Too Polite to be Human: Evaluating LLM Empathy in Korean Conversations via a DCT-Based Framework

Seoyoon Park[*], **Jaehee Kim**[*], **Hansaem Kim**[†]
Yonsei University, South Korea
{seoyoon.park, kim1016jh, khss}@yonsei.ac.kr

Abstract

As LLMs are increasingly used in global conversational settings, concerns remain about their ability to handle complex sociocultural contexts. This study evaluates LLMs' empathetic understanding in Korean—a high-context language—using a pragmatics-based Discourse Completion Task (DCT) focused on interpretive judgment rather than generation. Our dataset systematically varies in relational hierarchy, intimacy, and emotional valence, enabling fine-grained comparisons between proprietary/open-source LLMs and native Korean speakers. Most LLMs showed over-empathizing tendencies and struggled with ambiguous relational cues. Neither model size nor Korean fine-tuning significantly improved performance. Additionally, humans exhibit a nuanced understanding of social context and relational nuances, whereas LLMs rely on surface-level heuristics. These findings highlight the limitations of LLMs in sociopragmatic reasoning and introduce a scalable, culturally flexible framework for evaluating socially aware AI.

1 Introduction

With the rapid rise of large language models (LLMs), generating human-like text for tasks such as creative writing and ideation has become increasingly feasible. As a result, LLMs are now widely used in everyday conversations. However, despite the global popularity of English-centric models like GPT-4o and Claude, they often fall short in capturing complex cues such as context, relationships, mood, and emotional nuance. This limitation becomes particularly salient in empathy-driven interactions, where understanding goes beyond surface-level fluency. Empathy is inherently shaped by sociocultural norms, requiring not only appropriate expression but also the accurate interpretation of social meaning (He, 1991; Gladkova, 2010; Meiners, 2017).

To examine this challenge, we evaluate LLMs' capacity for empathetic understanding in Korean, a high-context language where relational nuance and social hierarchy are deeply embedded in linguistic form. Korean's systematic use of politeness strategies, situated between Japanese and Chinese in terms of structural regularity (Bak, 2018; Shin, 2021), provides both linguistic richness and analytic control for our study.

For evaluation, we introduce a pragmatics-based Discourse Completion Task (DCT) designed to assess LLMs' social judgment in empathetic scenarios. Drawing on existing corpora, we construct a dataset of dialogue prompts that vary in key situational factors, including relational hierarchy, intimacy, emotional valence, and conversational context. This enables a fine-grained comparison of proprietary and open-source LLMs across diverse scenarios. Figure 1 presents the DCT structure used to probe LLMs' sociopragmatic reasoning. This study addresses four research questions, with key findings as follows:

RQ1: Can LLMs empathize like humans?

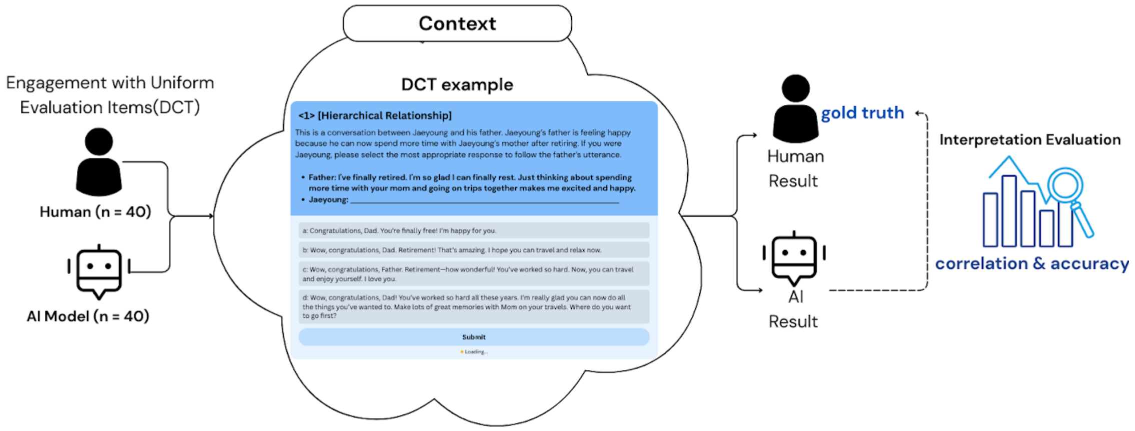
→ Most LLMs tended to over-empathize, using more intense expressions than humans.

RQ2: Does model size or fine-tuning improve empathy?

→ Neither Korean fine-tuning nor larger size consistently enhanced empathetic ability.

[*] These authors contributed equally.
[*] Corresponding author.

Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025), pages 79–92
July 31, 2025 ©2025 Association for Computational Linguistics



<1> [Hierarchical Relationship]

This is a conversation between Jaeyoung and his father. Jaeyoung's father is feeling happy because he can now spend more time with Jaeyoung's mother after retiring. If you were Jaeyoung, please select the most appropriate response to follow the father's utterance.

- **Father: I've finally retired. I'm so glad I can finally rest. Just thinking about spending more time with your mom and going on trips together makes me excited and happy.**
- **Jaeyoung: ____________________**

a: Congratulations, Dad. You're finally free! I'm happy for you.

b: Wow, congratulations, Dad. Retirement! That's amazing. I hope you can travel and relax now.

c: Wow, congratulations, Father. Retirement—how wonderful! You've worked so hard. Now, you can travel and enjoy yourself. I love you.

d: Wow, congratulations, Dad! You've worked so hard all these years. I'm really glad you can now do all the things you've wanted to. Make lots of great memories with Mom on your travels. Where do you want to go first?

Figure 1: Proposed Interpretation Evaluation Framework for Empathic Dialogue

RQ3: What shapes LLMs' empathic behavior most?

→ Social relationships influenced responses more than dialogue context, with intimacy having a greater impact than hierarchy. Models struggled in ambiguous relational settings.

RQ4: Why do LLM responses feel unnatural?

→ Despite fluency, LLMs often lacked awareness of context, relational nuance, and face-saving, leading to awkwardness.

