, 1997), we formalize emotional validation for spoken dialogue as three subtasks: validating response identification, validation timing detection, and validating response generation, providing a clinically grounded decomposition compatible with ERG pipelines.

C Base Corpora

EmpatheticDialogues (ED) contains 24.8k two-speaker conversations collected from crowdworkers imagining specific emotional situations (Rashkin et al., 2019). ESConv complements ED with 1,053 counselor-style dialogues in which trained volunteers comfort users facing real-life stressors (Liu et al., 2021). To further evaluate under the spoken dialogue scenario, we include the TUT Emotional Storytelling Corpus (TESC) (Oishi et al., 2021), a Japanese two-party, multi-turn corpus (247 sessions; ≈9.2 h) where close friends recount experiences under eight Plutchik emotion prompts (Plutchik, 2001).

Table 7 summarises the size, modality and interactional profile of the three corpora that constitute our base dataset. EmpatheticDialogues (ED) offers the broadest coverage with 24.8k English and 20k Japanese text conversations; yet these crowdsourced exchanges are succinct - just 4.3 and 2.0 turns on average, with roughly 15–25 tokens per utterance - because speakers were asked to recount short personal stories to a listener who responds empathetically. ESConv contributes 1.3 k expert-annotated emotional-support dialogues per language. Although an order of magnitude smaller than ED, each session resembles real counselling, spanning ≈ 14 turns; Japanese utterances are notably longer (≈ 22 tokens) than English (≈ 15), matching prior observations on script complexity in bilingual corpora. Finally, the TUT Emotional Storytelling Corpus (TESC) introduces the spoken modality with 247 transcribed sessions per language. The oral setting yields markedly longer utterances (≈ 35 tokens in English, 41 in Japanese) while keeping the turn budget concise at eight per dialogue.

D Translation Quality Evaluation

As part of our data construction method involved the automatic translation from the LLMs, it is neccessary to evaluate the quality of these synthetic data.

D.1 Objective Quality Evaluation

We first conducted an objective evaluation. Since we do not have a dataset that consist of reference that allow us to use the normal machine traslation metric like BLEU score, we decided to evaluate through the COMETKiwi metric (Rei et al., 2022). COMETKiwi is a reference-free model employs a

Dataset	Modality	#dialogue	#utterance	Average #word	Average #turns
EmpatheticDialogues					
-English	Text	24.8k	82k	15.2	4.31
-Japanese	Text	20k	80k	25	2
ESConv					
-English	Text	1.3 k	17.6k	15.19	13.62
-Japanese	Text	1.3 k	17.6k	21.84	13.62
TESC					
-English	Speech	247	3080	34.85	8
-Japanese	Speech	247	3080	41	8

Table 7: Statistics of the English and Japanese splits of the three base corpora employed in this study.

regression approach and is built on top of InfoXLM. It has been trained using direct assessments from WMT17 to WMT20, as well as direct assessments from the MLQE-PE corpus. It provides scores ranging from 0 to 1, where 1 signifies a perfect translation. We have evaluated on both utterance and response perspective, i.e. evaluate the translation quality for the generated utterance and generated response. As a result, The translations scored 0.8470 for utterances and 0.8479 for responses, indicating high overall quality for dataset construction.

D.2 Subjective Quality Evaluation

We also conducted another human assessment on top of our objective evaluation. We recruited three bilingual Japanese–English speakers through crowdsourcing and asked them to evaluate sampled translations using five criteria: *Accuracy*, which measures how closely the translation preserves the meaning of the source text; *Contextual Consistency*, which evaluates consistency of terminology and style across the dialogue; *Emotional Consistency*, which assesses whether the intended emotion and tone are preserved; *Fluency and Readability*, which measures how natural and easy the translation is for a native speaker of the target language; and *Cultural Appropriateness*, which evaluates whether the translation remains culturally suitable for the target audience. To prevent the biasness by the annotaters, we also calculated the inter-annotator agreement (IAA). The average scores are shown in Table 8. Across all dimensions, the translations received a mean score of 5.63, with the highest score on Fluency and Readability (5.72) and consistently strong ratings on emotion-related dimensions such as Emotional Consistency (5.61) and Cultural Appropriateness (5.64). These results complement the COMETKiwi scores and indicate that the translation workflow preserves not only semantic content, but also affective tone and cultural appropriateness.

