Multilingual, Not Multicultural: Uncovering the Cultural Empathy Gap in LLMs through a Comparative Empathetic Dialogue Benchmark

Woojin Lee[1,2*] **Yujin Sim**[2*] **Hongjin Kim**[1] **Harksoo Kim**[2†]

[1]Electronics and Telecommunications Research Institute, Republic of Korea
[2]Konkuk University, Republic of Korea
{writerwoody, drjin}@etri.re.kr[1]
{simuzin, nlpdrkim}@konkuk.ac.kr[2]

Abstract

Large Language Models (LLMs) demonstrate remarkable multilingual capabilities, yet it remains unclear whether they are truly multicultural. Do they merely process different languages, or can they genuinely comprehend the unique cultural contexts embedded within them? This study investigates this critical question by examining whether LLM's perception of emotion and empathy differs across linguistic and cultural boundaries. To facilitate this, we introduce the Korean Empathetic Dialogues (KoED), a benchmark extending the English-based EmpatheticDialogues (ED) dataset. Moving beyond direct translation, we meticulously reconstructed dialogues specifically selected for their potential for cultural adaptation, aligning them with Korean emotional nuances and incorporating key cultural concepts like 'jeong' and 'han' that lack direct English equivalents. Our cross-cultural evaluation of leading multilingual LLMs reveals a significant "cultural empathy gap": models consistently underperform on KoED compared to ED, struggling especially with uniquely Korean emotional expressions. Notably, the Korean-centric model, EXAONE, exhibits significantly higher cultural appropriateness. This result provides compelling evidence that aligns with the "data provenance effect", suggesting that the cultural alignment of pre-training data is a critical factor for genuine empathetic communication. These findings demonstrate that current LLMs have cultural blind spots and underscore the necessity of benchmarks like KoED to move beyond simple linguistic fluency towards truly culturally adaptive AI systems.[1]

1 Introduction

The rapid advancement of Large Language Models (LLMs) has ushered in an era of unprecedented multilingual capabilities. Following the breakthrough of GPT-4 (Achiam et al., 2023), models such as Claude 3 (Anthropic, 2024) and Llama 3 (Dubey et al., 2024) have demonstrated the ability to seamlessly switch between dozens of languages, producing fluent text that often rivals native speakers. Yet, a fundamental question remains: Does multilingual competence equate to multicultural understanding? Can an LLM trained predominantly on English data truly grasp the cultural nuances embedded in Korean conversations, or does it merely translate linguistic forms without internalizing cultural meaning?

This distinction between linguistic fluency and cultural comprehension becomes particularly evident in emotionally charged contexts. Cultural differences fundamentally shape how emotions are expressed, interpreted, and responded to in social interactions. As Jin et al. (2024) demonstrated, the same social situation can evoke entirely different cultural stereotypes and assumptions between Korean and American contexts. For instance, while drug use is stereotypically associated with low socioeconomic status in the United States, it correlates with high socioeconomic status in Korea—a cultural inversion that profoundly affects how empathetic responses should be formulated. This gap is also evident in subtle conversational acts. For instance, an empathetic response to someone's birthday in an American context, such as "I hope you get lots of gifts!", might be naturally replaced in a Korean context with the question, "Did you have seaweed soup this morning?". This question is not a literal inquiry about a meal but a deeply empathetic gesture that acknowledges the birthday and invokes a shared cultural tradition. An LLM lacking this cultural grounding risks misinterpreting the empathetic subtext, treating it as a literal, out-of-place question about soup and generating an inappropriate response. Such cultural variations extend beyond stereotypes to the very fabric of emotional expression and empathetic communication.

[*]Main contributors
[†]Corresponding author
[1]https://github.com/KUNLP/KoED

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 791–809

Current benchmarks for evaluating empathetic AI predominantly rely on English-centric datasets, most notably the EmpatheticDialogues (ED) dataset (Rashkin et al., 2019). While valuable for assessing empathy in US-centric cultural contexts, these resources fail to capture the rich tapestry of emotional expression across cultures. This limitation raises critical concerns: If we evaluate AI empathy solely through a Western cultural lens, are we inadvertently creating systems that perpetuate cultural blindness? As AI assistants become integral to global communication in counseling (Hu et al., 2025), education (Huang et al., 2025), and social support (Qiu and Lan, 2025), their ability to recognize and respond to culturally-specific emotional cues becomes not just desirable but essential.

To address this gap, we present Korean Empathetic Dialogues (KoED), a culturally-aware benchmark that extends beyond translation to capture authentic Korean emotional expression. Our approach mirrors the methodology pioneered by Jin et al. (2024), who demonstrated that direct translation of culturally-rooted benchmarks often fails to preserve essential cultural contexts. Following their lead, we meticulously analyzed the ED dataset and selected 1,280 dialogue samples (40 samples for each of 32 emotion categories) that could be meaningfully adapted to Korean cultural contexts while preserving their emotional core. Samples deeply rooted in American-specific cultural scenarios were excluded, ensuring that our adaptations remain authentic rather than forced.

Beyond adaptation, we recognized that certain emotional concepts central to Korean culture have no direct parallel in English-based datasets. To address this, we handcrafted an additional 80 dialogue samples (40 each) for two uniquely Korean emotions: 'jeong' (정; 情)—a deep emotional bond that develops through shared experiences and mutual care—and 'han' (한; 恨)—a complex emotion encompassing collective suffering, resilience, and unfulfilled longing rooted in Korean historical experience (Ka, 2010). These untranslatable emotions serve as critical test cases for evaluating whether LLMs can truly understand culture-specific emotional landscapes.

Our comprehensive cross-cultural evaluation reveals what we term the "cultural empathy gap" in current LLMs. Through systematic experiments comparing model performance on both ED and KoED datasets, we uncover three key findings: First, all evaluated multilingual models demonstrate significantly degraded performance on KoED compared to ED, with particularly pronounced difficulties in recognizing and responding to culture-specific emotions. This performance gap persists across model architectures and sizes, suggesting systemic limitations in how current LLMs process cultural context. Second, the Korean-centric LLM EXAONE (LG AI Research, 2024) exhibits markedly superior cultural appropriateness in its responses, despite not necessarily outperforming other models on general language tasks. This finding parallels observations in cross-cultural bias research and highlights the profound impact of culturally-aligned pre-training data on model behavior. Third, our analysis of culture-specific emotion recognition reveals that even with explicit descriptions, models struggle to perceive emotions like 'jeong' and 'han', achieving recognition rates of no more than 17.7% even under optimal conditions. This suggests that current training paradigms fail to instill the deep cultural knowledge necessary for authentic empathetic communication.

As Omdahl (2014) and Ma et al. (2020) noted, empathy is fundamental to numerous AI applications in mental health, education, and social support. A system that misinterprets cultural emotional cues risks providing not just inadequate but potentially harmful responses. Our work thus contributes to a growing body of research that challenges the assumption that multilingual capability automatically confers multicultural competence.

