

The Limits of ChatGPT's Pragmatic Interpretation: An Experimental Study of Conversational Implicature in Indonesian Language Conversations

Aris Badara¹, Muhammad Rifa'ie¹, Ruslan Yusoh², Ibrahim Salama³

¹ Department of Indonesian Language and Literature Education, Universitas Halu Oleo, Kendari, Indonesia

² Darussalam School, Narathiwat, Thailand

³ Dar Al-Thikr Schools, Jeddah, Kingdom of Saudi Arabia

Article Info

Article History:

Received

20 October 2026

Approved

24 October 2026

Published

25 October 2026

Keywords: ChatGPT, conversational implicatures, contextual sensitivity, pragmatics, minimal stimulus pairs

Abstract

This study evaluates ChatGPT's interpretative accuracy and contextual sensitivity in understanding conversational implicatures in Indonesian. A computational experimental design employed 80 minimal stimulus pairs, yielding 160 responses across five communication themes and four pragmatic phenomena. Each pair contained identical utterances presented in different situational contexts. Data were collected using GPT-6 at a 'High' reasoning level via Temporary Chat (unpersonalised). Two pragmatics experts independently assessed the responses using a 0–2 ordinal rubric. The analysis comprised Interpretation Accuracy (IA), Pairwise Joint Accuracy (PJA), descriptive comparisons between conditions, and qualitative examination of pragmatic inferences. IA reached 98.13% for Rater 1 and 91.25% for Rater 2, while PJA reached 96.25% and 82.50%, respectively. Exact agreement was 93.13%, with a weighted Cohen's kappa of 0.332, indicating limitations in scoring consistency. According to Rater 2, pairwise accuracy was highest for indirect criticism or evaluation (95.00%) and lowest for indirect refusal (65.00%). Accuracy was descriptively higher in Condition A (96.25%) than in Condition B (86.25%). Qualitative analysis identified instances of inferences exceeding contextual support and insufficient use of relevant contextual information. These findings highlight the importance of evaluating both interpretative accuracy and contextual justification, while acknowledging the sensitivity of performance estimates to rater judgments.

© Copyright 2026

INTRODUCTION

The development of Large Language Models (LLMs) has expanded artificial intelligence's ability to generate and interpret natural language for a variety of communication needs. Although models such as ChatGPT can produce coherent responses that resemble human communication (Rifa'ie, 2026), linguistic fluency does not always reflect the ability to understand implied communicative intent.

In everyday interactions, utterances often do not fully express their meaning through sentence structure. Instead, it is inferred from context, the relationship between speakers, and the purpose of the conversation. This ability is a key part of pragmatic competence, particularly for understanding conversational implicatures.

Hu et al. (2023) suggest that understanding non-literal language remains a challenge for language models, even though large-scale models can achieve high accuracy on a number of

✉ Corresponding author:

Email: arishadara71@yahoo.co.id

p-ISSN: 2252-6323

e-ISSN: 2721-4540

pragmatic tasks. Bojić et al. (2025) also show GPT-4's strong performance on dialogue-based pragmatic tasks, even outperforming the human participants on the measures used. However, success on specific tasks does not automatically guarantee accurate interpretation when an utterance's meaning changes because of differences in the communicative situation. Therefore, evaluating pragmatic ability needs to focus on the model's accuracy in integrating contextual information.

Contextual interpretation is increasingly relevant because conversational implicatures allow speakers to convey refusal, criticism, agreement, or sarcasm without directly expressing their intent. The same statement may carry different meanings depending on the time, circumstances, purpose and shared knowledge of the conversation participants.

Qiu et al. (2023), through three pre-registered experiments, found that ChatGPT has not yet demonstrated human-like flexibility when switching between semantic and pragmatic processing. The study also found no influence of communicative context on implicature inference, as is commonly reported in human language processing.

In light of these issues, Sravanthi et al. (2024) developed the Pragmatics Understanding Benchmark (PUB), comprising approximately 28,000 data points across 14 pragmatic tasks. An evaluation of nine models revealed significant variation in performance across pragmatic phenomena and gaps between the models' capabilities and humans'. This evidence suggests that success in understanding one form of implicature cannot be generalised to other forms. Consequently, the key question is whether ChatGPT can generate pragmatic interpretations, and whether these interpretations remain accurate when contextual cues either support or constrain the derivation of implicatures.

Several recent studies have developed more systematic approaches to evaluating LLMs' pragmatic capabilities. Park et al. (2024) introduced MultiPragEval, which comprises 1,200 question items in English, German, Korean and Mandarin based on the Co-operation Principle and Grice's four conversational maxims. Meanwhile, Arai and Ren (2026) developed DRInQ through

controlled contextual variations whilst maintaining the same question format. The results showed that a model's ability to generate plausible pragmatic scenarios does not necessarily correlate with its ability to infer the intended implicatures.

Li et al. (2026) subsequently introduced the PaCE benchmark with over 3,000 context-flip samples. They identified a tendency toward pragmatic hallucination, namely drawing pragmatic inferences beyond the support provided by the literal context. These three studies demonstrate the importance of context manipulation as an evaluation strategy. However, these studies have not directly addressed how ChatGPT interprets open-ended responses in Indonesian across various pragmatic phenomena and domains of social communication.

Furthermore, multiple-choice tasks in many benchmarks may limit observations of interpretative reasoning, uncertainty, and additional inferences that emerge in the model's responses. This gap underscores the need for open-ended, response-based evaluation that can identify both the accuracy and the limits of pragmatic inference.

In the Indonesian context, natural language processing resources have advanced, but evaluating contextual pragmatic abilities still requires special attention. Cahyawijaya et al. (2023), through the NusaCrowd initiative, compiled 137 datasets and 118 standardised data loaders to support research into the Indonesian language and regional languages. This development strengthens the availability of resources for a broader evaluation of language comprehension. Furthermore, Cahyawijaya et al. (2024) developed the Cendol model—a suite of generative models tailored for Indonesian and regional languages—and reported improved performance across various tasks. However, it still exhibited limitations in representing local knowledge and cultural values.

