Unveiling the Limits of Large Language Models in Inferring Pragmatic Meaning from Non-Verbal Responses

Sugyeong Eo[1], **Heuiseok Lim**[2*]
[1]Department of Software, Yonsei University Mirae Campus, Republic of Korea
[2]Department of Computer Science and Engineering, Korea University, Republic of Korea
s.eo@yonsei.ac.kr limhseok@korea.ac.kr

Abstract

Although large language models (LLMs) have shown considerable progress in pragmatic language understanding, prior research has focused mainly on their comprehension of verbal behavior. Nonetheless, non-verbal behavior remains a fundamental component of human communication, especially when deliberately utilized in isolation to convey indirect meanings. In this work, we present the first systematic evaluation of LLMs' ability to infer pragmatic meaning in dialogue consisting solely of non-verbal responses. We explore three research questions: (1) *Can LLMs recognize indirect intent conveyed through non-verbal responses?* (2) *When and how do LLMs fail to capture non-verbal intent?* (3) *How can we improve LLMs' ability to interpret non-verbal intent?*. Through the evaluation, we observe that LLMs struggle to infer underlying meaning from non-verbal responses, with accuracy dropping by up to 60% points compared to verbal ones. Further extensive analysis reveals a behavioral pattern in LLMs' interpretations of non-verbal behavior and demonstrates that in-context learning facilitates pragmatic inference.

1 Introduction

Recent years have demonstrated remarkable progress in large language models (LLMs), with strong generalization across diverse natural language processing tasks (Brown et al., 2020; Ruis et al., 2023b; Wei et al., 2022a,b). Building on this progress, an increasing number of studies have explored whether LLMs are capable of understanding non-literal language phenomena, such as implicature and indirect speech acts, which require sensitivity to cognitively grounded aspects of human communication (Grice, 1975; Austin, 1975; Searle, 1975; Hu et al., 2023; Lee et al., 2025). Notably, LLMs have demonstrated growing competence in pragmatic language understanding, extending their capabilities beyond surface-level understanding to the inference of implicit meaning (Wu et al., 2024; Park et al., 2024).

Model	Verbal	Non-Verbal	Δ
3B-Llama	0.96	0.37	-0.59
3.8B-Phi	1.00	0.41	-0.59
8B-Llama	0.88	0.37	-0.51
8B-Ministral	0.48	0.46	-0.02
7B-Qwen	0.94	0.47	-0.47
12B-Mistral-NeMo	0.96	0.38	-0.58
14B-Qwen	0.78	0.31	-0.47
14B-Phi	0.98	0.48	-0.50
32B-Qwen	0.90	0.58	-0.32
72B-Qwen	0.94	0.76	-0.18
70B-Llama	0.96	0.70	-0.26
GPT-4.1-mini	0.88	0.61	-0.27
GPT-4o	0.82	0.46	-0.36
GPT-o3	0.96	0.73	-0.23
Claude-3.7-Sonnet	0.94	0.54	-0.40
Human	0.99	0.91	-0.08

Table 1: Performance comparison of LLMs on verbal and non-verbal response settings. We report accuracy for both settings and their performance gap (Δ).

Despite these advances, existing research has primarily focused on assessing LLMs' ability to interpret verbal responses, with non-verbal behavior typically considered only in conjunction with verbal input. However, in real-world contexts, non-verbal elements constitute a fundamental component of communication and often operate independently as exclusive signals of communicative intent (Wharton, 2009). For instance, a person remaining silent after being asked "Am I disturbing you?" is pragmatically interpreted as an indirect indication that the speaker is causing a disturbance. Such non-verbal responses are deliberately employed to convey indirect intent. While these interactions are intuitive to humans, it remains an open question whether LLMs are capable of recognizing commu-

*Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45304–45316
July 2-7, 2026

nicative intent solely from non-verbal responses. Accurate understanding of intended non-verbal behaviors is especially important, as they are universally recognized communicative cues that play a vital role when spoken language is limited by linguistic boundaries (Burgoon et al., 2021).

In this work, to the best of our knowledge, we present the first systematic evaluation of LLMs in understanding responses composed solely of non-verbal elements within dialogue. We categorize three types of non-verbal responses and assess model performance: silence, facial expressions, and movements. To guide our evaluation, we establish the following research questions: (1) *Can LLMs recognize indirect intent conveyed through non-verbal responses?* (2) *When and how do LLMs fail to capture non-verbal intent?* (3) *How can we improve LLMs' ability to interpret non-verbal intent?*.

We conduct extensive experiments across six model families, covering parameter sizes from 3B to over 100 billion. Remarkably, although most LLMs achieve near-human performance in interpreting intent from verbal responses, their accuracy drops by as much as 60% points when processing responses composed solely of non-verbal signals. Performance is notably poor in the silence category, where LLMs achieve only 0.45 accuracy compared to 0.91 for human-level performance, indicating a clear failure to capture implicit intent. Further analysis indicates that the majority of errors stem from abstract and literal descriptions, which restrict interpretations to surface-level features and fail to align with the underlying communicative intentions. Further investigation into the potential for enhancing this capability reveals that few-shot in-context learning substantially enhances the alignment of surface-level cues with the underlying communicative intentions. Our contributions are threefold:

- To the best of our knowledge, this work provides the first systematic evaluation of LLMs' ability to infer underlying intent from dialogue responses composed exclusively of non-verbal signals.

- Comprehensive experiments spanning six prominent LLM families (3B–100B+) demonstrate that the models face persistent challenges in inferring underlying meaning, resulting in markedly diminished performance relative to their verbal counterparts.

- Further analysis reveals LLM behavioral patterns and shows that few-shot in-context learning enhances their ability to comprehend non-verbal intent.

2 Related Work

Pragmatic Language Understanding Pragmatics is the systematic study of meaning that arises from language use, encompassing key theoretical constructs such as implicature, indirect speech acts, and deixis (Grice, 1975; Searle, 1975; Levinson, 1983). Recent efforts have investigated whether LLMs handle such phenomena by proposing new benchmarks (Qi et al., 2023; Sravanthi et al., 2024; Park et al., 2024). Additionally, attempts have been made to enhance LLMs' pragmatic abilities, including modifying training objectives (Wu et al., 2024) or designing prompts that target specific pragmatic phenomena (Ruis et al., 2023a; Lee et al., 2025). In parallel, researchers have analyzed common failures in LLMs' pragmatic reasoning. For example, Hu et al. (2023) observe that LLMs often exhibit human-like error patterns, while Sravanthi et al. (2024) highlight particular difficulty with contrasting distractors. Prior studies mainly focus on verbal cues, with non-verbal behavior typically serving as contextual support for interpreting verbal responses.

