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

Anonymous ACL submission

Abstract

Emotional validation - explicitly acknowledging that a user's feelings make sense - has proven therapeutic value but has received little computational attention. We introduce the first three-stage framework for validation in dialogue systems, decomposing the problem 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. Finally, we benchmark GPT-4.1 nano and Llama-3.1 8B on validating response generation; few-shot prompting delivers the best balance between semantic fidelity, lexical diversity, and empathy-signal coverage, while chain-of-thought prompts increase diversity at the cost of precision.[1]

1 Introduction

Empathy is a cornerstone of effective human–computer communication because it nurtures trust, rapport, and sustained engagement in human-robot interaction (HRI) and conversational agents. Recent studies show that systems capable of modulating their empathic behavior in real time are better trusted and perceived as more helpful by users, underscoring the practical value of artificial empathy (Leite et al., 2013; Morris et al., 2018; Casas et al., 2021).

[1]All code, data, and models will be released upon acceptance.

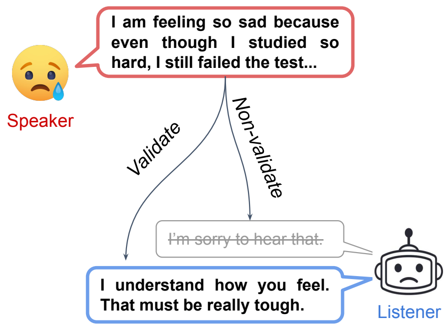

Figure 1: Examples of dialogues with validating response and non-validating response.

Research on empathetic dialogue has therefore focused on enriching response generation with a spectrum of socio-cognitive signals. Representative directions include leveraging commonsense reasoning (Sabour et al., 2022; Fu et al., 2023), extracting emotion causes (Gao et al., 2021), simulating users' emotional states (Majumder et al., 2020), and modeling speaker personality to tailor empathic style (Zhong et al., 2020; Cai et al., 2024; Fu et al., 2024). These techniques, often trained on large-scale resources such as EmpatheticDialogues (Rashkin et al., 2018), have markedly improved automatic and human judgments of empathy. Furthermore, the effectiveness of artificial empathy has been demonstrated across diverse applications, including education (Mendolia, 2023; Yusuf et al., 2025), marketing (Liu-Thompkins et al., 2022; Hanni-Vaara, 2022), and counselling (Trappey et al., 2022; Lee et al., 2023).

Yet conventional "I'm sorry to hear that" responses can still fall short for people who habitually suppress emotions or face chronic stressors. Psychotherapy literature highlights ***emotional validation*** - a communication technique to recognize, understand, and acknowledge others' emotional states, thoughts, and actions - as a deeper intervention that de-escalates negative affect and strength-

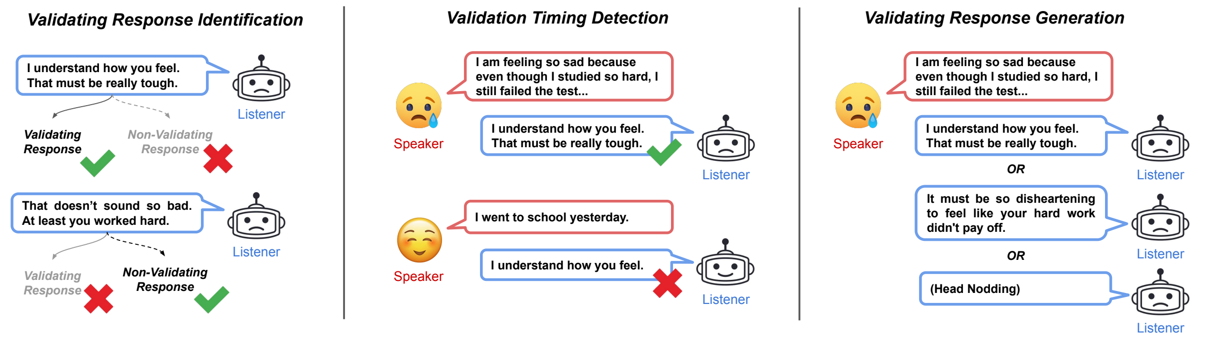

Figure 2: Emotional Validation can be subclassified into three subtasks: (1) Validating Response Identification: Identify whether a response is validating response. (2) Validation Timing Detection: Determine when should the user be validated by the system, and (3) Validating Response Generation: What kind of validating response should be generated by the system to provide emotional support to the user.

ens therapeutic alliance (Linehan, 1997). An example dialogues with validating responses and non-validating responses showed in Figure 1. Validating statements such as "It makes sense that you feel frustrated" reliably lower pain intensity in chronic-pain patients (Edlund et al., 2015), foster treatment adherence in youth mental-health journeys (Wasson Simpson et al., 2022), and predict positive emotional change in dialectical-behaviour therapy sessions (Carson-Wong et al., 2018), among other benefits (Lambie and Lindberg, 2016; Daniel, 2023).