This study reveals the limitations of current LLMs in sociopragmatic understanding through a DCT-based framework that evaluates their social judgment. Results show that LLMs often misread social cues, over-empathize, or fail to adjust appropriately. Additionally, our proposed DCT method is simple, flexible, and adaptable across languages and cultures, offering a foundation for evaluating socially aware LLMs and their interpretive competence.

Beyond implications for general LLM evaluation, our findings suggest that improving empathetic capabilities in LLMs can support beneficial applications such as mental health support, counseling, and socially aware virtual agents. At the same time, understanding how LLMs generate and modulate empathy is also critical for identifying potential risks. In particular, artificial expressions of empathy could be exploited in manipulative scenarios, such as voice phishing, where fabricated rapport is used to gain users' trust. This dual perspective highlights the importance of evaluating not only the fluency of LLMs but also the appropriateness and intent behind their social behaviors.

2 Related Works

2.1 Empathetic Dialogue Evaluation

Empathy has become a key focus in LLM-based dialogue research, with studies like Rashkin et al. (2019) and Kim et al. (2021) proposing models that infer emotional states and causes to generate empathetic responses. These works advanced empathy modeling through emotion-cause reasoning and lexical cues, showing strong results in human and automatic evaluations. Others, such as Lai et al. (2021) and Wu et al. (2024), conducted qualitative analyses. However, most prior work has emphasized how empathetic models could speak, overlooking the contextual appropriateness of their responses. Moreover, automatic metrics often miss sociocultural nuances, while qualitative methods, though richer, remain prone to subjectivity. These gaps highlight the need for evaluation methods that assess both fluency and context-sensitive judgment.

2.2 Discourse Completion Task

This paper assesses the empathy interpretation of LLMs using a pragmatics-based method, Discourse Completion Task (DCT). In general, a DCT presents a single or multi-turn dialogue, including discourse context, situational background, and speaker relationship. Through suggested choices or blanks, the test taker selects the most contextually appropriate response. This format efficiently evaluates pragmatic reasoning and socially appropriate empathy (Kasper & Rose, 2002; Walker, 2019). Widely used in cross-cultural pragmatics (Ogiermann, 2018) and increasingly in AI evaluation (Sperlich, 2016), the DCT here is

Step 1. Select dialogue samples from the raw dataset.
Step 2. Create enriched versions by paraphrasing the raw dialogues to enhance fluency and naturalness.
Step 3. Define the generation protocol for obtaining human and AI responses to both raw and enriched dialogues.
Step 4. Collect human responses: annotators generate responses for both versions, assuming the role of the empathizer. Reprocess the dialogues by explicitly assigning the roles of empathizer and empathized speaker (14–16 utterance turns per dialogue) to construct Dataset 1.
Step 5. Generate AI responses: use GPT-4o, Claude 3.5 Sonnet, and HyperClova to produce responses as empathizers, thereby constructing Dataset 2.

Table 1: Dataset construction methods

Sub-dataset	Recipient	Empathizer	Utterance length (avg.)
Dataset 1	Human (Paraphrased)	Human	7.6 words
Dataset 2		AI	40.8 words

Table 2: Empathic Dialogue Dataset: Role Assignment and Generation Methods.

adapted to compare human and LLM responses. Unlike prior work focused on generation, this approach highlights interpretive judgment, offering new insight into LLMs' sociopragmatic competence.

2.3 High-Context Languages

Hall (1959, 1976) classifies languages by context reliance: high-context languages, such as Korean, Japanese, and Chinese, depend on implicit cues, while low-context languages, like English, favor explicitness. In high-context cultures, empathy reflects relational closeness and hierarchy—overly emotional responses can feel intrusive. Thus, empathy is a socially regulated act, not just emotional expression (Fukushima & Haugh, 2014).

Korean features rich pragmatic strategies—honorifics, politeness norms, and 'nunchi,' a key skill for inferring emotional states and responding appropriately. Korean speakers judge empathy based on nuanced assessments of social distance hierarchy (Lee, 2022; Jung, 2023). LLMs must account for these cultural variables to generate contextually appropriate empathy in Korean.

2.4 Social Implications of Modeled Empathy

Recently, researchers have also begun to explore the broader social consequences of simulated empathy. On the positive side, empathetic LLMs show promise in areas such as mental health support, social companionship, and counseling assistance (Qiu & Lan, 2024; Ruosi, 2023; Naik et al., 2025). At the same time, concerns have emerged regarding the potential misuse of artificial empathy in manipulative settings—e.g., persuasive dialogue, deceptive persuasion, and phishing-like scenarios (Carrasco-Farre, 2024; Roy et al., 2024; Trinh et al., 2025). These studies highlight that beyond linguistic fluency, the perceived intent and appropriateness of empathetic responses are critical in ensuring safe and trustworthy interactions with LLMs. Our study contributes to this dual perspective by evaluating not only how empathetic a response sounds but also how well it aligns with the social norms and expectations of the dialogue context.

3 Dataset Construction

To compare human and AI response patterns in empathetic dialogue, we derived two sub-datasets by reorganizing the existing dataset, Korean Empathetic Dialogues (2022) from AIHub. Each sub-dataset comprises responses generated by three LLMs—GPT-4o, Claude, and HyperClova—as well as native Korean speakers. These responses were subsequently utilized to construct DCT items for direct comparative analysis. The original corpus comprises dialogues with 14 to 16 turns, each annotated with emotional labels, relational roles, and situational contexts. These dialogues were reprocessed via two complementary strategies: (i) paraphrasing to enhance fluency and plausibility and (ii) retaining the original utterances for baseline comparison. In both versions, LLMs and human annotators generated empathetic responses while explicitly assuming the role of the empathizer. LLMs were provided with prompts specifying the target emotion, interpersonal relationship, and situational context. Table 1 presents the construction workflow, and Table 2 summarizes the composition of each sub-dataset.