Both studies show progress in developing Indonesian language resources and models, but neither specifically evaluated ChatGPT's ability to understand implied meaning based on changes in conversational context. This issue is important because the same utterance can have different meanings depending on the communicative situation. These differences highlight that

interpreting implicatures requires understanding the relationship between utterances and situational information. Therefore, testing is required to measure the accuracy of ChatGPT's interpretation when the form of the utterance is retained whilst the communicative context is altered.

This need forms the basis for examining ChatGPT's contextual sensitivity—its ability to adjust its interpretation of utterances based on available situational information. This issue is relevant across various domains of communication, such as education, friendships, family, work and public services, which frequently involve indirect refusals, implicit criticism and the use of irony. These circumstances underscore the need for a controlled evaluation to identify the accuracy of ChatGPT's interpretations and its tendencies towards errors in pragmatic inference in Indonesian conversations.

In light of this gap, this study aims to evaluate ChatGPT's accuracy and contextual sensitivity in interpreting conversational implicatures in Indonesian. The testing used 80 pairs of minimal stimuli, yielding 160 responses, while keeping utterances identical across two situational conditions. This study measures interpretation accuracy both individually and in pairs, compares performance across communication themes and pragmatic phenomena, and identifies patterns of contextual inference errors. Using a quantitative approach and qualitative pragmatic analysis, this study is expected to provide an empirical understanding of ChatGPT's ability to adapt its interpretations to changes in context and to limit inferences to available information. Figure 1 presents the relationship between stimulus design, the interpretation process, and the evaluation of accuracy within the conceptual framework.

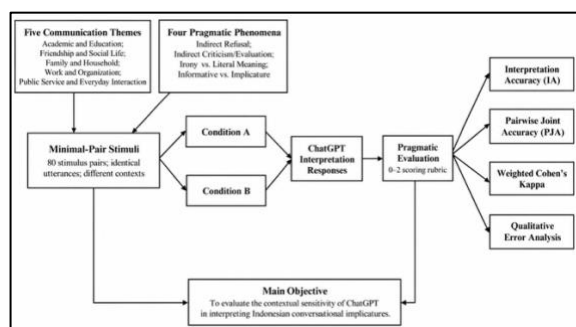


Figure 1. Conceptual Framework

This study is based on Grice's theory of conversational implicatures (1975) and Sperber and Wilson's Relevance Theory (1995), which view interpretation as an inferential process for linking utterances to communicative intent. A sentence's meaning does not fully determine the speaker's intention; therefore, comprehension requires consideration of intention, informativeness, and shared context (Sperber & Wilson, 2002; Frank & Goodman, 2012; Goodman & Frank, 2016). Research on scalar implicatures highlights the role of conversational context, the availability of alternatives, and the speaker's knowledge in forming interpretation (Breheny et al., 2006; Degen & Tanenhaus, 2015; Zhang et al., 2023). Probabilistic approaches also account for semantic uncertainty and expectations about unspoken alternatives (Bergen et al., 2016; Degen, 2023; Hu et al., 2023). This framework supports an evaluation based on contextual shifts, without assuming that all pragmatic phenomena share identical mechanisms.

Operationally, interpretative accuracy is defined as the alignment of a response with the reference interpretation supported by the stimulus, whilst contextual sensitivity refers to the alignment of interpretative changes with contextual changes. Over-inference occurs when a model prioritizes a conclusion without adequate support; context neglect occurs when relevant information is not utilized. However, additional information is not automatically an error as it may serve a communicative function, as demonstrated in research on referential expressions (Degen et al., 2020). Explicitly stated alternative mentions are uncertain and must therefore be distinguished from unfounded assertions. Although GPT-4's high pragmatic performance has been reported on specific tasks (Bojić et al., 2025), this study's assessment remains limited to output accuracy, use of context, and the appropriateness of the degree of certainty in the tested stimuli.

The research findings are expected to provide empirical evidence for the development of an assessment of LLMs' pragmatic competence in Indonesian, as well as to clarify the limits of the models' ability to distinguish between literal meaning and context-dependent implicatures..

METHODS

Research Design

This study employed a minimal-pair experimental design to evaluate ChatGPT's contextual sensitivity in interpreting conversational implicatures in Indonesian. We developed 80 stimulus pairs across two conditions (A and B) with identical utterances but different situational contexts, yielding 160 responses.

The stimuli were distributed evenly across five communication themes, namely: 1) academia and education, 2) friendship and social life, 3) family and household, 4) work and organisations, and 5) public services and everyday interactions. Each theme covers four pragmatic phenomena: indirect refusal, indirect criticism or evaluation, irony versus literal meaning, and informative utterances versus implicatures. This design draws on the principle of controlled contextual variation in evaluating conversational implicatures (Arai & Ren, 2026).

The stimuli are presented in written conversational Indonesian, using a relatively standard structure and vocabulary. The speech employs forms such as 'saya' and 'kamu', as well as the expressive marker 'wah', and situational vocabulary such as 'puskesmas', 'piket', and 'notula'. This variety is used to present short dialogues in academic, social, family, work, and public service contexts. The instrument is not designed to represent a specific regional dialect or ethnic group, and does not include the prosodic realizations of spoken conversation. Consequently, its linguistic scope is limited to written Indonesian dialogues in the specified situations.

The sociopragmatic dimension of the stimuli relates to inter-speaker relationships, social responsibilities, and interaction purposes. PKJ-001, for example, presents a request to attend a meeting within a superior-subordinate relationship, while PKJ-002 contains a request for task allocation by a coordinator. In KLG-001 and KLG-004, requests for assistance relate to the responsibility of caring for family members. Chojimah's (2015) study of Indonesian university students showed that social status is linked to the use of politeness strategies in refusal, meaning that the reasons given must be interpreted in conjunction with the relationship and

conversational context. However, information on hierarchy and closeness in this study serves as a backdrop for interpretation, rather than a systematically manipulated variable. Therefore, the research findings explain the pragmatic interpretation of the Indonesian-language stimuli used, without claiming to represent the entirety of Indonesian society's communication norms.

The research was carried out from June to September 2026 and comprised stimulus development, instrument validation, collection of ChatGPT responses, and data analysis.