This paper makes three primary contributions:

- We introduce KoED, a culturally-adapted empathetic dialogue benchmark comprising 1,360 dialogues that authentically reflect Korean emotional expression, including untranslatable culture-specific emotions.
- We provide empirical evidence of the "cultural empathy gap" through comprehensive evaluation of LLMs developed across diverse linguistic and cultural contexts, revealing consistent underperformance on culturally-adapted KoED dataset.
- We demonstrate that cultural alignment in pre-training data significantly impacts empathetic response generation, with implications for developing truly multicultural AI systems.

2 Related Work

2.1 Empathetic Dialogue: The Hidden Cultural Assumptions

The development of empathetic dialogue systems has been largely benchmarked on the English-based ED dataset (Rashkin et al., 2019), which has served as the de facto standard for both response generation and emotion recognition tasks. These systems are often grounded in psychological theories distinguishing between affective empathy (sensing another's feelings) and cognitive empathy (understanding the reason for those feelings) (Davis, 1983; Spaulding, 2017). While language models have achieved impressive performance in both dimensions—from mimicking emotional expressions (Majumder et al., 2020) to leveraging commonsense reasoning (Sabour et al., 2022; Zhou et al., 2022)—their development has been overwhelmingly confined to the US-centric cultural context in which ED was created.

The critical oversight in this line of research is the implicit assumption that cognitive empathy is culturally universal. Cognitive empathy requires interpreting situations based on a shared cultural understanding—a framework composed of culturally-specific scripts, norms, and implicit assumptions. As demonstrated in the context of social bias research (Jin et al., 2024), the interpretation of the same social situation can be inverted across cultures. Models trained solely on American contexts therefore lack the cultural "cognitive toolkit" to accurately engage in empathetic dialogue with users from different backgrounds. This inherent cultural specificity of cognitive empathy reveals that current benchmarks, by relying exclusively on ED, measure whether models have learned a specific set of empathetic scripts, rather than possessing a generalizable multicultural capability.

It is precisely this challenge that makes empathetic dialogue a uniquely powerful lens for investigating the multilingual-multicultural divide. While prior work has probed this gap through lenses such as explicit cultural knowledge (Zhou et al., 2024), translation and adaptation quality (Cao et al., 2024), or word-level semantic alignment (Dai et al., 2025), and some have benchmarked cross-cultural emotion understanding directly (Belay et al., 2025), our work argues that empathy offers a more rigorous test. It moves beyond assessing factual recall (what a model knows) or recognition (what emotion it perceives) to evaluating behavioral competence (how a model acts in nuanced social interactions). This is a high-stakes evaluation; in critical applications like AI-driven counseling and social support, a failure in cultural empathy is not a minor error but a potentially harmful one. Moreover, generating an appropriate empathetic response is a holistic task, requiring the model to synthesize cultural scripts, emotional signals, and linguistic fluency, thereby providing a more comprehensive measure of multicultural understanding than isolated metrics offer.

2.2 From Translation to Cultural Reconstruction

The inadequacy of direct translation for preserving cultural validity has catalyzed a paradigm shift in cross-cultural NLP. Ponti et al. (2020) showed that culturally-specific scenarios often lose meaning when literally translated, particularly in emotion-related tasks where cultural scripts fundamentally shape appropriate responses. This challenge has led to systematic cultural reconstruction methods.

Jin et al. (2024) pioneered this approach for Korean with their KoBBQ dataset, categorizing content as SIMPLY-TRANSFERRED, CONTEXT-MODIFIED, or SAMPLE-REMOVED. Their finding that approximately 30% of American scenarios had no Korean equivalent powerfully underscores the limitations of translation-based approaches. We extend this reconstruction paradigm to empathetic dialogue through KoED. We adopted this dual approach—selectively adapting transferable content and handcrafting new content for culture-specific concepts like 'jeong' and 'han' (Ka, 2010)—to ensure our benchmark rigorously tests for authentic cultural understanding, not artifacts of translation.

2.3 The Multilingual-Multicultural Divide in LLM Evaluation

Evaluating LLMs' nuanced capabilities presents formidable challenges, particularly for culturally-bound tasks like empathy. While automated evaluation methods have emerged (Liu et al., 2023a; Ye et al., 2024), even state-of-the-art models struggle with zero-shot empathy evaluation (Xu and Jiang, 2024). This difficulty is magnified when considering multicultural competence—an under-explored dimension that challenges the assumption that cross-lingual transfer ability confers multicultural understanding (Ahuja et al., 2023).

Recent research reveals a fundamental disconnect: LLMs trained predominantly on Western web data project Western-centric values even when oper-

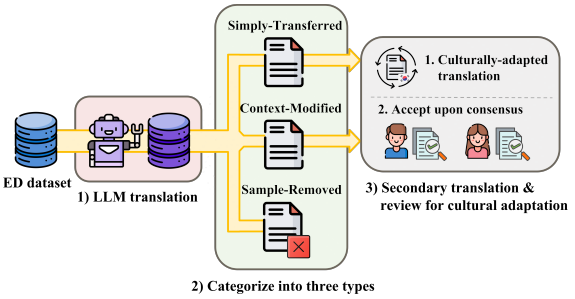

Figure 1: Overview of the KoED dataset construction process. The process consists of three main steps: (1) initial translation of the ED dataset using a large language model (LLM), (2) categorization of the translated content into three types—SIMPLY-TRANSFERRED, CONTEXT-MODIFIED, or SAMPLE-REMOVED, and (3) secondary translation and review for cultural adaptation. Final translations are accepted based on consensus among human reviewers.

ating in non-Western languages (Wang et al., 2024). This "data provenance effect" suggests that LLMs' worldview and cultural alignment are profoundly shaped by their pre-training data's geographic and cultural origins (Bisk et al., 2020). Our work investigates whether this effect extends to empathetic communication. Through the controlled comparison of ED and KoED, we provide the first systematic evaluation of whether the "cultural empathy gap", driven by data provenance, limits current LLMs' ability to engage in authentic cross-cultural empathetic dialogue.

3 KoED Dataset

Building on the cultural reconstruction paradigm established in cross-cultural NLP research, we present KoED—a benchmark designed to test whether LLMs can truly understand Korean emotional expression beyond surface-level translation.

3.1 Cultural Reconstruction and Annotation Pipeline

Our primary goal is to create a benchmark that allows for a fair and controlled comparison of empathetic capabilities across different cultural contexts. To achieve this, we developed a comprehensive pipeline (summarized in Figure 1) that integrates cultural reconstruction with nuanced emotional annotation.

The process began with an initial translation of the ED dataset (Rashkin et al., 2019) using GPT-4o. Following this, the first two authors, both native Koreans with expertise in both cultures, categorized each dialogue based on the cultural adaptation methodology pioneered by Jin et al. (2024): SIMPLY-TRANSFERRED, CONTEXT-MODIFIED, or SAMPLE-REMOVED.

