Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluating Large language models on Understanding Korean indirect Speech acts

Youngeun Koo

sarah8835@skku.edu

Sungkyunkwan University

Jiwoo Lee

Seoul National University

Dojun Park

Artificial Intelligence Institute of Seoul National University

Seohyun Park

Brain Humanities Lab of Seoul National University

Sungeun Lee

Seoul National University

Article

Keywords:

Posted Date: August 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-7946521/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Evaluating Large language models on Understanding Korean indirect Speech acts

Youngeun Koo[1,2,*], **Jiwoo Lee**[2,3], **Dojun Park**[2], **Seohyun Park**[4], **and Sungeun Lee**[2,3,4,*]

[1]Department of German Language and Literature, Sungkyunkwan University, Seoul, 03063, South Korea
[2]Artificial Intelligence Institute of Seoul National University (AIIS), Seoul, 08826, South Korea
[3]Department of German Language and Literature, Seoul National University, Seoul, 08826, South Korea
[4]Brain Humanities Lab, Seoul National University, Seoul, 08826, South Korea
[*]sarah8835@skku.edu, cristlo5@snu.ac.kr

ABSTRACT

Understanding the intention of an utterance is crucial in conversational communication. As conversational artificial intelligence models are rapidly being developed and applied in various fields, it is important to evaluate the LLMs' capability to understand the intentions of user's utterance. Speech act is a linguistic concept in pragmatics, the study of human language use, and can be simply understood as referring to the intention of an utterance. This study evaluates whether current LLMs can understand the intention of an utterance by considering the given conversational context, particularly in cases where the actual intention differs from the literal intention of the sentence, i.e. indirect speech acts. With a specific focus on Korean, a context-sensitive language, we construct evaluation datasets, consisting of three types of indirect speech acts based on Searle's speech act scheme. The dataset consists of two scenarios, where the same utterance conveys direct speech act and indirect speech act respectively, depending on the conversational context. We adopt two experimental setups: the conventional Multiple-Choice Question (MCQ) format for automatic evaluation, and the Open-Ended Question (OEQ) for detailed assessment by human experts. For thorough evaluation, we also conduct MCQ-based experiment with human participants to set gold standard. Our findings reveal that Claude3-Opus outperformed the other competing models, with 71.94% in MCQ and 65% in OEQ, showing a clear advantage. In general, proprietary models exhibited relatively higher performance compared to open-source models. Nevertheless, no LLMs reached the level of human performance. Most LLMs, except for Claude3-Opus, demonstrated significantly lower performance in understanding indirect speech acts compared to direct speech acts, where the intention is explicitly revealed through the utterance. This study not only performs an overall pragmatic evaluation of each LLM's language use through the analysis of OEQ response patterns, but also emphasizes the necessity for further research to improve LLMs' understanding of indirect speech acts for more natural communication with humans.

Introduction

ChatGPT, released in November 2022, demonstrated remarkable performance not only in traditional Natural Language Processing (NLP) tasks, such as categorization and translation, but also in more complex tasks, including advanced logical problem-solving and creative writing[1,2]. In response, numerous corporations have invested in developing their own models, resulting in the emergence of increasingly sophisticated models. With the increasing number of Large Language Models (hereinafter LLMs) being developed, the need for rigorous performance evaluation has become increasingly critical. These evaluations not only ensure alignment with standards deemed acceptable in contemporary fields but also help identify the strengths and weaknesses of each model[3].

Currently, benchmarks are widely employed as standard tools for assessing the capabilities of LLMs[4]. These benchmarks facilitate the evaluation process by providing standardized and automated task-specific datasets. However, the prevailing trend in current benchmarking solely focuses on tasks such as knowledge processing, logical reasoning, and computation[5,6]. As LLMs become increasingly integrated into the daily lives of the general public, their performance as personal agents gains paramount importance. This trend necessitates the development of models capable of understanding human language and comprehending user intentions to offer customized assistance effectively.

Pragmatics is a branch of linguistics that examines how language is used and understood within specific contexts and situations. Table 1 shows an example where the same utterance can be interpreted with different intentions depending on the context. For example, as shown in Table 1, the utterance "the elevator is out of order" can convey both a literal and an implied meaning depending on the context. Literally, it means that the elevator is currently non-functional. However, in a particular situation, it may carry an implied meaning, such as serving as a suggestion or direction to take an alternative route, like using the stairs. This example underscores the importance of context-dependent interpretation and illustrates how a speaker utilizes

language to convey intentions and implications based on the surrounding context. Simultaneously, in face-to-face interactions, the listener interprets the utterance's implications and the speaker's underlying intention. This highlights the essential role of pragmatic competence in facilitating effective communication and managing the dynamic interplay between speaker and listener.

Utterance	Intention Interpretation
"The elevator is out of order."	**Context 1:** A janitor walking by points where the stairs are with his/her finger. **Intention 1:** Suggesting to use the stairs instead of trying to use the elevator.
	Context 2: A customer in the store talks to the facility manager. **Intention 2:** Informing the facility manager about the inconvenience and demanding to solve the problem.

Table 1. Intention interpretation depending on the context

Despite the significant necessity of pragmatic competence in LLMs, there are relatively few evaluation studies specifically addressing this aspect. To bridge this gap, we develop a test set tailored to assess the pragmatic capabilities of LLMs, focusing on their ability to comprehend a speaker's intention conveyed directly or indirectly through an utterance within its contextual setting. The test suite consists of 240 question units written in Korean, designed to evaluate the models' proficiency in interpreting a speaker's intentions based on contextual cues. This focus is particularly significant in the Korean language, where shared context and background information play a crucial role in effective communication[7–9].[1] This evaluation framework is grounded in Austin and Searle's Speech Act Theory[11,12], which categorizes speech intentions into five distinct types: representatives, directives, commissives, expressives, and declaratives. We also assessed how well the LLMs comprehend speakers' indirect intentions, specifically directive, commissive, and expressive intentions realized in representative utterances.

Additionally, we conduct an experiment to evaluate human speech act comprehension capabilities, providing reference data for analyzing and comparing the performance of each LLM. This human reference serves as a critical point of comparison to assess the extent to which LLMs align with human-like understanding of speech acts.

To summarize, the primary contributions of this paper are:

- **Development of a Speech Act Comprehension Test Set**: A specialized test set designed to evaluate LLMs' ability to understand and interpret three different categories of speech acts within contextual frameworks.[2]
- **Collection of Human Test Data**: Human performance data collected to serve as a benchmark for comparison against LLMs' performance, providing a reference point, i.e. Gold Standard, for evaluation.
- **Systematic Evaluation and Analysis of LLMs' Performance**: A thorough evaluation and analysis of the LLMs' capabilities, focusing on their pragmatic competence and alignment with human-like understanding.

Related Work

Speech Act Theory and Indirect Speech Acts

It is originally considered by positivist philosophers that every sentence we utter in our everyday life describes or constates something, and thus can always be discerned as true or false. Austin[11] argued against this concept and categorized our language usage into two types. The first type is the 'constative utterance', which can be understood as the same as the traditional idea of a statement. On the other hand, in the case of the second type of language usage called the 'performative utterance', sentences have no truth-value. Instead, they become performative acts themselves. For example, uttering "I bet 10 dollars on that" can be considered as an act of 'betting', and it cannot be determined whether the sentence itself is true or false. In other words, with performative utterances, we don't describe the acts we do, but perform the very act itself. We can understand this concept of language usage as 'Speech Acts'.