Despite these evidences, computational work on emotional validation remains nascent. Existing studies rely on hand-crafted phrase lists, making annotation brittle and language-specific, and have so far been evaluated almost exclusively in Japanese (Pang et al., 2023, 2024b). They also treat validation as a monolithic label, leaving unanswered questions about *how* to identify validation, *when* to validate, and *what* to generate to express contextually appropriate validating responses.

To address these gaps, in this study, we make four key contributions as below:

1. **Task formalisation**: We propose a three-stage framework that decomposes emotional validation into *validating response identification*, *validation timing detection*, and *validating response generation*, providing clear sub-tasks for future benchmarks.

2. **Multilingual corpus**: We release the first open-source, semi-automatic verified multilingual dialogue corpus annotated for validation phenomena, enabling cross-lingual evaluation beyond prior Japanese-only efforts.

3. **MEGUMI**: We introduce *Multilingual Emotion-aware Gated Unit for Mutual Integration (MEGUMI)*, which fuses monolingual emotional cues with multilingual semantic representations to detect validation timing more accurately.

4. **EmoValidBench**: We present the first benchmark for validating response generation, providing evaluation scripts, LLM baselines, and empathy-oriented metrics to enable standardized comparisons across future models.

2 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, 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 2.

2.1 validating response Identification

The first requirement is to decide whether a system utterance is, in fact, validating. Mis-labeling brings risk: inappropriate reassurance or pseudo-empathy can increase user distress[2] or alienation (Breslau et al., 1998). Linguistic studies of dialogue acts provide methodological precedents, showing that

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

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 successful hybrid annotation pipelines in emotion research (Canales et al., 2016; Fonteyn et al., 2024)

2.2 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 et al., 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 the dialogue context, enabling models such as our MEGUMI architecture to decide whether the upcoming turn warrants validation.

2.3 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 Benchmark, the first benchmark for validating response generation. It pairs each user turn that requires validation with evaluation scripts that measure semantic fidelity, lexical diversity, and empathy-signal coverage (see Subsection 5.2). We report strong baselines using instruction-tuned large language models (GPT-4.1 nano, Llama-3.1 8B) under zero-shot, few-shot, and chain-of-thought prompting, which establishes a common test bed for future modeling efforts.

3 Dataset Construction

We begin with two publicly available English datasets that emphasize affective support. EmpatheticDialogues (ED) contains 24.8 k two-speaker conversations elicited via crowd workers who imagined themselves in specific emotional situations (Rashkin et al., 2018). ESConv complements ED by focusing on longer, counselor-style sessions: 1 053 multi-turn dialogues in which trained volunteers comfort users facing real-life stressors (Liu et al., 2021). To enable cross-lingual evaluation, we also implement Japanese ED (Sugiyama et al., 2021), which is a 20k two-speaker pseudo dialogue written by the crowdworker. Meanwhile, as the ESConv does not have any available Japanese version dataset, we produced one using a GPT-4-based workflow. We prompted GPT-4.1-mini[3] with professional-translator instructions, then post-edited any literal or culturally awkward renderings. Combining the English and Japanese versions of ED and ESConv, we formed the main dataset used in this study, Multilingual-Empathetic Dialogue Emotional Support Conversation (M-EDESConv) dataset.

To further evaluate our task in a spoken dialogue scenario, We add the TUT Emotional Storytelling Corpus (TESC) (Oishi et al., 2021), a Japanese two-party, multi-turn spoken-dialogue dataset in which close friends recount personal experiences under eight Plutchik emotion prompts (Plutchik, 2001). TESC comprises 247 sessions ($\approx$ 9.2 h). We translate the whole dialogue into everyday English with the same GPT-4 workflow used for Japanese ESConv, yielding multilingual transcripts suitable for the experiments in this study. We refer the multilingual version of this dataset as M-TESC in this study. The summary of all six datasets in this study is presented in the table in Appendix A, and the translation prompts used are shown in Appendix C.1 and C.2.