Instrument Validity

Content validity was analysed using the Content Validity Index (CVI) based on a matrix of eight sets of assessments by eight Indonesian pragmatics experts holding PhDs, with relevance rated on a scale of 1–4. Scores of 3 and 4 were categorised as relevant for calculating the Item-Level Content Validity Index (I-CVI), the Scale-Level Content Validity Index based on Average (S-CVI/Ave), and Universal Agreement (S-CVI/UA), with an acceptance criterion of I-CVI $\geq$ 0.875.

Data Collection Procedure

Data collection was carried out using ChatGPT GPT-6 with a High reasoning level via the Temporary Chat feature, set to Unpersonalised. Each condition was tested independently in a new conversation session to limit the influence of memory, personalised instructions and conversation history on the model's responses. The question consistently asked was: "*Based on this situation, how do you interpret B's response to A? Explain briefly based on the available information.*" Each stimulus was tested once without providing any feedback. All 160 responses were documented by stimulus code, communication theme, pragmatic phenomenon, and experimental condition.

Response Assessment

ChatGPT's responses were evaluated using a three-category ordinal rubric: incorrect (0), partially correct (1) and correct (2). Two pragmatics experts evaluated ChatGPT's responses: Evaluator 1 (E1) and Evaluator 2 (E2). The agreement between the evaluation results was

analysed using the percentage of agreement and weighted Cohen's kappa..

Data Analysis Techniques and Tools

The data were tabulated in Microsoft Excel, and statistical analysis was carried out in Python via Jupyter Notebook using the pandas, NumPy, scikit-learn, SciPy, statsmodels, and Matplotlib libraries. We calculated interpretation accuracy (IA) using a metric adapted from classification accuracy and the pragmatic evaluation approach for LLMs (Sravanthi et al., 2024; Wu et al., 2024). IA is defined as the percentage of responses scoring 2 out of all responses assessed. Meanwhile, this study operationally defined Pairwise Joint Accuracy (PJA) based on the concept of contrast consistency (Zhang et al., 2023) and controlled context variation evaluation (Arai & Ren, 2026). PJA measures the percentage of stimulus pairs that receive a score of 2 in both conditions. Both indicators are formulated as follows:

$$IA = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(s_i = 2) \times 100\%$$

$$PJA = \frac{1}{M} \sum_{j=1}^M [\mathbb{I}(s_{j,A} = 2) \cdot \mathbb{I}(s_{j,B} = 2)] \times 100\%$$

Note.

- IA Interpretation Accuracy.
 PJA Pairwise Joint Accuracy.
 N Total number of responses ($N=160$).
 M Number of stimulus pairs ($M=80$).
 s_i The interpretation score for response (i) is 0, 1, or 2.
 $s_{j,A}$ Score under Condition A for pair (j)
 $s_{j,B}$ Score under Condition B for pair (j).
 $\mathbb{I}(\cdot)$ Indicator function; returns 1 if the condition is met and 0 if it is not.

IA and PJA were analysed as a whole and according to five communication themes and four pragmatic phenomena. The accuracy distribution for Conditions A and B was examined descriptively using the percentage of correct responses and paired cross-tabulations, with a score of 2 categorized as correct and a score of 0 or 1 as not fully correct. The qualitative analysis examined the relationship between model interpretation, situational cues, and target implicatures, including the adequacy of contextual support and the use of available information.

RESULTS AND DISCUSSION

Research findings

Stimulus characteristics and content validity of the instrument

The results show that all 80 pairs of ChatGPT responses achieved a relevance agreement of 8/8, with an I-CVI value of 1.000 for each pair.

Table 1. Content validity results

Communication themes	Pairs (n)	Agreement Relevance	I-CVI
Academic and Educational	16	8/8	1,000
Friendships and Social Life	16	8/8	1,000
Family and Household	16	8/8	1,000
Work and Organisations	16	8/8	1,000
Public Services and Everyday Interactions	16	8/8	1,000
Overall	80	8/8	1,000

Note. I-CVI = Item-Level Content Validity Index. The I-CVI value in the 'Overall' column is the average of 80 items. S-CVI/Ave = 1.000; S-CVI/UA = 1.000. The acceptance threshold for the eight assessors is I-CVI ≥ 0.875 .

All pairs met the content relevance threshold based on the scoring system used. However, agreement on relevance is not the same as accuracy in terms of pragmatic interpretation..

Score Distribution and Agreement on Response Evaluation

Two pragmatics experts evaluated 160 ChatGPT responses using the same 0–2 ordinal rubric. E1 awarded full marks to 157 responses (98.13 per cent), whilst E2 awarded full marks to 146 responses (91.25 per cent). A total of 149 responses (93.13 per cent) received identical scores from both raters, with a weighted Cohen's kappa of 0.332. The discrepancy between the high percentage of exact agreement and the relatively low kappa coefficient is attributable to the concentration of ratings in the highest score category and marginal differences in distribution between the two raters.

Table 2. Agreement on scores and weighted Cohen's kappa by theme

Communication themes	Agreement (n/N)	Agreement t (%)	Weighted κ
----------------------	-----------------	-----------------	-------------------

Academic and Educational	30/32	93,75	0,000
Friendships and Social Life	30/32	93,75	0,000
Family and Household	30/32	93,75	0,000
Work and Organisations	29/32	90,63	0,368
Public Services and Everyday Interactions	30/32	93,75	0,636
Overall	149/160	93,13	0,332

The kappa coefficient was zero for the first three themes because E1 scored all responses as 2. This meant the observed weighted agreement equalled the expected agreement based on the marginal distribution of the ratings.

Interpretation Accuracy Based on Theme

Interpretation accuracy is calculated as Interpretation Accuracy (IA), the percentage of responses scoring 2. Success in both conditions within a single pair is calculated using Pairwise Joint Accuracy (PJA), which is the percentage of pairs that achieved full marks in Conditions A and B. Based on the E2 assessment, the overall IA reached 91.25 per cent (146/160), whilst the PJA reached 82.50 per cent (66/80).