Searle[13] delved further into Speech Act Theory and categorized speech acts into five classes: representatives, directives, commissives, expressives, and declaratives. The main intention of representative speech acts is to convey or describe something that the speaker believes to be true. For directive speech acts, the intention is to make the listener perform (or not perform) a

[1] Hall[7] proposes a spectrum of context dependency levels which has low- and high-context systems at each end. In low-context communication, the listener knows very little and must be told practically everything, whereas in high-context communication, the listener is already 'contextualized' and does not need to be given much background information[10]. According to Bhagat et al.[8], nations such as Japan, China, Greece and Spain have those of high-context cultures and those Asian countries show high-context driven communication styles[9].

[2] The test set is publicly available on our GitHub repository at https://anonymous.4open.science/r/LLM-Indirect-SpeechAct-44D1/.

certain actions. Commissive speech acts involve the speaker committing to a future action. Expressive speech acts express the speaker's emotions about a particular situation. Declarative speech acts create a new state or situation.

Based on the context in which an utterance is made, we can interpret a sentence and the intention of the speaker differently. Take a look at the following utterance: "The ice is thin". Without a context, it can simply be understood as a description of thin ice, hence as an assertive speech act. However, imagine a situation in which several children are attempting to step on a frozen pond. The same utterance could be interpreted as an act of warning, preventing the children from stepping on the thin ice and falling into the cold water, thus as a directive speech act. We can differentiate between (direct) speech acts and indirect speech acts, which are defined by Searle as utterances in which a certain speech act is performed by performing a different class of speech act. In the example, a directive speech act of 'advisory warning' (do not step on the ice) is performed indirectly through an assertive speech act of 'fact presentation' (the ice is thin).

As we can see, to understand a speaker's intention from an utterance, it is crucial to interpret a sentence based on the surrounding context. Considering the existing trend of LLM evaluations through the leaderboards[14], which focus primarily on performance in tasks assessing knowledge and understanding explicit literal meanings, evaluating the pragmatic competence of LLMs using the framework of Speech Act Theory can be a meaningful in-depth evaluation of the linguistic performance of LLMs.

Evaluation of human understanding ability of indirect speech acts

The evaluation tests of human pragmatic abilities are primarily developed and employed to determine how close the pragmatic competence of the test subjects is to that of general individuals. Most of the assessments focus on individuals with communicative challenges[15–18], L1-speaking children and L2 learners[19].

Arcara & Bambini[15] developed a test for the Assessment of Pragmatic Abilities and Cognitive Substrates (APACS), targeting individuals who have acquired communicative deficiencies. The test comprises the following six tasks in Italian language: Interview, Description, Narratives, Figurative Language 1/2, and Humor. The first task includes autobiographical questions. The Description task involves describing elements of ten everyday pictures. The Narratives task includes comprehension questions about the global topic, specific elements and figurative expressions of a certain story. Figurative Language 1 consists of multiple-choice questions about five idioms, metaphors, and proverbs each. Figurative Language 2 involves open-ended questions. The Humor task includes multiple-choice questions where participants select the appropriate ending (funny, straightforward, unrelated ending) for seven brief stories. The consistency and retest reliability of the test were secured and proved by conducting sample tests with 119 healthy participants representing the general population.

Kim et al.[16] developed and conducted a test to evaluate the ability of the students with mild intellectual disabilities to understand speech acts and indirect speech acts. It is composed of multiple-choice questions, in which certain speech acts are presented in a context that includes two or three sentences. There are three answer options: 'correct interpretation', 'wrong, literal interpretation' and 'wrong, context-based interpretation'. Eighteen questions are about indirect speech acts, with half being interrogative sentences and the other half declarative sentences. Another nine questions are about general speech acts, five of which are interrogative sentences and four are declarative sentences. The test was shown to be reliable by producing significantly different results between subjects with varying levels of intellectual challenges.

Other cases of evaluating human understanding of speech act include the tests targeted at children with Asperger syndrome[17], right-brain impaired patients[18] and school-aged children[19].

In a similar vein, we developed and employed a speech act comprehension test specifically targeted at LLMs. The main purpose of this test is to determine how well the individual LLMs understand speech acts, with the potential to compare their competence to that of typical human individuals in further research involving human test participants.

Evaluation of LLMs' understanding ability of indirect speech acts

Recently, many services are being developed using LLMs in various fields. To assess whether these services can successfully communicate with users and provide requested services, studies on LLMs' ability to understand the intention of an utterance are increasing in each field[20–23]. In particular, there are many studies using LLMs in language education. Among them, Han et al.[20] analyzed whether ChatGPT, a LLM-based chatbot, can understand the EFL (English as foreign language) learners' intention of utterances, which learners enter into ChatGPT while writing English essays. They evaluated whether ChatGPT can accurately classify user input into 13 predefined intent categories such as 'Request for Translation', 'Request for Information', 'Statement', and 'Acknowledgement'. Wang et al.[21] examined whether ChatGPT understands which of the talk moves, such as 'asking for more information' and 'making a claim', correspond to utterances that occur during class, and compared the performance with BERT-based models.

Moreover, whether LLMs generate appropriate sentences based on a correct understanding of speech acts is dealt with great interest nowadays. Many studies focus on instructing LLMs to generate adequate responses for certain objectives, such as data augmentation and automatic annotation, and evaluating their performance[23–25].

Pragmatic Evaluation of LLMs for Korean

Eo et al.[26] evaluated GPT-family models on their ability to solve riddles in Korean, a task that demands high creativity and an understanding of language-specific nuances. The results demonstrated that while GPT-4 achieves the highest scores, generally, the models struggle with this task, scoring below 10% in both EM and F1 scores. This highlights their challenges in tasks that require a deep understanding of the subtleties of human language.

Nam et al.[27] tested Kakao Mini, an AI speaker, for its communicative performance in Korean based on real AI-human conversations in a multi-turn dialogue setup. They leveraged the framework of Gricean conversational theory for the evaluation. The results showed that the maxim of relation was the most frequently flouted by the tested model, indicating significant room for improvement in achieving natural communication capacity. Park et al.[28] evaluated various LLMs, including Korean-specific ones, for their pragmatic competence in Korean, based on Gricean conversational theory. The study assessed whether the models can accurately infer meanings implied by the context, similar to human inference. The findings revealed that while GPT-4 is notable, HyperCLOVA X, a model tailored for Korean, exhibits superior performance on Korean-specific questions. Building on this, Park et al.[29] expanded the test suite in size and scope to include other flagship models and additional languages, such as English, Chinese, and German.

Methods

Dataset for Evaluation

To evaluate whether LLMs can understand indirect speech acts, we constructed a dataset consisting of 240 items. Based on Searle's scheme of speech act categories given in Table 2, we have set three types of indirect speech acts as research objects. **ReDi** refers to the case where a declarative sentence, which is a 'Representative speech act' as on the surface information, performs a 'Directive speech act' as an illocutionary act. Likewise, **ReCo** and **ReEx** refer to the case where a declarative sentence, which is a 'Representative speech act' as in the surface information, performs a 'Commissive speech act' and 'Expressive speech act' respectively as an illocutionary act. Each indirect speech act type contains 80 items.

Category	Examples
Representative (Assertive)	The earth is round. This jacket looks good on you.
Directive	Please pass me the salt. Stop running around!
Commissive	I promise to finish it by today. I'll help you with your homework.
Expressive	I'm so sorry! Welcome to my house!
Declarative	I pronounce you husband and wife. You are fired!

Table 2. Speech Act Categories of Searle[12]

This study aims to evaluate whether LLMs understand and distinguish that the same utterance can be interpreted with different meanings depending on the context. Therefore, we organized all 80 items for each type of indirect speech act into 40 pairs, where the same utterance performs direct and indirect speech act. Examples of our dataset are given in Table 3.