3.1 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 inspired by large-scale emotion datasets (Yang

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

et al., 2012). In the first stage, we manually labeled roughly 3000 utterances per source ($\approx 2\%$ 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. We fine-tune the *xlm-roberta-large* model (Conneau et al., 2019a) 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. The classifier achieves a macro-average F1 score of 85.28 and an F1 score of 86.67 for the minority (validation) class on the manually annotated test set. As part of our ablation study, we compare this model against several baselines, including a random baseline, multilingual BERT (mBERT)[4], LLaMA 3.1 8B[5], and GPT-4.1-nano[6]. The comparative results are presented in Table 7 in Appendix B.

Using the trained classifier, we proceed to annotate the full dataset. 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 thresholds of 0.90 for ED and 0.95 for ESConv to align the automated annotations with the original distribution.

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 6.

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

Table 1: Distribution of dataset in validation and non-validation

4 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 two complementary information streams - what the user is saying (semantic content) and how they are feeling (affective cues). The proposed Multilingual Emotion-aware Gated Unit for Mutual Integration (MEGUMI) architecture, shown in Diagram 3, 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.

4.1 Validation Timing Detection Modal

Semantic backbone The core encoder of MEGUMI is XLM-RoBERTa-large (1024 h units), chosen for its strong zero-shot transfer across 100+ languages. We freeze its parameters to preserve multilingual lexical knowledge and to curb computational cost, passing only the [CLS] token representation to downstream modules.

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[7] 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[8]. The emotion [CLS] vector from the relevant channel is concatenated with the frozen semantic [CLS].

Emotion-enhanced multilingual attention As both English and Japanese cues are present in the training batches, we apply an emotion-enhanced multilingual attention block inspired by the Multimodal Transformer (Tsai et al., 2019). The module projects one lingual's concatenated vector as query

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

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

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

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

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

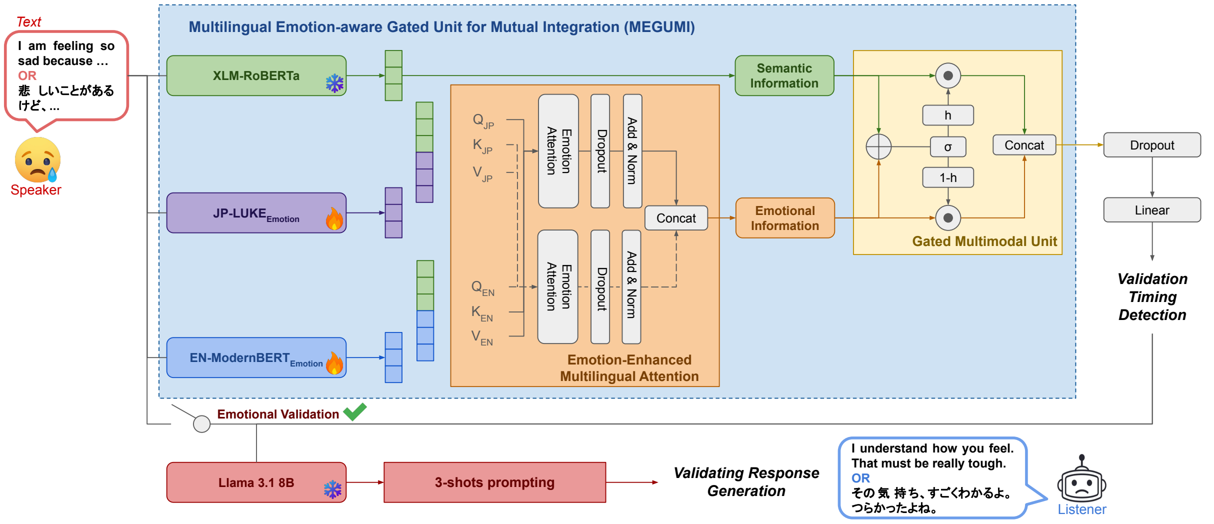

Figure 3: Overall proposed architecture in this study. We proposed a Multilingual Emotion-aware Gated Unit for Mutual Integration (MEGUMI) for the Validation Timing Detection Task.

and the other as key/value, computes scaled dot-product attention, and returns two residual-normed streams. With the existence of both Japanese and English emotion channels, it allows MEGUMI to learn latent alignments between semantic patterns (e.g., "I lost my job") and affective priors (e.g., fear vs. anger) from a multilingual perspective.