Table 3. Interpretation Accuracy and Pairwise Joint Accuracy by communication theme

Communication themes	IA E1 (%)	IA E2 (%)	PJA E1 (%)	PJA E2 (%)
Academic and Educational	100,00	93,75	100,00	87,50
Friendships and Social Life	100,00	93,75	100,00	87,50
Family and Household	100,00	93,75	100,00	87,50
Work and Organisations	96,88	87,50	93,75	75,00
Public Services and Everyday Interactions	93,75	87,50	87,50	75,00
Overall	98,13	91,25	96,25	82,50

Based on E2, the first three themes yielded the highest IA (93.75 per cent) and PJA (87.50 per cent). Meanwhile, the themes 'Work and Organisation' and 'Public Services and Everyday Interactions' achieved an IA of 87.50 per cent and a PJA of 75.00 per cent. The difference between the two indicators shows that accurate individual interpretations do not always translate into success in both stimulus-pair conditions..

Accuracy of Interpretation Based on Pragmatic Phenomena

Interpretation accuracy also varies across four pragmatic phenomena. Based on E2, indirect criticism or evaluation has the highest IA (97.50 per cent), while indirect refusal has the lowest IA (82.50 per cent). The differences are more pronounced in PJA, with a pairing success rate of 95.00 per cent for indirect criticism and 65.00 per cent for indirect rejection.

Table 4. Accuracy of interpretation based on pragmatic phenomena

Pragmatic phenomena	Appropriate responses (n/N)	IA (%)	Successful pairs (n/N)	PJA (%)
Indirect refusal	33/40	82,50	13/20	65,00
Indirect criticism/evaluation	39/40	97,50	19/20	95,00
Irony versus literal meaning	36/40	90,00	16/20	80,00
Informative versus implicature	38/40	95,00	18/20	90,00
Overall	146/160	91,25	66/80	82,50

Indirect rejection yielded seven partial responses, more than any other phenomenon. By contrast, indirect criticism or evaluation yielded only one partial response. This difference indicates variations in accuracy depending on the type of pragmatic inference required, although the size of the subgroups limits the drawing of broader conclusions.

Comparison of Accuracy Between Condition A and Condition B

A paired analysis of the E2 scores showed that 77 out of 80 responses in Condition A achieved a full score, compared with 69 out of 80 responses in Condition B. A total of 66 pairs were correct in both conditions, whilst 11 pairs were correct only in Condition A and three pairs were correct only in Condition B.

Table 5. Accuracy of interpretation for Conditions A and B

Kondisi A	B correct	B not entirely correct	Total
A correct	66	11	77
A not entirely correct	3	0	3
Total	69	11	80

Note. Accuracy is categorised on a binary scale: a score of 2 = correct; a score of 0 or 1 = not entirely correct.

The accuracy distribution for Conditions A and B was examined descriptively using the percentage of correct responses and paired cross-tabulations, with a score of 2 categorized as correct and a score of 0 or 1 as not fully correct. The qualitative analysis examined the relationship between model interpretation, situational cues, and target implicatures, including the adequacy of contextual support and the use of available information.

Discussion

This study shows that ChatGPT interprets conversational implicatures in Indonesian relatively well, but its performance is not entirely consistent when identical utterances appear in different situational contexts.

Based on the findings, Interpretation Accuracy (IA) reached 91.25 per cent, while Pairwise Joint Accuracy (PJA) stood at 82.50 per cent. The difference between these two indicators highlights the importance of distinguishing between the accuracy of individual responses and the success of interpretation at the pair level. These findings align with Hu et al. (2023), who identified high pragmatic ability in large-capacity language models but also noted limitations with certain phenomena, particularly those involving breaches of social expectations.

Sravanthi et al. (2024), using the Pragmatics Understanding Benchmark (PUB), also showed performance variation across pragmatic tasks and a gap between models and humans. In contrast, Bojić et al. (2025) reported that GPT-4 outperformed human participants' scores on the dialogue-based pragmatic evaluation they used. These three studies indicate that a single accuracy metric cannot adequately capture the pragmatic capabilities of LLMs when detached from task characteristics, response formats, and assessment mechanisms. Therefore, the results of this study are best understood as evidence of interpretative ability under the tested experimental conditions, rather than as proof of human-level pragmatic competence.

Conceptually, the distinction between individual and paired accuracy relates to whether the model truly integrates context when

constructing meaning. Grice's Co-operativism (1975) explains that speakers and addressees use co-operative assumptions to infer unstated intentions through the relationship between utterances, conversational goals and communicative maxims. Meanwhile, Sperber and Wilson's (1995) Relevance Theory frames utterance interpretation as an inferential process that accounts for contextual information and cognitive relevance.

Both frameworks provide a basis for distinguishing between meanings that can reasonably be inferred from context and interpretations that are merely speculative. Zhang et al. (2023) developed the concept of 'contrast consistency' to evaluate the consistency of a model's accuracy on model inputs subject to controlled changes, while Arai and Ren (2026) used contextual variations while maintaining the form of the question to test conversational implicatures.

The successful interpretation of 66 out of 80 stimulus pairs indicates that ChatGPT can generate responses consistent with the reference interpretation under both conditions for most stimuli. However, Pairwise Joint Accuracy (PJA) merely represents combined accuracy at the pair level and does not fully explain how the model adjusts its interpretation when contextual information changes. Therefore, this quantitative measurement needs to be supplemented with a qualitative analysis of the shifts in meaning produced under Conditions A and B. This approach aligns with Arai and Ren (2026), who emphasise the importance of controlled contextual variation in evaluating a model's ability to infer conversational implicatures. By examining the responses in pairs, it is possible to determine whether ChatGPT genuinely adjusts its communicative intent based on context or merely generates interpretations that appear generally relevant.

The importance of examining context-dependent interpretation is illustrated by stimulus pair AKD-001, presented in Figure 2. Both conditions contained the same exchange between Speakers A and B. Speaker A asked, "*Kamu ikut seminar besok pagi?*" ("Are you attending the seminar tomorrow morning?"), to which Speaker B replied, "*Besok saya ada ujian praktik*" ("I have

a practical exam tomorrow”). Although the utterances remained identical, the situational contexts differed, providing a basis for examining whether ChatGPT adjusted its interpretation according to the available contextual information.