Criterion	Average	IAA
Accuracy	5.65	0.392
Contextual Consistency	5.52	0.407
Emotional Consistency	5.61	0.263
Fluency and Readability	5.72	0.318
Cultural Appropriateness	5.64	0.339
Overall	5.63	0.344

Table 8: Human evaluation results for translation quality. Three bilingual Japanese–English annotators rated the translated data on a Likert scale across five dimensions.

E Validating Response Identification

E.1 Training Hyperparameters

We fine-tune the *xlm-roberta-large* model (Conneau et al., 2020) with a learning rate of 1×10^{-5}, a batch size of 64, and train for 20 epochs. Evaluation is performed every 200 steps, using the Adam optimizer with L2 regularization (weight decay of 0.01). Early stopping is applied with a patience of 5 epochs. To ensure high precision for the validation class, we apply a confidence threshold of 0.75 during inference.

E.2 Comparative Baselines

As part of our ablation study, we compare the fine-tuned *xlm-roberta-large* model against several baselines, including a random baseline, multilingual BERT (mBERT)[8], LLaMA 3.1 8B[9], and GPT-4.1-nano[10].

E.3 Result and Discussion

Table 9 reports the performance of several automatic classifiers that were used to propagate validating response labels from a manually annotated seed set to the full M-EDESConv corpus. All scores are averaged over five random initialisations; we

[8] https://huggingface.co/google-bert/bert-base-multilingual-cased

[9] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[10] https://openai.com/index/gpt-4-1/

	M-EDESConv				Human-validated subset			
	Overall	Target Class			Overall	Target Class		
	M-F1	Precision	Recall	F1-Score	M-F1	Precision	Recall	F1-Score
Random Baseline	48.38	32.45	50.13	39.40	51.23	51.25	51.75	51.48
mBERT	74.30	70.24	76.32	73.15	72.96	71.30	77.00	74.04
Llama 3.1 8b								
- *Zero-shot*	41.04	34.37	85.61	49.04	52.13	54.14	87.60	66.92
- *3-shot*	32.18	33.82	**97.88**	50.27	46.58	53.26	98.00	69.01
- *LoRA*	32.96	37.16	97.02	53.74	44.05	51.67	93.00	66.43
GPT 4.1 Nano								
- *Zero-shot*	57.98	42.71	85.19	56.89	72.06	73.33	69.60	71.36
- *3-shot*	66.57	50.36	74.60	60.13	61.90	59.57	84.00	69.71
XLM-RoBERTa	**85.28**	**80.66**	93.64	**86.67**	82.43	78.76	89.00	83.57
Human	No Applicable				85.10	80.94	92.00	86.12

Table 9: Results of validating response identification task [%]. The left block reports results on the M-EDESConv dataset for automatic annotation, and the right block reports performance on the human-validated subset.

present both macro-averaged metrics (capturing overall label balance) and scores for the minority validate class, which is the critical signal for downstream tasks.

The fine-tuned XLM-RoBERTa model clearly emerges as the most reliable annotator. It achieves 85.3% macro F1 and 86.7% target-class F1, outperforming the next-best baseline (mBERT) by roughly eleven points on each metric. Precision gains (+10.4 pp over mBERT) indicate fewer false positives, while the high recall of 93.6% demonstrates that the model rarely misses genuine validating utterances - essential for minimising label noise. Instruction-tuned LLMs require careful prompting to approach Pre-trained Language Model (PLM) performance. In zero-shot mode, GPT-4.1 nano delivers respectable macro F1 (58.0%) but suffers from low precision (42.7%) on the validation class, leading to many spurious positives. Providing three in-context examples narrows the gap (macro F1 = 66.6%), yet precision (50.4%) still trails far behind XLM-RoBERTa. Llama-3.1 8B exhibits a complementary error profile: three-shot prompting attains the highest recall in the table (97.9%) but collapses precision to 33.8%, effectively labelling almost every response as validating and therefore offering little discriminative value.