Understanding Non-verbal Cues in Language Models Non-verbal behavior refers to communicative signals that are conveyed without the use of words (Knapp et al., 1972; Hall et al., 2019). Such signals include facial expressions, movements, prosody, and other behavioral cues that play a vital role in expressing intent during communication. Given the communicative significance of non-verbal behavior, recent work has explored incorporating these signals into language models. For instance, Lee et al. (2023) develop an empathetic LLM by conditioning it on non-verbal cues. Other approaches leverage emojis as affective indicators to infer emotional states from text (Felbo et al., 2017). Hakami et al. (2023) examine the role of facial expressions in detecting sarcasm. These studies highlight that non-verbal cues enrich pragmatic interpretation when combined with verbal input. This study departs from previous approaches by focusing exclusively on scenarios where verbal utterances are entirely absent, and meaning must be inferred purely from non-verbal signals.

Catetory	Options	Content
Question		Choose the option that most appropriately interprets the context and underlying intent of the conversation below. Assume that both parties faithfully engage in the conversation.
Silence	Conversation	A: Can you spare some time this weekend? I'm looking for someone to help me move. B: ...
	Option-1	B is indirectly rejecting A's request by deliberately remaining silent.
	Option-2	A is asking B for help with moving, which is the act of relocating to a different place.
	Option-3	A is asking B for help with moving, but B has nothing to say and didn't respond.
	Option-4	In response to A's request, B is expressing a positive intention through silence.
	Option-5	A is asking a question, and B is not responding to it.
Facial Expressions	Conversation	A: You know you can't miss our team dinner again, right? B: 🙄
	Option-1	B looked up after A spoke. Maybe there was something on the ceiling.
	Option-2	B seems to be avoiding the team dinner through evasive actions.
	Option-3	B expresses the intention to participate in the gathering by raising their eyes after A spoke.
	Option-4	A is talking about team dinners, which help build team spirit.
	Option-5	A is asking a question, and B is showing only a facial expression in return.
Movements	Conversation	A: Starting today, write 10 pages of the report every day. B: (Points to the broken laptop.)
	Option-1	B agrees with A's statement but did not mention it directly, considering social propriety.
	Option-2	B, not fully understanding A's statement, pointed to a broken laptop to change the subject.
	Option-3	B picked up the broken laptop to fix it.
	Option-4	B indirectly expresses that they cannot do it because their laptop is broken, in response to A's statement.
	Option-5	A is asking a question, and B shows only an action in return.

Table 2: An example presenting the three evaluation categories of silence, facial expressions, and movements. The answer is highlighted in yellow.

3 Methodology

3.1 Problem Scope

This study focuses on scenarios where responses consist solely of non-verbal behaviors. While paralinguistic cues such as pitch and amplitude are part of non-verbal communication, we exclude them from consideration as they mainly operate in conjunction with verbal behavior. To evaluate pragmatic competence, we instruct the model to *"choose the option that most appropriately interprets the context and underlying intent of the conversation."*. We further include the instruction *"Assume that both participants are faithfully engaged in the conversation"* in the prompts, ensuring that non-verbal responses are intentionally structured to convey communicative intentions (Ekman and Friesen, 1969). The evaluation examines how well LLMs infer intended meaning solely from non-verbal cues, without verbal utterances.

3.2 Three Categories of Non-verbal Behavior

Grounded in the theoretical foundations outlined in Appendix A, this study categorizes non-verbal behavior into facial expressions and bodily movements, while additionally conceptualizing silence as a distinct category owing to its absence of overt communicative signals.

Silence Silence is not merely the absence of speech, but rather constitutes a deliberate and strategic communicative act whose meaning must be inferred from context (Johannesen, 1974; Jaworski, 1992; Lane et al., 2002; Bruneau, 1973; Jensen, 1973). This serves as a communicative signal, conveying implicit meanings such as refusal, deliberate avoidance, and tacit agreement. For instance, silence following a request often implies rejection, whereas silence in response to a question involving sensitive information typically indicates avoidance. Although it carries little explicit information, silence plays a critical role in conversational pragmatics as a powerful contextual signal.

Facial expressions Facial expressions serve as a visual symbolic form of non-verbal expression, often conveying emotions, intentions, or pragmatic nuances (Dresner and Herring, 2010; Hayati et al., 2019). Despite their ubiquity in everyday communication, understanding the meaning of a facial expression requires contextual understanding. For example, the facial expression " 😅 " (smiling face with sweat) is often used to express embarrassment, or polite avoidance. Similarly, the facial expression " 🙄 " (face with rolling eyes) typically conveys annoyance or disbelief, even though no verbal cue is provided. In this way, facial expressions, while visually simple, encode rich pragmatic meaning that must be inferred through contextual reasoning.

Movements Movements include gestures and postures, function as non-verbal cues that reflect a speaker's psychological or interpersonal stance, playing a crucial role in conveying implicit intentions (Patterson, 2012; Hall et al., 2019; Wharton, 2009). For instance, movements such as looking away may suggest intentional avoidance or discomfort, while scratching one's head often implies confusion, hesitation, or disagreement. The interpretation of such movements is highly context-dependent, as identical behaviors can convey different pragmatic meanings depending on situations.

3.3 Evaluation Protocol

We evaluate LLMs' ability to interpret non-verbal responses using a set of multiple-choice questions, each consisting of five answer options. Given a prompt comprising a `[question]`, `[a dialogue context with a non-verbal response]`, and `[five candidate options]`, the model is tasked with selecting the interpretation that most accurately reflects the conversational context and the underlying communicative intent.