For the dialogues selected for inclusion in KoED, we deliberately employed a "negotiated agreement" methodology performed by the two co-first authors, rather than a standard multi-annotator voting system. This two-annotator process is better suited for the nuanced, interpretive task of cultural reconstruction, leveraging the authors' deep expertise in both Korean and American cultures. We argue that for capturing subtle cultural meaning, a process requiring in-depth discussion until 100% consensus is achieved on every single sample is more rigorous and reliable than a simple tie-breaker vote. This consensus-driven process consisted of three stages: First, each author independently reconstructed half of the dialogues (16 emotion categories each). This initial reconstruction involved not only rewriting the dialogue to ensure cultural authenticity but also simultaneously performing multi-label emotion annotation. Recognizing that single labels often fail to capture emotional complexity, we extended the original annotation by assigning supplementary labels to reflect the mixed feelings present in the conversation. Second, the authors conducted a cross-check review, where each validated the reconstructions and multi-label annotations done by the other. During this phase, both primary and supplementary emotion labels were often revised to better align with the new emotional trajectory of the reconstructed dialogues. Finally, any disagreements were resolved through a joint review session to reach a final, unified consensus on every dialogue, ensuring the naturalness, cultural appropriateness, and annotation validity of all 1,280 adapted dialogues.

3.2 Handcrafting for Culture-Specific Emotions

To create unambiguous test cases for concepts entirely absent from the source data, we also handcrafted an additional 80 dialogues. These dialogues are intentionally single-label, designed to test the uniquely Korean emotions of 'jeong' and 'han' (Ka, 2010). This design allows for a focused evaluation of a model's ability to recognize these specific, culturally-rich concepts without the confounding factor of mixed emotions. All materials, including the original ED dataset, are used under the CC BY-NC 4.0 license.

3.3 Benchmark Characteristics

The final KoED benchmark, a product of our reconstruction and annotation pipeline, comprises 1,360 high-quality dialogues. The dataset is balanced across 34 emotion categories—the 32 original ED emotions plus the two uniquely Korean emotions, 'jeong' and 'han'—with 40 dialogues per category. This balanced design ensures that our cross-cultural evaluation is robust and not skewed by any particular emotion.

A key feature of KoED is its emotional granularity. Our multi-label annotation approach resulted in 555 unique combinations of primary and supplementary emotions, providing a rich representation of the complex, mixed feelings often found in authentic conversations. This stands in contrast to single-label benchmarks and allows for a more nuanced assessment of a model's emotional intelligence.

Crucially, KoED is designed as a benchmark solely for evaluation, not for model training. As such, it is presented as a single dataset without predefined train/dev/test splits. This encourages a standardized, zero-shot evaluation setting, allowing for a direct and fair comparison of different models' abilities to navigate a new cultural context. To provide a clear picture of our methodology, we present concrete examples for each of the three reconstruction categories: a SIMPLY-TRANSFERRED case (Figure 4), a pivotal CONTEXT-MODIFIED case (Figure 5), and a culturally dissonant SAMPLE-REMOVED case (Figure 6).

4 Experiment

To empirically investigate the "cultural empathy gap" and test our hypothesis regarding the "data provenance effect", we designed a series of experiments centered on a controlled, cross-cultural comparison. In this section, we detail our experimental setup, the models selected to represent diverse cultural provenances, and the metrics designed to quantify empathetic and cultural competence.

4.1 Experimental Setup

Models To empirically test our hypothesis regarding the "data provenance effect", we strategically selected a suite of open-source models with distinct, verifiable data provenances. We selected Meta-Llama-3.1-8B-Instruct (Dubey et al., 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) as representatives of models developed within a predominantly Western context. This choice is grounded in their data composition: their pre-training corpora are heavily English-dominant, with over 95% of Llama 3's pre-training data being English, according to its official model card. As a point of contrast representing a non-Western but major linguistic sphere, we included Qwen2-7B-Instruct (Qwen Team, 2024), a model whose training was informed by a large-scale multilingual corpus with a significant Chinese component. Critically, to isolate the effect of deep, target-culture alignment, we included EXAONE-3.0-7.8B-Instruct (LG AI Research, 2024), a model distinguished by its pre-training on one of the world's largest high-quality Korean corpora, including 45 million professional documents. This curated selection provides a controlled setup to compare Western-centric, East Asian-aligned, and Korean-specialized models. To benchmark these models against the state-of-the-art, Claude-3.5-sonnet (Anthropic, 2024) was included as a high-performance baseline for the response generation task. Separately, GPT-4o (OpenAI, 2024) was employed exclusively as our automated evaluator to score the quality of all generated responses.

Generation Setting For empathetic response generation, all models were set to a temperature of 1 for output diversity and limited to a maximum of 256 tokens. All experiments were conducted on a single RTX 4090 GPU.

Methodology Our experimental methodology employs a unified, two-stage prompting pipeline for both tasks: emotion recognition and empathetic response generation. This approach, adapted from Qian et al. (2023), allows us to first test a model's cultural perception and then evaluate its subsequent behavioral response based on that perception. *Stage 1: Emotion Recognition (Cultural Perception Task).* In the first stage, we test a model's ability to interpret the emotional context within a specific cultural framework. Instead of relying on a model's internal (and potentially biased) knowledge, we provide it with the complete list of 34 emotion labels from our benchmark, including the culturally-specific concepts of 'jeong' and 'han'. The model's task is to select the most relevant emotion(s) from this explicit list. The accuracy of this selection serves as the result for our emotion recognition experiment. *Stage 2: Empathetic Response Generation (Behavioral Response Task).* In the second stage, the model is tasked with generating an empathetic response. Critically, this response is conditioned

Model | Dataset	ED (English-based)		KoED (Korean-adapted)	
	Without *jeong*, *han*	With translated *jeong*, *han*	Without *jeong*, *han*	With *jeong*, *han*
Meta-Llama-3.1-8B-Instruct	34.53%	30.74%	32.19%	28.53%
Mistral-7B-Instruct-v0.3	39.38%	36.40%	33.52%	31.10%
Qwen2-7B-Instruct	32.73%	30.74%	25.70%	22.57%
EXAONE-3.0-7.8B-Instruct	31.41%	29.12%	30.31%	26.18%
Claude-3.5-sonnet	46.17%	42.94%	41.72%	38.31%
Average	36.84%	33.99%	32.69%	29.34%

Table 1: Performance comparison of multilingual models on the emotion recognition task across the ED and KoED datasets. We test under two conditions: (1) "Without *jeong*, *han*": Models were evaluated on the 1,280 dialogues of the 32 shared emotions, with only these 32 emotions provided as recognition candidates. (2) "With *jeong*, *han*": Models were evaluated on all 1,360 dialogues, with the full list of 34 emotions (including descriptions for *jeong* and *han*, translated for the ED condition) provided as recognition candidates. This comparison reveals the impact of introducing culture-specific concepts on model performance.

on the emotion(s) it identified in Stage 1. This allows us to evaluate how a model's initial cultural interpretation directly influences its subsequent empathetic behavior. To ensure we are evaluating cultural adaptation rather than a model's default behavior, we prepend a clear instruction to every prompt, adapted from Wang et al. (2024):

Task Definition: This is a Korean/English empathetic dialogue task. . . .

This unified pipeline provides a rigorous framework to assess the full cycle of multicultural empathy: from perception to action. The complete prompt structure is detailed in Appendix A.

