



How does GPT-4.1 comprehend conversational implicatures? Reasoning with contextual alternatives in discourse frames^{*}

Vincent Xian Wang^{**} · Lily Lim

(University of Macau · Macao Polytechnic University)

Wang, Vincent Xian and Lily Lim. 2025. How does GPT-4.1 comprehend conversational implicatures? Reasoning with contextual alternatives in discourse frames. *Linguistic Research* 42(Special Edition): 29-58. This study examines the pragmatic competence of GPT-4.1, specifically its ability to recover meaning conveyed by Levinsonian generalised conversational implicatures (GCIs) in the context of interaction with general users. Through nine open-ended queries, we evaluated whether GPT-4.1 can infer implicated meaning as predicted by Levinson's Q-, M-, and I-principles, and investigated the mechanisms underlying its performance. Our findings indicate that GPT-4.1 effectively recovers implicated meanings consistent with theoretical predictions. Analysis of the model's responses reveals a reliance on evidence-based, context-sensitive interpretation of lexical items and their alternatives, rather than on the direct application of abstract pragmatic principles. While the chatbot demonstrates notable competence in processing implicated meaning, it primarily operates through contrastive meaning analysis of lexical alternatives in context, backed by large-scale statistical associations, rather than explicit rule-based reasoning. These results lend support to a growing body of evidence suggesting that LLMs approach human-like performance in pragmatic inference, while also highlighting areas for further research into their ability to use abstract pragmatic rules. (University of Macau · Macao Polytechnic University)

Keywords GPT-4.1, large language model, conversational implicature, discourse frame, heuristics

* The research is funded by the Research grant number RP/BELL-01/2022 of Macao Polytechnic University.

** First and corresponding author

1. Introduction

The rapid development of LLMs powering chatbots such as ChatGPT, Llama, Gemini, Claude, Grok, DeepSeek, and Kimi has brought artificial intelligence to a level that approaches, and some argue, may surpass human intelligence in certain domains. However, human intelligence encompasses not only linguistic and reasoning competence, but also the ability to convey and interpret implied meanings beyond what is explicitly stated. Assessing LLMs' ability to comprehend implied meaning in natural language use is crucial, yet research in this area remains limited.

In response to the need for research in the specific area, we formulated our research questions as follows:

- a) Does GPT-4.1 recover meaning conveyed by Levinsonian generalised conversational implicatures (GCIs)?
- b) With what mechanism does GPT-4.1 recover the meaning conveyed by the GCIs?
- c) Does GPT-4.1 apply Levinson's Q-, M- and I-principles to recover meaning conveyed by the GCIs?

The first research question examines whether the LLM is able to recover the implicated meaning in full. The second question investigates the means by which the chatbot recovers implicated meaning. The final question probes whether the LLM is able to (spontaneously) apply pragmatic principles in order to recover implicated meaning.

2. Literature survey

We survey the literature on theoretical constructs of conversational implicatures and also the research studies relating to LLMs and implicatures.

2.1 Conversational implicatures

Our scholarly understanding of conversational implicatures primarily stems from

Grice's (1975) seminal work on the Cooperative Principle. Grice argued that interlocutors communicate not only by what is said, but also by what is implied. He named the latter what is *implicated*, and the inferred meaning *implicature*. Grice believed that interlocutors observe the Cooperative Principle (CP) when communicating with each other, and he proposed that CP consists of four maxims – Quality, Quantity, Relation, and Manner (ibid). Grice reasoned that flouting one of the maxims engenders implicature.

Grice's notion of non-natural meaning (or meaning-*nn*) – i.e. meaning intended by a speaker by his/her utterance to create an effect on a hearer and to be recognised by the hearer – was adopted by Levinson (e.g. 1983: 16) as a basic concept in his neo-Gricean account of pragmatics, focusing on generalised conversational implicatures (GCIs: ibid; Levinson 2000). Levinson proposed a framework consisting of three principles – i.e. the Q(uality)-, the M(anner)-, and the I(nformativeness)-principle – to explain the distinct mechanisms of three types of GCI – i.e. Q-, M-, and I-implicatures (ibid). The principles predict clear and accountable implicatures at work in conversations, accounting for a layer of 'meaning potential', argued by Levinson (ibid), with which speakers can say less and convey more in meaning-*nn* to ease the 'bottleneck' of communication.

We constructed our battery of tests in accordance with Levinson's three types of implicature (cf. Section 3.1). Levinson's three principles for GCIs are meticulously introduced in Huang's (2014) textbook on pragmatics, which is cited as follows:

The Q-principle (ibid: 50):

Speaker's maxim:

Do not provide a statement that is informationally weaker than your knowledge of the world allows, unless providing a stronger statement would contravene the I-principle.

Recipient's corollary:

Take it that the speaker made the strongest statement consistent with what he knows

The M-principle (ibid: 62):

Speaker's maxim:

Indicate an abnormal, non-stereotypical situation by using marked

expressions that contrast with those you would use to describe the corresponding normal, stereotypical situation.

Recipient's corollary:

What is said in an abnormal way indicates an abnormal situation, or marked messages indicate marked situations.

The I-principle (ibid: 57):

Speaker's maxim: the maxim of minimization

'Say as little as necessary', that is, produce the minimal linguistic information sufficient to achieve your communicational ends, (bearing the Q-principle in mind).

Recipient's corollary: the rule of enrichment.

2.2 LLMs and implied meaning

There is only a small array of studies that investigate LLMs' competence to process implied meaning. One of the earliest and frequently cited studies was conducted by Hu et al. (2022), using a range of scenarios to test humans' and LLMs' comprehension of pragmatic meaning. The scenarios were designed to contain a rather wide range of events in which implied meaning was at work – e.g. indirect speech acts, (flouting a) Gricean maxim, irony, metaphor, coherence of discourse. Multiple-choice tests were administered, and among the LLMs, the highest accuracy was achieved by OpenAI's text-davinci-002, with accuracy rates close to humans', while GPT-2 performed markedly poorer compared to humans and text-davinci-002. From the results, Hu et al. (ibid) predicted that pragmatic competence would *emerge* in the LLMs even if the models do not explicitly establish a representation of the mental state of the (human) agent in communication. Like Hu et al. (ibid), Cho and Mook (2024) observed that GPT-2 exhibited processing difficulties in inferring pragmatic implicatures in context.