Q.	relations	recipient - empathizer	Dialogue Context & sentiment polarity
Q1	hierarchy	father-child	Father's retirement (pos)
Q2	hierarchy	mother-child	A mysterious lump found on mother's neck (neg)
Q3	hierarchy	child-mother	Child receives good grades at school (pos)
Q4	hierarchy	child-father	Child moves on to a new school and parts ways with friends (neg)
Q5	hierarchy	same age friends	goes out for a family dinner after a long time (pos)
Q6	hierarchy	same age friends	Discovers that a junior had lied (neg)
Q7	intimacy	distant	Receives first business card after joining the company (pos)
Q8	intimacy	distant	Attending an English academy but not seeing improvement (negative)
Q9	intimacy	not much close	Upgraded to the latest smartphone model (pos)
Q10	intimacy	not much close	Blind date partner suddenly stops contacting (neg)
Q11	intimacy	very intimate	Receives incentive at work (pos)
Q12	intimacy	very intimate	Unrequited crush gets a girlfriend (neg)

Table 3: Combinations of features for DCT question design.

a: Concise human response (avg. 9.1 words)
b: Standard human response (avg.14.1 words) – We picked human response from Dataset1 randomly.
c: Enhanced human response (avg.25.3 words)
d: AI-generated response (avg.42.8 words) – We picked generated response from Dataset2 randomly.
e: etc. (generated appropriate response)

Table 4: Features and average length of DCT question choices.

<Hierarchy setting>
This conversation takes place between {interlocutor + relationship}. {Interlocutor} is experiencing {situation & polarity}. What would you say in response to this?
{Dialogue} \n {choices} \n reason: ________________________

<Intimacy setting>
This conversation is between you and a {distant | not very close | very intimate} {interlocutor}. Currently, {interlocutor} is experiencing {situation & polarity}. What would you say in response? And why did you choose to say that?
{Dialogue} \n {choices} \n reason: ________________________

Table 5: Basic prompt for DCT. Full prompts are in Appendix A and Appendix B.

4 Experimental Settings

4.1 DCT-based task Setup

This study extends beyond evaluating empathy generation and introduces a DCT-based task to assess LLMs' social interpretation and judgment in empathy contexts. Therefore, we constructed 12 DCT items from reprocessed dialogues, varying in three key factors: relationship type (hierarchy vs. intimacy), situational context, and emotional polarity (Table 3). Each item comprised a single-turn prompt extracted from a dialogue instance that necessitated an empathetic response. LLMs and human participants selected the most contextually appropriate response under identical conditions and provided justifications, allowing analysis of their sociocultural reasoning.

The hierarchy condition reflects asymmetrical power relations, typically entailing the use of honorifics. We defined two relationship types: child-to-parent (hierarchical) and friend-to-friend (non-hierarchical). Respecting intimacy, we categorized it into three levels—distant, moderately close, and very close — between non-hierarchical relations. Empathetic scenarios were created by combining these relationships with specific dialogue contexts and the recipient's sentiment state, yielding diverse, context-rich stimuli.

As shown in Table 4, each DCT item included five options (a–e). Option b was a human-generated response; a and c were its shorter and longer variants, modified by researchers. d was an AI-generated response, and e allowed free input. Options were ordered from shortest to longest (a–e), with empathic intensity generally increasing with length. Both humans and AI provided justifications for their choices, enabling reasoning analysis. DCT prompts were presented in two formats based on relational context, as shown in Table 5.

proprietary	Global	Claude3.5 Sonnet, GPT4o
	Korean	HyperClova, Solar 10.7B
open source	Multilingual	Qwen2.5 7B, LLaMA 3.1 8B/70B, LLaMA3.2 3B
	Korean fine-tuned	Qwen2.5 7B-KO, LLaMA 3.1 8B-KO, LLaMA 3.1-70B-KO, LLaMA3.2 3B-KO, EXAONE 3.5-7.8B

Table 6: LLM models using in Experiments. We compared between 1) proprietary and open source models, 2) multilingual-korean specified models, and 3) small and large open source models. Model details in Appendix C.

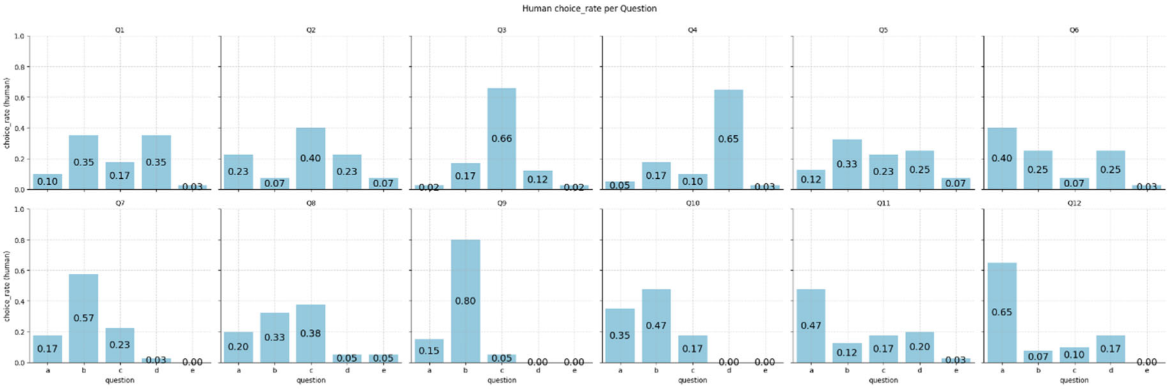


Figure 2: Humans' DCT choice rates per question.

4.2 Human Baseline

Using the DCT from Section 4.1, we collected responses and rationales from 40 Korean native speakers and generated 40 LLM responses per model for comparative analysis. All human participants were Korean natives and non-experts in linguistics, thereby contributing responses that reflect intuitive and naturalistic language use. The sample was balanced across age groups (20s–60s) and gender (54% female, 46% male).