Distractors are carefully designed to represent four incorrect types: (1) Misinterpretations, which involve misunderstanding the intended meaning of the response (Sravanthi et al., 2024); (2) Literal explanations, which rely on literal readings while disregarding the pragmatic context (Hu et al., 2023); (3) Lexically coherent but semantically divergent responses, which display surface-level lexical similarity but lack contextual relevance (Yue et al., 2024; Zheng et al., 2021); and (4) Abstract explanations, which consist of vague descriptions that even fail to address the core topic of the dialogue (Peters and Chin-Yee, 2025).

The representation scheme for non-verbal responses is formulated on the basis of two principal considerations. To enable rigorous evaluation of an LLM's understanding of non-verbal behavior, we clearly separate the form of non-verbal responses from verbal ones and define an independent response format. At the same time, in order to encompass both general conversational contexts and web chat environments, response types are specified to accommodate both forms of communication. We represent responses in the silence category as '...' (Wu et al., 2025; Li et al., 2017), facial expressions using Unicode emojis (Felbo et al., 2017; Hakami et al., 2023; Hu et al., 2017), and movements as descriptive actions enclosed in parentheses (Dirik et al., 2021; Chen et al., 2022). Descriptions of the detailed dataset construction, quality control, and statistics are provided in Appendix C, while an illustrative example is presented in Table 2.

4 Experiments

This section reports a systematic evaluation of LLMs' ability to interpret non-verbal behavior, structured around three guiding research questions: (1) Can LLMs recognize indirect intent conveyed through non-verbal responses? (2) When and how do LLMs fail to capture non-verbal intent? (3) How can we improve LLMs' ability to interpret non-verbal intent?

4.1 Evaluation Setup

We employ publicly available instruction-tuned language models to assess their ability to infer implicit intentions conveyed through non-verbal cues. Each input is framed as a multiple-choice question comprising five candidate options, and the model is tasked with selecting the most appropriate option. To better reflect real-world scenarios, all experiments are conducted in a zero-shot setting.

We conduct comprehensive experiments across 13 models with six different model families. The HuggingFace checkpoints used in this study are provided as follows:

- **Qwen**: 7B (`Qwen/Qwen2.5-7B-Instruct`), 14B (`Qwen/Qwen2.5-14B-Instruct`), 32B (`Qwen/Qwen2.5-32B-Instruct`), 72B (`Qwen/Qwen2.5-72B-Instruct`)
- **Mistral**: 8B (Ministral) (`mistralai/Ministral-8B-Instruct-2410`), 12B (NeMo) (`mistralai/Mistral-Nemo-Instruct-2407`)
- **Llama**: 3B (`meta-llama/Llama-3.2-3B-Instruct`), 8B (`meta-llama/Llama-3.1-8B-Instruct`), 70B (`meta-llama/Llama-3.3-70B-Instruct`)
- **Phi**: 3.8B (`microsoft/Phi-3-mini-4k-instruct`), 14B (`microsoft/phi-4`)
- **API-based models**: `GPT-4.1-mini`, `GPT-4o`, `GPT-o3`, `Claude-3.7-Sonnet`

To ensure consistency across experiments, we uniformly apply a decoding temperature of 0.0 for deterministic outputs and limit the maximum input length to 512 tokens for all models. All evaluations are conducted on four NVIDIA A6000 GPUs (48GB). With regard to the human evaluation setup for comparison with model scores, six undergraduate students are recruited and asked to complete the tasks under identical experimental conditions[1].

[1]Annotators were compensated at rates exceeding the

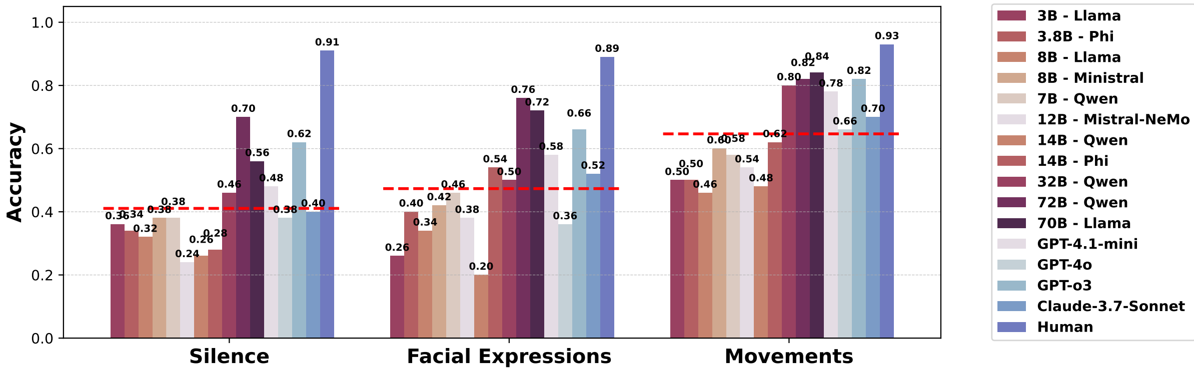

Figure 1: Evaluation of LLM performance across three non-verbal response categories: silence, facial expressions, and movements. The red dotted line represents the average performance of all models within the category.

Since the dataset encompasses everyday social interactions, no restrictions are placed on the annotators' areas of expertise. In addition, as the dataset includes globally recognized communicative practices, annotators are recruited from diverse national backgrounds (American, Korean, Vietnamese, and Bangladeshi), all of whom are proficient in English. The inter-annotator agreement is 0.7924 as measured by Krippendorff's alpha, demonstrating substantial agreement among the annotators.

4.2 RQ1. Can LLMs recognize indirect intent conveyed through non-verbal responses?

Performance of LLMs across categories Figure 1 reports the performance of language models across three categories of non-verbal responses. Among the three categories, the movement category shows the highest average accuracy of 0.67, followed by facial expressions with 0.51 and silence with 0.45. While the performance on movement-based cues appears relatively promising, it remains substantially below the human-level accuracy of 0.93, highlighting the considerable gap between model and human comprehension of implicit intent. The notably low performance in the silence condition is attributed to communicative richness, which poses greater challenges for inference by LLMs. In contrast to movements and facial expressions, the silence category conveys little to no explicit information, substantially increasing the difficulty of pragmatic inference for language models. The findings indicate that, while humans are adept at interpreting subtle social cues, LLMs face difficulties, especially when the response provides minimal explicit information.