Echoing Hu et al. (2022) and Cho and Mook (2024), a subsequent study revealed that GPT-3 was less flexible than humans in switching between pragmatic and semantic interpretations in communicative events in which implicatures are in effect (Qiu et al. 2023). In addition, Cong (2024) investigated LLMs' competence in interpreting

Manner implicatures and discovered that the models still fell short of an ability to interpret sophisticated contextualised cues, and only GPT-4 scored well in reasoning tests.

Only until more recently did researchers observe significant improvements in LLMs' pragmatic competence. Strachan et al. (2024) discovered that GPT-4 models achieved human-like competence in identifying indirect requests, even outperforming humans at times. Additionally, Bojić, Kovačević, and Čabarkapa (2025) investigated LLMs' ability to interpret context-specific cues and implied meaning using dialogues designed based on Gricean maxims, revealing that GPT-4 scored (4.80) higher than humans (4.55). We need to point out that the results of relevant studies cannot be compared in a simple manner, because tests used in the experiments are of varied levels of difficulty, and also different contexts of discourse are introduced in the tests.

In terms of research methodology, Wu et al. (2024) raised a challenge against the multiple-choice test format, which has been predominantly used in assessing LLMs' pragmatic competence (e.g. Hu et al. 2022; Hu et al. 2025). Wu et al. (2024) pointed out that multiple-choice tests do not allow the LLMs' reasoning processes to be revealed, and thus propose the use of free-form responses to give room to LLMs to elaborate their reasons in open-ended answers. We believe their observation is valid, and there is an urgent need for in-depth interaction with LLMs to gather qualitative data of their extensive responses. This would provide new insights into the reasoning mechanisms of LLMs.

The pragmatic competence demonstrated by LLMs is closely linked to scholarly interest in the mechanisms by which these models process implicated meanings. To date, LLMs' mechanism of reasoning has been a burgeoning field of research, particularly concerning LLMs' use of (linguistic) rules. Cho et al. (2025) tested LLMs' capability in reasoning with phonological rules that had not been recorded in the literature and discovered ChatGPT has achieved a considerable level of competence in this respect, with GPT-4o outperforming GPT-4. Research is scarce on LLMs' reasoning mechanism of implicatures. Yue et al. (2024) found that LLMs are less able to explain their reasoning, while GPT-4 performed the best among the LLMs in their experiments.

We hold that reasoning for implicated meaning heavily relies on interlocutors' identifying and comparing alternative words and expressions in the relevant context of discourse. Levinson's Q- and M-principles explicitly articulate the interplay between

the stronger and the weaker alternates and also between the marked and the unmarked (Levinson 2000; Huang 2014). Research literature also underscores the crucial role of pragmatic alternatives. Acton and Hunt (2025) argued that pragmatic alternatives contribute to engender not only implicatures but also social meaning. In addition, Hu et al. (2023) used experimental evidence to demonstrate that the strength of interlocutors' expected alternatives in terms of meaning in context robustly predicts their interpretation of Q-scalar implicature.

Given LLMs' outstanding strengths in identifying lexical and discursive meaning in multiple layers of context, the present study places emphasis on examining pragmatic alternatives that LLMs identify and compare in contexts of discourse for recovering implicated meaning.

3. Method

3.1 Design of tests

We designed our battery of tests on three types of implicature according to Levinson's framework on GCIs (cf. Section 2.1). There are nine tests in total, with three for each type of implicature. Each test contains a sentence or short discourse, followed by questions that probe the LLM's comprehension of the implicated meaning and its reasons.

Our questions now seek the chatbot's opinion on what *hearers* of the sentence would think. The questions are purposely formulated *not* to directly ask what the chatbot 'thinks' the sentence means. This is because our pilot study revealed that GPT-4.1 tended to 'play safe' by providing broader rather than more specific answers when our questions directly solicited its 'own' opinion. By contrast, the chatbot would answer more specifically when being asked what hearers would think.

3.2 The language model and the host interface

We used GPT-4.1 (model of 2025-04-14) in our experiments, which is a new and representative language model of OpenAI. Poe.com notes: "OpenAI's latest flagship model with significantly improved coding skills, long context (1M tokens), and

improved instruction following. Supports native vision, and generally has more intelligence than GPT-4o.”

We accessed GPT-4.1 via the Users Interface (UI) of Poe.com (<https://poe.com/GPT-4.1>) from 19 to 28 July 2025. Our experiment utilised UI rather than API (Application Programming Interface) to access the chatbot, and we were therefore unable to set parameters for the LLM, such as temperature. We also understand that Poe.com may fine-tune the version of the language model without explicit notification. However, accessing the chatbot via UI allows us to assess GPT-4.1 as it performs and fine-tunes itself for general users rather than expert programmers. The experiment condition, therefore, assimilates the chatbot’s competence for the majority of users. The particulars indicated at Poe.com are: Powered by OpenAI: gpt-4.1-2025-04-14, OFFICIAL.

Poe.com is one of the largest hosts of the most updated LLMs for regular users across the globe, and is representative in the market (Zhao et al. 2025). LLMs at Poe.com have been studied in the literature, e.g. on their creativity and comprehension of linguistic subtleties compared to humans (Molares-Cardoso et al. 2024).

3.3 Test administration and data analysis

Each test was administered as a new query to the chatbot, either in a new thread or by cleaning the context to avoid interference of previous queries, except the occasions in which follow-up questions were purposely asked. We read the chatbot’s suggestion of follow-up questions at the end of each query to detect any signs of continued reasoning. We also designed our own (follow-up) questions to engage the chatbot in an expert role or to apply a specific pragmatic principle.

We examined GPT-4.1’s answers in terms of whether the chatbot had recovered the implicated meaning-*nn* and by what means. We paid particular attention to the pragmatic alternatives it considered and discussed, comparing the alternatives with the expected alternatives in theoretical accounts of GCIs. We investigated the chatbot’s logic of reasoning and the contexts of discourse it introduced to match specific pragmatic alternatives. Our case study is exploratory and qualitative in nature.

4. Results

We present the results of GPT-4.1’s performance in responding to our battery of tests concerning the Q-, M- and I-implicatures, respectively, in this section.

4.1 Q-implicature

GPT-4.1 drew conclusions that successfully recover implicated meanings of our three tests on Levinson’s Q-implicatures, while the reasons provided by the chatbot are carefully evaluated in context in this section.