4.3 Models

To consider diverse model characteristics, we evaluated both proprietary and open-source LLMs, ranging from large conversational agents to smaller models (Table 6). The proprietary models comprised globally deployed systems such as GPT-4o, Claude 3.5 Sonnet, and Korean-specialized models like HyperClova and Solar 10.7B, allowing us to investigate the impact of language-specific tuning. Among the open-source models, we focused on the Qwen and LLaMA series, which have multilingual capabilities and vary in model size. Especially LLaMA, we considered training versions. These settings aim to examine the effects of scale and recency on performance. Additionally, we included Korean-specific open-source models such as EXAONE and fine-tuned LLaMA variants (denoted -KO) to assess whether Korean-specific pretraining enhances empathetic performance in Korean dialogue settings.

Each model was provided with a standardized system instruction and performed the DCT under zero-shot conditions, using randomized seeds to replicate the conditions applied to human participants. Additionally, each model underwent 40 runs, enabling a comparison of response variability and consistency.

5 Results

5.1 Quantitative Analysis

Figure 2 presents the distribution of response patterns among human participants. Participants tended to prefer longer responses in hierarchical scenarios (Q1–Q4) and shorter responses in intimacy-based scenarios (Q7–Q12). In scenarios characterized by weaker hierarchical relations (Q5, Q6), shorter responses were also preferred. Similarly, higher levels of intimacy (Q11, Q12) resulted in more concise replies. In instances where the empathy recipient held a lower social status (Q3, Q4), participants strongly favored longer responses. This reflects sociocultural norms in Korean discourse, wherein higher-status speakers are expected to convey not only empathy but also guidance or consolation. In intimacy-based relationships, participants preferred shorter responses (options a and b), with response length modulated by the degree of interpersonal closeness; stronger relational ties were associated with more concise replies. These findings suggest that in close relationships, empathy is conveyed more through

Model	Spearman's	Rank	Model	Spearman's	rank
Claude3.5	**0.52**	1	LLaMA 3.1 (70B) KO	0.11	7
LLaMA 3.2 (3B)	0.29	2	LLaMA 3.1 (70B)	0.09	8
HyperClova	0.24	3	LLaMA 3.2 (3B) KO	0.05	9
LLaMA 3.1 (8B)	0.19	4	EXAONE	0.04	10
LLaMA 3.1 (8B) KO	0.13	5	Qwen2.5 (7B)	-0.05	11
GPT4o	0.12	6	Qwen2.5 (7B) KO	-0.06	12

Table 7: Spearman's correlation (Human-LLM) ranks in model level. Claude 3.5 Sonnet had highest correlation with human response tendencies.

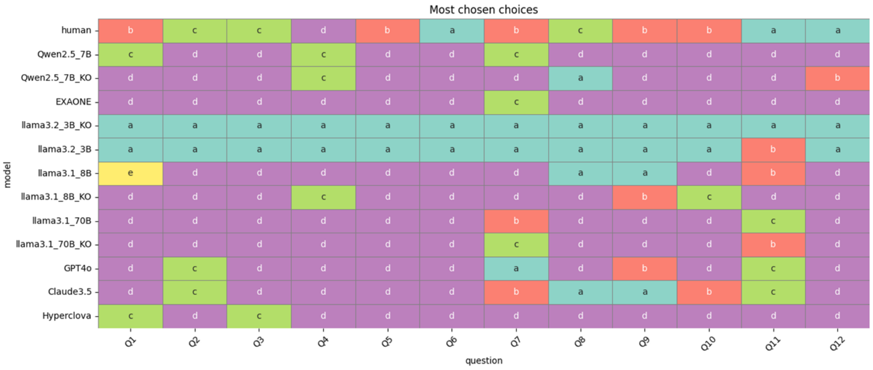


Figure 3: Most chosen answers among humans and LLMs.

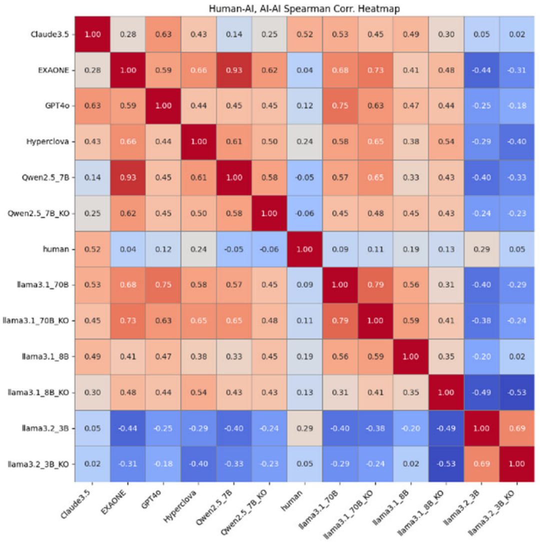


Figure 4: Spearman's correlations in model levels.

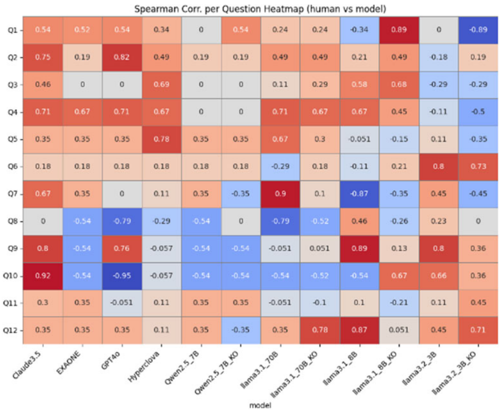


Figure 5: Spearman's correlations in question levels.(Human-LLM)

implicit, contextually grounded cues than through response length. Dialogue context and sentiment polarity had minimal impact on response selection, as empathy judgments remained largely consistent across both positive (odd-numbered) and negative (even-numbered) scenarios.

Figure 3 presents the most frequently selected response option per item, as chosen by human participants and LLMs. Except for LLaMA 3.2, most models exhibited a stronger preference for option d (longer responses) relative to human participants, indicating a general tendency toward over-empathizing. Claude 3.5 demonstrated the highest degree of variability across items, whereas open-source models produced more consistent response patterns.