Notably, even the widely used GPT-4o demonstrates subpar performance in these settings. It achieves accuracies of 0.38, 0.36, and 0.66 in the silence, facial expressions, and movements, respectively, which is comparable to the performance observed in open-source 14B models. These suggest that, despite their advanced reasoning and factual competence, even highly capable LLMs struggle to interpret cues composed solely of non-verbal behavior. Even the thinking model GPT-o3 achieves performance comparable to that of the Llama 70B model, yet both fall significantly short of human-level accuracy, highlighting the challenges.

legally mandated minimum wage in the respective countries.

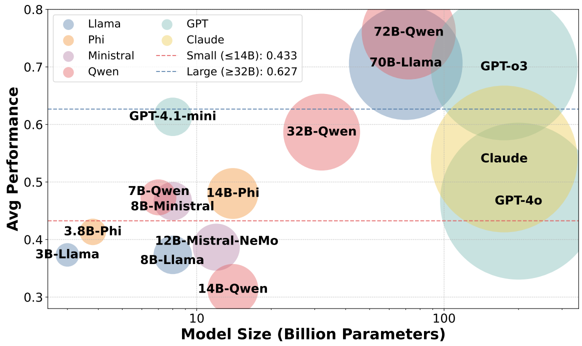

Figure 2: Distribution of average performance across LLMs of varying parameter scales

The effect of parameter scale on model performance Figure 2 visualizes the average performance of each model across the three non-verbal categories. The results show that models in the 3B to 14B range exhibit relatively lower performance, whereas larger models with 32B and 70B parameters tend to achieve higher accuracy. In particular, models from the Llama, Qwen, and Phi families display a consistent trend of improved performance as parameter size increases. These findings imply that scaling up parameter size enhances

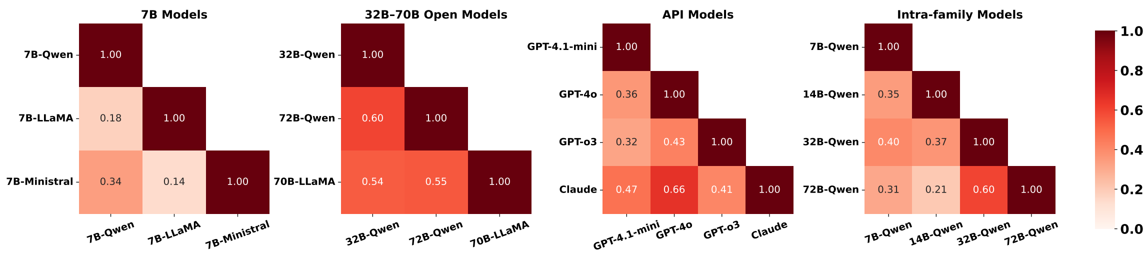

Figure 3: Heatmap of prediction correlations across language models. The heatmap visualizes response-level agreement between models, grouped by model family and parameter size.

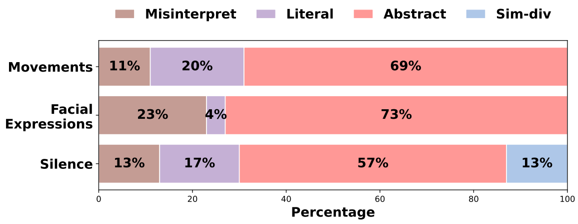

Figure 4: Distribution of errors across four distractor types: misinterpretation, literal explanation, abstract explanation, and lexically coherent but semantically divergent response (sim-div). The analysis is based on the results of Llama-70B, Qwen-32B, and GPT-4o.

a model's capacity to infer intentions from non-verbal responses. However, this pattern does not extend to API-based models. GPT-4.1-mini surpasses GPT-4o, and Claude and GPT-4o underperform relative to their scale. These observations suggest that, while increasing model size within the same architecture generally leads to improved performance, additional external factors may also play a non-negligible role in shaping model effectiveness.

Correlational patterns in model predictions To better understand the behavioral similarity among language models, we analyze prediction correlations based on response-level outputs, grouped by model family and parameter scale. The results depicted in Figure 3 show that 7B-scale models exhibit less consistent response patterns, suggesting that they lack sufficiently consolidated knowledge for interpreting non-verbal behavior. Models with 32B to 70B parameters tend to exhibit increased positive correlations, suggesting a degree of shared underlying reasoning strategies. However, this also implies that such models show similar error patterns, pointing to potential shared limitations in their pragmatic inference capabilities.

We observe that intra-family prediction correlations tend to increase with parameter size, further supporting the finding that larger models exhibit more aligned interpretive behavior when processing non-verbal signals. In conclusion, the predictive tendencies are more strongly influenced by parameter scale than by shared architectural similarity.

Evaluating model performance in verbal and non-verbal response settings The ability of LLMs to infer contextual intent from non-verbal-only responses remains limited, particularly in categories such as silence and facial expressions. These findings raise the possibility that the difficulty does not lie solely in the type of communication, but rather in a broader limitation of LLMs to capture indirect and implicit contextual cues. To further explore this hypothesis, we conduct an additional experiment using a verbal setting, in which the responses are indirect but explicitly verbalized. An example is: "`A: Can you spare some time this weekend? I'm looking for someone to help me move. B: Oh, my legs hurt,`" where speaker B indirectly declines the request by offering an excuse rather than a direct refusal.

As shown in Table 1, LLMs demonstrate notably strong performance under the verbal condition. Even relatively small models, such as those with 3B or 8B parameters, achieve near-human performance, with the exception of the 8B-Ministral. These results indicate that LLMs correctly infer intent when conveyed through explicit verbal responses. In contrast, performance in the non-verbal condition shows a significant decrease. Models with 3B parameters exhibit performance drops of up to 0.59, and even models with 32B to 70B parameters report a reduction of up to 0.32. This degradation is also observed in API-based models. In comparison, human performance remains nearly consistent across both settings, suggesting that current LLMs still have room for improvement.