AKD-001-A

Situasi: A mengajak B menghadiri seminar besok pagi. B wajib mengikuti ujian praktik di laboratorium sepanjang waktu seminar dan tidak diperbolehkan meninggalkan ujian.

A: "Kamu ikut seminar besok pagi?"
B: "Besok saya ada ujian praktik."

Pertanyaan: Berdasarkan situasi tersebut, bagaimana Anda memahami respons B terhadap A? Jelaskan secara singkat berdasarkan informasi yang tersedia.

Respons B dapat dipahami sebagai penolakan secara tidak langsung terhadap ajakan A untuk mengikuti seminar. Meskipun B tidak secara eksplisit mengatakan "tidak bisa", informasi bahwa B harus mengikuti ujian praktik pada waktu yang bersamaan dan tidak diperbolehkan meninggalkan ujian menunjukkan bahwa B tidak dapat menghadiri seminar tersebut.

Secara pragmatik, respons B mengandung implikatur percakapan, yaitu menyampaikan ketidakhadiran melalui penjelasan alasan, bukan melalui penolakan langsung.

AKD-001-B

Situasi: A mengajak B menghadiri seminar besok pagi. B mempunyai ujian praktik pada sore hari setelah seminar selesai dan telah menyiapkan waktunya pada pagi hari.

A: "Kamu ikut seminar besok pagi?"
B: "Besok saya ada ujian praktik."

Pertanyaan: Berdasarkan situasi tersebut, bagaimana Anda memahami respons B terhadap A? Jelaskan secara singkat berdasarkan informasi yang tersedia.

Respons B merupakan jawaban tidak langsung terhadap ajakan A. Pernyataan "Besok saya ada ujian praktik" dapat mengisyaratkan keraguan atau keengganan untuk mengikuti seminar karena B perlu mempersiapkan ujian.

Namun, karena ujian berlangsung sore hari dan B memiliki waktu luang pada pagi hari, respons tersebut tidak dapat dipastikan sebagai penolakan. Berdasarkan informasi yang tersedia, B belum menyatakan secara jelas apakah akan mengikuti seminar atau tidak.

Figure 2. Changes in ChatGPT's Responses to Differences in Contextual Stimuli

In Condition A (AKD-001-A), B was required to attend a practical examination in the laboratory throughout the seminar and was not permitted to leave the examination. ChatGPT interpreted B's utterance as "*penolakan secara tidak langsung*" ("an indirect refusal") and concluded that "*B tidak dapat menghadiri seminar tersebut*" ("B cannot attend the seminar"). This interpretation was consistent with the situational context, as the examination directly prevented B from attending the seminar. The response therefore reflected an appropriate inference of the intended communicative meaning rather than a merely literal interpretation of B's statement.

In Condition B (AKD-001-B), the examination was scheduled for the afternoon, after the seminar had ended, and B had already set aside time in the morning. Under these circumstances, ChatGPT acknowledged that "*respons tersebut tidak dapat dipastikan sebagai penolakan*" ("the response cannot be conclusively interpreted as a refusal"). However, it also suggested that B's statement

"*dapat mengisyaratkan keraguan atau keengganan untuk mengikuti seminar karena B perlu mempersiapkan ujian*" ("could suggest hesitation or reluctance to attend the seminar because B needed to prepare for the examination"). Although this alternative interpretation was expressed as a possibility rather than a definite conclusion, the context did not establish that B needed additional preparation time or was reluctant to attend

This difference indicates that the model can use temporal information to adjust its interpretation of implicatures in line with changes in context. However, ChatGPT still suggests that B may be reluctant to attend the seminar because they need to prepare for an exam, even though the context does not support this assumption. This finding demonstrates that the ability to identify changes in communicative intent is not always accompanied by the ability to restrict inferences to the available information. Consequently, ChatGPT's contextual sensitivity must be understood through two complementary aspects: the accuracy of interpretative adjustment and the control of pragmatic inferences not supported by the context.

ChatGPT's tendency to retain possible implicatures despite limited contextual support, as observed in AKD-001-B, aligns with the findings of Arai and Ren (2026) regarding the imbalance between a model's ability to generate plausible pragmatic interpretations and its ability to determine the implicature most appropriate to the context. This similarity suggests that recognising possible implied meanings does not necessarily come with the ability to evaluate the adequacy of the situational information underpinning those interpretations. Thus, the limitations of ChatGPT's pragmatic interpretation relate not only to the recognition of indirect meaning, but also to the accuracy with which context is used to determine whether an inference can be upheld.

This issue also relates to findings from previous research on the processing of implicatures and the comprehension of non-literal meaning by large language models. Research on general conversational implicatures and scalar implicatures reports that ChatGPT has not yet demonstrated human-like flexibility in adapting semantic and pragmatic processing, nor has it shown equivalent sensitivity to how

communicative context influences the interpretation of implicatures. Similar limitations are discussed in research on the comprehension of humour, metaphors, and text-based inference, which indicates that the ability to generate language resembling human language use is not always accompanied by accuracy in interpreting non-literal meaning (San Pietro et al., 2023; Al-Asadi, 2026). From a sociopragmatic perspective, Godwin-Jones (2025) also highlights the distinction between mastery of linguistic forms and pragmatically appropriate strategies, and the ability to adapt utterances to the situation and the interlocutor's perspective. This body of findings reinforces the notion that an interpretation which appears linguistically plausible is not necessarily the most justifiable one based on the communicative context.

In this study, this issue appeared not only in individual responses but also in variations in interpretation accuracy across pragmatic phenomena. This variation suggests that ChatGPT's success in interpreting implicatures is related to the characteristics of inference and the sufficiency of contextual information required for each phenomenon. Therefore, the subsequent discussion focuses on the phenomena with the highest levels of inaccuracy to identify the conditions under which ChatGPT tends to retain the possibility of an implied meaning even when situational information does not adequately support it.