To complement frequency-based analyses, we employed Spearman's rank correlation (Table 7) to evaluate the alignment between human and LLM responses at both the overall and item-specific levels. Claude 3.5 achieved the highest correlation with human responses (ρ = 0.52), followed by the

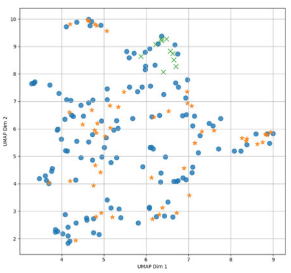
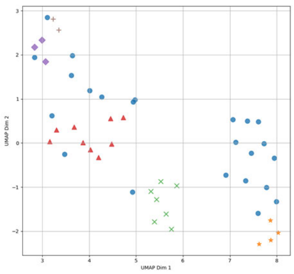
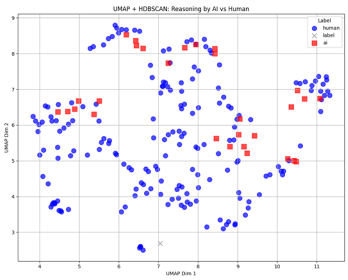
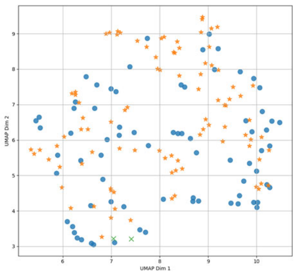
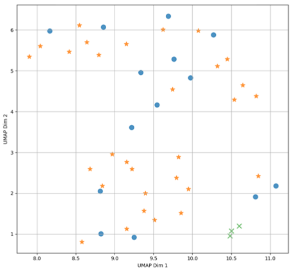
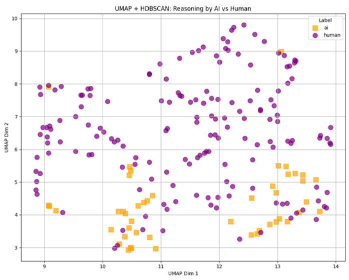
	Human	LLMs	Overall
H			
I			

Table 8: Results of HDBSCAN in sentence embedding level. The sentence embedding distribution for humans is more dispersed, whereas that of LLMs is more constrained. In the 'Human' and 'LLMs' columns, circles indicate noise points, while other shapes represent clusters. In the 'Overall' column, circles denote human embeddings, and squares denote LLMs'. 'H' means hierarchy relationship, and 'I' is intimacy relationship.

LLaMA 3.2 (ρ = 0.29), which demonstrated a consistent preference for shorter replies, aligning more closely with human selection patterns. Notably, LLaMA 3.2 exhibited a low correlation with other models, which may be attributed to its distinctive preference for shorter responses (Figure 4).

Contrary to expectations, model-level correlations indicated that neither model size nor Korean specialization through fine-tuning significantly enhanced alignment with human empathetic responses. These findings imply that language-specific fine-tuning alone may have a limited impact on the development of generalizable empathetic behavior in LLMs. Question-level analysis (Figure 5) revealed stronger human–LLM alignment in hierarchical scenarios than in those based on intimacy. Within intimacy-based items, correlations were lowest in moderately close conditions and highest in very intimate settings.

Dialogue context and sentiment polarity had little impact on correlation. These findings suggest that LLMs exhibit higher alignment with human responses in hierarchical contexts but encounter greater difficulty in interpreting interpersonal distance within intimacy-based scenarios. When relational closeness was ambiguous, models demonstrated weaker contextual understanding, while clearer boundaries improved alignment.

Overall, LLMs exhibited lower response variability and a systematic bias toward longer responses, in contrast to the more balanced patterns observed in human participants. These tendencies highlight their limited grasp of the dynamics between hierarchy and intimacy, as well as their difficulty in adjusting empathic expression appropriately to contextual demands—particularly to response length.

5.2 Qualitative Analysis

To complement the quantitative results, we conducted a semantic analysis of human and AI-selected responses (Table 8). The responses were embedded using a Korean fine-tuned Sentence-BERT model[1], which was used in inference mode without any additional fine-tuning. Semantic clusters were then identified using HDBSCAN (McInnes et al., 2017), a non-parametric clustering algorithm robust to noise and capable of discovering variable-density clusters. We set 4 as the minimum cluster size. To interpret each cluster, we applied TF-IDF to extract representative lexical features, revealing characteristic patterns of empathetic reasoning associated with human and model responses.

This multi-stage analysis was motivated by three considerations. First, similar surface expressions may encode distinct pragmatic meanings in human

[1] snunlp/KR-SBERT-V40K-klueNLI-augSTS

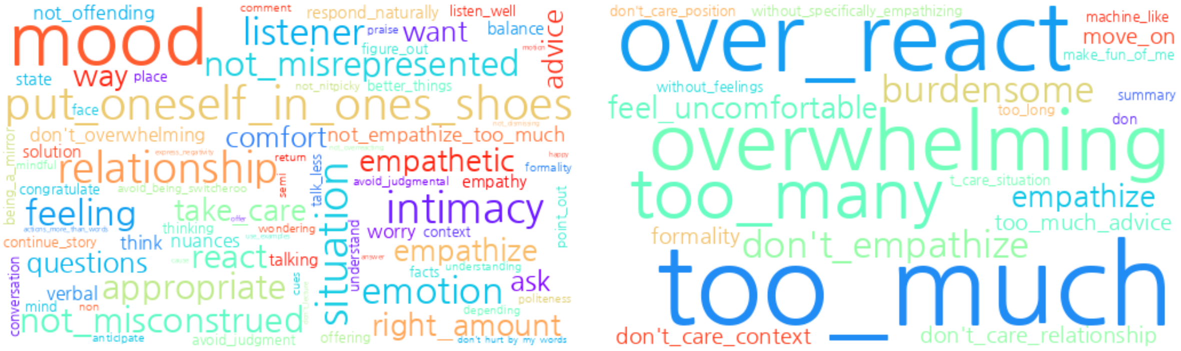


Figure 6: Word clouds summarizing human perspectives on empathic dialogue.

versus LLM-generated responses, necessitating sentence-level semantic comparison. Second, the open-ended nature of the responses precluded the use of predefined categories, making HDBSCAN an appropriate choice. Third, TF-IDF enabled the identification of salient lexical features within each cluster, thereby capturing diverging patterns of empathic emphasis between humans and LLMs.