Evaluation Metrics A major challenge in evaluating empathetic responses is their subjective nature, which makes objective measurement difficult. To address this, we developed a holistic evaluation framework, leveraging LLM-based assessment methodologies (Liu et al., 2023b; Ye et al., 2024). The framework is centered on five key indicators, each rated on a 5-point Likert scale:

1. **Explorations (EX)**: How deeply does the model explore the interlocutor's situation and emotions?

2. **Interpretations (IP)**: How accurately does the model understand and interpret the other party's emotions and situation?

3. **Emotional Reactions (ER)**: How appropriate is the model's emotional response to the situation?

4. **Evoked Emotion Alignment (EEA)**: How similar is the model's emotional response to that of a typical person?

5. **Cultural Appropriateness (CA)**: How well does the model reflect the norms of the specified [Korean/English] culture and maintain grammatical integrity?

The EX, IP, and ER metrics are adapted from the multifaceted framework of Sharma et al. (2020), while the EEA metric follows the concept from Huang et al. (2023). Our novel Cultural Appropriateness (CA) metric was specifically designed to assess alignment with cultural norms and serves as our primary metric for quantifying the "cultural empathy gap". The detailed scoring rubrics for these indicators are specified in Figures 9 through 13. Evaluation was performed by both an automated evaluator (GPT-4o) and a human panel. The panel consisted of four native Korean speakers (two men, two women), aged between 25 and 29, all holding at least a bachelor's degree from diverse academic backgrounds spanning the social sciences, natural sciences, and engineering. This demographic diversity helps ensure that the evaluations are not biased by a single disciplinary perspective. All human participants were compensated ($0.1 per data point), participated voluntarily, and provided informed consent for their anonymized ratings to be used in this research. To ensure reliability, both human and automated evaluators used the identical, detailed rubric provided in Appendix B.

4.2 Emotion Recognition

As a foundational test of cultural perception—the first stage of empathy—we conducted an emotion recognition experiment. The goal was to quantify whether models' ability to identify emotions is consistent across different cultural contexts.

Experimental Design. We evaluated five models on both ED and KoED under two conditions, using our multi-label annotations as the ground truth for both datasets to ensure a fair comparison. (1) Shared Emotions: We used the 1,280 dialogues of the 32 shared emotions. (2) All Emotions: We used all 1,360 dialogues, which involved adding English translations of our 80 handcrafted 'jeong' and 'han' dialogues to the ED set. The task required models to predict a single, most representative emotion for each dialogue. A prediction was marked as correct if it matched any of the ground-truth labels (either primary or supplementary), accommodating the multi-label nature of our dataset. This task was executed using the first stage of our prompting pipeline, modified to output only one emotion.

Results and Interpretation. The results, detailed in Table 1, reveal a consistent and significant trend. Across both conditions, all models achieved higher accuracy on the US-centric ED dataset than on the culturally-reconstructed KoED dataset. The average accuracy dropped from 36.84% (ED) to 32.69% (KoED) in the shared emotions condition, and this gap widened slightly from 33.99% (ED) to 29.34% (KoED) when culture-specific concepts were introduced.

This performance degradation on the culturally-adapted dataset provides the initial empirical evidence for the "cultural empathy gap". The fact that this gap exists even when evaluating shared emotions suggests that the issue transcends simple vocabulary; models struggle to interpret the nuances of how these emotions are expressed within a Korean cultural context. The consistent underperformance across all models points to a systemic limitation in their current multicultural capabilities.

4.3 Empathetic Response Generation

Model	EX	IP	ER	EEA	CA
Llama-3.1	2.58	2.85	2.80	2.77	2.94
Mistral	2.67	2.41	2.33	2.19	2.04
Qwen-2	2.47	2.42	2.25	2.22	2.14
EXAONE	2.84	3.37	3.34	3.40	3.71
Claude-3.5-sonnet	**4.14**	**4.54**	**4.41**	**4.45**	**4.64**

Table 2: Human evaluation results for empathetic response generation on the KoED dataset. The scores represent the average ratings given by four evaluators (rounded to two decimal places) on a 5-point Likert scale across five evaluation dimensions.

Moving from cultural perception to action, we next evaluated the models' ability to generate empathetic responses. The experiment followed the two-stage pipeline described in our methodology, where models first identified emotions (Stage 1) and then generated a response conditioned on their own perception (Stage 2).

Evidence of the "Cultural Empathy Gap". The results from both automated and human evaluations provide decisive evidence of the "cultural empathy gap" in response generation. The automated evaluation scores (Figure 2) show a clear trend: all open-source models suffered a significant performance degradation when moving from the US-centric ED to the culturally-reconstructed KoED. This drop was particularly pronounced on our primary metric, Cultural Appropriateness (CA), confirming that models struggle to align their responses with Korean interactional norms. The human evaluation results on KoED (Table 2) reveal these difficulties in greater detail: not only Western-developed models such as Llama-3.1 and Mistral, but also Qwen-2, which was developed in China — an East Asian cultural context — scored significantly lower on CA compared to other metrics.

Support for the "Data Provenance Effect" Hypothesis. Furthermore, the results offer compelling support for our "data provenance effect" hypothesis. As shown in both automated and human evaluations, the Korean-centric model, EXAONE, consistently outperformed other open-source models on the KoED dataset. While Claude-3.5-sonnet remained the top performer overall, EXAONE's score on the crucial CA metric (3.71 in human evaluation) was dramatically higher than that of its open-source peers and approached the level of the state-of-the-art commercial model. This outcome strongly suggests that pre-training on a large corpus of culturally-relevant data is a critical factor in a model's ability to produce genuinely empathetic and culturally appropriate responses.

Isolating the Cultural Factor from Linguistic Proficiency. A critical consideration is whether the observed performance gap on KoED stems from a lack of cultural understanding or simply weaker Korean linguistic proficiency. The unique performance of EXAONE provides compelling evidence for the former. Despite being a 7.8B model, comparable in scale to Llama-3.1 (8B) and Mistral (7B), its Cultural Appropriateness (CA) score of 3.71 in human evaluations dramatically surpasses its peers (Llama: 2.94, Mistral: 2.04). If linguistic skill at a given model scale were the primary determinant, their performances should have been