This tendency is most evident in indirect refusal, which yields the lowest PJA of 65.00 per cent, compared with indirect criticism or evaluation (95.00 per cent), informative utterances versus implicatures (90.00 per cent), and irony versus literal meaning (80.00 per cent). The low accuracy in this phenomenon may be related to the inferential demands involved in distinguishing between reasons that genuinely hinder the fulfilment of a request and information that merely conveys the speaker's circumstances. Based on Grice's Co-operativeness Principle (1975), the interpretation of indirect refusal requires a relevant connection between the utterance, the communicative intent, and the conversational context. This issue is evident in the KLG-002-B stimulus data, where the morning shift had ended several hours before the family dinner. Yet,

ChatGPT still stated, "B might feel tired after the shift or have other reasons that have not been mentioned." Whilst such possibilities are socially conceivable, the context does not provide sufficient evidence to serve as a basis for interpreting a refusal. This tendency is conceptually linked to Qiu et al. (2023), who identified limitations in ChatGPT's flexibility in adapting its semantic and pragmatic processing to implicature tasks.

The adequacy of such contextual cues is also evident when the model must distinguish irony from literal evaluation. In stimulus AKD-012, the utterance "*Cepat sekali!* - That was quick!" is used in two different administrative service situations. When the service took nearly four hours for a process that usually takes five minutes, ChatGPT in Condition A concluded that "Response B is a form of sarcasm or irony." This interpretation is consistent with the discrepancy between the positive evaluation and the negative factual circumstances. However, when the service was completed in five minutes as expected in Condition B, ChatGPT still stated that "it cannot yet be determined whether B was speaking sincerely or sarcastically."

From the perspective of Relevance Theory, interpreting irony requires cues that support an understanding of the speaker's attitude toward the expressed meaning (Sperber & Wilson, 1995). In Condition B, the absence of a clear contradiction between the utterance and the situation meant that the available context better supported a literal interpretation. However, ChatGPT still retains the possibility of irony even though the contextual cues are not strong enough to support that interpretation. This finding suggests that the issue is not merely an inability to recognise irony, but a limitation in determining the interpretation that best fits the communicative situation. This tendency aligns with Wen and Tian (2025), who identified difficulties for ChatGPT-4o in understanding verbal irony, particularly when interpretation requires consideration of the speaker's emotional tone and communicative intent. These findings extend the pattern observed in indirect negation. In both phenomena, interpretation accuracy depends on the ability to determine when an implicature is justified and when the interpretation must be restricted.

The emergence of similar patterns across several pragmatic phenomena suggests that additional inferences need to be examined as a broader characteristic of errors (Sterken & Kirkpatrick, 2024). In this regard, Li et al. (2026) introduced the concept of ‘pragmatic hallucination’ via the PaCE benchmark, which refers to a model’s tendency to over-interpret literal context to the point of producing conclusions not supported by the information provided. This pattern resembles the responses AKD-001-B, KLG-002-B, and AKD-012-B, where ChatGPT identified the main meaning relatively accurately but still offered weak alternative interpretations. A similar tendency is also evident in PLY-014-B, where the model states that “the complaints desk may be near or integrated with the information desk.” In fact, the context describes only the positions of the information desk and the complaints desk on the building floor plan, without establishing a spatial relationship between them.

These errors do not stem from an inability to recognise the objects in question, but from adding relationships that lack a contextual basis. Although these results align with the characteristics of pragmatic hallucination described by Li et al. (2026), this study did not examine the influence of training or model alignment; consequently, the mechanisms underlying these errors cannot be directly equated.

Whilst previous cases showed a tendency to add information that was not available, the PKJ-008-B stimulus reveals the opposite form of inaccuracy: a failure to utilise information that had, in fact, been explicitly stated. In this stimulus, the context explains that the meeting minutes were completed in full and on time, while B states, “The minutes are complete.” ChatGPT then concludes that this response does not provide information about the timeliness of the minutes’ completion, even though the situational description already includes this information. These findings indicate that the model has not yet fully integrated information from the utterance and the context in its pragmatic evaluation.

Within Grice’s framework, the informativeness of an utterance must be interpreted by taking into account the knowledge available within the communicative event, not merely the elements explicitly stated.

Consequently, the pragmatic inaccuracies in this study can be categorised into two tendencies: over-inference about unavailable information and disregard for information that is actually available. Both tendencies show that managing contextual information is a more fundamental issue than merely recognising implicature forms.

These distinct error characteristics also highlight the importance of using an open-ended response format in pragmatic evaluation. Wu et al. (2024) highlight the limitations of multiple-choice-based assessment in capturing socio-pragmatic reasoning and propose directly assessing free-response answers, which aligns more closely with human judgement. This approach matters because responses with a correct main conclusion may still include additional inferences that conflict with the context.

In this study, the difference in the assessment results between E1 and E2 reflects this issue. The IA based on E1 reached 98.13 per cent, whilst that based on E2 was 91.25 per cent. This difference indicates that estimates of interpretative accuracy may be influenced by the scope of the scoring criteria, particularly when the model produces answers that combine the primary interpretation with speculative alternatives. As Wu et al. (2024) note, pragmatic reasoning in a social context does not always have a single definitive answer formulation. Therefore, evaluating open-ended responses needs to balance the accuracy of the reference interpretation with the possibility of alternative interpretations that remain justifiable in context.

Once the characteristics of errors and the influence of assessment criteria were identified, the analysis expanded to examine their distribution across various domains of communication. The PJA matrix in Figure 3 shows that indirect rejection in the themes of family and work each reached 50.00 per cent, whilst irony in public service also achieved 50.00 per cent. Conversely, some combinations of themes and phenomena achieved full accuracy (100.00 per cent). This pattern indicates that interpretation success is not uniformly distributed across all stimulus categories. Sravanthi et al. (2024) reported heterogeneity in LLM performance across various pragmatic phenomena, whilst Park et al. (2024), through MultiPragEval, emphasised the

importance of evaluating a model's inferential capabilities across diverse linguistic contexts.