Table 1. Test (Q1): prompt and answer

Prompt	GPT-4.1’s answer (abridged by the investigators for conciseness)
(Q1) “John tried to learn swimming last summer.” Questions: From the sentence, would hearers think John learned how to swim last summer? Why would they think so?	No, hearers would not think that John necessarily learned how to swim last summer just from the sentence. Why? The verb “ tried ” indicates <i>attempt</i> rather than <i>success</i> . When someone says “John tried to learn swimming,” it means John made an effort to learn, but it doesn’t tell us whether he succeeded. • “Tried to X” usually implies that X may not have happened. e.g. “She tried to open the door.” (Maybe could not open it.) Implicature Hearers often infer that if John had succeeded, the speaker would simply say: • “John learned to swim last summer.” By choosing “tried to learn,” the speaker <i>implies</i> that John may not have learned to swim. Conclusion Hearers would NOT think John learned to swim last summer, because “tried to learn” suggests that he may have attempted but did not necessarily succeed.

In response to test (Q1), the LLM refuted the position expressed in our prompt that hearers would think John learned swimming last summer. The reasons provided by GPT-4.1 are twofold. First, semantically, ‘tried’ entails attempts rather than success. Second, borrowing the term ‘implicature’ from pragmatics, the LLM reasoned that,

for hearers, a speaker would say “John learned to swim last summer” if he had succeeded; and if the speaker chose ‘tried to learn’ instead of ‘learned’, he or she implies that “John may **not** have learned to swim” (cf. the Implicature section in Q1).

The second reason looks like the logic expressed in Levinson’s Q-principle (or Horn scale), but close comparisons of the two reveal the former still has crucial differences from the latter:

- a) GPT-4.1 did not treat ‘tried’ and ‘learned’ as the *stronger* and the *weaker* members of a set of alternatives in order.
- b) GPT-4.1’s reasoning did not rely on the logic of the Q-principle that hinges on the order (strong to weaker). By contrast, the Q-principle does as follows: a speaker should *not* use the weaker alternate to formulate a weaker statement than the speaker’s knowledge allows, in compliance with the Quantity Maxim of Austin’s Cooperative Principle. And, logically, the use of informationally weaker alternate (‘tried’) implies that the condition for making the strong statement (‘John learned to swim’) is not met (cf. Huang 2014: 50ff).
- c) GPT-4.1’s reasoning leads to a statement of implicature – “John may not have learned to swim” – not as clear as the Q-implicature “John failed”.

GPT-4.1’s reasoning mainly lies in its understanding of the meaning that ‘*tried to learn*’ conventionally infers in discourses, compared to that of ‘learned’. The chatbot effectively recovered the implicated meaning, but there is little evidence that the LLM utilised pragmatic rules to do it.

Test (Q2a) was designed to contain Q-clausal conversational implicature. The matrix clause of the sentence uses the verb ‘believes’, which is a weaker form in relation to ‘knows’, the stronger form. The use of the weaker form suggests that John *does not know* his son has passed the test for sure, because otherwise he would use ‘knows’, the stronger form (Huang 2014: 52).

Table 2. Test (Q2a): prompt and answer

Prompt	GPT-4.1's answer (abridged by investigators)
(Q2a) "John believes his son has passed the driving test." Questions: Based on this sentence, would hearers think John's son may have passed the test, or possibly he may not have? Why would they think so?	Based on the sentence, hearers would think: John's son may or may not have passed the test — the outcome is uncertain. Why would they think so? • The verb "believes" signals that this is John's opinion or assumption, not a confirmed fact. • The sentence doesn't say "John's son has passed the test" (which would state it as a fact). • When someone says, "X believes Y," it often means there is some doubt or lack of confirmation about Y. • Hearers infer that the speaker does not know for sure whether John's son has passed; they only know what John thinks. Summary: Hearers would think it's possible John's son passed, but it's also possible he didn't. The use of "believes" introduces uncertainty.
(follow-up question offered by GPT-4.1) What if the sentence used "knows" instead of "believes"?	If the sentence is: "John knows his son has passed the driving test." Would listeners think John's son may not have passed? No. Listeners would strongly assume that John's son has, in fact, passed the test. Why? • The verb "knows" implies that the statement is true and that John has reliable information or evidence. • "John knows X" <u>presupposes</u> that X is a fact. • The speaker isn't just reporting John's opinion, but presenting the passing of the test as something certain and established. In short: With "knows," listeners would take it as a fact that John's son passed the driving test.

GPT-4.1 arrived at the same conclusion as predicted by Q-clausal implicature – i.e. John's son *may or may not* have passed the test. The LLM provided reasons that are hinged on the meaning of 'believes' – it is about opinion rather than fact, and using 'believes' in a discourse leads hearers to "infer that the speaker does not know for sure". The argument makes sense, although it does not rely on the Q-principle or Q-clausal implicature, more specifically.

The chatbot offered one of the follow-up questions that goes – “What if the sentence used ‘knows’ instead of ‘believes’?”. The LLM was hereby contrasting ‘believes’ with ‘knows’ to test whether they infer different meanings in the discourse. This is worth noting because the pair of alternates plays a crucial role in Q-clausal implicatures. For example, Huang (2014: 52) uses the term ‘verbal doublet’ for the pair of verbs, in which ‘know’ is the stronger alternate while ‘believe’ is the weaker. Asking the follow-up question led the chatbot to answer that “John knows X *presupposes* that X is a fact” and therefore “hearers would take it as a fact that John’s son passed the driving test”. It effectively used the pragmatic concept of *presupposition*, although it did not revisit the question of ‘believe’ with the new light (see Q2a-follow-up).

Given that it can take up to two query-answer loops for the LLM to begin employing pragmatics concepts to resolve the problem in (Q2a), the investigators were curious about the LLM’s full potential and wondered whether engineering our prompt would enable the chatbot to exercise fuller linguistic capacities in the first place – for example, devising a prompt that invites GPT-4.1 to act like a linguistics expert. The prompt in (Q2b) includes an additional line that requires LLM to enact the role of a professor of linguistics, and was launched in a new thread to avoid interference with the context.