As shown in Table 8, human responses exhibited a broader distribution in the embedding space than AI responses, reflecting greater semantic diversity and sensitivity to contextual nuance. Both human and AI embeddings were more dispersed in intimacy-based scenarios than in hierarchical ones, indicating that empathy judgments are more complex when social boundaries are less defined.

To explore reasoning differences, we examined a TF-IDF analysis on justification texts within each cluster. In the hierarchy condition, human responses included TF-IDF terms such as "appropriate" and "suitable," reflecting efforts to tailor empathy to the context. In contrast, AI responses featured surface-level labels and generic empathy terms, suggesting a limited ability to interpret context. Similarly, in intimacy scenarios, humans used terms like "close friend" and "not close" to calibrate responses, while AIs again relied on prompt-derived, generic vocabulary. These findings indicate that humans adjust their empathy in response to social closeness and the context of dialogue. In contrast, LLMs struggle to adjust empathetic intensity in response to relational subtlety, particularly in socially ambiguous contexts.

5.3 Human Views on Empathy and LLM Responses

As part of the qualitative analysis, we asked 40 Korean speakers to identify what matters most in empathetic dialogue and the reasons why LLMs' responses are perceived as awkward. As shown in Figure 6, most emphasized inferring emotions and relational stance through subtle cues—captured by the Korean concept of 'nunchi,' a key social skill for appropriate empathy. While LLM responses were fluent and affectively appropriate, participants often found them "unnatural" or "robotic." Word cloud analysis of participant feedback revealed frequent mentions of overdone expressions (e.g., over_react), lack of contextual awareness (e.g., don't_care_context), and poor perspective-taking (e.g., burdensome). These results suggest that genuine empathy requires more than fluency—it depends on adapting to relational and situational contexts, which LLMs still struggle to achieve.

These findings prompt a key question: What defines social competence in LLMs? True social capability requires more than fluent output—it demands sensitivity to relational dynamics, context, emotional tone, and interactional roles. Unlike humans, who adjust their empathy based on subtle cues and social impact, current LLMs lack this calibration. This phenomenon calls for a shift from agent-like models to interactional beings that respond contextually and socially.

6 Conclusions

This study introduces a DCT-based framework to evaluate the capacity of LLMs for socially appropriate empathy, focusing on interpretive judgment rather than generative fluency. By examining responses in Korean, we reveal that LLMs often over-empathize or misread social cues, particularly in intimacy-based or relationally ambiguous contexts. Contrary to expectations, model size and language-specific fine-tuning had minimal effect on performance. Semantic and TF-IDF analyses further show that human speakers modulate empathy based on subtle relational and emotional factors, whereas LLMs rely on surface-level patterns and reactive strategies.

These findings underscore the need for LLMs to move beyond agent-like behavior and toward socially responsive communication. In particular, our framework can inform both the development of empathy-driven applications—such as virtual counseling or companionship—and the detection of manipulative misuse, where artificial empathy may be used to exploit users' trust in high-stakes settings such as voice phishing or persuasive dialogue. In this way, our study contributes to a broader understanding of how empathy should be calibrated, interpreted, and evaluated in socially deployed AI systems.

The proposed DCT framework is simple, flexible, and generalizable, offering a valuable foundation for future research on culturally grounded and socially competent AI. Future work should address current limitations by incorporating more diverse social scenarios (e.g., teacher-student, workplace, stranger interactions) and extending the framework to multi-turn dialogues that better reflect the dynamics of real-world empathy. In addition, building datasets enriched with discourse-level features, such as relationship type, emotion cause, and social distance, will be crucial for developing models aligned with the sociocultural norms of high-context languages like Korean.

7 Limitations

This study does not propose new training models or fine-tuning techniques to improve LLM performance directly. While it analyzes the rationale behind response choices to hedge the black-box nature of the models, it does not identify the exact causes of the observed response biases. Nevertheless, by examining the current limitations of conversational AI in understanding social meaning and introducing a multi-layered evaluation approach centered on social appropriateness and pragmatic judgment, the study offers a foundational contribution to the future design of socially aware language models.

8 Ethics Review

All human participant responses were collected with informed consent. Participants were recruited anonymously and voluntarily with no personally identifiable information recorded. The study did not involve any vulnerable populations and adhered to standard ethical research practices. The use of publicly available datasets was conducted in compliance with their respective usage licenses and privacy policies.

Acknowledgements

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2025-02215142, Development of Pseudonymization Technology for Suspected Criminal Information) and the Yonsei University Research Fund of 2024 (2024-22-0317)

References

He, Z. 1991. Pragmatic Empathy in Daily Verbal Communication. Beijing: Foreign Language Teaching and Research Press.

Gladkova, Anna. 2010. Sympathy, compassion, and empathy in English and Russian: A linguistic and cultural analysis. Culture & Psychology, 16(2), 267-285.

Meiners, Jocelly G. 2017. Cross-cultural and interlanguage perspectives on the emotional and pragmatic expression of sympathy in Spanish and English. The pragmeme of accommodation: The case of interaction around the event of death, 319-348.

Hae Jeong Bak. 2018. Educational Method for Korean Honorification By Comparing With Japanese Language. The Education of Korean Language and Culture 12(1) 1-27. 10.31827/EKLC.2018.12.1.1

Rim Shin. 2021. A study on honorific expression education plan in korean language education focus on chinese korean learners. Matster's thesis. Shilla University. Busan.

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.

Kim Hyunwoo, Byeongchang Kim and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. https://arxiv.org/abs/2408.157872109.08828.

Lai Yuanyuan Eleni Lioliou and Panos Panagiotopoulos. 2021. Understanding Users' switching Intention to AI-Powered Healthcare Chatbots. ECIS.

Wu Shenghan Wynne Hsu and Mong-Li Lee. 2024. EHDChat: A Knowledge-Grounded Empathy-Enhanced Language Model for Healthcare Interactions. Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024).