As shown in Table 3, each item consists of context and utterance. Through the utterance in item #75, speaker intends to simply provide information through a declarative statement. In its context where the conversation partner is inquiring about the weather, the utterance "it's nice outside" performs the speech act of 'providing information'. However, in item #76, the utterance "it's nice outside" in the given context does not primarily perform the speech act of 'providing information', which is typically associated with declarative statements. Instead, it implicitly, indirectly performs 'suggesting' to go outside. This is an example of **ReDi** where assertive utterance has been uttered to perform the directive speech act.

The context for item #97 involves a question about the dinner menu and the utterance presents response to this inquiry, making the item #97 an example of a direct speech act. In contrast, in item #98, the example of **ReCo**, the mother indirectly performs a 'promise' through a declarative utterance, that she will prepare beef steak as a dinner menu for Younghee. Here, the utterance conveys an indirect speech act by performing a commitment through declarative utterance.

Similarly, the context for item #171 asks about the dinner menu and the utterance responds to this question, making the item #171 another example of a direct speech act. Despite that, item #172 is an example of **ReEx** because the utterance 'Cheolsu is a chef' is intended to praise Cheolsu for cooking like a professional chef, rather than to state that Cheolsu's job is a chef.

ID	Type	Example (English Translation)
75	ReDi-Direct	철수와 영희는 부부이다. 영희는 밖에 나갈 준비를 하고 있고, 준비를 마친 철수에게 영희는 오늘 날씨가 어떤지 물었다. 이에 철수는 다음과 같이 말했다. *'밖에 날씨가 좋네'*. (Cheolsu and Younghee are a married couple. Younghee, preparing to go out, asked Cheolsu, who was already ready, about the weather. Cheolsu said, *'It's nice outside'*.)
76	ReDi-Indirect	철수와 영희는 부부이다. 두 사람은 오랜만에 여유로운 주말을 보내게 되었다. 그러나 철수가 계속 쇼파에만 누워있는 것을 본 영희는 철수에게 다음과 같이 말했다. *'밖에 날씨가 좋네'*. (Cheolsu and Younghee are a married couple. They finally had a relaxing weekend after a long time. Noticing Cheolsu lounging on the couch all day, Younghee said, *'It's nice outside'*.)
97	ReCo-Direct	영희는 엄마에게 내일 저녁에 무엇을 먹을지 물었다. 이에 엄마는 다음과 같이 말했다. *'내일 저녁 메뉴는 소고기야'*. (Younghee asked her mother what they would eat for dinner tomorrow. Her mother said, *'Tomorrow's dinner menu is beef steak'*.)
98	ReCo-Indirect	영희는 엄마에게 소고기를 먹고 싶다고 말했다. 이에 엄마는 다음과 같이 말했다. *'내일 저녁 메뉴는 소고기야'*. (Younghee told her mother that she wanted to eat beef steak. Her mother said, *'Tomorrow's dinner menu is beef steak'*.)
171	ReEx-Direct	영희는 민수에게 철수의 직업이 무엇인지 물었다. 이에 민수는 다음과 같이 대답했다. *'철수는 요리사야'*. (Younghee asked Minsu what Cheolsu's job is. Minsu replied, *'Cheolsu is a chef'*.)
172	ReEx-Indirect	철수가 집들이를 해서 집들이 음식을 만들었다. 철수가 만든 음식을 먹고, 한 친구가 다음과 같이 말했다. *'철수는 요리사야'*. (Cheolsu hosted a housewarming party and cooked food for his friends. After tasting the food Cheolsu had made, one of the friends said, *'Cheolsu is a chef'*.)

Table 3. Examples of Dataset

LLMs under Evaluation

In our study, we compared 12 different LLMs divided into two categories. Table 4 lists eight proprietary models accessed via API and four open-source models, each around the size of ten billion parameters, for which we accessed the model weights directly. Among the proprietary models, we included two versions of GPT[30], three versions of Claude[31], and three versions of Mistral [3]. We excluded Gemini by Google from our scope due to limited API access. The open-source models feature Llama3-8B (hereinafter Llama3)[32], Qwen1.5-14B (hereinafter Qwen)[33], Solar-10.7B[34], and T3Q-ko-solar [4]. Solar-10.7B (hereinafter Solar) is developed by Upstage, a Korean company, while T3Q-ko-solar represents a fine-tuned version of Solar, which was ranked at the top of the Open Ko-LLM Leaderboard[35] as of May 17, 2024.

[3] https://mistral.ai/
[4] https://huggingface.co/chihoonlee10/T3Q-ko-solar-dpo-v7.0

Type	Model	Version
Proprietary	GPT-3.5	turbo-0125
	GPT-4	turbo-2024-04-09
	Claude3-Haiku	haiku-20240307
	Claude3-Sonnet	sonnet-20240229
	Claude3-Opus	opus-20240229
	Mistral-small	small-2402
	Mistral-medium	medium-2312
	Mistral-large	large-2402
Open-Src.	Llama3-8B	Instruct
	Qwen1.5-14B	1.5-14B-Chat
	Solar-10.7B	Instruct-v1.0
	T3Q-ko-solar	dpo-v7.0

Table 4. Overview of Evaluated Models

Experimental Setup and Evaluation Metrics

We used two different setups for our experiment to assess LLMs' ability to understand indirect speech acts: multiple-choice questions (MCQs) and open-ended questions (OEQs). Examples of prompts given to the models for each MCQ and OEQ experiment are shown in Table 5.

Type	Prompt	Prompt (English Translation)
MCQ	철수와 영희는 부부이다. 영희는 밖에 나갈 준비를 하고 있고, 준비를 마친 철수에게 영희는 오늘 날씨가 어떤지 물었다. 이에 철수는 다음과 같이 말했다. '*밖에 날씨가 좋네*'. 위 발화가 갖는 가장 적절한 의도를 고르세요. 1) 진술 2) 제안 3) 추측 4) 유도	Cheolsu and Younghee are a married couple. Younghee, preparing to go out, asked Cheolsu, who was already ready, about the weather. Cheolsu said, *'It's nice outside'*. Choose the most appropriate intention of the above utterance. 1) State 2) Suggest 3) Assume 4) Induce
OEQ	철수와 영희는 부부이다. 영희는 밖에 나갈 준비를 하고 있고, 준비를 마친 철수에게 영희는 오늘 날씨가 어떤지 물었다. 이에 철수는 다음과 같이 말했다. '*밖에 날씨가 좋네*'. 위 발화가 갖는 가장 적절한 의도를 서술하세요.	Cheolsu and Younghee are a married couple. Younghee, preparing to go out, asked Cheolsu, who was already ready, about the weather. Cheolsu said, *'It's nice outside'*. Describe the most appropriate intention of the above utterance.

Table 5. Examples of Prompts

In the MCQ experiment type, we attempted to evaluate whether LLMs can select among the given options what the speech act of the utterance is, given the context and utterance. Four options are presented for LLMs to select from. One is the correct speech act type, another is the 'opposite type', and the others are two randomly selected speech act types. 'Opposite type' refers to direct speech act type when the test item is regarding indirect speech act, and vice versa. By having direct speech act as one of the options for indirect speech act context and utterance, and vice versa, we seek to evaluate whether LLMs can distinguish that the utterance can perform either indirect speech act or direct speech act depending on the given context.

For MCQ experiment, the performance of LLMs is calculated by the 'agreement rate', whether the option selected by LLMs is same with the correct answer. Considering that the response of the LLMs may randomly change each time, we ran through

three trials for each model and test units, for robust and reliable assessment. Therefore, the performance of each model in MCQ setting is the percentage of correct answers among all responses, obtained on average by three trials for each of the 240 test units.