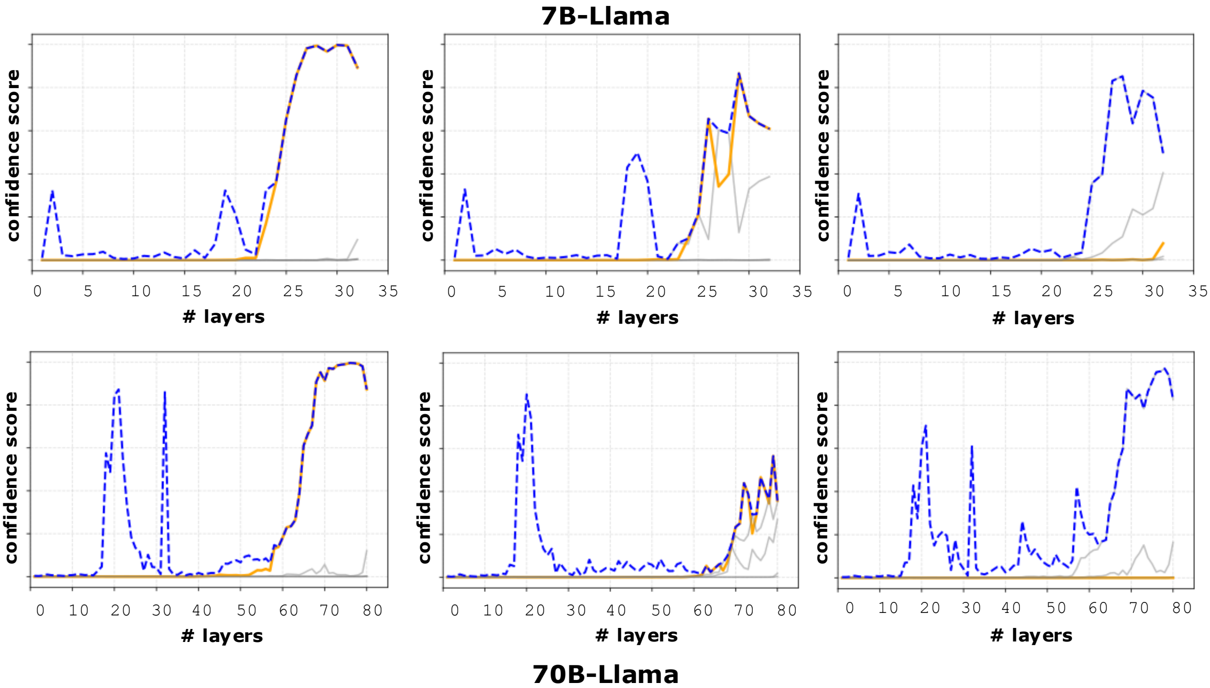

Figure 5: Layer-wise probability analysis via logit lens. The blue dashed line shows the top-1 probability across layers, with the correct option in orange and the distractor options in gray.

4.3 RQ2. When and how do LLMs fail to capture non-verbal intent?

Error analysis Our focus shifts to a deeper exploration of the error patterns that emerge in LLMs' interpretation of non-verbal responses. We analyze the selection behavior of LLMs across four predefined distractor types, and the corresponding error distributions are presented in Figure 4. The results reveal that abstract explanations constitute the predominant source of error, accounting for more than half of the cases across all categories. Even when explicitly instructed to select the option that best reflects the dialogue context and underlying communicative intent, the models often default to narrowly oversimplified descriptions. This pattern reflects a tendency of the models to favor broadly applicable yet minimally variable responses (Peters and Chin-Yee, 2025), limiting their ability to engage in deeper inferential reasoning. A related error type is literal interpretation, in which the responses become granular but still fail to capture the intended meaning. Taken together, these findings suggest that the primary source of the models' limited pragmatic competence lies in their focus on shallow and surface-level representations over the recognition and reasoning of latent intentions embedded in dialogue. In contrast, as shown in Table 1, the models exhibit comparatively strong performance on verbal responses, suggesting that the core issue is not an inherent deficit in pragmatic reasoning itself but rather insufficient grounding for non-verbal signals. Meanwhile, misinterpretation accounts for approximately 11% to 23% of all errors and is predominantly elicited by contrastive distractors, which is consistent with previous findings reported by Sravanthi et al. (2024).

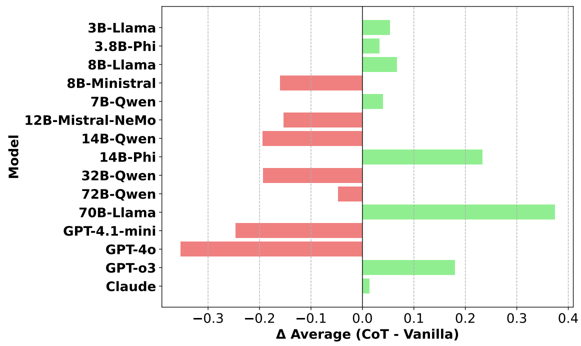

Figure 6: Average performance change after applying CoT prompting, computed by subtracting each model's vanilla score from its corresponding CoT score

Behavioral analysis of the model To advance a deeper analytical understanding, we investigate the internal behavioral patterns of LLMs. We aim to examine from which layer the model tends to attend to pragmatic cues, and how internal pat-

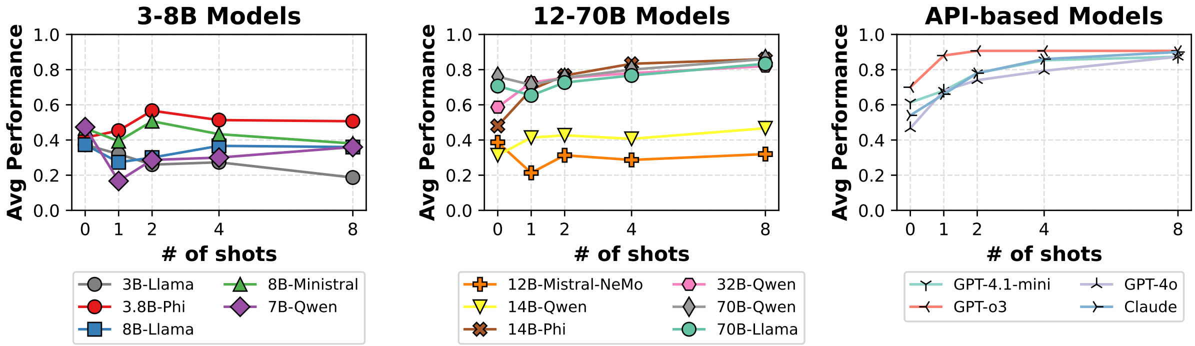

Figure 7: Average model accuracy across varying numbers of few-shot examples. Models are reported into three groups based on parameter size and access type.