Kasper Gabriele and Kenneth R. Rose. 2002. Pragmatic development in a second language. Language learning .

Walker Chad. 2019. L1 and L2 Korean evidential use: Using the discourse completion task (DCT). Language facts and Perspective 46 31-55.;

Ogiermann Eva. 2018. Discourse completion tasks. Methods in pragmatics 10. 229-255.

Sperlich Darcy Jaiho Leem and Eui-Jeen Ahn. 2016. The interaction of politeness systems in Korean learners of French. Proceedings of the 30th Pacific Asia Conference on Language Information and Computation. Waseda University.

Edward T. Hall. 1959. The Silent Language. Doubleday New York.

Edward T. Hall. 1976. Beyond Culture. Anchor Press Garden City NY.

Fukushima Saeko and Michael Haugh. 2014. The role of emic understandings in theorizing im/politeness: The metapragmatics of attentiveness empathy and anticipatory inference in Japanese and Chinese. Journal of Pragmatics 74. 165-179.

Lee Hye-Yong. 2022. A Proposal for a Politeness Theory Based on Korean Sociocultural Context. Korean Semantics 78 383–409.

Jung Ji-Hoon. 2023. A Study on the Factors and Patterns of Politeness Judgment: Focusing on Interactional Politeness. Proceedings of the Discourse and Cognitive Linguistics Society of Korea Conference 191–202.

Qiu, Huachuan, and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. https://arxiv.org/abs/2408.15787.

Ruosi Shao. 2023. An Empathetic AI for Mental Health Intervention: Conceptualizing and Examining Artificial Empathy. In Proceedings of the 2nd Empathy-Centric Design Workshop (EmpathiCH '23). Association for Computing Machinery, New York, NY, USA, Article 4, 1–6. https://doi.org/10.1145/3588967.3588971

Naik Aditya, Thomas Jovi,Sree Teja, Reddy Himavant. 2025. Artificial Empathy: AI based Mental Health. https://arxiv.org/abs/2506.00081.

Carrasco-Farre, C. 2024. Large language models are as persuasive as humans, but how? About the cognitive effort and moral-emotional language of LLM arguments. https://arxiv.org/abs/2404.09329.

Sayak Saha Roy, Poojitha Thota, Krishna Vamsi Naragam, & Shirin Nilizadeh. 2024. From chatbots to phishbots?: Phishing scam generation in commercial large language models. In 2024 IEEE Symposium on Security and Privacy (SP) (pp. 36-54). IEEE.

Trinh, Quang Minh, Samiha Zarin, and Rezvaneh Rezapour. 2025. Master of Deceit: Comparative Analysis of Human and Machine-Generated Deceptive Text. In Proceedings of the 17th ACM Web Science Conference 2025 (pp. 189-198).

McInnes, Leland, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. J. Open Source Softw., 2(11), 205.

Datasets

Korean Empathetic Dialogues (2022) from AIHub https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=data&dataSetSn=71305

Appendix A Example of whole DCT Prompt - Hierarchy

You are a native speaker of Korean.
You are about to participate in a survey designed to explore how you prefer to express empathy, depending on the hierarchical and interpersonal relationship between you and the listener.
Please respond to the survey according to the instructions provided below.

1. Survey Instructions

Carefully read the relationship between the speaker and the listener described in each prompt.
After reading the listener's statement, choose the most appropriate response from options a to d.
※ If none of the options seem appropriate, select "e. Other" and write your own response.
Briefly explain the reason for your selected or written response.

2. Consider the Social relationship to Complete the Dialogue (2 factors)

Hierarchy: How much higher in status is the other person compared to you?
Intimacy: How close are you to the other person?

- ✓ Distant acquaintance: Someone you've only met once or twice, or barely know (not someone you dislike or have conflict with)
- ✓ Not much close friend: A friend you see occasionally, such as in a business or professional setting
- ✓ Very close friend: A best friend with whom you've had a long-standing, close relationship

<Dialogue>

<1> [위계 관계] 이 대화는 재영이와 아버지의 대화입니다.
<1> [Hierarchy] This conversation is between Jaeyoung and his father.
현재 재영이의 아버지는 퇴임 후 어머니와 함께 시간을 보낼 수 있어 기쁜 상태입니다.
Jaeyoung's father is currently feeling happy because he can now spend more time with his wife after retirement.
당신이 재영이라면 아버지의 말을 듣고 이어서 할 말로 가장 적절한 것을 골라주세요.
If you were Jaeyoung, please choose the most appropriate response following your father's statement.
그리고 <1>에 이와 같이 응답한 이유는 무엇인가요?
Also, please explain why you responded this way in <1>.

아버지: 드디어 정년 퇴직이야. 이제 마음껏 쉴 수 있어 기쁘다.
Father: I'm finally retiring. I'm happy that I can now rest as much as I want.
네 엄마랑 여행도 다니고 오손도손 그렇게 지낼 생각하니까 벌써 신나고 설레는 거 있지?
Thinking about traveling with your mom and spending peaceful time together already makes me excited.

재영(Jaeyoung): ____

a. 축하 드려요 아빠. 드디어 자유시네요 저도 기뻐요.
a. Congratulations, Dad. You're finally free! I'm happy for you too.
b. 와 축하해요 아빠. 정년 퇴직이라니! 너무 멋저요. 이제 여행도 다니시면서 편하게 지내세요.
b. Wow, congratulations, Dad! Retirement—how amazing! Now you can travel and relax.
c. 와 축하 드려요 아버지. 정년 퇴직이라니 너무 멋져요. 그동안 고생 많으셨죠? 이제 여행도 다니시면서 편하게 지내세요. 사랑해요.
c. Wow, congratulations, Father. Retirement is wonderful. You've worked so hard all this time. Now, enjoy your travels and take it easy. I love you.
d. 와, 정말 축하드려요, 아빠! 그동안 고생 많으셨어요. 이제는 여유롭게 하고 싶으신 것들 하실 수 있어서 정말 다행이에요. 엄마랑 여행 다니시면서 좋은 추억 많이 만드세요. 어디부터 가고 싶으세요?
d. Wow, huge congratulations, Dad! You've been through so much. I'm so glad you can now do what you want at your own pace. Make lots of great memories traveling with Mom. Where would you like to go first?

e. 기타: ____

e. Other: ____

Appendix B Example of whole DCT Prompt - Intimacy

You are a native speaker of Korean.
You are about to participate in a survey designed to explore how you prefer to express empathy, depending on the hierarchical and interpersonal relationship between you and the listener.
Please respond to the survey according to the instructions provided below.