In the OEQ experiment type, the responses generated by each LLM were qualitatively assessed by human evaluators. The LLM's responses to the OEQ test were scored on a scale of 1 to 5, in the aspect of pragmatic intent and literal meaning of an utterance. A score of 5 denotes a 'Correct Answer,' indicating that the LLM accurately grasped both the intent and meaning of the utterance. A score of 4, labeled as an 'Acceptable Answer,' applies when the LLM correctly understood either the intent or the meaning, while the other was interpreted in somewhat awkward manner. A score of 3, or 'Partial Answer,' is assigned when the LLM correctly understood either the intent or the meaning, but misinterpreted the other. A score of 2, termed a 'Incorrect, but Partially Acceptable Answer,' indicates that the LLM misunderstands one of the aspects, and even shows only an unnatural, awkward understanding of the other. A score of 1 represents an 'Incorrect Answer,' assigned when the LLM fails to comprehend both the intent and meaning. Each test unit was evaluated by three human evaluators, and the final score for OEQ test response was calculated as the average of their scores.[5] For ease of comparison with MCQ scores, OEQ scores were rescaled, ranging from 0 to 100.

To establish a gold standard, we collected human reference data by conducting the same MCQ test on humans which was held for LLMs.[6] The participants for the experiment were recruited from the general public, who are 20-30 years old and completed the regular South Korean education curriculum with South Korean nationality. To ensure representativeness, those who majored in linguistics or related disciplines were excluded. A total of 52 participants took part in the experiment, and their detailed demographic information is given in Table 6. According to Table 6, there is no significant difference in the performance score of understanding speech act between subjects based on gender, age, education level, major, etc. Hence, this indicates that the human reference data obtained in this study are sufficiently representative.

Group	Attribute	Number of Participants	Performance
Gender	Male	24	77.12
	Female	28	78.08
Age	20-24	29	78.01
	25-30	23	77.94
Academic Background	Undergraduate Students	30	78.4
	Undergraduate Graduates	21	76.39
	Not Applicable	1	72.92
Major	Humanities	12	76.25
	Social Sciences	9	77.41
	Educations	2	80.63
	Natural Sciences	4	79.79
	Engineering	14	76.67
	Arts	3	79.72
	Medical Science	2	78.13
	Others	6	79.38

Table 6. Demographic Information on Experiment Participants

Result

Quantitative Analysis on LLM Performance

Overall Performance

Table 7 illustrates the performance of each LLM on the MCQ test. In general, proprietary models outperform open source models. Among the proprietary models, Claude3-Opus achieved the highest accuracy with a score of 71.94, with GPT-4 and Mistral-large closely following with 71.39 and 69.58, respectively. Flagship models generally outperform lightweight models. This result confirms that models trained with a greater number of parameters excel in speech intention inference task, which requires an understanding of subtle contextual nuances. Claude3-Haiku showed the lowest performance among both the

[5]Intraclass correlation coefficient score rated 0.715, based on ICC3k Average fixed raters metrics.

[6]Survey was conducted by 'Google Forms' and the following link was sent to the experiment participants: https://forms.gle/qTRcRYNrBgs5xz577. Questionnaire was written in Korean, including the guideline and all survey questions.

proprietary models and Claude family models. In contrast, another lightweight model, Mistral-small, showed no significant difference compared to other models in family.

Solar demonstrated the highest performance among open-source models, achieving a score of 65.0. However, compared to the top-performing model, Claude3-Opus, it falls short by 7 points and is comparable to Claude3-Sonnet, Mistral-small, and Mistral-medium. T3Q-ko-solar, which follows Solar, is a model fine-tuned based on Solar, but its score dropped by nearly 5 points. While fine-tuning led to improved performance in the Ko-LLM benchmark tasks[35], it was found that performance decreased in context-based inference.

Notably, the human score for speech act comprehension was 77.64, which was clearly higher than all LLMs' scores. Human participants showed a total score of 77.5, achieving a higher performance than Claude3-Opus, which had a total score of 71.94. This indicates that humans' capability of understanding speech acts still surpasses the highest-performing LLM, Claude3-Opus.

To conduct an in-depth analysis on the models' ability to understand indirect speech acts, a selected subset of models was re-evaluated using OEQ tasks, which requires generating answers without giving pre-defined options. First, the top-performing models each from proprietary and open source type, Claude and Solar, were selected. Next, one of the low performing model from the proprietary models (GPT-3.5) along with an open model which exhibits similar performance to it (Llama3) were chosen for further evaluation. As previously explained, the LLM responses in the OEQ test were evaluated manually by human evaluators.

Type	Model	Performance
Proprietary	GPT-3.5	50.14
	GPT-4	71.39
	Claude3-Haiku	33.75
	Claude3-Sonnet	64.31
	Claude3-Opus	**71.94**
	Mistral-small	65.56
	Mistral-medium	66.67
	Mistral-large	69.58
Open-Src.	Llama-3-8B	49.44
	Qwen-14B	40.14
	Solar-10.7B	**65**
	T3Q-ko-solar	59.86
Human		**77.64**

Table 7. Scores on MCQ Test

Table 8 presents the OEQ scores by each LLM. Similar to the MCQ task, Claude3-Opus showed the best performance among the models, Solar following behind. The subtle difference in performance between GPT-3.5 and Llama3 in the MCQ test is also observed in the OEQ test, with only a 3-point difference. Interestingly, the rank of average accuracy among the models was consistent in MCQ and OEQ test, i.e., in the order of Claude3-Opus, Solar, GPT-3.5, and Llama3. This was likewise observed in terms of the stability of the models. Figure 1 shows box plots that illustrate the score distributions of each LLM in the OEQ task. Claude3-Opus showed not only the highest accuracy, as in Table 8, but also demonstrates highest stability, given that the Interquartile Range (IQR), the difference between Q1 (25% percentile) and Q3 (75% percentile), is the smallest. Considering the accuracy score and the stability of the model, it is found that the level of speech act comprehension is superior in the order of Claude3-Opus, Solar, GPT-3.5, and Llama3.

Type	Model	Performance
Proprietary	GPT-3.5	78.88
	Claude3-Opus	90.37
Open-Src.	Llama-3-8B	75.64
	Solar-10.7B	86.24