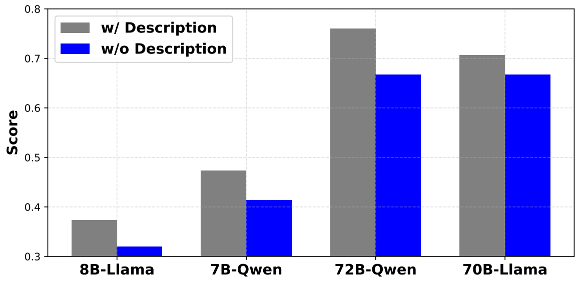

Figure 8: Comparison of model performance with vs. without context descriptions

terns emerge during the prediction process. To this end, we adopt the logit lens (Nostalgebraist, 2020; Halawi et al., 2024), which facilitates layer-wise analysis of prediction dynamics within the model. We compute the token probability distribution at each layer l by extracting the hidden state corresponding to the final token position, denoted as $\mathbf{h}_{-1}^{(l)} \in \mathbb{R}^d$. After applying layer normalization and the language modeling head, the resulting distribution is obtained as: $\mathbf{p}^{(l)} = \text{Softmax}\Big(\mathbf{W}_{\text{LM}} \cdot \text{LayerNorm}(\mathbf{h}_{-1}^{(l)}) + \mathbf{b}_{\text{LM}}\Big)$. For a fixed candidate set $T = \{t_1, ..., t_5\}$ representing the tokenized options 1 to 5, we extract the confidence assigned to each candidate token $P_t^{(l)} = \mathbf{p}^{(l)}[t]$. We further extract top-1 prediction at each layer, defined as the token with the highest probability and its corresponding confidence score $t_{\text{top1}}^{(l)} = \arg\max_t \mathbf{p}^{(l)}[t], \quad P_{\text{top1}}^{(l)} = \max_t \mathbf{p}^{(l)}[t]$.

As illustrated in Figure 5, confidence scores for the five candidate options remain uniformly low in the early layers and begin to rise significantly in the middle to later layers. This trend suggests that the model's ability to interpret and infer context is primarily established in the deeper layers, where fine-grained semantic representations are formed.

The top two subfigures illustrate correct predictions where the answer choice aligns with the top-1 probability. The model assigns sharply higher confidence to the correct option while effectively marginalizing alternative choices, indicating a clear resolution of the decision. In contrast, the bottom two subfigures illustrate incorrect predictions, revealing distinct error patterns. The model confidently assigns high probability to the wrong option, reflecting a complete misinterpretation of pragmatic cues. In the middle two figures, the model distributes low confidence across multiple candidates. Compared to high-confidence predictions, these examples exhibit delayed convergence and heightened instability in confidence patterns, particularly in the deeper layers. The results indicate that, much like human respondents, the model exhibits patterns of hesitation, confident correctness, and misinterpretation that depend on its reading of pragmatic cues.

4.4 RQ3. How can we improve LLMs' ability to interpret non-verbal intent?

Chain-of-Thoughts prompting This study investigates the potential to enhance the model's capacity for interpreting non-verbal intent. To this end, we apply Chain-of-Thought (CoT) prompting strategy to facilitate step-by-step reasoning. As shown in Figure 6, most models demonstrate a trend in which the magnitude of performance degradation exceeded that of improvement, with GPT-4o notably declining by more than 0.3. We find that GPT-4o frequently favors literal interpretations, suggesting a limited ability to understand contextual cues even under CoT prompting. In contrast, Llama-70B shows the most significant improvement, with consistent gains observed across most Llama and Phi models. Mistral and Qwen families generally show performance degradation, indicating that the effect of CoT varies by model family.

In-context learning We extend our investigation to few-shot prompting, a well-established approach for enhancing in-context learning. As shown in Figure 7, in-context learning leads to notable performance gains, particularly among API-based models. GPT-o3, in particular, shows substantial improvement, suggesting that the model internalizes relational patterns from in-context examples well. The 70B model also showed progressively enhanced performance as the number of shots increased. Smaller models (3B to 7B) exhibit slight performance declines, while the 14B model shows only marginal improvement. These findings indicate that larger models possess the representational capacity necessary to benefit from few-shot prompting, facilitating a more proper interpretation of non-verbal responses.

4.5 Evaluating model performance without context description

Our original prompt includes an explicit instruction "`Assume that both parties faithfully engage in the conversation.`", guiding the model to interpret non-verbal responses as intentional rather than incidental. To approximate real-world conditions, we ablate this contextual description and compare the results, as shown in Figure 8. All four evaluated models, spanning two families and parameter scales, show decreased performance. These findings suggest that, in real-world use cases where the models are required to operate solely based on the given input without access to any additional context or few-shot examples, inferring intent from non-verbal content becomes substantially more challenging.

5 Conclusion

Interpreting indirect meaning embedded in non-verbal behavior constitutes a critical step in advancing the pragmatic competence of LLMs. This study provided the first systematic evaluation of LLMs in interpreting communicative intent conveyed exclusively through non-verbal responses. Experimental results indicated a significant decline in performance under non-verbal conditions relative to verbal settings. A detailed analysis further revealed recurrent error types and behavioral patterns. Moreover, few-shot in-context learning was shown to enhance performance, indicating that LLMs acquire pragmatic mappings when supplemented with additional examples. By introducing a new dimension of pragmatic reasoning, this work extends the scope of LLM evaluation to encompass non-verbal communication.

Limitations

This study is subject to several limitations. First, the evaluation is conducted exclusively in English and relies on pragmatic inferences that are generally assumed to hold across cultures. We do not examine how models interpret non-verbal behavior across different languages. Extending this work to a multilingual setup would offer valuable insights into language-specific model behavior and generalization.