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 and multimodal emotion recognition (Arevalo et al., 2017). A sigmoid gate h decides, per sample, how much of the emotion-projected vector versus the semantic-projected vector to pass forward, producing a 768-d fused representation:

$$z = \mathbf{h} z_{semantic} + (\mathbf{1} - \mathbf{h}) z_{emotion}$$

The fused vector is fed to a dropout–linear softmax head for the binary labels validate/ non-validate. To counter class imbalance we weight the cross-entropy loss by inverse-frequency factors computed from the training split. Reproducibility is ensured via deterministic seeds, and all components except the text encoder are fine-tuned.

4.2 Validation Timing Detection Result

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.

4.2.1 Baselines

We benchmarked against (i) a random classifier, (ii) two fine-tuned multilingual language models, mBERT and XLM-RoBERTa (Conneau et al., 2019a), and (iii) instruction-tuned large language models: Llama-3.1.1 8B-Instruct and GPT-4.1 nano, each in zero-shot and 3-shot prompt-engineering regimes, refer to Appendix C.3. Larger 70B Llama variants were excluded owing to real-time memory constraints.

4.2.2 Evaluation Metrics

In the context of everyday conversation, when a system chooses to validate matters more than how often it does so. Consequently, we treat target-class precision - the proportion of predicted validate turns that truly warrant validation - as the principal metric. 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 averages across both classes to ensure that performance on the majority non-validate class is not neglected.

	M-EDESConv						M-TESC					
	Marco Average			Target Class			Marco Average			Target Class		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
Random Baseline	50.20	50.21	49.23	36.45	50.35	42.30	50.26	50.29	49.03	34.41	50.02	40.77
mBERT	62.07	62.63	59.10	45.16	74.15	56.19	53.28	53.39	53.29	38.29	41.35	39.76
XLM-RoBERTa	62.78	63.13	59.03	45.19	76.96	57.01	55.29	55.74	55.10	40.24	50.00	44.60
Llama 3.1 8b												
- *Zero-shot*	57.33	52.81	37.18	37.72	93.75	53.79	56.84	54.04	40.76	36.31	89.58	51.68
- *3-shot*	57.36	52.40	35.69	37.49	**94.84**	53.73	55.93	51.38	31.53	34.81	**96.29**	51.13
GPT 4.1 Nano												
- *Zero-shot*	58.42	57.74	51.68	41.43	79.25	54.41	56.71	57.31	54.04	39.95	66.46	49.90
- *3-shot*	58.87	56.39	46.75	40.02	87.04	54.83	**58.34**	**57.65**	50.14	39.01	80.93	**52.65**
MEGUMI (Ours)	**63.94**	**65.02**	**63.71**	**51.07**	66.11	**57.62**	56.86	57.36	**56.89**	**41.44**	48.70	44.78

Table 2: Results of validation timing detection task in multilingual setting [%]

	Criteria			Macro Average			Target Class		
	EE	EEMA	GMU	Precision	Recall	F1-Score	Precision	Recall	F1-Score
XLM-RoBERTa	-	x	x	56.78	55.43	47.29	39.60	82.42	53.50
+ Mono-EN	EN	x	x	57.21	57.40	57.27	45.10	48.23	46.61
+ Mono-JP	JP	x	x	57.99	58.63	56.86	43.97	62.88	51.75
+ Multi-Concat	Both	x	x	62.73	63.34	59.80	46.75	73.86	57.26
+ Multi-EEMA	Both	v	x	62.83	63.85	62.48	49.70	65.31	56.45
MEGUMI (Ours)	Both	v	v	63.94	65.02	63.71	51.07	66.11	57.62

Table 3: Ablation results for validation-timing detection, showing the impact of adding Emotion Embedding (EE), Emotion-Enhanced Multilingual Attention (EEMA), and the Gated Multimodal Unit (GMU) on both macro-average and target-class precision, recall, and F1-score [%].

4.2.3 Results

Table 2 shows that our MEGUMI lifts precision to 51.07% - a relative improvement of +5.88% over the best baseline (45.19% from the XLM-RoBERTa) on the target validation class. Moreover, the MEGUMI model attains a macro-F1 of 63.71%, exceeding the strongest traditional baseline (mBERT) by +4.61 and the XLM-RoBERTa by +4.68. While GPT-4.1 nano and Llama 3.1 8B exhibit recall above 87% in 3-shot mode, their precision collapses to 40%, corroborating evidence that zero-shot LLMs over-predict minority classes. MEGUMI therefore offers a superior balance, validating when appropriate rather than validating always. As an additional understanding on the model performance, we also reported the result for monolingual tasks in Appendix D.