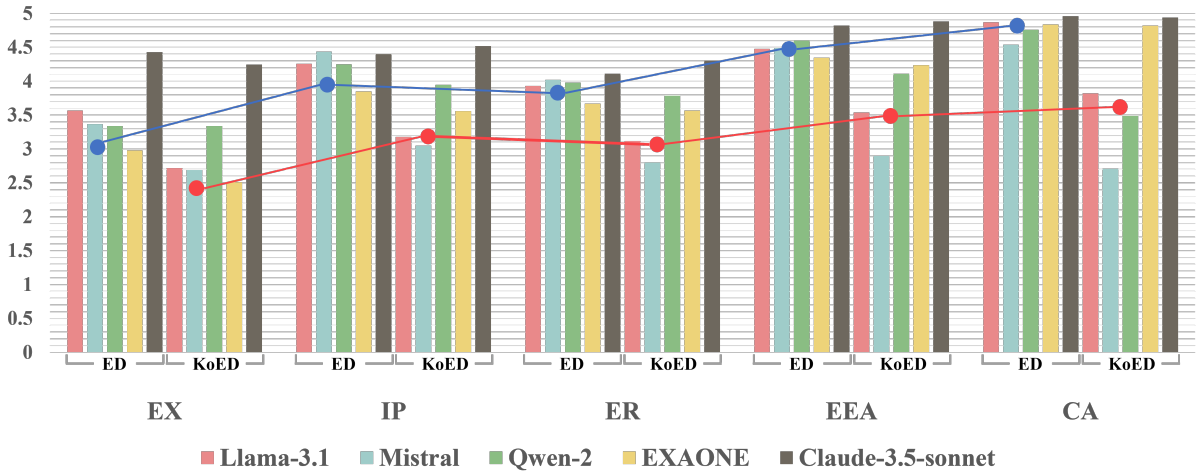

Figure 2: Qualitative performance of each language model in generating empathetic responses. The figure presents the average scores assigned by GPT-4o to the responses generated by various models on two datasets: the original ED dataset (indicated by blue dots) and the culturally adapted KoED dataset (indicated by red dots). These scores reflect the models' ability to produce responses that are both emotionally and culturally appropriate, highlighting differences in performance across the two datasets.

similar. EXAONE's anomalous success on a metric specifically designed to measure cultural alignment, bringing it closer to the state-of-the-art Claude-3.5-sonnet (4.64) than to its own size-class, strongly suggests that the "data provenance effect" is not merely a linguistic artifact. It is a distinct cultural advantage conferred by pre-training on a deeply aligned corpus.

Implications of Scale on the Cultural Empathy Gap. The significant performance gap between the open-source models and the proprietary baseline, Claude-3.5-sonnet, warrants discussion. On one hand, the superior performance of a much larger model like Claude-3.5-sonnet may suggest that at a massive scale, models are exposed to enough diverse data to begin approximating cultural norms, thereby shrinking the cultural gap through superior general reasoning. However, our findings with EXAONE offer a crucial nuance. The fact that a modest 7.8B model can achieve a Cultural Appropriateness score that dramatically surpasses its similarly-sized peers and significantly closes the gap with Claude indicates that direct cultural alignment is a more efficient, and perhaps more effective, path to genuine cultural competence than scale alone. While scale may be a mitigator, it does not appear to be a panacea for the cultural empathy gap; culturally-aligned data remains a key ingredient.

Evaluation Reliability. To validate our evaluation process, we measured the consistency of our human ratings. The inter-evaluator agreement was very high, with an Intraclass Correlation Coefficient (ICC(2,4)) of 0.884 (Shrout and Fleiss, 1979). Additionally, the Spearman correlation coefficient between the human panel's average scores and the GPT-4o automated scores was 0.649 (Spearman, 1961), indicating a moderate-to-strong positive correlation and affirming the reliability of our automated evaluation as a proxy for human judgment. Overall, these findings confirm that multicultural empathy requires more than just linguistic fluency; it necessitates a deep, culturally-grounded understanding—one that is significantly influenced by a model's training background and cannot be explained by linguistic capability alone.

5 In-Depth Analysis: Probing the Limits of Cultural Comprehension

Having established the existence of a "cultural empathy gap", we now conduct an in-depth analysis to probe its underlying reasons. We focus on the most challenging test cases in our benchmark—the culturally-specific emotions of jeong and han—to investigate how LLMs process concepts that lack direct parallels in their primary training data.

5.1 Can Explicit Prompts Teach Culture?

To test whether a model's cultural knowledge deficit can be compensated for by explicit instruction, we designed a controlled experiment. We evaluated five models' recognition accuracy for jeong and han under four prompting conditions, each providing a different level of contextual information:

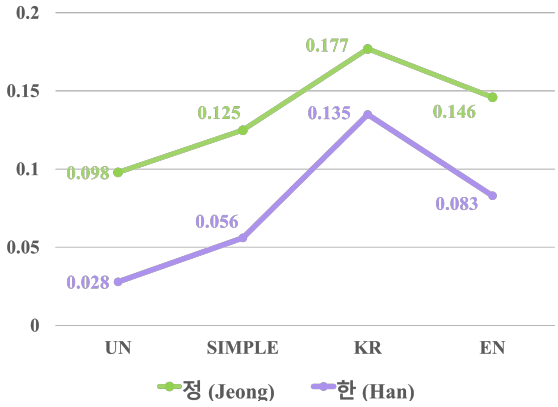

Figure 3: Comparative analysis of recognition accuracy for the Korean-specific emotions “jeong” and “han” under varying annotation conditions. The figure shows how models’ average performance improves with increasing contextual information—from no description (UN) to simple descriptions (SIMPLE), and further to detailed explanations in Korean (KR) and English (EN), with the most notable gains observed under the KR condition.

Undescribed (UN): Providing only the words, testing the model’s pre-existing knowledge.

Without Description (UN):

정, 한

Translation:

Jeong, Han

Simply-described (SIMPLE): Providing a minimal tag to identify the concepts as culturally unique.

With Korean Description (SIMPLE):

정 (한국 고유의 감정),
한 (한국 고유의 감정)

Translation:

Jeong (Unique Korean emotions),
Han (Unique Korean emotions)

Korean-described (KR): Providing a detailed, official definition in the original language from the Standard Korean Language Dictionary.

With Korean Description (KR):
정 (느끼어 일어나는 마음 혹은 사랑이나 친근감을 느끼는 마음), 한 (몹시 원망스럽고 억울하거나 안타깝고 슬퍼 응어리진 마음)

English-described (EN): Providing an English translation of the KR definition using DeepL.

With English Description (EN):
Jeong (A feeling that arises in one’s heart, or a feeling of love or affinity),
Han (A feeling of bitter resentment, injustice, pity, or sadness)

The results, shown in Figure 3, reveal a clear pattern: providing more context improves performance. Accuracy is near-zero under the UN condition, underscoring that these concepts are not part of the models’ inherent knowledge base. Performance significantly improves as descriptions are added, with the most substantial gains occurring under the KR (native language) condition. This highlights the importance of providing cultural knowledge in its original linguistic form.

However, the most critical finding lies in the performance ceiling. Even under the optimal KR condition, the peak recognition accuracy remains strikingly low (17.7% for jeong and 13.5% for han). This suggests a fundamental limitation: while explicit prompting can provide surface-level pattern matching cues, it fails to instill the deep, contextual understanding required to genuinely comprehend concepts like jeong and han. The models appear to be performing a sophisticated form of lexical matching rather than true cultural reasoning. This finding challenges the notion that cultural competence can be simply "injected" at inference time and points towards the necessity of incorporating such knowledge during the pre-training phase itself.