Table 3. Test (Q2b): prompt and answer

Prompt	GPT-4.1’s answer (abridged by investigators)
(Q2b)	Certainly! Let’s break down the sentence:
“John believes his son has passed the driving test.”	Step 1: Presupposition vs. Implicature
Questions:	• Presupposition: What is assumed to be true by the speaker. • Implicature: What is suggested or implied, but not directly stated.
Based on this sentence, would hearers think John’s son may have passed the test, or possibly he may not have? Why would they think so?	Step 2: Analysis of the Sentence The verb “believes” is a verb of propositional attitude. When we say “John believes X”, we are attributing a belief to John, but crucially, we are not asserting that X is true ourselves. What do hearers infer?
Please answer as a professor explaining to college students majoring in linguistics.	• Hearers would think: It is possible that John’s son has passed the driving test, and it is also possible that he has not.
	Why do hearers think so?
	• The sentence does not entail (does not guarantee) that John’s son has passed.

-
- The **truth of "his son has passed the driving test" is not presupposed.**
 - The use of "believes" marks the proposition as **John's mental state**, not as a fact.

Step 3: Linguistic Explanation

Intensional Contexts

- "Believes" creates an **intensional context**: the truth of the embedded clause is not asserted by the speaker.
- In intensional contexts, **reference is opaque**: we cannot substitute truth for belief.

Example for Comparison

- "John *knows* his son has passed the driving test."
→ Here, most would infer the son did pass, because "knows" presupposes the truth of the complement.
- "John *believes* his son has passed the driving test."
→ Here, no such presupposition exists.

Pragmatic Reasoning

- If the speaker simply wanted to report the fact, they would have said, "John's son **has passed** the driving test."
- By using "believes," the speaker signals **uncertainty** about the fact itself, foregrounding John's belief rather than the event.

Conclusion

Hearers would think that John's son **may have passed the test**, but it is also possible that **he may not have**. This is because the sentence attributes only a belief to John, not a fact about the world. The speaker remains non-committal regarding the actual outcome.

This revised prompt led the chatbot to immediately utilise *pragmatics* terminology ('presupposition') and linguistic concepts (e.g. 'propositional attitude verb', 'intensional context') to structure a much lengthier and rather technical answer. (Note: it also defined 'implicature' in Step 1 but did not use it in the analysis). The conclusion is the same as that of (Q2a), but it now stands on two reasons. The first is that 'believes' is a *propositional attitude verb* that denotes John's mental state rather than a fact. This is the same argument raised in (Q2a) ('[it] signals John's opinion or assumption...'), while, in (Q2b), an additional concept 'intensional context' was brought in to extend a theoretical ground that 'believes' is about a mental state. The second reason is that 'believes' does not presuppose the truth (about the driving-test

result), while, by contrast, ‘knows’ presupposes it. Both reasons sound valid.

However, like in (Q1), GPT-4.1 still did not implement the logic of the Q-principle in (Q2b). It used ‘knows’ simply to illustrate that ‘believes’ lacks the pragmatic capacity of presupposing. In fact, the LLM analysed the two verbs individually, rather than considering them in a stronger-weaker relation in the structure ‘John knows/believes (a proposition *p*)’. Absent in the LLM’s reasoning is the logical argument that the use of ‘believes’ implies John is not in a position to use ‘knows’ (the stronger alternate), because he does not know about the test result for sure. The reasoning in (Q2b) lies in the meaning analysis of ‘believes’ in contrast with ‘knows’ – i.e. believes’ introduces epistemic uncertainty that the speaker does not know for sure, while ‘knows’ entails the speaker’s commitment to the truth. It does not draw the strengths from pragmatic rules to bring the effect of Q-clausal implicature to light.

Test (Q3a) was included in our test battery because it can be readily resolved using Levinson’s Q-principle, particularly its simplified formulation “What is not said is not the case” (Huang 2014: 51). The principle may be applied as follows: the fact the quotation does *not* say anything about human stupidity implicates that human stupidity is *not* (something “I’m not sure about”). That means human stupidity is something I feel certain about.

Table 4. Tests (Q3a) and (Q3b): prompts and answers

(Q3a) “Two things are infinite: The universe and human stupidity. And I’m not sure about the universe.” Questions: For hearers of the quotation, what is the speaker’s position about human stupidity? How can the hearers get it if the speaker does not say it explicitly?	Let’s break down the quotation and the questions: 1. What is the speaker’s position about human stupidity? The speaker is expressing the opinion that human stupidity is limitless or infinite—even more so than the universe, whose infinity is actually uncertain. 2. How can hearers get this position? Hearers can understand the speaker’s position through several clues: • Comparison: The speaker compares human stupidity to the universe, which is commonly thought of as vast or infinite. By <u>putting them side by side</u> [underlines are ours], the speaker emphasizes the enormity of human stupidity. • Irony/Humor: The second sentence (“And I’m not sure about the universe.”) uses irony. The speaker questions the infinity of the universe—but not of human stupidity. This contrast highlights that the speaker is even more certain about the endlessness of human stupidity than about the universe itself.
---	--

	• Implication: The structure and wording imply this belief strongly through suggestion and contrast. In summary: Hearers infer the speaker’s position through the clever use of comparison and irony. The <u>retorical structure</u> [underlines are ours] leads the listener to understand that the speaker sees human stupidity as truly boundless.
(Q3b) Investigators’ follow-up question: Does Levinson’s Q-heuristic help to analyse this sentence?	Yes, Levinson’s Q-heuristic is very useful for analyzing this sentence! What is Levinson’s Q-heuristic? The Q-heuristic is often summed up as: “What isn’t said, isn’t.” Or more precisely, if a speaker doesn’t make a stronger statement, hearers infer that the stronger statement doesn’t hold. Applying the Q-heuristic to the Quotation Let’s break it down: 1. Explicit Statement: The speaker states that both the universe and human stupidity are infinite. 2. Qualification: The speaker immediately qualifies (casts doubt on) the infinity of the universe: “And I’m not sure about the universe.” 3. Q-heuristic Effect: By <i>not</i> doubting or qualifying the infinity of human stupidity, the speaker strongly suggests that human stupidity is definitely infinite—even more so than the universe. According to the Q-heuristic: If the speaker doesn’t mention doubt about human stupidity, the hearer infers that the speaker is <i>certain</i> about it. Conclusion Levinson’s Q-heuristic helps explain how hearers understand the speaker’s position: • Because the speaker expresses doubt only about the universe and not about human stupidity. • The Q-heuristic fills in what is left unsaid

The LLM arrived at the identical conclusion as predicted by the Q-principle for (Q3a) – i.e. the speaker is certain about human stupidity. The LLM elaborated on the *retorical structure* of the quotation as a key to delve into. GPT-4.1 observed that juxtaposing two things that were considered to be infinite – i) the universe and ii) human stupidity – while commenting on the first *only*, a contrast has been created.