Without additional fine-tuning, the same checkpoints were evaluated on M-TESC, a spontaneous spoken-dialogue corpus. All systems suffer domain drift, yet MEGUMI remains top with 56.9% macro F1 and the highest validation-precision (41.4%). These findings underline the benefit of cross-lingual pre-training for speech-text transfer, observed previously for XLM-RoBERTa-style encoders (Conneau et al., 2019b).

4.2.4 Ablation Study

To disentangle architectural choices we incrementally removed (i) Emotion Embeddings (EE), (ii) the Emotion-Enhanced Multilingual Attention (EEMA), and (iii) the Gated Multimodal Unit (GMU). The overall ablation study result shown in Table 3.

From the ablation study, we found that adding either monolingual emotion channel (+Mono-EN, +Mono-JP) raises target Precision by 4-7% over text-only, confirming that language-specific affect encoders inject useful priors. In addition, simple concatenation of both channels adds a further +2.5%, but replacing it with EEMA yields an additional +2.7% precision by aligning semantics and affect bidirectionally. Last but not least, incorporating the GMU lifts precision another +1.4%, showing that dynamic gating helps the model suppress noisy or redundant cues.

5 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. This section details the benchmark design, experimental protocol, automatic met-

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

rics, and baseline results.

5.1 Benchmark construction

From the M-EDESConv corpus we extract every user utterance whose gold timing label is validate = true. 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[9], while Japanese inputs are segmented by MeCab + UniDic[10] 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), and Zero-shot CoT ("Let's think step by step" preamble) (Kojima et al., 2022), see Appendix C.4. No model parameters were updated.

5.2 Evaluation metrics

To comprehensively assess validating response generation, we employ a suite of complementary metrics that capture semantic fidelity, lexical diversity, and empathetic signal presence.

Semantic Fidelity. We utilize BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2019) to evaluate the semantic alignment between generated responses and reference texts. BLEU measures n-gram overlap, providing insights into surface-level similarity, while BERTScore leverages contextual embeddings to assess deeper semantic correspondence, reporting precision, recall, and F1 scores.

Lexical Diversity. To quantify the diversity of generated language, we calculate Distinct-1 (D1) and Distinct-2 (D2) (Li et al., 2015) , which represent the ratios of unique unigrams and bigrams to the total number of tokens. Higher values indicate a broader range of lexical choices, reflecting more varied and potentially more engaging responses.

Empathetic Signal Coverage. Inspired by prior work on empathetic communication (Lee et al., 2022; Fu et al., 2024), we incorporate three categories of empathetic signals: *IP* (interpretations), *EX* (explorations), and *ER* (emotional reactions). Specifically, *IP* represents expressions of acknowledgments or understanding of the interlocutor's emotion or situation. *EX* represents expressions of active interest in the interlocutor's situation; *ER* represents expressions of explicit emotions. For the English version, we follow the official annotation schema[11] and apply three RoBERTa (Liu et al., 2019) based classifiers to identify whether a response implies a certain signal individually. For Japanese, we translate the English corpus using the NLLB-200-3.3B model (Costa-Jussà et al., 2022)[12], followed by classification with LUKE[13] based models. For the multilingual setting, we use XLM-RoBERTa (Conneau et al., 2019a) followed by two linear layers as a unified classifier across languages. Evaluations results of the empathetic signal predictors is summarized in the Table 4.

5.3 Results

Table 5 presents scores averaged over five seeds.

Semantic fidelity: 3-shot GPT-4.1 leads on BLEU for both languages (13.31 in English, 22.9 in Japanese, 17.8 in Multilingual) and attains the top BERT-F across splits, suggesting that minimal in-context examples suffice to anchor the model to reference phrasing.

Lexical diversity: CoT increases Distinct-2 by +2–3% relative to 3-shot in every language, corroborating prior reports that reasoning traces encourage richer wording (Wei et al., 2022). However, BLEU falls by ≈ 3%, indicating a diversity–fidelity tension.

Empathy coverage: Few-shot prompts yield the best balanced profile: GPT-4.1 3-shot scores 77.6 IP and 60.8 ER in Japanese, outperforming zero-shot recall without the over-generation seen in CoT. Llama-3.1 trails by 2% in ER but matches IP on English.