6 Conclusion

In this paper, we challenged the notion that multilingual competence equates to multicultural understanding, testing this in the nuanced, high-stakes domain of empathetic dialogue. By introducing KoED, a benchmark meticulously reconstructed—not translated—from the English ED dataset, we created a controlled environment for cross-cultural evaluation. Our experiments provided direct empirical evidence of a "cultural empathy gap", revealing that LLMs fail to transfer empathetic capabilities from a US-centric to a Korean context. Crucially, we demonstrated this gap is not a mere linguistic artifact but is strongly linked to the "data provenance effect," as evidenced by the Korean-centric model EXAONE. Its superior cultural appropriateness, despite its modest scale, cannot be explained by language proficiency alone and suggests that authentic multicultural empathy is a direct, and more efficiently achieved, function of culturally-aligned pre-training, not simply an emergent property of scale. Ultimately, our work provides both a new benchmark (KoED) and a clear methodology for probing the cultural limitations of LLMs. We have shown that to build truly empathetic AI, the field must move beyond culturally monolithic paradigms, prioritizing deep cultural alignment over mere linguistic fluency.

Limitations

While this study provides strong empirical evidence for the "cultural empathy gap", we acknowledge several limitations that define the boundaries of our claims and offer important avenues for future research.

First, and most critically, our experimental design, which compares model performance on an English benchmark (ED) versus a Korean one (KoED), inherently conflates cultural and linguistic factors. A performance drop on KoED could be attributed to a model's failure to grasp Korean cultural nuances (our central claim), but also, in part, to weaker general linguistic proficiency in Korean. While we cannot fully disentangle these variables, our analysis in Section 4.3 provides strong evidence that the cultural component is a decisive factor. The Korean-centric model EXAONE, despite its modest scale, achieved a Cultural Appropriateness score far exceeding its similarly-sized peers and approaching the state-of-the-art. This result cannot be explained by linguistic proficiency alone and strongly supports our conclusion that a genuine "cultural empathy gap" exists, driven by the cultural alignment of pre-training data.

Second, our evidence for the "data provenance effect" hinges on a single, albeit powerful, case study of EXAONE. While EXAONE's dramatically superior performance in cultural appropriateness provides compelling support for our hypothesis, it does not constitute definitive proof of a universal principle applicable to all non-Western models. A more comprehensive study is needed to test the generalizability of our findings. This would involve a broader comparative analysis including other culturally-specialized models developed for different non-Western contexts (e.g., models specialized for Japanese, Arabic, or Hindi) to see if a similar "home-ground advantage" in cultural competence consistently emerges.

Third, while our cultural reconstruction methodology is a significant advance over direct translation, KoED is not a substitute for a corpus created natively from the ground up. The original situations and emotional prompts from the ED dataset still provide the thematic scaffolding for KoED. Consequently, our benchmark may not capture certain types of empathetic conversations or emotional scenarios that are unique to the Korean experience and would only arise in a natively generated corpus. The development of such large-scale, natively-created empathetic dialogue datasets for diverse cultures represents a critical and necessary next frontier for the field, and our work highlights the urgent need for such resources.

Acknowledgments

This work was supported by Institute of Information Communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190004, Development of semisupervised learning language intelligence technology and Korean tutoring service for foreigners), and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00553041, Enhancement of Rational and Emotional Intelligence of Large Language Models for Implementing Dependable Conversational Agents).

References

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

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

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. *arXiv preprint arXiv:2503.10688*.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. *Preprint*, arXiv:2004.10151.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Xunlian Dai, Li Zhou, Benyou Wang, and Haizhou Li. 2025. From word to world: Evaluate and mitigate culture bias via word association test. *arXiv preprint arXiv:2505.18562*.

Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

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

He Hu, Yucheng Zhou, Juzheng Si, Qianning Wang, Hengheng Zhang, Fuji Ren, Fei Ma, and Laizhong Cui. 2025. Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling. *Preprint*, arXiv:2505.15715.

Haoyu Huang, Tong Niu, Rui Yang, and Luping Shi. 2025. RAM2C: A liberal arts educational chatbot based on retrieval-augmented multi-role multi-expert collaboration. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 448–458, Abu Dhabi, UAE. Association for Computational Linguistics.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Yohan Ka. 2010. Jeong-han as a korean culture-bound narcissism: Dealing with jeong-han through jeong-dynamics. *Pastoral Psychology*, 59(2):221–231. Copyright - Springer Science+Business Media, LLC 2010; Document feature - ; Last updated - 2024-03-21.

LG AI Research. 2024. Exaone 3.0 7.8b instruction tuned language model. *arXiv preprint arXiv:2408.03541*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

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*.

Becky Lynn Omdahl. 2014. *Cognitive appraisal, emotion, and empathy*. Psychology Press.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Yushan Qian, Weinan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6516–6528, Singapore. Association for Computational Linguistics.

Huachuan Qiu and Zhenzhong Lan. 2025. PsyDial: A large-scale long-term conversational dataset for mental health support. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21624–21655, Vienna, Austria. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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.

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

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

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Shannon Spaulding. 2017. Cognitive empathy. In *The Routledge handbook of philosophy of empathy*, pages 13–21. Routledge.

Charles Spearman. 1961. The proof and measurement of association between two things.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Zhichao Xu and Jiepu Jiang. 2024. Multi-dimensional evaluation of empathetic dialog responses. *Preprint*, arXiv:2402.11409.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. Flask: Fine-grained language model evaluation based on alignment skill sets. *Preprint*, arXiv:2307.10928.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2022. Case: Aligning coarse-to-fine cognition and affection for empathetic response generation. *arXiv preprint arXiv:2208.08845*.

Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li, and Daniel Hershcovich. 2024. Does mapo tofu contain coffee? probing llms for food-related cultural knowledge. *arXiv preprint arXiv:2404.06833*.

A Generation and Recognition Prompt

This section details the unified prompt template used for both of our primary tasks: emotion recognition and empathetic response generation. The prompt is designed as a versatile, two-stage process, illustrated in Figure 7. The crucial modification lies in the instruction for Stage 1 (Emotion Inference), depending on the specific task being performed:

- **For the Emotion Recognition Task (Section 4.2):** The model is instructed to analyze the dialogue and select *only a single* most representative emotion from the predefined list of 34. The model's process stops here for this task.
- **For the Empathetic Response Generation Task (Section 4.3):** The model is instructed to select *up to four* candidate emotions from the same list to build a rich understanding of the context.

Stage 2 (Response Generation) is executed only for the response generation task. The model's objective is to generate the final Listener response based on the emotion(s) it inferred in Stage 1. To standardize this task across dialogues of varying lengths, the following procedure is applied:

- If a conversation context ends on a Speaker's turn (an odd number of utterances), the model is prompted to generate a new, subsequent Listener response for the next turn.
- If a conversation context already ends on a Listener's turn (an even number of utterances), the final ground-truth utterance is removed. The model is then tasked with regenerating this utterance based on the preceding context and its own emotional inference from Stage 1.

This versatile two-stage process allows us to use a consistent framework to first test perception, then action.