This contrast allows the speaker to humorously and ironically express the opinion that the second is not something worth doubting.

The LLM continued its line of argument in its self-provided follow-up questions. One of the questions goes – “Can you provide more examples illustrating this logical implication?”. Asking the question allowed the chatbot to provide five examples illustrating the structure of discourse in (Q3a) – that is, “how implying certainty about one thing by expressing doubt about the other works”, articulated by the chatbot. Three of the illustrative examples are presented as follows:

- Two things are spicy: wasabi and black pepper. And *I’m not sure* about black pepper.
- Two things are reliable: my old car and my best friend. And *I’m not sure* about my old car.
- Two things are difficult: quantum physics and assembling IKEA furniture. And *I’m not sure* about quantum physics. [italics by the investigators]

The LLM exhibits its strength in extracting the rhetorical structure from the quotation, identifying the logical moves in the structure, and formulating new examples modelling it. While Levinson’s Q-principle is at work in (Q3a), the rhetorical structure identified by GPT-4.1 functions to highlight the effect. In the discourse of the quotation, what is unsaid is not something from nowhere, but is precisely one of two things claimed as infinite at the beginning of the discourse. The speaker’s commenting on *only* one of the two sounds abrupt and incomplete, forcing hearers to search out what is meant about the unsaid.

The investigators agree that the rhetorical feature identified is salient in the quotation (Q3a), and this invites readers to search for the speaker’s intended meaning. However, it is crucial to understand that the rhetorical structure itself produces no prediction on what the speaker thinks about what is *not* said. It is the Q-principle that is at work, engendering the prediction – that is, what is not said is *not*. The rhetorical structure highlights the effect of Q-implicature, rather than generating the implicature. The Q-principle, therefore, is always the underlying reason that explains why hearers get the implied meaning. Again, like in (Q1) and (Q2), there is no evidence that the chatbot has included Levinson’s Q-principle as a part of its reasoning.

We have seen that GPT-4.1 explained how implied meaning is signalled in (Q3a)

by focusing on its rhetorical structure. The investigators were interested in knowing whether the LLM is able to look into the quotation through the lens of Levinson’s theoretical framework. The investigators posed a follow-up question in (Q3b), directly asking GPT-4.1 whether Levinson’s Q-heuristic helps to analyse the sentence. GPT-4.1 answered by effectively drawing on Levinson’s Q-principle – more precisely, utilising the simplified version of Q-heuristic – to reach the same conclusion as drawn in (Q3a). That is, human stupidity is definitely infinite, even more so than the universe.

The LLM is therefore capable of conducting its own observations in context to draw a conclusion, as in (Q3a), as well as applying an established theoretical framework to arrive at the same conclusion when being so required (Q3b). The LLM performed both tasks with valid logical reasoning.

In summary, the LLM has recovered the implicated meaning in our battery of tests on Q-implicature. The chatbot ably identified the keywords or salient rhetorical structure that contribute to the implicated meaning, and introduced alternative expressions to compare. This allows the LLM to tease out the distinct meaning the keywords or expressions infer in discourses, leading to the right conclusions about the implicated meaning. However, the chatbot did not (spontaneously) apply pragmatic principles in its reasoning, although it can aptly apply them when being required to do so.

4.2 M-implicature

We found the LLM effectively recovered the meaning conveyed by the marked expressions in our battery of tests (M1 to M3) on M-implicatures, identifying useful unmarked counterparts in the examples and using them as references.

Test (M1) was designed based on Levinson’s (2000: 138) discussion on M- and I-heuristics. The unmarked expression ‘John *could* solve the problem’ is notably more concise compared to the marked form ‘John *had the ability to* solve the problem’. The former implies ‘(... and) he did’, while the latter signals otherwise – that is, ‘(... and) he didn’t’.

GPT-4.1 aptly drew the conclusion that hearers would infer that John did *not* solve the problem. The chatbot gave reasons based on the understanding that the semantic meaning of ‘had the ability to’ is about potential or capacity rather than

the actual action. The LLM identified two alternative expressions to compare with ‘had the ability to solve’: i) ‘solved the problem’, which states the completed action, and ii) ‘was able to solve’, which conventionally conveys that the action was done. GPT-4.1 is therefore effective in identifying most pertinent alternates in relation to the expression in question and comparing them with the latter, in a very similar way to contrasting the marked with the unmarked forms as Levinson’s M-principle functions. The LLM distinguished the phrase ‘had the ability to’ from the two alternates in terms of whether the action (‘solving the problem’) was performed, and this is in line with the prediction of Levinson’s M-principle that the marked form implicates meaning in the complement of the denotative meaning expressed by the unmarked form (Huang 2014: 62).

In addition, it is worth noting that one of the follow-up questions posed by the LLM introduces one more alternate (‘could’) for comparison. The question reads:

What subtle differences exist between “*could*” and “had the ability to”? [italics are ours]

Note ‘could (solve the problem)’ is the unmarked form discussed by Levinson (2000: 138) in contrast with ‘had the ability to’. We can now observe that GPT-4.1 has effectively located at least two relevant alternates for ‘had the ability to’ – i.e. ‘could’ and ‘was able to’ – and has compared the alternates (in addition to the straight statement ‘John solved the problem’) to determine their nuanced semantic meaning and pragmatic implicatures.

Moreover, the chatbot reasoned that ‘had the ability to’ typically indicates *possibility* rather than *actuality*, and can also be used to present a contrast between the two – that is, it was within John’s capacity to solve the problem, but (unfortunately) he did not. GPT-4.1 noted that this type of contrast *often* occurs, possibly making this claim based on its training data, which contains ‘had the ability to’ being frequently used for a contrast at the discourse level.

The LLM solved the problem in (M1) by pinpointing the specific meaning of ‘had the ability to’ in relation to its alternates and also in terms of function in a discourse. However, the chatbot did not address the concept of markedness, nor did it point out prolixity in ‘had the ability to’, although these are the key factors articulated in Levinson’s M-principle. Even though we revised the prompt by adding the line

of answering as a professor explaining to linguistics-major students, GPT-4.1 still did not use Levinson’s M-principle (nor Grice’s Manner maxim) to reason. Only when the investigators posed a straightforward follow-up question – i.e. ‘can you utilise Levinson’s heuristic/s or Grice’s maxim of Manner to analyse it?’ – Did the LLM apply both Levinson’s M-heuristic and Grice’s maxim of Manner to compare the sentence in question (M1) with the sentence ‘John solved the problem’, using the latter as the unmarked form as a reference to carry out its contrastive analysis.