[9] https://github.com/luismsgomes/mosestokenizer

[10] https://taku910.github.io/mecab/

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

[12] https://huggingface.co/facebook/nllb-200-3.3B

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

	Semantics				Diversity		Empathy		
	BLEU	$BERT_{Pre}$	$BERT_{Rec}$	$BERT_{F1}$	D1	D2	ER	IP	EX
English									
Llama 3.1 8b									
- *Zero-shot*	13.20	87.74	89.04	88.36	4.35	23.94	62.37	76.11	68.73
- *3-shot*	13.18	87.86	89.11	88.45	4.51	24.93	62.01	76.73	67.59
GPT 4.1 Nano									
- *Zero-shot*	12.72	87.78	88.92	88.33	5.02	25.60	62.73	75.65	69.10
- *3-shot*	13.32	87.89	89.22	88.53	4.59	22.88	60.71	75.52	67.75
- *CoT*	12.70	87.34	88.96	88.13	4.77	27.01	62.37	75.79	67.86
Japanese									
Llama 3.1 8b									
- *Zero-shot*	18.22	88.17	90.24	89.15	5.23	23.67	54.73	72.07	61.38
- *3-shot*	19.76	89.22	89.98	89.55	5.79	24.97	54.67	74.27	61.10
GPT 4.1 Nano									
- *Zero-shot*	19.84	88.52	90.66	89.54	4.56	18.40	53.34	76.77	61.62
- *3-shot*	22.92	89.82	90.60	90.16	5.00	20.33	57.70	77.60	60.83
- *CoT*	13.91	86.90	89.47	88.14	7.00	27.74	49.90	78.09	61.35
Multilingual									
Llama 3.1 8b									
- *Zero-shot*	15.56	87.82	89.60	88.67	4.88	23.96	52.90	70.94	67.14
- *3-shot*	15.80	89.11	89.03	89.03	5.52	25.97	54.66	71.54	69.59
GPT 4.1 Nano									
- *Zero-shot*	16.14	88.06	89.75	88.87	4.80	22.37	51.95	73.17	69.97
- *3-shot*	17.77	88.61	89.97	89.26	4.58	21.42	52.32	72.29	68.76
- *CoT*	13.44	87.15	89.19	88.13	4.83	24.53	50.72	73.30	71.21

Table 5: Validating response generation results across English, Japanese, and multilingual settings [%]

6 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. This model achieves substantial gains in precision and macro F_1 over strong baselines, and generalises effectively from written chat to spontaneous speech.

We further explored the capabilities of off-the-shelf LLMs in generating validating responses, showing that careful prompt design - particularly few-shot exemplars - yields the best trade-off between surface overlap, lexical diversity, and empathy-signal coverage. Our findings underline that while LLMs can produce plausible validating utterances, the balance between creative expression and clear acknowledgment of emotion remains sensitive to the choice of prompting strategy.

Looking ahead, our work opens multiple directions for future research. First, extending emotional validation to a broader range of languages - including dialectal and culturally nuanced variants - would enhance the applicability of validation-aware systems across diverse global populations. Second, incorporating non-verbal modalities such as prosody, gesture, and facial expression, alongside multimodal pretraining, could enable more naturalistic and contextually appropriate validation behaviours (Linehan, 1997). Finally, deploying validation-capable agents in real-world interactive settings through embodied conversational agents or robots. Such deployments would enable the assessment of perceived authenticity, trustworthiness, and therapeutic benefit across a variety of scenarios, including emotional support (Erel et al., 2022), interviews (Pang et al., 2024a), and attentive listening tasks (Lala et al., 2017). Further, agent embodiment and appearance - ranging from screen-based virtual characters (Lee, 2023) to physically humanoid androids (Kawahara, 2019; Pang et al., 2025) - may modulate user perceptions of validation and should be systematically explored.

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. Third, our validating response generation experiments rely exclusively on automatic metrics and empathy-signal classifiers; without human judgements of perceived empathy, naturalness, and user satisfaction, we cannot fully gauge the real-world effectiveness or potential unintended effects of generated replies.

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.

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*.

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. *arXiv preprint arXiv:2204.12191*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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.

Hadas Erel, Denis Trayman, Chen Levy, Adi Manor, Mario Mikulincer, and Oren Zuckerman. 2022. Enhancing emotional support: The effect of a robotic object on human–human support quality. *International Journal of Social Robotics*, 14(1):257–276.

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.

Päivi Hanni-Vaara. 2022. Human or nonhuman agent?: Experiences of empathy in a digital customer tourism journey. In *Empathy and Business Transformation*, pages 231–245. Routledge.