Second, non-verbal behaviors can be perceived through various modalities. We focus solely on textual input, since the language model is ultimately responsible for interpreting the speaker's intent and context, regardless of the modality through which non-verbal signals are received. Nevertheless, incorporating multimodal inputs in future work will enhance the applicability of LLMs to real-world human–AI interactions.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425) and this work was supported by the Commercialization Promotion Agency for R&D Outcomes(COMPA) grant funded by the Korea government(MSIT)(2710086166) and this work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by MSIT (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI) and this work was supported by IITP-ICT Creative Consilience Program grant funded by MSIT(IITP-2026-RS-2020-II201819).

References

John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

Ray L Birdwhistell. 1952. *Introduction to kinesics:(An annotation system for analysis of body motion and gesture)*. Department of State, Foreign Service Institute.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Thomas J Bruneau. 1973. Communicative silences: Forms and functions. *Journal of communication*, 23(1):17–46.

Judee K Burgoon, Valerie Manusov, and Laura K Guerrero. 2021. *Nonverbal communication*. Routledge.

Cheryl L Carmichael and Moran Mizrahi. 2023. Connecting cues: The role of nonverbal cues in perceived responsiveness. *Current Opinion in Psychology*, 53:101663.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

M. Danesi. 2006. Kinesics. In Keith Brown, editor, *Encyclopedia of Language Linguistics (Second Edition)*, second edition edition, pages 207–213. Elsevier, Oxford.

Alara Dirik, Hilal Donmez, and Pinar Yanardag. 2021. Controlled cue generation for play scripts. *CtrlGen: Controllable Generative Modeling in Language and Vision Workshop at NeurIPS 2021*.

Eli Dresner and Susan C Herring. 2010. Functions of the nonverbal in cmc: Emoticons and illocutionary force. *Communication theory*, 20(3):249–268.

Paul Ekman and Wallace V Friesen. 1969. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Shatha Ali A. Hakami, Robert Hendley, and Phillip Smith. 2023. ArSarcasMoji dataset: The emoji sentiment roles in Arabic ironic contexts. In *Proceedings of ArabicNLP 2023*, pages 208–217, Singapore (Hybrid). Association for Computational Linguistics.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2024. Overthinking the truth: Understanding how language models process false demonstrations. In *The Twelfth International Conference on Learning Representations*.

Judith A Hall, Terrence G Horgan, and Nora A Murphy. 2019. Nonverbal communication. *Annual review of psychology*, 70(2019):271–294.

Shirley Anugrah Hayati, Aditi Chaudhary, Naoki Otani, and Alan W Black. 2019. What a sunny day : Toward emoji-sensitive irony detection. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 212–216, Hong Kong, China. Association for Computational Linguistics.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Tianran Hu, Han Guo, Hao Sun, Thuy-vy Nguyen, and Jiebo Luo. 2017. Spice up your chat: the intentions and sentiment effects of using emojis. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 102–111.

Adam Jaworski. 1992. *The power of silence: Social and pragmatic perspectives*. Sage Publications.

J Vernon Jensen. 1973. Communicative functions of silence. *ETC: A Review of General Semantics*, pages 249–257.

Richard L Johannesen. 1974. The functions of silence: A plea for communication research. *Western Journal of Communication (includes Communication Reports)*, 38(1):25–35.

Mark L Knapp, Judith A Hall, and Terrence G Horgan. 1972. *Nonverbal communication in human interaction*. Thomson Wadsworth.

Robert C Lane, Mark G Koetting, and John Bishop. 2002. Silence as communication in psychodynamic psychotherapy. *Clinical psychology review*, 22(7):1091–1104.

Joshua Lee, Wyatt Fong, Alexander Le, Sur Shah, Kevin Han, and Kevin Zhu. 2025. Pragmatic metacognitive prompting improves LLM performance on sarcasm detection. In *Proceedings of the 1st Workshop on*

Computational Humor (CHum), pages 63–70, Online. Association for Computational Linguistics.

Yoon Kyung Lee, Yoonwon Jung, Gyuyi Kang, and Sowon Hahn. 2023. Developing social robots with empathetic non-verbal cues using large language models.(2023). In *the 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–5.

Stephen C Levinson. 1983. *Pragmatics*. Cambridge university press.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Steven McCornack and Joseph Ortiz. 2022. *Choices & connections: An introduction to communication*. Macmillan Higher Education.

Nostalgebraist. 2020. Interpreting gpt: the logit lens, 2020.

Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, and Sungeun Lee. 2024. MultiPragEval: Multilingual pragmatic evaluation of large language models. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 96–119, Miami, Florida, USA. Association for Computational Linguistics.

Miles L Patterson. 2012. Nonverbal behavior: A functional perspective.

Uwe Peters and Benjamin Chin-Yee. 2025. Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12(4):241776.

Peng Qi, Nina Du, Christopher Manning, and Jing Huang. 2023. PragmatiCQA: A dataset for pragmatic question answering in conversations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6175–6191, Toronto, Canada. Association for Computational Linguistics.

Laura Eline Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023a. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Laura Eline Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023b. Large language models are not zero-shot communicators.

John R Searle. 1975. Indirect speech acts. In *Speech acts*, pages 59–82. Brill.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Jacqueline Urakami and Katie Seaborn. 2023. Nonverbal cues in human–robot interaction: A communication studies perspective. *ACM Transactions on Human-Robot Interaction*, 12(2):1–21.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Tim Wharton. 2009. *Pragmatics and non-verbal communication*. Cambridge University Press.

Baoqin Wu, Muhammad Afzaal, and Dina Abdel Salam El-Dakhs. 2025. 'yet his silence said volumes': a pragmatic analysis of conversational silence in rapport management. *Cogent Arts & Humanities*, 12(1):2451490.

Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. 2024. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22583–22599, Miami, Florida, USA. Association for Computational Linguistics.

Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. 2024. Do large language models understand conversational implicature- a case study with a Chinese sitcom. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1270–1285, Taiyuan, China. Chinese Information Processing Society of China.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online. Association for Computational Linguistics.

A Categorizing Non-verbal Behavior

Non-verbal communication involves the transmission of messages through behavioral signals (Mc-

Cornack and Ortiz, 2022). This transmits meaning even when people encounter language barriers, making their comprehension within conversational contexts crucial (Burgoon et al., 2021).