Table 5. Test (M1): prompt and answer

Prompt	GPT-4.1’s answer (abridged by investigators)
(M1) “John had the ability to solve the problem.” Questions: Would hearers of the sentence infer John solved the problem or he did not? And Why?	Hearers of the sentence would generally infer that John did not actually solve the problem , but <i>could have</i> done so if he had tried or chosen to. Why? • The phrase “had the ability to” focuses on John’s <i>potential</i> or <i>capacity</i> to solve the problem, not on whether he actually did it. • If the speaker wanted to say that John <i>did</i> solve the problem, they would likely say “John solved the problem” or “John was able to solve the problem” (the latter often implies successful completion, especially in context). • In English, using “had the ability to” typically highlights the <i>possibility</i> rather than the <i>actuality</i>. • Often, such phrasing hints at a contrast: John <i>could</i> have solved it, but for some reason, he <i>did not</i>. In conclusion: Hearers would infer John did not solve the problem, because the sentence only comments on his potential or capability, not his actual action.

Test (M2a) contains an unusual expression ‘timekeeping apparatus’, which was designed to sound notably different from its unmarked counterpart (‘watch’). According to Levinson’s M-heuristic, hearers take what is said in an *abnormal* way as an indication of an *abnormal* situation (Huang 2014: 62). This predicts that hearers should notice the markedness in ‘timekeeping apparatus’ and consider it referring to something other than a watch, thinking the speaker purposely avoids the easy-to-use form (‘watch’).

The LLM concluded that hearers would not immediately assume ‘a timekeeping apparatus’ refers to a watch. According to GPT-4.1, the reason lies in a) its odd

wording and b) its ambiguity, that is, a range of timekeeping devices may potentially be a candidate. The chatbot further considered different frames of context in which hearers would think ‘a timekeeping apparatus’ refers to differently – for example, ‘a watch’ in a business meeting and ‘a (mobile)phone’ or ‘smartwatch’ in the frame of modern life.

We can observe that GPT-4.1 focuses on the analysis of lexical alternates and their referencing potential, also were taken into account were typical frames of context in which the expression can be used. Nevertheless, the chatbot did not apply pragmatic rules to solve the problem – for example, Levinson’s M-principle that predicts a marked form would cause hearers to expect something unconventional.

Table 6. Tests (M2a) and (M2b): prompts and answers

Prompt	GPT-4.1’s answer (abridged by investigators)
(M2a)	Not necessarily. Most hearers would not immediately assume that "a timekeeping apparatus" specifically refers to a watch.
“I carry a timekeeping apparatus to tell the time.”	Why?
Question:	• Unusual Wording: "Timekeeping apparatus" is a very formal, technical, or even old-fashioned way to describe something that tells the time. People almost never use this phrase in everyday speech. • Ambiguity: "Apparatus" could refer to any device that keeps time, including: ◦ A wristwatch ◦ A pocket watch ◦ A clock (even a small travel clock) ◦ A stopwatch ◦ A phone (since phones tell time) ◦ Any other device designed to keep time
Would hearers think this apparatus is a watch?	Listener Reaction
And why?	• Context Matters: If you are in a setting where watches are common (e.g., a business meeting), some might guess "watch". • Default Guess: Most modern hearers, if pressed to guess, might think of a phone or a smartwatch, since those are common time-telling devices today. • Literal Meaning: Without more context, the phrase is too vague to mean "watch" specifically.
	Conclusion
	Hearers would not automatically think of a watch. They would likely be unsure and might ask for clarification, or picture any device that can keep time.
(M2b)	Certainly! Let’s break down the sentence:

“I carry a timekeeping apparatus to tell the time.”

Question:

Would hearers think this apparatus is a watch?

And why?

Please answer as a professor explaining to college students majoring in linguistics.

“I carry a timekeeping apparatus to tell the time.”

1. Literal Meaning and Lexical Choice

The phrase “timekeeping apparatus” is a formally descriptive, almost technical way of referring to something that tells time. In everyday English, people overwhelmingly use more specific terms: *watch*, *clock*, or *phone*.

2. Pragmatics and Grice’s Maxims

In pragmatics, we often refer to **Grice’s Maxims**—principles that guide how people communicate effectively.

- **Maxim of Manner:** Be clear and avoid unnecessary obscurity.
- **Maxim of Quantity:** Provide just enough information—no more, no less.
- **Maxim of Relation:** Be relevant.

Using the term “**timekeeping apparatus**” violates the Maxim of Manner, [underlines in GPT-4.1’s answers are ours] because it’s unnecessarily vague and formal. A native speaker would typically say “watch” or “phone” or perhaps even “clock,” depending on the context.

3. Hearer’s Inference

Given the formality and ambiguity, a hearer will likely **not immediately assume “watch”**. Here’s why:

- **Ambiguity:** “Timekeeping apparatus” could refer to many things: a wristwatch, a pocket watch, a clock, a stopwatch, or a phone.
- **Social Usage:** In modern society, the most common portable timekeeping devices are watches and smartphones. If the speaker is talking about something they *carry* with them, these two are most likely.
- **Formality Triggers:** The unusual phrasing might make hearers suspect the speaker is being deliberately vague, technical, or perhaps even mysterious—hinting at something less conventional.

4. Default Assumptions

In the absence of further context, most hearers will narrow their guesses based on **prototypicality**: What do people most commonly carry to tell the time? For many, the answer is a **watch** or a **phone**. However, because “apparatus” sounds a bit old-fashioned or technical, it might push hearers slightly more toward something mechanical, like a watch, rather than a smartphone (which is usually called just a “phone”).

5. Conclusion

Would hearers think this apparatus is a watch?

- **Not with certainty.** They might consider a watch as a likely candidate, but they would also entertain other possibilities (like a phone or even something unusual).
-

-
- **The vagueness of “timekeeping apparatus” prevents hearers from confidently assuming it’s a watch.** The phrasing sounds odd and would likely prompt a follow-up question: “What do you mean by ‘timekeeping apparatus’? Do you mean a watch?”