Table 8. Scores on OEQ Test

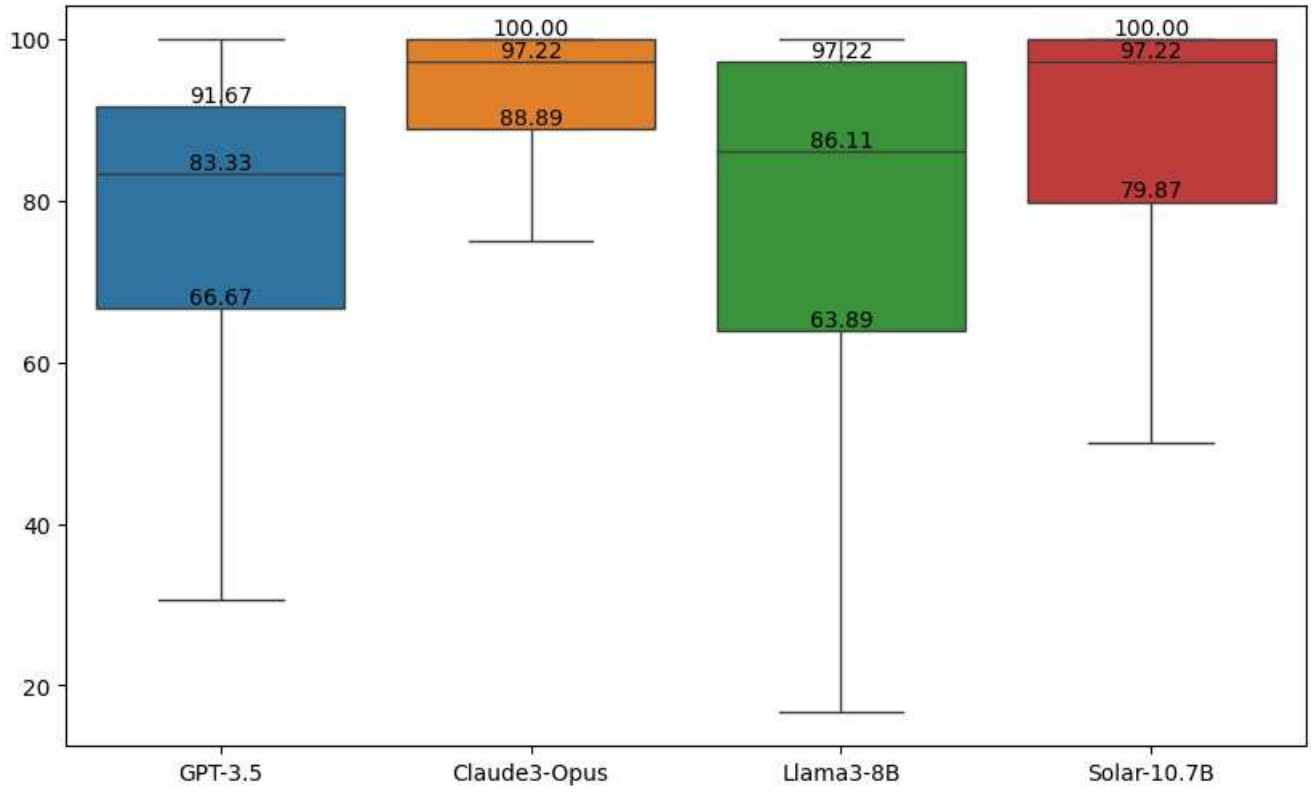


Figure 1. OEQ Score Distribution by each LLM

Directive vs. Commissive vs. Expressive speech act

Scores on MCQ test across three indirect speech act types, **ReDi**, **ReCo**, and **ReEx**, are listed in Table 9. The values in bold represent the maximum values, while the underscored values represent the minimum values. As shown in the Table 9, all models achieved highest accuracy for **ReEx**, where declarative utterance intends expressive speech act. Most items for **ReEx** indirect speech act, in our dataset, consist of utterances using idiomatic expressions. For example, the Korean expression '꿀맛이야 (It's honey flavored)' is often used to indicate that something is very delicious. Conversely, '개밥이야 (It's a dog food)' refers that the quality of a certain food is low or unappetizing to eat. In Korea, a dog's food is often seen as unappealing because it consists of various kibbles mixed together roughly.

Declarative sentences with expressive indirect speech act in our dataset were constructed manually by humans. Therefore, test items include natural expressions of indirect speech acts used by humans, and consequently, score for human on **ReEx** indirect speech act type was the highest compared to **ReDi** and **ReCo** analyzed in Table 9. However, not only humans but all LLMs had the highest score for **ReEx** among the three types of indirect speech acts. This suggests that LLMs learned a considerable amount of Korean idiomatic expressions and reflect the tendency of human showing the highest scores for **ReEx**.

In directive indirect speech sentences(**ReDi**), which have the intent to demand something, both humans and all LLMs performed the least. To be specific, declarative sentences that describe facts or objective information were difficult to understand as requests, compared to the sentences that describe one's opinion or subjective information. For example, a staff member could say, "It's sold out" to customers waiting in line outside a restaurant, to suggest the customers to go to another restaurant. This example is a test sample that all 52 experiment participants and 12 LLMs failed to interpret correctly. All of them understood this utterance as 'providing information' rather than 'suggesting'. However, in the test sample where the speaker pointed to the air conditioner and said, "It's too cold," most of the experimental participants (51 out of 52) and LLMs (10 out of 12) analyzed it correctly as 'requesting' rather than 'claiming'. Likewise, when a declarative sentence describes facts rather than asserting opinions, it is challenging to interpret it as a directive speech act. Given that nearly 63% of the evaluation data in this study consists of **ReDi** type utterances that depict factual information, both LLMs and human evaluators demonstrated the lowest comprehension in **ReDi** indirect speech act.

Another key finding from Table 9 is that humans show relatively small deviations across different types of indirect speech acts, whereas the majority of LLMs exhibit larger deviations. Especially lightweight models showed a great difference in scores between indirect speech act types.

Type	Model	Performance		
		ReDi	ReCo	ReEx
Proprietary	GPT-3.5	19.17	40.00	**50.8**
	GPT-4	38.33	45.83	**57.5**
	Claude3-Haiku	18.33	23.33	**35.83**
	Claude3-Sonnet	23.33	53.33	**65**
	Claude3-Opus	63.33	73.33	**87.50**
	Mistral-small	38.33	45.83	**57.5**
	Mistral-medium	25	47.5	**65**
	Mistral-large	37.5	54.17	**75**
Open-Src.	Llama-3-8B	25.83	20.83	**45**
	Qwen-14B	14.17	25	**46.83**
	Solar-10.7B	32.5	38.33	**65.83**
	T3Q-ko-solar	20	43.33	**59**
Human		72.45	76.92	**90.77**

Table 9. Scores on MCQ Across Indirect Speech Act Types

Direct vs. Indirect speech act

Table 10 describes how well LLMs understand direct and indirect speech acts, respectively. Most LLMs show significantly lower accuracy in indirect speech acts, compared to direct speech acts. While other models exhibited a significant score gap of 31.64%p, on average, between direct and indirect speech acts, only Claude3-Opus, which ranked first for highest performance, showed balanced accuracy for direct and indirect speech acts. The score difference is only 5.55%p and Claude3-Opus even showed higher accuracy in indirect speech acts.

In the case of humans, scores for both direct and indirect speech acts were relatively high and balanced. In particular, the comprehension of indirect speech acts was slightly higher, and therefore, Claude3-Opus was the only model that showed similar tendencies to humans.

Type	Model	Performance	
		Direct	Indirect
Proprietary	GPT-3.5	**63.61**	36.67
	GPT-4	**84.72**	58.06
	Claude3-Haiku	**41.94**	25.83
	Claude3-Sonnet	**81.39**	47.22
	Claude3-Opus	69.17	**74.72**
	Mistral-small	**83.89**	47.22
	Mistral-medium	**87.50**	45.83
	Mistral-large	**83.61**	55.56
Open-Src.	Llama-3-8B	**68.33**	30.56
	Qwen-14B	**51.67**	28.61
	Solar-10.7B	**84.44**	45.56
	T3Q-ko-solar	**78.89**	40.83
Human		75.22	**80.05**

Table 10. Scores on MCQ for Direct and Indirect Speech Act

Error Analysis

Figure 2 demonstrates the distribution of which incorrect answer options each model selected during MCQ task, informing the tendency of incorrect answers chosen by LLMs. This figure aims to reveal whether LLMs have difficulty understanding speech acts in a given context, or whether LLM cannot understand the input itself, i.e., language expression. Figure 2 shows the ratio

of two types of errors, '(In)direct' and 'Random', among the total errors. This shows which type of error LLM makes more frequently. The MCQ type among the test data constructed in this study consists of four options: one is the type of speech act when an utterance is understood as a direct speech act, another is the type of speech act when an utterance is understood as an indirect speech act, and the rest are two types of random speech acts. '(In)direct' is when an utterance that is actually a direct speech act is misunderstood as an indirect speech act or vice versa, and 'Random' is when one of the two random options is misunderstood.