These results motivated our choice of the XLM-RoBERTa classifier for corpus-wide pseudo-labelling. To mitigate residual noise we retained only predictions with confidence 0.90, yielding a class distribution that closely mirrors the manually annotated subset (described in 2.1). Although LLMs currently lag behind supervised PLMs for this task, their high recall could still prove useful in an ensemble or active-learning setting - an avenue we leave for future work.

F Human Evaluation on Automatic Annotation

As 94.11% of M-EDESConv was labeled automatically, we conducted an additional human evaluation study to quantify the reliability of the pseudo-labels. For each language (English and Japanese), we randomly sampled 100 automatically annotated responses, balanced across classes (50 validate and 50 non-validate), and collected three native-speaker judgments via crowdsourcing on whether each response was validating. The human judgments agreed with the automatic labels in 85.0% of cases on average. Agreement between individual annotators and the automatic labels ranged from 83–88% for English and 84–87% for Japanese. We also measured inter-annotator agreement among the human raters. The average pairwise Cohen's κ was 0.752 overall (English: 0.740/0.744/0.706; Japanese: 0.819/0.820/0.681), indicating substantial agreement.

Table 9 further compares model performance on the manually verified sample. Among the automatic methods, fine-tuned XLM-RoBERTa performs best, achieving 82.43 macro-F1 and 83.57 validation-class F1, outperforming mBERT and the compact LLM baselines. Its performance is also closest to human performance (85.10 macro-F1; 86.12 validation-class F1), and thus, shown the effectiveness of choosing XLM-RoBERTa as the backbone for the automatic annotation in this study.

G Annotated Dataset Details

To preserve label distribution consistency with the manually annotated subset, we analyze the prediction confidence scores across each sub-dataset. Based on this analysis, we set confidence thresh-

olds of 0.90 for ED and 0.95 for ESConv to align the automated annotations with the original distribution. As a result, the total dataset comprises 98.6k responses, including 81k from EmpatheticDialogues and 17.6k from ESConv. Among these, 5,805 responses were manually annotated, where 1,204 validating and 1,663 non-validating (total: 2,867) from EmpatheticDialogues, and 680 validating and 2,258 non-validating (total: 2,938) from ESConv. This means manual annotations account for 5.89% (5,805) of the dataset, while the remaining 94.11% (92,795 responses) were annotated automatically using our trained validation classifier.

Languages	Models	Traits	Acc.	BA.	F1
English	RoBERTa	ER	84.76	84.13	84.70
		IP	84.12	85.35	84.23
		EX	94.81	92.46	94.86
Japanese	LUKE Japanese	ER	73.74	72.64	73.52
		IP	79.09	79.29	79.22
		EX	88.82	77.37	88.27
Both	XLM-RoBERTa	ER	77.88	76.66	77.61
		IP	81.28	82.13	81.42
		EX	91.82	85.08	91.69

Table 10: Evaluation results of the empathetic signal predictors. Acc., BA., and F1 refer to accuracy, balanced accuracy, and weighted F1 score, respectively.

H Experimental Settings in Validation Timing Detection

We fine-tuned all models on the M-EDESConv corpus with a learning rate of 1×10^{-5}, a batch size of 64 (with gradient accumulation over 8 steps), and a 20-epoch cap. To regularise training we combined L2 normalization (weight decay rate of 0.01) with early stopping after five stagnant validation checks. Validation was run every 250 steps and the best checkpoint was selected by F1. Five random seeds were used throughout to mitigate variance.