Key takeaway for linguists:

When the speaker deviates from common terms, hearers may struggle to arrive at a single, confident interpretation.

The prompt in (M2b) has the additional line about acting as a professor to linguistics-major students, and was launched in a new thread. This enabled the chatbot to apply Grice’s Manner maxim, and arrived at a valid observation that ‘timekeeping apparatus’ is a violation of the said maxim. The chatbot followed the logic of the Gricean Cooperative Principle, predicting that hearers would feel uncertain about what the unusual term refers to, and suspect something less conventional because the speaker’s use of such an unusual term is probably deliberate.

However, the LLM moved on to develop a section on Default Assumptions, which is unnecessary, though. It discussed what hearers would consider based on prototypicality – i.e. ‘watch’ or ‘phone’ under various considerations – and this line of argument pertains to I-implicature in Levinson’s (2000) framework. Levinson (ibid: 39ff, 138; Huang 2014: 65) pointed out that M-implicature prevails over I-implicature. Given that ‘timekeeping apparatus’ is a marked expression, hearers would consider marked situations rather than prototypical ones, and therefore, the I-principle does not apply. This is the reason that the section on Default Assumptions is unnecessary, and it distracts the logic flow.

GPT-4.1 concluded in (M2b) that hearers might consider it a watch, although with uncertainty, and they would also need to seek clarification. The first half of the conclusion is in conflict with the prediction of Grice’s Manner maxim as well as Levinson’s M-principle – i.e. flouting the said maxim engenders the implicature that the speaker refers to something other than a watch.

To summarise, the LLM effectively applied the Gricean maxim to analyse the sentence in (M2b), and, accordingly, delivered a valid theoretical prediction that hearers would expect something unconventional. Nevertheless, its discussion on the Default Assumptions side-tracked the logic flow and is unnecessary. It failed to recognise that the M-principle is prioritised over the I-principle and proceeded to a conclusion that is inconsistent with the logic flow, and defies the predication of

pragmatic theory. This causes serious doubt on whether the LLM is able to logically and consistently apply abstract rules such as pragmatic principles in natural language processing.

Test (M3) is a classic example of markedness expressed by double negation. Double negation is normally taken as weaker than the direct positive form (Huang 2014: 63).

Table 7. Test (M3): prompt and answer

Prompt	GPT-4.1's answer (abridged by investigators)
(M3) "John is not unaware of the problem." Questions: For hearers of the sentence, does it sound stronger or weaker compared to "John is aware of the problem"? And why?	Let's break it down: 1. Comparing the Sentences • "John is aware of the problem." ◦ This is a direct, positive statement. ◦ It straightforwardly asserts that John knows about the problem. • "John is not unaware of the problem." ◦ This is an example of a litotes—an understatement using double negatives. ◦ The literal meaning is the same: John <i>is</i> aware. ◦ However, the phrasing is more cautious and indirect. 2. How It Sounds to Hearers • Strength of Assertion: ◦ "John is aware..." sounds stronger and more direct. ◦ "John is not unaware..." sounds weaker or less forceful. • Why? ◦ The double negative introduces a sense of hedging, suggesting that while John has <i>some</i> awareness, it might not be complete or fully engaged. ◦ It can imply "He knows about it to some extent," or "At least he isn't totally oblivious." ◦ It avoids the straightforwardness of "He is aware," making it feel less committed or less enthusiastic. 3. Implications and Pragmatics • The "not unaware" construction can sometimes be used to be polite, diplomatic, or to downplay the level of awareness. • It may also suggest that awareness is present, but maybe not as strong as if you simply said "He is aware."

Summary Table			
Statement	Strength	Tone	Implication
John is aware of the problem.	Strong	Direct	Firm awareness
John is not unaware of the problem.	Weaker	Indirect/Cautious	Some or qualified awareness
In conclusion: "John is not unaware of the problem" sounds weaker and more indirect than "John is aware of the problem," because it uses a double negative that often implies only partial or qualified awareness, rather than strong, direct awareness.			

The LLM concluded that the sentence with double negation sounds weaker than a direct positive statement, giving the reasons including indirectness, politeness, and the speaker's lower commitment. The reasons are relevant factors that motivate the use of double negatives in discourses. The chatbot considered double negation a type of hedging, which is reasonable. Nevertheless, like in (M1) and (M2), the chatbot failed to point out that double negation exhibits *markedness* compared to the direct positive form.

In summary, in response to our tests on M-implicature, GPT-4.1 did not identify prolix or unconventional expressions as *marked*, nor did it apply Levinsonian M-principle. It only utilised Grice's Maxim of Manner once when acting as a linguistic professor in (M2b), but did not point out what implicature was engendered as the maxim was flouted. It aptly retrieved relevant alternative expressions to tease out the distinct meaning of the marked expressions (though it does not label them as marked) to recover the implicated meaning.

4.3 I-implicature

The LLM is effective in solving our tests on I-implicature by retrieving prototypical properties in the relevant frame of discourse.

Test (I1) involves pragmatic enrichment of the conjunction 'and' (cf. Huang 2014: 58, 'conjunction buttressing'). GPT-4.1 associated the sentence with the relevant social frame and attributed 'that' with the function of signalling the sequence of the two events according to the order they occur in the sentence.

Test (I2) examines the stereotypical sense of 'drink' in the frame of traffic offence

and traffic-law enforcement (cf. *ibid*: 59, ‘frame-based inference’). The LLM had no problem locating the context of the discourse and defined the meaning of ‘drink’ accordingly in this frame.

Test (I3) probes the chatbot’s ability to specify what ‘safe’ technically entails in this frame of context (i.e. ‘frame-based inference’) and also identify what entities are referred to as safe. The LLM aptly pointed out that ‘safe’ refers to both users’ health and the utensil’s integrity, without mentioning the microwave, though, which practically is not an article of immediate safety concern here. In addition, the chatbot listed a variety of common hazards that the utensil was designed to prevent, demonstrating LLM’s capacity to locate tangible knowledge concerning technicality, health and safety, from which contextually enriched typical meanings of ‘safe’ were derived.