This study's findings complement both approaches by offering observations across five domains of communication in Indonesian. However, each theme-phenomenon combination consists of only four pairs; consequently, a single pair that does not achieve full accuracy can change the PJA by 25 percentage points. Consequently, the differences between cells in Figure 3 represent a descriptive overview of the distribution of success rates, rather than evidence that one domain of communication is intrinsically more difficult than another. This distributional diversity matters for developing Indonesian language proficiency assessments for LLMs.

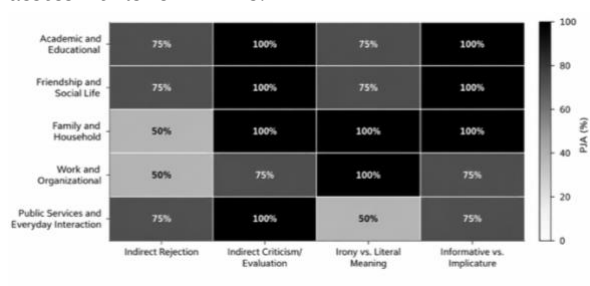


Figure 3. The PJA Matrix based on Communication Themes and Pragmatic Phenomena

Cahyawijaya et al. (2024), through Cendol, show that improvements in model performance across various Indonesian language tasks have not yet fully overcome limitations in representing knowledge and local cultural values. Although this study did not specifically measure ChatGPT's conversational implicatures, the results confirm that success in general linguistic tasks is not sufficient to represent all aspects of communicative competence.

In this study, stimuli covering academic, social, family, work and public service communication provide a variety of situations to evaluate the relationship between utterances and contextual information. This contribution is primarily methodological, as the study did not compare ChatGPT with other models, did not specifically test cultural knowledge, and did not use a corpus of natural conversations representing the diversity of Indonesian speakers. Consequently, these findings are best regarded as preliminary evidence for evaluating implicatures in the Indonesian language context.

In addition to variations by phenomenon and theme, comparing the two experimental conditions provides information on the distribution of inaccuracies at the pair level. Based on the E2 assessment, Condition A yielded an accuracy of 96.25 per cent, whilst Condition B achieved 86.25 per cent. Eleven pairs achieved a full score only in Condition A, whilst three pairs achieved a full score only in Condition B.

The accuracy distribution shows more responses without a full score in Condition B, but this pattern cannot yet be attributed to any single inferential requirement. Contextual changes across pairs include the removal of situational constraints, shifts in evaluative focus, and shifts in support for literal or non-literal interpretations. Furthermore, support for the target implicature is not always found in the same version of the stimulus. Consequently, the aggregate comparison between A and B illustrates the distribution of interpretative success, while explaining contextual sensitivity requires examining the relationship between contextual changes and the model's responses at the pair level. This distinction limits conclusions to the accuracy of the output within the instrument used, without equating differences between conditions with differences in the ability to recognize and disavow implicatures.

The dependence of results on assessment criteria also needs to be considered through an analysis of the consistency of the scoring procedures. The percentage of exact agreement between E1 and E2 reached 93.13 per cent, whilst the weighted Cohen's kappa was 0.332. The difference between these two measures reflects the concentration of scores in the 'full accuracy' category, which influences the agreement calculation after accounting for chance.

Overall, ChatGPT's ability to interpret conversational implicatures can be understood through two complementary inferential requirements: the ability to construct context-supported implied meanings and the ability to withhold interpretations that lack adequate support. The first requirement is evident in the model's success in adjusting its interpretations based on temporal conflicts and discrepancies between verbal evaluations and factual circumstances. The second requirement is evident in responses that still retain assumptions of

rejection or situational relationships without sufficient contextual evidence.

This distinction clarifies why a high IA score does not necessarily indicate the same level of inferential success across all stimulus pairs. Theoretically, these findings are consistent with the importance of context in Grice's theory of implicature (1975) and Sperber and Wilson's Relevance Theory (1995). Empirically, these findings complement those of Arai and Ren (2026) on controlled contextual variation, Wu et al. (2024) on the evaluation of open-ended responses, and Li et al. (2026) on excessive pragmatic inference.

The main contribution of this study is preliminary evidence that evaluating ChatGPT's pragmatic capabilities in Indonesian must consider not only the accuracy of the dominant meaning, but also the adequacy of the contextual basis for the resulting interpretation. Nevertheless, this study remains limited by several methodological factors. The testing utilised only a single GPT-6 configuration with a High reasoning level via Temporary Chat (unpersonalised) and a single response for each condition; consequently, the stability of the results' reliability across repetitions and generalisation across models cannot yet be determined. The instruments were designed as controlled experimental stimuli rather than derived from natural conversation; consequently, they do not yet fully represent the complexity of spontaneous interactions involving prosody, interpersonal relationships, and broader shared information.

Consequently, further research should prioritise repeated testing and explicit coding of the directional accuracy of interpretative shifts within each pair. This development is necessary to evaluate contextual sensitivity based on the model's success in responding to stimuli and its ability to adapt and consistently constrain pragmatic inferences.

CONCLUSION

This study indicates that ChatGPT shows high accuracy in interpreting conversational implicatures in Indonesian, with an Interpretation Accuracy (IA) of 91.25 per cent and a Pairwise Joint Accuracy (PJA) of 82.50 per cent based on

preliminary assessments. Accuracy varied across the four pragmatic phenomena, with the highest performance in indirect criticism or evaluation (PJA 95.00 per cent) and the lowest in indirect refusal (PJA 65.00 per cent). Qualitative analysis identified two error tendencies: pragmatic inferences that go beyond contextual support and the disregard of relevant situational information.

The research findings show that ChatGPT's success in producing pragmatic interpretations consistent with the reference is not always accompanied by inferential constraints based on available contextual information. These patterns emerged from a qualitative examination of responses that included additional interpretations and disregarded situational information.

This study provides preliminary empirical evidence on the accuracy of interpretations, success rates across stimulus pairs, variations in performance across communication themes and pragmatic phenomena, and the characteristics of inferential errors. Theoretically, the findings reinforce the relevance of situational context in forming conversational implicatures; methodologically, the study offers an approach to evaluating the pragmatics of Indonesian-language LLMs through controlled context manipulation based on minimal stimulus pairs. Instrument validation and independent assessment by pragmatics experts are required to strengthen the validity of the results and determine their applicability to broader communicative contexts..