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.

Tatsuya Kawahara. 2019. Spoken dialogue system for a human-like conversational robot erica. In *9th International Workshop on Spoken Dialogue System Technology*, pages 65–75. Springer.

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.

Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. 2017. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 127–136.

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.

Akinobu Lee. 2023. MMDAgent-EX.

Yoon Kyung Lee, Yoonwon Jung, Gyuyi Kang, and Sowon Hahn. 2023. Developing social robots with empathetic non-verbal cues using large language models. *arXiv preprint arXiv:2308.16529*.

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. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Marsha M Linehan. 1997. Validation and psychotherapy.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

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*.

Yuping Liu-Thompkins, Shintaro Okazaki, and Hairong Li. 2022. Artificial empathy in marketing interactions: Bridging the human-ai gap in affective and social customer experience. *Journal of the Academy of Marketing Science*, 50(6):1198–1218.

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. *arXiv preprint arXiv:2010.01454*.

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.

Tracy A Mendolia. 2023. Empathetic chatbot: Enhancing medical education with artificial intelligence. *Immersive Learning Research-Practitioner*, pages 86–90.

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, Divesh Lala, Mikey Elmers, Koji Inoue, and Tatsuya Kawahara. 2024a. Human-like embodied ai interviewer: Employing android erica in real international conference. *arXiv preprint arXiv:2412.09867*.

Zi Haur Pang, Yahui Fu, Divesh Lala, Mikey Elmers, Koji Inoue, and Tatsuya Kawahara. 2025. Does the appearance of autonomous conversational robots affect user spoken behaviors in real-world conference interactions? *arXiv preprint arXiv:2503.13625*.

Zi Haur Pang, Yahui Fu, Divesh Lala, Keiko OCHI, Koji INOUE, and Tatsuya KAWAHARA. 2023. Prediction of validating response from emotional storytelling corpus. In *人工知能学会全国大会論文集 第 37 回 (2023)*, pages 2O5OS2a03–2O5OS2a03. 一般社団法人人工知能学会.

Zi Haur Pang, Yahui Fu, Divesh Lala, Keiko Ochi, Koji Inoue, and Tatsuya Kawahara. 2024b. Acknowledgment of emotional states: Generating validating responses for empathetic dialogue. *arXiv preprint arXiv:2402.12770*.

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. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Shanequa S Roscoe-Nelson, MHA Silvera PhD, A Geoffrey, and 1 others. 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.

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.

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.

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. 2021. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. *Preprint*, arXiv:2109.05217.

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

Amy JC Trappey, Aislyn PC Lin, Kevin YK Hsu, Charles V Trappey, and Kevin LK Tu. 2022. Development of an empathy-centric counseling chatbot system capable of sentimental dialogue analysis. *Processes*, 10(5):930.

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.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. *arXiv preprint arXiv:2012.04080*.

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.

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.

Habeeb Yusuf, Arthur Money, and Damon Daylamani-Zad. 2025. Pedagogical ai conversational agents in higher education: a conceptual framework and survey of the state of the art. *Educational technology research and development*, pages 1–60.

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

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

Appendix

A Base Corpora

Table 6 summarises the size, modality and interactional profile of the three corpora that constitute our base dataset. EmpatheticDialogues (ED) offers the broadest coverage with 24.8 k English and 20 k 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.

B Validating Response Identification

Table 7 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 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 3.1). 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.

C Prompts

C.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.

C.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.

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 6: Statistics of the English and Japanese splits of the three base corpora employed in this study

	Macro Average			Target Class		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Random Baseline	50.00	50.00	48.38	32.45	50.13	39.40
mBERT	74.33	74.51	74.30	70.24	76.32	73.15
Llama 3.1 8b						
- Zero-shot	54.85	53.39	41.04	34.37	85.61	49.04
- 3-shot	61.03	52.77	32.18	33.82	**97.88**	50.27
GPT 4.1 Nano						
- Zero-shot	64.49	65.04	57.98	42.71	85.19	56.89
- 3-shot	67.21	69.57	66.57	50.36	74.60	60.13
XLM-RoBERTa	**86.42**	**85.30**	**85.28**	**80.66**	93.64	**86.67**

Table 7: Results of validating response identification task (automatic annotation) [%]

C.3 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.

C.4 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.

D Monolingual validation timing Detection Results

Table 8 compares English-only and Japanese-only models on the M-EDESConv test split as well as on the out-of-domain spoken M-TESC corpus. Three observations stand out.