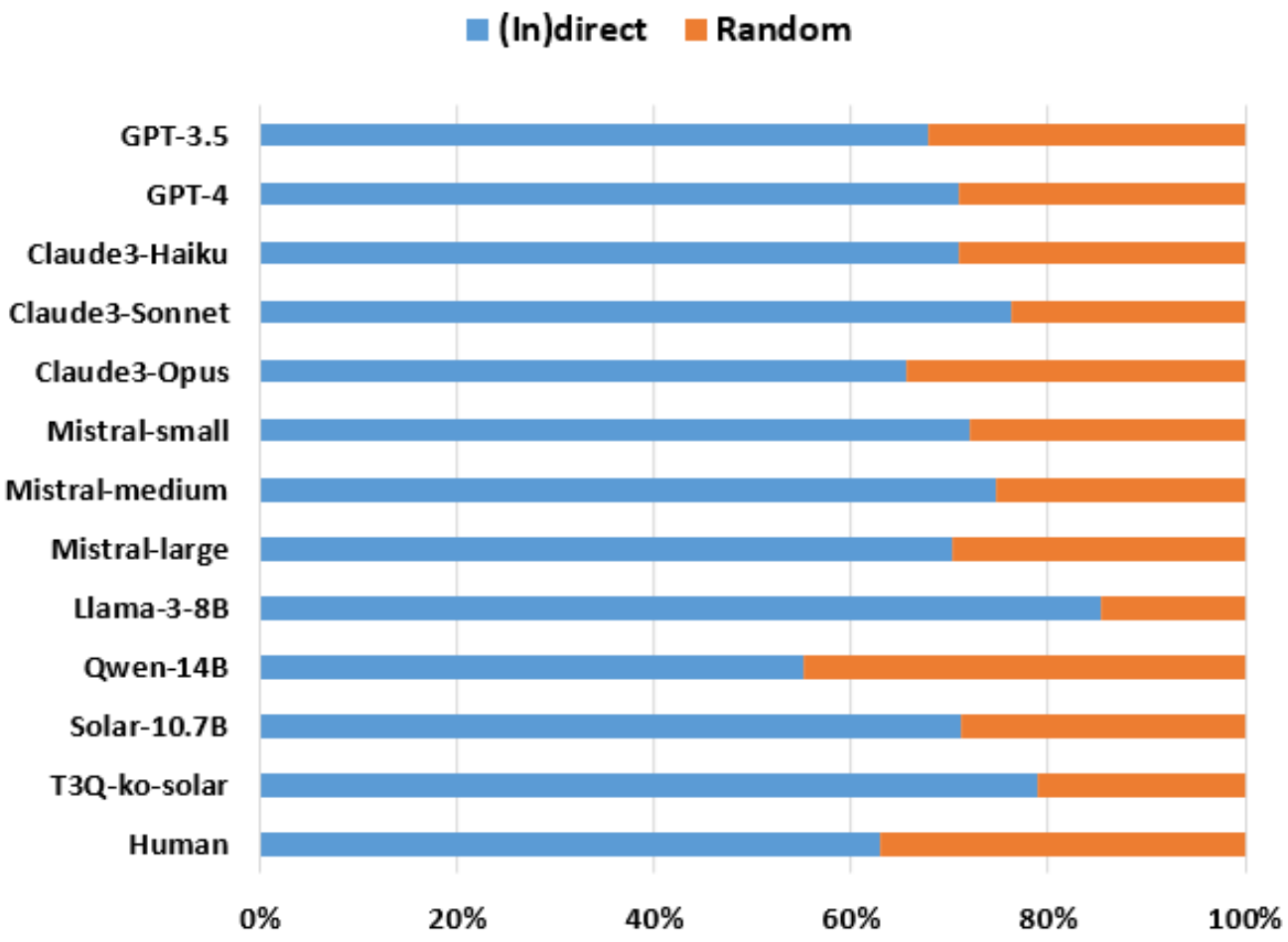


Figure 2. MCQ Error Choice Distribution by each LLM

As can be seen in Figure 2, most models made errors by confusing direct and indirect speech acts rather than selecting random options. Except for the Qwen-14.B model, which has a significantly low accuracy and thus does not make much sense to analyze the tendency of error types, we can see that the flagship models, such as Claude3-Opus, Mistral-large, and top open-source model, Solar, made fewer mistakes caused by confusion between direct and indirect. Humans were the least likely to confuse direct and indirect, and predominantly excelled at context-dependent interpretation, compared to LLMs.

Qualitative Analysis on LLM Performance

The ranking of average accuracy among the models was consistent in OEQ and MCQ result, i.e., in the order of Claude3-Opus, Solar, GPT-3.5, and Llama3. Therefore, we also conducted a qualitative analysis of LLM responses in the MCQ and OEQ tests, for a thorough evaluation of the linguistic competence of each model.

Explanatory Capabilities of LLMs in MCQ Task

In the MCQ test, we evaluated and scored each model's interpretative capabilities, specifically assessing whether the model selected the correct option number from the answer choices. However, given the generative nature of LLMs, their ability to produce relevant explanations and contextual information should also be considered when assessing overall performance[36].

Overall, Open-source models adhered closely to prompt instructions, generating responses limited to the answer choice and its corresponding number. In the few instances where they deviated, the responses contained irrelevant contents straying away from the correct answer. Notably, T3Q-ko-solar-10B distinctively demonstrated this tendency, generating not only the option number but also additional comments and explanations — regardless of their correctness — in 94 out of a total of 240 cases.

The other models — Llama3, Solar, and Qwen — formulated their answers in alignment with the prompt instructions, correctly following the directive 'to choose one of the options.' Notably, Qwen exhibited a distinctive answer pattern, in some cases providing only the specific number of the selected option, without including the name of each option's element.

Proprietary models demonstrated distinct tendencies in performance and response patterns associated with their respective developing companies. The three Mistral models, regardless of the accuracy rate, strictly adhered to prompt instructions by selecting and generating only the option number and its name. Similarly, GPT-3.5, as a prominent proprietary model alongside GPT-4 and Claude3-Opus, produced answers well-suited to MCQ formats, demonstrating a comparable adherence to prompt instructions, similar to open-source models. In contrast, GPT-4 and Claude models often went beyond simply selecting an option, frequently providing additional explanations to elucidate the inference process for users. This behavior suggests a more user-oriented approach in GPT-4 and Claude, emphasizing transparency in reasoning alongside task performance.

Response Pattern by LLMs in OEQ Task

In the open-ended question test set, we evaluated the overall generative capabilities of two proprietary and open-source language models. Through an analysis of the responses, we identified common answer patterns and specific challenges where each model encounters difficulties in interpreting the intended meaning of speech within a given context.

Llama3 exhibited a predominant error tendency by generating random, irrelevant tokens in nearly all of its responses, many of which included unreadable emoticons and emojis. Furthermore, although all prompts in the test set were presented in Korean, the majority of Llama3's responses — an average of 185 out of 240 (77 percent) — were generated in English. This suggests that the model is possibly overfitted to English prompts, assuming users are native English speakers without knowledge of the Korean language. This behavior indicates a potential misalignment with the language context specified in the prompts' instructions.

Another open-source model, Solar, exhibited a similar error pattern to Llama3, though with fewer occurrences. A unique but non-dominant characteristic of Solar was its tendency to generate its own answer choices, such as (A), (B), and (C), and then select an answer from these, mimicking the response format of an MCQ test. This behavior occurred an average of 17 times across three repeated responses, deviating from the typical approach humans use for similar tasks. This suggests that Solar may be overfitted to MCQ test sets that mirror benchmark formats, likely as a strategy to optimize its performance on leaderboard evaluations. This answer pattern appears to be closely associated with its inclination to generate one or a few definitive answers. Furthermore, Solar exhibited a tendency to produce Chinese Hanja characters, which are no longer commonly used in contemporary Korean language, reflecting a possible bias or outdated language modeling in its training.