I Empathetic Signal Coverage Training Details

For English, we follow the official annotation schema[11] and fine-tune three RoBERTa models (Liu et al., 2019), each followed by a linear classifier, to determine whether a response conveys each signal. For Japanese, we translate the English corpus by GPT-4.1-mini with temperature set to 0.1, and then classify the signals using a LUKE model[12] followed by two linear layers. In the multilingual setting, we employ XLM-RoBERTa (Conneau et al., 2020) with two linear layers as a unified classifier across languages.

We set the batch size to 32, the learning rate to 2×10^{-5}, and the dropout rate in the linear layers to 0.1, and train for up to seven epochs with early stopping after three epochs without improvement. The evaluation results of the empathetic-signal predictors are summarized in Table 10.

Model	Text
Utterance	Yeah they must practice a lot. I would be afraid of getting trampled
Ground Truth	Me too, that would be terrible.
Llama 3.1 8B	I can understand that. I would feel the same way.
GPT-4.1 Nano	That makes sense, it's understandable to feel worried about that.

Table 11: Comparative case studies between different LLM baselines.

J Case Studies

The case studies in Table 11 illustrate the same pattern discussed in Section 4.3. Model outputs are often close to the gold responses in meaning, as reflected by high BERTScore, even when surface forms differ. For example, for "Yeah they must practice a lot. I would be afraid of getting trampled," Llama 3.1 8b's "I can understand that. I would feel the same way" and GPT-4's "That makes sense; it's understandable to feel worried about that" both preserve the validating intent of the gold response, "Me too, that would be terrible."

K Prompts

K.1 English-Japanese translation

You are a professional Japanese translator. For the following English utterance, please translate into natural, spoken-style daily Japanese as a native speaker would say. You should avoid literal word-for-word renderings.

K.2 Japanese-English translation

You are a professional English translator. For the following Japanese utterance, please translate into natural, spoken-style daily English as a native speaker would say. You should avoid literal word-for-word renderings.

[11] https://github.com/behavioral-data/Empathy-Mental-Health

[12] https://huggingface.co/studio-ousia/luke-japanese-base

K.3 Validating Response Identification

Definition of validation: Validation is a communication technique, where we recognize, understand, and acknowledge others' emotional states, thoughts, and actions.

Please classify the response into validating or non-validating response. Return validate if it is a validating response and non-validate if it is a non-validating response.

Followed by the three examples dialogues with validating response, and another three examples dialogues with non-validating response in each language, respectively.

K.4 Validation Timing Detection

Definition of validation: Validation is a communication technique, where we recognize, understand, and acknowledge others' emotional states, thoughts, and actions.

Please classify each utterance into whether a validating response should be generated. Return validate if needed to generate a validating response and non-validate if not necessary to generate (meaning that it will generate a non-validating response)

Followed by the three examples dialogues with validating response, and another three examples dialogues with non-validating response in each language, respectively. Let's think step by step.

K.5 Validating Response Generation

Definition of validation: Validation is a communication technique, where we recognize, understand, and acknowledge others' emotional states, thoughts, and actions.

You should act as a listener, in speech conversations. Please generate a validating response for the given utterances from the speaker. The generated response should be a validating response. You should only respond with a validating response, excluding other information (without Listener:).

Followed by three examples dialogues with validating response in each language, respectively. Let's think step by step.

K.6 LLM-as-judge Evaluation

System: You are an expert evaluator of EMOTIONAL VALIDATION in supportive conversations. Score the response for how well it validates the user's feelings. Be concise, neutral, and avoid offering new counseling. Use the NURSE and OARS principles. Return a structured output only.

User: Evaluate how well the `RESPONSE` validates the user's emotions.

`USER:` {user utterance}
`RESPONSE:` {candidate response}

Rubric:

- `acknowledges_feelings`: names or mirrors the user's feelings
- `accurate_reflection`: reflects the user's emotion and context without distortion
- `respect_nonjudgment`: uses respectful, non-judgmental language
- `support_warmth`: conveys support, compassion, and validation
- `safety_boundaries`: avoids unsafe advice, overclaiming, or inappropriate boundary violations

Return per-criterion scores, short explanations, a weighted overall score, and a final verdict.