Supervised PLMs outperform prompted LLMs on written chat. For both languages, fine-tuned XLM-RoBERTa offers the best macro F1 on M-EDESConv (62.2% EN, 60.8% JA) and achieves the highest validation-class F1 among the baseline encoders. The gains come mainly from higher precision: 47–48%, compared with 42–45% for BERT variants. This confirms that domain-specific fine-tuning is still advantageous when timing accuracy is critical.

Prompted LLMs trade precision for recall. Instruction-tuned models such as GPT-4.1 nano and Llama-3 8 B show markedly different error profiles. With three in-context examples, both LLMs boost recall beyond 95% in English and Japanese, but precision drops to the mid-30% range, driving overall F1 well below that of supervised PLMs. Zero-shot prompting moderates this effect, yet pre-

	M-EDESConv						M-TESC					
	Marco Average			Target Class			Marco Average			Target Class		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
English												
Random Baseline	50.47	50.51	49.40	35.83	50.59	41.95	50.14	50.15	48.84	34.29	50.27	40.76
BERT	62.75	63.54	59.74	45.16	74.15	56.19	53.28	53.39	53.29	38.29	41.35	39.76
ModernBERT	62.07	62.63	59.10	44.94	75.02	56.27	53.05	53.31	52.80	37.67	46.77	41.73
XLM-RoBERTa	**63.51**	**64.74**	**62.22**	47.24	70.38	**56.57**	55.22	55.73	54.83	39.94	52.09	45.21
Llama 3.1 8b												
- *Zero-shot*	57.43	54.09	40.90	37.55	90.54	53.08	57.95	57.07	49.07	38.55	81.64	**52.37**
- *3-shot*	58.52	52.24	33.75	36.45	**96.47**	52.91	58.40	52.58	34.15	35.41	**95.93**	51.73
GPT 4.1 Nano												
- *Zero-shot*	59.41	59.71	55.36	42.53	74.43	54.13	55.44	56.03	53.74	39.29	60.80	47.73
- *3-shot*	58.92	59.61	56.69	43.19	68.49	52.98	55.37	55.96	53.70	39.24	60.57	47.62
MEGUMI (Ours)	61.76	62.71	61.67	**48.40**	60.99	53.97	**58.65**	**58.28**	**58.41**	**45.07**	41.56	43.24
Japanese												
Random Baseline	49.86	49.85	49.01	37.23	50.09	42.71	50.38	50.42	49.22	34.54	49.77	40.77
BERT	61.51	61.43	57.35	44.83	76.57	56.61	56.60	57.33	55.67	40.90	58.56	48.16
ModernBERT	61.77	59.17	51.15	42.02	86.48	56.67	55.01	55.35	51.51	38.18	66.92	48.62
XLM-RoBERTa	62.25	62.98	60.76	47.52	69.73	56.56	54.60	55.00	54.39	39.45	49.05	43.73
Llama 3.1 8b												
- *Zero-shot*	56.43	50.87	30.98	37.79	97.82	54.51	58.48	51.26	29.75	34.74	98.21	51.33
- *3-shot*	61.51	50.69	29.22	37.69	99.44	54.67	**64.74**	50.66	27.07	34.46	**99.85**	51.23
GPT 4.1 Nano												
- *Zero-shot*	55.97	54.70	47.03	40.28	81.51	53.91	57.68	57.83	52.48	39.63	74.68	51.78
- *3-shot*	58.25	53.61	39.67	39.28	**99.94**	55.22	60.02	56.24	43.38	37.53	91.41	**53.21**
MEGUMI (Ours)	**63.67**	**64.22**	**61.20**	**48.83**	75.87	**59.42**	57.49	**58.37**	**57.09**	**41.35**	55.84	47.51

Table 8: Results of validation timing detection task in monolingual setting [%]

cision remains 8–10 pp lower than XLM-R. Thus, without additional control signals, LLMs tend to "over-validate," echoing the multilingual findings in the main paper.

MEGUMI narrows the domain gap. When restricted to a single language, our MEGUMI detector retains its advantage on spontaneous speech. On M-TESC it delivers the highest macro F1 in both English (58.4%) and Japanese (57.1%), outperforming all monolingual baselines by 2–3 pp despite only moderate gains on M-EDESConv. The improvement stems primarily from better precision in the noisier spoken setting (45.1% EN, 41.4% JA), suggesting that MEGUMI's gated fusion of semantic and affective cues remains beneficial even without cross-lingual context.