Non-verbal behavior has been categorized into distinct types in prior studies. For example, Birdwhistell (1952) and (Danesi, 2006) define kinesics as the study of non-verbal behavior and categorize it into gestures, facial expressions, eye behavior, touch, and posture. Carmichael and Mizrahi (2023) delineate four primary channels: touch, vocal tone, facial expressions, and bodily gestures. Urakami and Seaborn (2023) classify human sensory channels by identifying body movements, gestures, and facial cues as components of the visual channel, and vocalizations and sounds as belonging to the auditory channel. Wharton (2009) propose a more streamlined taxonomy, reducing non-linguistic cues to facial expressions and gestures.

Building upon this theoretical foundation, we broadly categorize the non-verbal behaviors into bodily movements and facial expressions. Vocalic paralinguistic cues are excluded from the present analysis, as the study focuses on response scenarios involving solely non-verbal behaviors. In particular, facial expressions encompass movements of the facial muscles, such as eyebrow raises, eye rolling, and changes in the corners of the mouth. The face, as a key component of the visual channel, serves as a powerful means of information transmission. Movements focus on bodily actions rather than facial ones, including body gestures, body postures, proxemics, haptics, and other actions. We further introduce a silence category based on its unique informational characteristics. Although silence, which is the absence of any expression, action, or speech, inherently contains no information, humans paradoxically interpret it within conversations to convey meanings such as approval, displeasure, or agreement depending on the context (Bruneau, 1973; Johannesen, 1974; Jensen, 1973). Namely, silence serves to convey meaning despite its lack of explicit information, justifying its classification as a distinct category.

By categorizing non-verbal behaviors into these *three distinct groups of silence, facial expressions, and movements*, this study offers a novel investigation into how effectively LLMs can interpret their implications when these behaviors serve as the sole form of response in a conversation.

B Detailed Problem Setting

While the problem scope is introduced in the main text, this section offers a more detailed elaboration. In addition to the setting where the dialogue response consists solely of non-verbal behaviors, we establish two fundamental premises. First, we acknowledge that certain non-verbal behaviors may occur due to non-interactional factors (e.g., fatigue, distraction). To control for this, each question explicitly specifies that both participants are actively engaged in the conversation, constraining the non-verbal behaviors to occur solely as responses to speaker A's utterances. Namely, we incorporate the following elements into the question, eliminating ambiguity about intentionality: "Assume that both parties faithfully engage in the conversation".

Second, literal explanations of the conversational context are not inherently incorrect. However, we restrict the setting to ensure that the non-verbal responses contain latent meanings, in order to assess whether LLMs possess pragmatic competence. To make this setting clearly, we explicitly instruct the model to select the response that *best interprets the context and underlying intent of the conversation*, as follows: "Choose the option that most appropriately interprets the context and underlying intent of the conversation below."

C Dataset Construction

Source Data For rigorous evaluation, we construct a held-out evaluation set that shares no samples with any existing training data. Adopting the approaches of Cui et al. (2020) and Hu et al. (2023), we generate 50 self-collected situational prompts based on the empirically grounded observations of real-world interactions and GPT-4o outputs. These scenarios are designed to reflect pragmatically rich contents that commonly occur in daily lives. Each response in the dialogue is annotated with multiple behavioral categories, including silence, facial expressions, movements, and verbal behavior. By systematically varying only the response category, we aim to examine how different forms of non-verbal signals influence LLM interpretation within identical contexts.

Distractor Configuration To further ensure the difficulty of the benchmark, we construct multiple carefully designed distractor options for each example. The distractors are designed to capture typical pitfalls or failure patterns in pragmatic interpretation, including the following:

- *Misinterpretations*, where the intended meaning is reversed or inaccurately inferred (e.g., interpreting an approving smile as disapproval) (Sravanthi et al., 2024).
- *Literal explanations*, in which the understanding of the conversational context is correct but the response is described merely as a surface-level behavior (e.g., "B smiled" or "B remained silent"), without attributing any pragmatic meaning, and in some cases, the non-verbal behavior is disregarded entirely (Hu et al., 2023).
- *Lexically coherent but semantically divergent responses*, which share surface-level lexical similarity with the original context while lacking contextual relevance (Yue et al., 2024; Zheng et al., 2021).
- *Abstract explanations*, which provide vague or generalized descriptions that lack contextual specificity and fail to address the core topic of the dialogue (e.g., stating only that "A asks a question and B responds with a behavior") (Peters and Chin-Yee, 2025).

Quality Control We implement a rigorous quality control procedure on the generated dataset, which is conducted in two stages: automated assessment by LLMs followed by human cross-checking. In the LLM evaluation stage, the following criteria are employed: (1) Clarity: The dialogue should be clear, grammatical, and immediately understandable. (2) Correctness: The dialogue context and the designated answer must be properly aligned. (3) Difficulty: Among multiple plausible choices, the most appropriate response must be selected in accordance with the question's intent. Based on these criteria, the LLMs evaluate quality scores on a 0–3 scale and provide a justification for each score: (1) 0 points: Poor dataset; both the dialogue and the options are entirely flawed. (2) 1 point: Major errors; the answer is not clearly correct. (3) 2 points: Minor errors, but the correct answer is clearly identifiable. (4) 3 points: Well-constructed dataset.

To enhance fairness, two different models (GPT-4o and GPT-4.1-mini) are employed for evaluation, and samples that consistently receive scores of 0 or 1 from both models are extracted. Subsequently, two human annotators review these samples, incorporate revisions informed by the LLM's reasoning, and perform a cross-check to finalize the dataset.

Category	# Samples	Avg. DL	Avg. OL	Avg. DL-Token	Avg. OL-Token
Silence	50	68.26	70.79	14.3	13.14
Facial Expressions	50	68.22	84.92	14.24	15.2
Movements	50	92.68	74.8	17.22	13.74
Verbal response	50	96.2	80.5	19.12	14.35

Table 3: Dataset statistics for a pragmatics-driven probing dataset. DL and OL denote dialogue length and option length, respectively.

Table 3 presents the statistics of the constructed dataset. We plan to release the dataset under the CC BY-SA 4.0 license, which permits redistribution and adaptation, provided that appropriate credit is given and derivative works are distributed under the same license.