B Evaluation Prompt and Rubrics

This section outlines the prompts used for both automated (GPT-4o) and human evaluation of the generated responses. The model's performance was assessed using five key metrics: Explorations (EX), Interpretations (IP), Emotional Reactions (ER), Evoked Emotion Alignment (EEA), and Cultural Appropriateness (CA). Figure 8 shows the common task instructions provided for each metric. Figures 9, 10, 11, 12, and 13 detail the specific 5-point Likert scale rubrics used for scoring. These same rubrics were provided to the human evaluators to ensure consistency during the qualitative assessments.

KoED Data Sample: A Prime Example of SIMPLY-TRANSFERRED

conv_id: `"hit:497_conv:994"`
Emotion Labels: `["anticipating", "excited"]`
Situation: *I can't wait for the weekend. I am going to Disney World in Florida.*

Dialogue Reconstruction Highlighting Cultural Adaptation

Turn 1 Speaker A:
[Original ED]: *I cant wait for next weekend. I am going to **Disney World** here in **Florida**.*
[KoED]: 다음 주말이 너무 기다려져. **용인 에버랜드**에 가기로 했거든.

Turn 2 Speaker B:
[Original ED]: *That sounds super fun, I love florida and orlando its quite awesome.*
[KoED]: 오, 좋겠다. 용인 볼거리 진짜 많지. **T-익스프레스** 타려고? **푸바오**는 볼 거야?

Turn 3 Speaker A:
[Original ED]: *Yes it is. I live in Orlando, but rarely go to the attractions so its long overdue.*
[KoED]: 응, 맞아! 제대로 즐겨야지. 사실 나 용인 살면서도 에버랜드 자주 안 갔거든. 그래서 이번엔 오랜만에 가는 거야.

Turn 4 Speaker B:
[Original ED]: *Yea thats fair, I could imagine if they are so near by you wouldnt care about them.*
[KoED]: 맞아, 가까이 있으면 오히려 잘 안 가게 되더라. 나도 집 근처 놀이공원 1년에 한 번 갈까 말까야.

Figure 4: An example of a SIMPLY-TRANSFERRED dialogue. This adaptation focuses on direct, word-level cultural substitution of named entities to maintain local relevance, preserving the original topic and trajectory. This is a common pattern for substituting entities (e.g., 'Uber' → 'Kakao Taxi'). In this example, the core topic (visiting an amusement park) is preserved in Turn 1 by replacing "Disney World in Florida" with its culturally equivalent counterpart, "용인 에버랜드" (Everland in Yongin). The subsequent turns demonstrate a parallel conversational structure that logically follows this substitution: in Turn 2, the discussion about "Orlando" naturally pivots to famous attractions at Everland, "T-익스프레스" (T-Express) and "푸바오" (Fubao). The core conversational logic (e.g., "living nearby but rarely visiting" in Turn 3) and the overall trajectory **remain identical to the original**. This contrasts with the CONTEXT-MODIFIED example (Fig. 5), where introducing a new cultural element ("seaweed soup") **fundamentally alters the dialogue's path** rather than simply translating the existing one.

KoED Data Sample: A Prime Example of CONTEXT-MODIFIED

conv_id: `"hit:46_conv:92"`
Emotion Labels: `["excited", "anticipating"]`
Situation: *Today is my birthday. I'm excited about messages and gifts.*

Dialogue Reconstruction Highlighting Cultural Adaptation

Turn 1 Speaker A:
[Original ED]: *Today is my birthday! I am 34!*
[KoED]: 오늘 내 생일이야! 서른넷 됐어!

Turn 2 Speaker B:
[Original ED]: *Happy Birthday!!!! What are you going to do for it?*
[KoED]: 생일 축하해!!!! 오늘 뭐 할 거야?

Turn 3 Speaker A:
[Original ED]: *Nothing today since its Thursday, hopefully I can **bar hop** this weekend. Waiting for gifts today.*
[KoED]: 평일이라 오늘은 조용히 보내고, 주말에 친구들이랑 **술자리** 가질까 해. 지금은 **카톡** 축하 메시지만 기다리는 중.

Turn 4 Speaker B: *(Pivotal Cultural Modification)*
[Original ED]: *I hope you get lots and lots! [of gifts]*
[KoED]: 아, 그렇구나. 생일을 2번 즐길 수 있네! 그럼 지금 카톡 알림음 들릴 때마다 설레겠다.ㅎㅎ **그나저나 아침에 미역국은 먹었어?**

Turn 5 Speaker A: *(Conversation Diverges Based on New Context)*
[Original ED]: *me too. I love goodies.*
[KoED]: 아직인데, 오늘 이따 회사 구내식당에서 마침 미역국이 나오더라고.

Turn 6 Speaker B:
[Original ED]: *That's great! Looks like your company is taking care of you. Happy birthday again!*
[KoED]: 회사에서 챙겨주는 것 같네! 기분 좋겠다. 다시 한 번 축하하고 생일 잘 보내~

Figure 5: An example of a CONTEXT-MODIFIED dialogue. The reconstruction aims to preserve the original conversation's core emotional theme (excitement, anticipation) and trajectory, while going beyond simple localization. For instance, in Turn 3, bar hop—a specific activity of moving between pubs, which is less common in Korea—is culturally adapted to 술자리 (suljari), the more resonant and general concept of a social drinking gathering. The pivotal modification, however, occurs in Turn 4, where the generic English wish for gifts is replaced with a culturally-specific Korean question: "Did you have seaweed soup (미역국) this morning?". Eating seaweed soup is a deep-rooted birthday tradition in Korea. **This single change fundamentally alters the conversational context while staying true to the initial empathetic goal**, as seen in the subsequent turns (5 and 6), demonstrating a true cultural adaptation rather than a mere translation.

KoED Data Sample: A Prime Example of SAMPLE-REMOVED

conv_id: `"hit:104_conv:208"`
Emotion Labels: `["angry"]`
Situation: *some guys shot my neighbour and ran into the woods*

Original Dialogue (Filtered during KoED Curation)

Turn 1 Speaker A:
[Original ED]: *i just moved to this neighborhood and some dumb criminals shot one of my neighbors and ran into the woods!*

Turn 2 Speaker B:
[Original ED]: *Thats not good.* ***Do you own a gun?***

Turn 3 Speaker A:
[Original ED]: *I do! I want to be able to protect my son*

Turn 4 Speaker B:
[Original ED]: *That is always number one goal.*

Figure 6: An example of a SAMPLE-REMOVED dialogue. This conversation was entirely excluded from the KoED dataset as its core topic is culturally dissonant and does not align with common Korean sentiment or daily experience. The dialogue, which begins in Turn 1 with a report of a neighborhood shooting, pivots in Turn 2 to a discussion about personal gun ownership ("Do you own a gun?"). This topic, while prevalent in some cultures (like the U.S.), is highly sensitive and largely irrelevant in the South Korean context, where civilian gun ownership is extremely rare and strictly regulated. A direct translation would be unrelatable, and a cultural adaptation (e.g., changing the weapon) would either escalate the situation unnaturally or trivialize the original's context. **To maintain the dataset's cultural integrity and relatability, such samples were filtered and removed during the curation process**. Notably, because these samples were discarded before our main annotation phase, they did not undergo the subsequent multi-emotion labeling applied to the final KoED corpus and thus retain their original single emotion label.