REFERENCES

- Al-Asadi, N.S. (2026). A pragmatic study of indirectness in artificial intelligence's generated language. *International Journal of Social Science and Human Research*, 09(01), 551–556. <https://doi.org/10.5281/zenodo.18334628>
- Arai, H. J., & Ren, X. (2026). DRInQ: Evaluating conversational implicature with controlled context variation. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 34594–34611). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.acl-long.1597>
- Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics*

- and *Pragmatics*, 9, Article 20. <https://doi.org/10.3765/sp.9.20>
- Bojić, L., Kovačević, P., & Čabarkapa, M. (2025). Does GPT-4 surpass human performance in linguistic pragmatics? *Humanities and Social Sciences Communications*, 12, Article 794. <https://doi.org/10.1057/s41599-025-04912-x>
- Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, 100(3), 434–463. <https://doi.org/10.1016/j.cognition.2005.07.003>
- Cahyawijaya, S., Lovenia, H., Aji, A. F., Winata, G., Wilie, B., Koto, F., Mahendra, R., Wibisono, C., Romadhony, A., Vincentio, K., Santoso, J., Moeljadi, D., Wirawan, C., Hudi, F., Wicaksono, M. S., Parmonangan, I., Alfina, I., Putra, I. F., Rahmadani, S., ... Purwarianti, A. (2023). NusaCrowd: Open source initiative for Indonesian NLP resources. *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 13745–13818). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.868>
- Cahyawijaya, S., Lovenia, H., Koto, F., Putri, R. A., Dave, E., Lee, J., Shadieq, N., Cenggoro, W., Akbar, S. M., Mahendra, M. I., Putri, D. A., Wilie, B., Winata, G. I., Aji, A. F., Purwarianti, A., & Fung, P. (2024). Cendol: Open instruction-tuned generative large language models for Indonesian languages. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 14899–14914). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.796>
- Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, 9, 519–540. <https://doi.org/10.1146/annurev-linguistics-031220-010811>
- Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, 39(4), 667–710. <https://doi.org/10.1111/cogs.12171>
- Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to “overinformative” referring expressions. *Psychological Review*, 127(4), 591–621. <https://doi.org/10.1037/rev0000186>
- di San Pietro, C. B., Frau, F., Mangiaterra, V., & Bambini, V. (2023). The pragmatic profile of ChatGPT: Assessing the communicative skills of a conversational agent. *Sistemi Intelligenti*, 35(2), 379–399.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998. <https://doi.org/10.1126/science.1218633>
- Godwin-Jones, R. (2025). ChatGPT, second language pragmatics, and critical AI literacy. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 274–286). Iowa State University Digital Press. <https://doi.org/10.31274/isudp.2025.211.14>
- Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11), 818–829. <https://doi.org/10.1016/j.tics.2016.08.005>
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 41–58). Academic Press. https://doi.org/10.1163/9789004368811_003
- Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2023). A fine-grained comparison of pragmatic language understanding in humans and language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4194–4213). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.230>
- Hu, J., Levy, R., Degen, J., & Schuster, S. (2023). Expectations over unspoken alternatives predict pragmatic inferences. *Transactions of the Association for Computational Linguistics*, 11, 885–901. <https://doi.org/10.1162/tacl.a.00579>
- Li, Z., Tian, Y., Lan, T., Li, J., Duo, Z., Gao, G., & Su, X. (2026). Born pragmatic, trained to hallucinate? Quantifying the origins of contextual bias in LLMs via the PaCE benchmark. *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 19211–19235). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.findings-acl.959>
- Park, D., Lee, J., Park, S., Jeong, H., Koo, Y., Hwang, S., Park, S., & Lee, S. (2024). MultiPragEval: Multilingual pragmatic evaluation of large language models. *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP* (pp. 96–119). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.genbench-1.7>
- Qiu, Z., Duan, X., & Cai, Z. (2023). Does ChatGPT resemble humans in processing implicatures?

- Proceedings of the 4th Natural Logic Meets Machine Learning Workshop* (pp. 25–34). Association for Computational Linguistics.
<https://aclanthology.org/2023.naloma-1.3/>
- Rifa'ie, M. (2026). When ChatGPT “feels like family”: Serial mediation pathways to reliance and critical use. *Turkish Online Journal of Distance Education*, 27(4), 23–41.
<https://dergipark.org.tr/en/pub/tojde/article/1835504>
- Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Blackwell.
- Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind & Language*, 17(1–2), 3–23. <https://doi.org/10.1111/1468-0017.00186>
- Sravanthi, S., Doshi, M., Tankala, P., Murthy, R., Dabre, R., & Bhattacharyya, P. (2024). PUB: A pragmatics understanding benchmark for assessing LLMs’ pragmatics capabilities. *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 12075–12097). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2024.findings-acl.719>
- Sterken, R. K., and Kirkpatrick, J.R. (2024). Conversational Alignment With Artificial Intelligence in Context. *Philosophical Perspectives* 38, no. 1: 38, 89–102.
<https://doi.org/10.1111/phpe.12205>
- Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence*, 7, Article 100255.
<https://doi.org/10.1016/j.caeai.2024.100255>
- Wen, X. & Tian, Y. (2025). Understanding ironic utterances: A comprehensive examination of ChatGPT-4o. *Intercultural Pragmatics*, 22(2), 259–283. <https://doi.org/10.1515/ip-2025-2004>
- Wu, S., Yang, S., Chen, Z., & Su, Q. (2024). Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 22583–22599). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2024.emnlp-main.1258>
- Zhang, Z., Bergen, L., Paunov, A., Ryskin, R., & Gibson, E. (2023). Scalar implicature is sensitive to contextual alternatives. *Cognitive Science*, 47(2), Article e13238.
<https://doi.org/10.1111/cogs.13238>
- Zhang, Z., Yu, W., Ning, Z., Ju, M., & Jiang, M. (2023). Exploring contrast consistency of open-domain question answering systems on minimally edited questions. *Transactions of the Association for Computational Linguistics*, 11, 1082–1096.
<https://doi.org/10.1162/tacl.a.00591>