While achieving a score comparable to Solar on the OEQ test set, GPT-3.5's responses were written in Korean and included significantly fewer random or unreadable outputs than those of Solar. GPT-3.5 consistently provided one or a few decisive answers that demonstrated an accurate understanding of the intended meaning of the speech. These subtle yet notable differences between Solar and GPT-3.5 highlight the importance of evaluating not only the comprehension abilities but also the generative quality of LLMs. Such evaluations provide a more comprehensive and accurate assessment of their overall performance. Claude3-Opus, which achieved the highest score among the models, demonstrated its advanced capabilities by offering multiple detailed explanations regarding speech intentions in specific contexts, often using a list format. Additionally, it provided supplementary information to enhance the user's understanding of the answers. This tendency reflects the model's high-level language capabilities, enabling it to perform sophisticated analyses and provide nuanced interpretations of complex linguistic tasks.

Discussion

Indirect Expression Comprehension of LLMs

Park et al.[28] and Park et al.[29] analyzed LLMs' ability to understand conversational implicature. They analyzed how much LLM understands conversational implicature according to Grice's four conversational maxims: quality, quantity, relation, and manner[37].[7] Implicature is one of the main research topics of pragmatics along with speech acts, and conversational implicature is closely related to speech acts, which is the subject of this paper, in that it refers to the meaning of a sentence that is not explicitly expressed in an utterance but can be understood from a given conversational context. Accordingly, in order to compare LLMs' speech act understanding and LLM's conversational implicature understanding, this paper analyzed the correlation between performances, as shown in Table 11.

[7]Conversational implicature refers to the meaning that is implied in a conversation, which is not directly stated but inferred based on context, shared knowledge, and conversational norms.

The analysis results showed that the correlation between speech acts and conversational implicature was 0.67, indicating a moderately high correlation. As such, speech acts and conversational implicature are both close pragmatic topics in that they are pragmatic interpretations of utterances that take context into account. On the other hand, having a moderate correlation implies that there are subtle differences between two linguistic phenomena in their underlying mechanisms[11,38]. In the case of conversational implicature, all four types have a common underlying principle, i.e. Grice's cooperative principle[37], but in the case of speech acts, each type of speech act, such as directive, expressive, and commissive speech acts, has different felicity conditions, a set of criteria that must be satisfied for a particular speech act to be successfully performed.[8]

According to Table 11, compared to direct speech acts, the understanding of indirect speech acts is more correlated with that of implicatures in the sense of implicit comprehension, showing a correlation score of 0.73. In particular, the understanding of indirect speech acts is more correlated with implicit comprehension caused by the maxim of relation.

As discussed in section 4.A.b., the relatively low comprehension of directive indirect speech acts by LLMs can be attributed to their strong reliance on the literal meaning of sentences. When the literal meaning is sufficiently interpretable, LLMs are inclined to interpret the utterance as a direct speech act based on literal meaning rather than recognizing its indirectness. However, when the analysis of the literal meaning fails to yield an acceptable interpretation, LLMs attempt to infer deeper indirect intentions beyond the surface-level literal meanings. This aligns with the findings of Park et al.[28], who analyzed the capability of LLMs to understand conversational implicature. According to the paper, most LLMs could easily understand implicatures arising from violations of the maxim of quality, that is, when the truth value of an utterance deviates from reality. In other words, when the literal meaning of a sentence differs from the conversational context or world knowledge, the LLMs were able to capture it and analyze the hidden implicatures beneath the surface meaning of the sentence. This suggests that LLMs with large-scaled training data possess exceptional ability in analyzing the literal meaning of sentences. However, this tendency may pose challenges when performing pragmatics tasks, such as indirect speech acts.

Model	**Speech Act**	**Direct SA**	**Indirect SA**	**Impl.**	**Quantity Impl.**	**Quality Impl.**	**Relation Impl.**	**Manner Impl.**
GPT-3.5	50.14	63.61	36.67	51.11	66.67	52.78	42.89	53.61
GPT-4	71.39	84.72	58.06	65	83.89	82.22	70	75.28
Claude3-Haiku	33.75	41.94	25.83	56.57	67.78	58.89	43.33	56.67
Claude3-Sonnet	64.31	81.39	47.22	62.22	81.67	67.22	54.44	66.39
Claude3-Opus	71.94	69.17	74.72	81.11	88.89	88.89	81.11	85
Mistral-small	65.56	83.89	47.22	57.22	57.78	54.44	35	51.11
Mistral-medium	66.67	87.5	45.83	61.11	69.24	72.22	62.22	66.25
Mistral-large	69.58	83.61	55.56	61.11	71.11	61.11	52.22	61.39
Llama-3-8B	49.44	68.33	30.56	54.44	68.89	44.44	45.56	53.33
Qwen-14	40.14	51.67	28.61	52.22	61.67	56.11	43.33	53.33
Solar-10.7B	65	84.44	45.56	58.33	65.56	62.22	51.11	59.31
Kendall Speech Act	1.00			**0.65**	0.67	0.51	0.64	0.53
Kendall Direct Speech Act		1.00		0.31	0.31	0.11	0.27	0.16
Kendall Indirect Speech Act			1.00	0.73	0.66	0.65	0.7	0.59

Table 11. Correlation Score between Speech Act and Conversational Implicature in Park et al.[29] (SA: Speech Act, Impl.: Implicature)

Context-dependent Language Comprehension of LLMs

We conducted a comparative analysis of the evaluation results using PUB, a pragmatic benchmark[39], to assess the effectiveness of our test set in evaluating the speech act comprehension capabilities of LLMs. One of the most prominent English test sets was selected for comparison due to the relative scarcity of comprehensive test sets available in the Korean language. PUB primarily focuses on implicatures, presuppositions, and deictic expressions in its test sets. Given that speech acts are most closely aligned with conversational implicatures as pragmatic mechanisms, we compared our test's scores with those of PUB that specifically evaluate the comprehension of implicated meanings in a dialogue. To ensure consistency, GPT-3.5 was selected as the comparison model, as it was evaluated in both our study and PUB.

[8]This concept, introduced by J.L. Austin and further developed by John Searle, plays a crucial role in determining whether a speech act is valid and effective within its context.

The test scores of GPT-3.5 are as follows (Implicature Natural Language Inference test scores are excluded as they do not pertain to conversational implicatures.):

- Direct/Indirect Response Classification (PUB): 80.2
- Implicature Recovery (PUB): 78.13
- Response Classification without Implied Meaning (PUB): 58.18
- Speech Act Comprehension (Ours): 50.14

The 'Speech Act Comprehension' test score (50.14) is comparable to that of 'Response Classification with Implied Meaning' test (58.18). Both tasks involve a brief context and employ a multiple-choice question format aimed at identifying the speaker's intention. This similarity in structure and objective suggests that these two tasks are closely related in nature. Consequently, this alignment ensures that the test set is a reliable and valid tool for assessing speech intention comprehension. As a test designed solely to determine whether a specific response is direct or indirect, the 'Direct/Indirect Response Classification' test is relatively straightforward. It is, therefore, unsurprising that GPT-3.5 achieved its highest score in this task. However, the high score in the 'Implicature Recovery' test is noteworthy, given the absence of descriptive context. It is speculated that the context constructed through multi-turn dialogue may have facilitated GPT-3.5's ability to comprehend the implied meaning of the last utterance.

Conclusions

In this study, we evaluated whether LLMs appropriately comprehend the speaker's intended meaning in a given conversational context. Through this, we investigated whether LLM, which has recently gained popularity for its high language ability, also has pragmatic ability, which is considered to be the highest level of language ability. To this end, we conducted an evaluation experiment by constructing a total of 240 Korean test items for three types of indirect speech acts.

Our findings revealed that all LLMs yet showed a lower level of speech act comprehension, compared with humans. In both MCQ and OEQ settings, Claude3-Opus surpassed all other models and it was the only model to show relatively higher understanding of indirect speech acts, similar to humans. Furthermore, this study pointed out the possibility of overfitting to English data in the Llama3 model, since it provided English responses even though the prompt was in Korean. In the case of Solar model, it was observed to be predominantly overfitted to MCQ typed leaderboard benchmarks, creating unnecessary four choices even for OEQs, which is somewhat different from how humans solve OEQ-typed problems. In contrast, Claude3-Opus demonstrated high-level, sophisticated linguistic abilities, providing plausible and correct answers.