Prompt Template for Empathetic Response Generation

Task Definition:
This is a/an {Korean / English} empathetic dialogue task: The first worker (Speaker) is given an emotion label and writes his own description of a situation when he has felt that way. Then, Speaker tells his story in a conversation with a second worker (Listener). The emotion label and situation of Speaker are invisible to Listener. Listener should recognize and acknowledge others' feelings in a conversation as much as possible.

Guideline Instruction:
Now you play the role of Listener, please give the corresponding response according to the existing context. You only need to provide the next round of response of Listener.

List of 34 Emotions:
Afraid (두려움), Angry (화남), Annoyed (짜증남), Anticipating (기대됨), Anxious (불안함), Apprehensive (염려됨), Ashamed (부끄러움), Caring (보살핌), Confident (자신감), Content (만족함), Devastated (충격받음), Disappointed (실망함), Disgusted (역겨움), Embarrassed (당황함), Excited (흥분됨), Faithful (충실함), Furious (격노함), Grateful (감사함), Guilty (죄책감), Hopeful (희망적), Impressed (감명받음), Jealous (질투남), Joyful (기쁨), Lonely (외로움), Nostalgic (향수에 젖음), Prepared (준비됨), Proud (자랑스러움), Sad (슬픔), Sentimental (감상적), Surprised (놀람), Terrified (겁에 질림), Trusting (신뢰함), 정(한국 고유의 정서), 한(한국 고유의 정서)

1. Do not use any emotion terms other than the 34 basic emotions listed above.
2. Combinations or mixtures of emotions are allowed.
3. Select up to 4 emotions that best describe the Speaker's emotional state.

Dialogue:
[dialogue]

Stage 1:
1. Analyze the given dialogue to identify the Speaker's complex emotional state.
2. Specify the identified emotions using multiple labels from the 34 emotions listed above. Select all that apply, with no minimum or maximum limit.
(STOP HERE. Do NOT proceed to steps 3 and 4 yet. Only identify the emotion at this stage.)

Stage 2:
Identified Emotions: [identified emotions]
Generate the next {Korean / English} empathetic response based on the identified emotions.

Figure 7: Prompt template for empathetic response generation and emotion identification. This two-stage process instructs the model first to analyze the dialogue and select relevant emotions from a predefined list of 34 emotions—including culturally specific emotions such as 'jeong' and 'han'—and then to generate an empathetic response based on the identified emotions.

You will be given one response for one dialogue.
Your task is to rate the response based on the criteria provided.
Please make sure you read and understand these criteria carefully. Refer back to them as needed during your evaluation.

Dialogue:
[dialogue]

Empathetic Response:
[empathetic response]

Please give feedback on the listener's responses. Also, provide the listener with a score on a scale of 1 to 5 for the {criterion}, where a higher score indicates better overall performance. Make sure to give feedback or comments for the {criterion} first and then write the score for the {criterion} .

Response Format:
Feedback: [Your feedback here]
Score: [1-5]

Figure 8: Prompt template for evaluating empathetic responses. This template provides detailed guidelines for assessing a response based on five key metrics: Explorations, Interpretations, Emotional Reactions, Evoked Emotion Alignment, and Cultural Appropriateness.

Score 1: No attempt to explore the interlocutor's emotions or experiences, showing no additional inquiry or interest.
Score 2: General attempts at exploration are made, but they are not specific or fail to delve deeply into the interlocutor's situation.
Score 3: Somewhat specific exploration is attempted, but it lacks depth or completeness.
Score 4: The response makes a fairly specific and clear attempt to explore the interlocutor's feelings or experiences.
Score 5: The response makes a very specific and thoughtful attempt to understand the interlocutor's emotions and experiences deeply.

Figure 9: Evaluation rubric for Explorations (EX) in empathetic response assessment. This rubric specifies the criteria for evaluating how thoroughly the model explores the dialogue context and the speaker's emotional state.

Score 1: The response does not show any acknowledgment or interpretation of the interlocutor's feelings or experiences.
Score 2: The response acknowledges the interlocutor's situation or feelings but does so at a very surface level.
Score 3: The response shows some understanding of the interlocutor's feelings or experiences, but it lacks depth.
Score 4: The response provides a fairly deep interpretation of the interlocutor's emotions and experiences, showing considerable understanding.
Score 5: The response offers a very deep and clear interpretation, fully conveying an understanding of the interlocutor's feelings and experiences.

Figure 10: Evaluation rubric for Interpretations (IP) in empathetic response assessment. This rubric outlines the criteria for assessing the model's accuracy in understanding and interpreting the speaker's emotions and context.

Score 1: The response lacks any expression of emotional reactions, warmth, compassion, or concern.
Score 2: Emotional reactions are present but are either unclear or insufficiently warm or compassionate.
Score 3: The response shows some emotional reaction but lacks sufficient depth or warmth.
Score 4: The response expresses a significant level of emotional reaction, showing warmth, concern, and deep compassion for the interlocutor's situation.
Score 5: The response shows a very deep and warm emotional reaction, displaying strong empathy and concern for the interlocutor's situation.

Figure 11: Evaluation rubric for Emotional Reactions (ER) in empathetic response assessment. This rubric details the criteria used to evaluate the appropriateness of the model's emotional response to the conversation.

Score 1: The response fails to recognize the interlocutor's emotional state and may respond with an inappropriate emotional tone, potentially exacerbating negative feelings.
Score 2: The response minimally recognizes the interlocutor's emotional state but reacts insufficiently, resulting in little to no noticeable change in the interlocutor's emotions.
Score 3: The response acknowledges the emotional state and partially addresses it, but falls short of a fully appropriate reaction, leading to a moderate change in the interlocutor's emotional state.
Score 4: The response accurately recognizes the emotional state and effectively addresses it, alleviating some of the negative feelings or appropriately balancing overly excited emotions.
Score 5: The response demonstrates highly accurate recognition of the emotional state, perfectly aligns its response, and effectively modulates the emotional state, providing clear emotional relief or balance when needed.

Figure 12: Evaluation rubric for Evoked Emotion Alignment (EEA) in empathetic response assessment. This rubric defines the criteria for comparing the model's generated emotional response to those typically observed in human responses.

Score 1: The response is very disconnected from language culture and contains significant grammatical errors, leading to potential cultural misunderstandings.
Score 2: The response somewhat misaligns with language culture and has some grammatical issues, with insufficient reflection of cultural context and expression.
Score 3: The response generally aligns with language culture and is grammatically correct with no major issues, but some expressions or grammatical usage may feel slightly awkward or incomplete.
Score 4: The response mostly aligns with language culture and is grammatically appropriate, reflecting cultural context and expression well.
Score 5: The response perfectly aligns with language culture and is grammatically accurate and natural, fully reflecting traditions, customs, and social expectations.

Figure 13: Evaluation rubric for Cultural Appropriateness (CA) in empathetic response assessment. This rubric presents the criteria for evaluating how well the model's response reflects the cultural context and maintains linguistic accuracy.