1. Survey Instructions

Carefully read the relationship between the speaker and the listener described in each prompt.
After reading the listener's statement, choose the most appropriate response from options a to d.
※ If none of the options seem appropriate, select "e. Other" and write your own response.
Briefly explain the reason for your selected or written response.

2. Consider the Social relationship to Complete the Dialogue (2 factors)

Hierarchy: How much higher in status is the other person compared to you?
Intimacy: How close are you to the other person?

- ✓ Distant acquaintance: Someone you've only met once or twice, or barely know (not someone you dislike or have conflict with)
- ✓ Not much close friend: A friend you see occasionally, such as in a business or professional setting
- ✓ Very close friend: A best friend with whom you've had a long-standing, close relationship

<Dialogue>

<8> [친소 관계] 이 대화는 당신과 안 친한 지인과의 대화입니다.

<8> [Intimacy] This conversation is between you and a distant acquaintance.

현재 당신의 지인은 영어학원을 계속 다니는데 실력이 늘지 않아서 슬프고 화가 난 상태입니다.

Your acquaintance is feeling sad and frustrated because their English skills haven't improved despite attending a language academy for a while.

지인의 말을 듣고 당신이 이어서 하고 싶은 말을 골라주세요.

Please choose what you would like to say in response to their statement.

그리고 <8>에 이와 같이 응답한 이유는 무엇인가요?

Also, explain why you responded that way in <8>.

지인: 영어 학원을 다닌지 벌써 반 년이 다 되어 가는데도 아직도 영어로 자기 소개도 못 해요. 학원 순 엉터리 아니에요? 돈만 버린 거 같아서 너무 화가 나네요.

Acquaintance: It's been almost six months since I started going to the English academy, but I still can't even introduce myself in English. Isn't the academy totally useless? I feel like I just wasted my money, and it makes me so angry.

나(You): ____

a. 반 년이나 다녔는데 실력이 안 늘어서 속상하셨구나. 그래도 조금만 더 꾸준히 해 보시는 게 어때요? 실력은 곧 늘 거예요.

a. You must feel upset that your skills haven't improved even after six months. Still, how about sticking with it just a bit longer? Your skills will improve soon.

b. 에고 속상하셨구나. 저도 그런 적 있어서 무슨 마음인지 알아요. 완전 속상하죠. 그래도 조금만 더 꾸준히 해 보시는 게 어때요? 실력은 곧 늘 거예요. 힘 내요!

b. Oh no, that must be frustrating. I've been through that too, so I know how it feels. It's really upsetting. Still, how about continuing just a bit more? You'll get better soon. Hang in there!

c. 에고 괜찮아요? 많이 속상하셨나 보네요. 저도 그 마음 알 거 같아서 완전 공감 돼요. 정 그러면 학원을 옮겨보시는 게 어때요? 학원이 문제가 있는 거 같아요. 너무 우울해하지 마시고 조금만 더 힘 내봐요.

c. Are you okay? You must have been really upset. I think I understand how you feel—I totally empathize. If that's the case, how about trying a different academy? It seems like this one might not be working. Don't be too discouraged. Just hang in there a little longer.

d. 아이고, 반 년이나 노력했는데 아직 성과가 안 보이면 정말 속상하겠네요. 학원에 대한 기대가 컸을 텐데 그런 결과가 나오니까 화가 날 수밖에 없죠. 자기 소개 같은 기본적인 부분도 못 배운 것 같다면, 학원의 수업 방식이 기대와 잘 맞지 않았던 걸 수도 있어요. 혹시 방법을 바꿔서 다른 학원을 알아보거나, 자기 주도 학습 방식으로 연습해 보는 건 어때요? 짧게라도 매일 자기소개를 연습하거나, 간단한 문장들을 반복하는 것도 도움이 될 거에요.

d. Oh dear, after working hard for six months with no visible results, it must be really upsetting. You probably had high hopes for the academy, so it's only natural to feel angry about the outcome. If you haven't even learned basic things like self-introductions, the teaching method might not have been a good fit. Maybe try a different academy or switch to a more self-directed learning approach? Even practicing short self-introductions daily or repeating simple sentences could really help.

e. 기타: ________________________________

e. Other: ________________________________

Appendix C Model details

Open /closed	Language	Name	Model version (URL)	note
proprietary	Global	Claude3.5 Sonnet	claude-3-5-sonnet-20241022	
		GPT4o	gpt-4o-2024-11-20	
	Korean	HyperClova	HCX-003	
		Solar 10.7B	solar-mini-250123	
open source	Multi-lingual	Qwen2.5 7B	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct	Models are from hugging-face
		LLaMA3.1 8B	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct	
		LLaMA 3.1 70B	https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct	
		LLaMA 3.2 3B	https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct	
	Korean fine-tuned	Qwen2.5 7B -KO	https://huggingface.co/beomi/Qwen2.5-7B-Instruct-kowiki-qa	
		LLaMA 3.1 8B -KO	https://huggingface.co/SEOKDONG/llama3.1_korean_v1.1_sft_by_aidx	
		LLaMA 3.1 70B -KO	https://huggingface.co/Bllossom/llama-3.2-Korean-Bllossom-3B	
		LLaMA 3.2 3B -KO	https://huggingface.co/Saxo/Linkbricks-Horizon-AI-Korean-llama3.1-sft-dpo-70B	
		EXAONE 3.5 -7.8B	https://huggingface.co/LGAI-EXAONE/EXAONE-3.5-7.8B-Instruct	