While our study provided a comprehensive evaluation of LLMs' ability to understand speaker's intention, to compare with the mainstream benchmarks, test sets in high-resource languages like English are needed. This study constructed an evaluation dataset grounded in linguistic theory, i.e., Searle's speech act scheme, and laid an optimal foundation for evaluating LLMs' ability to understand utterance intention. Our future research aims to develop data sets in other languages, such as English, Chinese or German, to validate the robustness of the data sets and to compare the speech act comprehension capabilities of LLMs across different languages.

Moreover, we will explore alternative experimental settings by having LLMs engage in simulated conversations with humans, to evaluate whether they can accurately understand utterance intentions and generate appropriate responses. This study assessed LLMs by having them either to select (MCQ) or explain (OEQ) the intention behind a given utterance considering the given conversational context. However, our future study will focus on evaluating LLMs in natural, conversational settings. Dialogue is an authentic communication format of human language use. LLMs' abilities could be examined under two settings: having them generate the next utterance considering the given multi-turn conversation as the input prompt ('Multi-turn Prompting') or having them exchange one utterance at a time to evaluate whether the LLM accurately understands the intent behind each utterance and generates a natural response ('Single-turn Prompting').

Data availability

The datasets generated and/or analysed during the current study are available in the anonymous GitHub repository: https://anonymous.4open.science/r/LLM-Indirect-SpeechAct-44D1/.

References

1. Sudirjo, F., Diantoro, K., Al-Gasawneh, J. A., Azzaakiyyah, H. K. & Ausat, A. M. A. Application of chatgpt in improving customer sentiment analysis for businesses. *J. Teknologi Dan Sistem Informasi Bisnis* **5**, 283–288 (2023).
2. Orrù, G., Piarulli, A., Conversano, C. & Gemignani, A. Human-like problem-solving abilities in large language models using chatgpt. *Front. artificial intelligence* **6**, 1199350 (2023).

3. Chang, Y. *et al.* A survey on evaluation of large language models. *ACM Transactions on Intell. Syst. Technol.* **15**, 1–45 (2024).
4. Fourrier, C., Habib, N., Lozovskaya, A., Szafer, K. & Wolf, T. Open llm leaderboard v2. *https://huggingface.co/collections/open-llm-leaderboard/open-llm-leaderboard-2-660cdb7601eba685243 1fffc* (2024).
5. Clark, P. *et al.* Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457* (2018).
6. Hendrycks, D. *et al.* Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).
7. Hall, E. T. Beyond culture. *Gard. City* (1976).
8. Bhagat, R. S., Kedia, B. L., Crawford, S. E. & Kaplan, M. R. Cross-cultural issues in organizational psychology: Emergent trends and directions for research in the 1990s. *Int. review industrial organizational psychology* **5**, 55–99 (1990).
9. Servaes, J. Cultural identity and modes of communication. In *Communication yearbook 12*, 383–416 (Routledge, 2012).
10. Hall, E. T. & Hall, M. R. *Understanding Cultural Differences* (Inter-Cultural Press Inc., New York, 1990).
11. Austin, J. L. *How to do things with words* (Harvard university press, 1975).
12. Searle, J. R. A classification of illocutionary acts1. *Lang. society* **5**, 1–23 (1976).
13. Searle, J. R. *Expression and meaning: Studies in the theory of speech acts* (Cambridge University Press, 1979).
14. Guo, Z. *et al.* Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736* (2023).
15. Arcara, G. & Bambini, V. A test for the assessment of pragmatic abilities and cognitive substrates (apacs): Normative data and psychometric properties. *Front. psychology* **7**, 70 (2016).
16. Kim, W., Eom, Y. & Lee, G. A study on the communication of students with mild intellectual disability seen in speech acts. *J. Intellect. Disabil.* **19**, 99–119 (2017).
17. Seo, S. An analysis of the domestic experimental research on asperger‘s disorder children. *Asia-pacific J. Multimed. Serv. Convergent with Art, Humanit. Sociol.* **9**, 527–535 (2019).
18. Jang, H. & Choi, H. Comprehension of indirect speech acts in patients with right brain damage. *Commun. Sci. & Disord.* **15**, 581–591 (2010).
19. Oh, S. Correlation between children's pragmatic language checklist and conversation observations. *Commun. Sci. Disord.* **23**, 109–118 (2018).
20. Han, J. *et al.* Recipe4u: Student-chatgpt interaction dataset in efl writing education. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 13666–13676 (2024).
21. Wang, D. *et al.* Can chatgpt detect student talk moves in classroom discourse? a preliminary comparison with bert. In *Proceedings of the 16th International Conference on Educational Data Mining* (2023).
22. Loukas, L., Stogiannidis, I., Diamantopoulos, O., Malakasiotis, P. & Vassos, S. Making llms worth every penny: Resource-limited text classification in banking. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, 392–400 (2023).
23. Bouzaki, A. G. *Enhancing intent classification via zero-shot and few-shot ChatGPT prompting engineering: generating training data or directly detecting intents*. Ph.D. thesis, Master thesis. NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS (2023).
24. Ostyakova, L., PetukhovaO, K., Smilga, V. & Zharikova, D. Linguistic annotation generation with chatgpt: a synthetic dataset of speech functions for discourse annotation of casual conversations. In *Proceedings of the International Conference Dialogue* (2023).
25. Yu, D., Li, L., Su, H. & Fuoli, M. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *Int. J. Corpus Linguist.* (2024).
26. Eo, S. *et al.* A study on proficiency in solving riddles of large language model. In *Annual Conference on Human and Language Technology*, 25–30 (Human and Language Technology, 2023).
27. Nam, Y., Chung, H. & Hong, U. Language artificial intelligences' communicative performance quantified through the gricean conversation theory. *Cyberpsychology, Behav. Soc. Netw.* **26**, 919–923 (2023).
28. Park, D., Lee, J., Jeong, H., Park, S. & Lee, S. Pragmatic competence evaluation of large language models for korean. In *Proceedings of the 38th Pacific Asia Conference on Language, Information, and Computation (PACLIC 38)* (2024).

29. Park, D. *et al.* Multiprageval: Multilingual pragmatic evaluation of large language models. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, 96–119 (2024).

30. Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

31. Anthropic. The claude 3 model family: Opus, sonnet, haiku. *https://api.semanticscholar.org/CorpusID:268232499* (2024).

32. Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

33. Bai, J. *et al.* Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).

34. Kim, D. *et al.* Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166* (2023).

35. Park, C. *et al.* Open ko-llm leaderboard: Evaluating large language models in korean with ko-h5 benchmark. *arXiv preprint arXiv:2405.20574* (2024).

36. Khatun, A. & Brown, D. G. A study on large language models' limitations in multiple-choice question answering. *arXiv preprint arXiv:2401.07955* (2024).

37. Grice, H. P. Logic and conversation. *Syntax semantics* **3** (1975).

38. Searle, J. R. Speech acts: An essay in the philosophy of language. *Camb. Univ.* (1969).

39. Sravanthi, S. L. *et al.* Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. In *Annual Meeting of the Association for Computational Linguistics* (2024).

Funding

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2023-00274280).

Author contributions statement

Y.K. designed and performed the experiments, constructed the test dataset, and analyzed the result. J.L. reviewed related works, constructed the test dataset, and conducted theoretical analysis. D.P. performed the experiments and assisted with statistical analysis. S.P. constructed the test dataset. S.L. designed and supervised the research. All authors reviewed the manuscript.