Table 8. Tests (I1) to (I3): prompts and answers

Prompt	GPT-4.1’s answer (abridged by investigators)
(I1) “They got married and had a child.” Questions: For hearers of the sentence, what is the order of the two events? And why?	For hearers of the sentence “They got married and had a child,” the typical interpretation is that the events happened in the order presented: 1. They got married 2. They had a child This is because in English, when two events are connected with “and,” the default assumption is that the events are presented chronologically, unless there is information suggesting otherwise.
(I2) “Drink and drive” Questions: For hearers of the phrase, what does drink refer to? And why?	For the phrase “drink and drive” , the word “drink” typically refers to drinking alcohol , not just any kind of drink. Explanation • The phrase <i>“drink and drive”</i> is a set expression in English that means to operate a vehicle after consuming alcoholic beverages. • For hearers, <i>drink</i> in this context does not mean drinking any liquid (e.g., water, juice), but specifically refers to alcoholic drinks. • This is due to cultural and legal associations: “drinking and driving” laws, warning signs, and public safety messages always use “drink” to mean “alcohol”.
(I3) “The utensil is microwave-oven safe” Question:	For hearers of the sentence, the word ‘safe’ refers to the utensil’s ability to be used in a microwave oven without causing harm or damage . Specifically, “safe” here means:

For hearers of the sentence, what does ‘safe’ refer to?

- **The utensil will not be damaged or destroyed by being used in a microwave oven** (e.g., it won’t melt, warp, or catch fire).
- **Using the utensil in a microwave oven will not cause harm to the user** (e.g., it won’t release toxic substances, emit sparks, or otherwise create a hazard).

In summary, ‘safe’ refers to the utensil’s compatibility with microwave ovens in a way that **avoids both damage to the utensil and risk to the user.**

5. Discussion

We can now answer our research questions outlined in the Introduction section, and envisage subsequent studies.

- a) Does GPT-4.1 recover meaning conveyed by Levinsonian generalised conversational implicatures (GCIs)?

GPT-4.1 effectively recovered the implicated meaning in our battery of tests, drawing valid conclusions in the first query of each test (without being required to act out a specific role or to apply pragmatic principles). The chatbot’s conclusions articulated the intended meaning as predicted by Levinson’s three principles for GCIs.

The results of our study lend support to recent studies (e.g. Bojić, Kovačević, and Čabarkapa 2025; Yue et al. 2024) that ChatGPT-4 has achieved high competence in comprehending implied meaning in discourse, even comparable to human competence in various ways.

- b) With what mechanism does GPT-4.1 recover the meaning conveyed by the GCIs?

Although we cannot gain direct access to the mechanism adopted by GPT-4.1 to recover the meaning in GCIs, we have investigated the chatbot’s messages in terms of the linguistic materials it brings in and the logic flow it develops, gaining insights into LLM’s approach to retrieve implicated meaning.

The LLM solves our test problems by primarily retrieving the meaning of the keywords and expressions in context, including their semantic meanings and typical functions in a discourse. For example, GPT-4.1 examined ‘tried’ (Q1) and ‘had the ability’ (M1a) in context and pointed out that the former typically means ‘tried without success’ while the latter ‘could have been done but did not’. GPT-4.1 exhibited its strengths in locating the meaning of keywords in typical contexts, and its resourcefulness probably derives from its training with an enormous amount of linguistic data in meaningful contexts. For instance, GPT-4.1 states that ‘tried to X’ usually [underlines are ours] implies that X may not have happened in (Q1). This claim may have a statistical basis.

In addition, the LLM’s analysis of meaning is heavily *contrastive*, particularly for retrieving Q- and M-implicatures, because it constantly identifies alternative words and expressions in relation to the keywords in question in order to tease out the distinct meaning of the keywords in question. The set of alternates typically consists of the stronger and the weaker items in a scalar continuum (from which Q-implicatures arise), and marked and unmarked items (that engender M-implicatures), though GPT-4.1 did not label them as the stronger/weaker or the un/marked. The chatbot’s contrastive analysis of lexical and pragmatic meanings in relevant frames of discourse led to valid conclusions about the implicated meaning.

In summary, we found GPT-4.1’s reasoning in our tests primarily evidence-based and anchored in contrastive analysis of lexical meaning in context, rather than abstractly rule-based.

- c) Does GPT-4.1 apply Levinson’s Q-, M- and I-principles to recover meaning conveyed by the GCIs?

There is little evidence that GPT-4.1 solves problems by directly applying pragmatic principles. Although its contrastive analysis of the alternative words and expressions looks somewhat like the logic of the Q- or the M-principle, the LLM analysed each alternate individually without establishing the relation between the strong and the weak alternate, nor that between the marked and the unmarked. It did not label the alternates as the strong or the weak by identifying the scale or order, nor as marked versus unmarked by commenting on proximity. It did not use the logic of the Q- or the M-principle, nor predicate Q- or M-implicatures accordingly.

GPT-4.1 at the present stage is not an able rule-applier. One of our tests (M2b) strongly suggests that the LLM can run into an inconsistent flow of logic. GPT-4.1 was unable to work out the order of priority as predicted by pragmatic theories – e.g. to determine that the M-principle precedes the I-principle – in order to solve the problem. To what extent LLMs are able to handle abstract pragmatic rules in processing natural language deserves attention in future research.

We would argue that GPT-4.1 solved the problems in our tests mainly by effective meaning analysis in multiple-layered contexts, including typical frames of discourse. The language model primarily worked *locally*, examining the keywords and the salient rhetorical structure, identifying expected alternatives, and matching typical contexts with the alternatives. It does not seem to work with abstract pragmatic rules to solve the problem, at least not in the first place. For example, it can rely on contextualised meaning analysis of ‘tried to learn’ and ‘learned’ to resolve our test (Q1), without the need to apply an abstract pragmatic rule (e.g. the Q-principle) to reach the same conclusion. The strength of expectedness of pragmatic alternatives in context was observed as a robust predictor for humans’ interpretation of Q-implicature (Hu et al. 2013). This may be a factor contributing to the LLM’s successful processing of implicated meaning in context, since LLMs are trained models that can effectively compute the relations between expected alternatives with statistical backing.

6. Conclusion

This study assessed the pragmatic abilities of GPT-4.1 in recovering meanings conveyed by Levinsonian generalised conversational implicatures (GCIs), focusing on its performance in real-world user interactions. Our findings show that GPT-4.1 reliably identifies and contrasts relevant lexical alternatives in context, leading to plausible interpretations of implicated meaning. The chatbot’s analysis is primarily grounded in contextualised contrastive meaning analysis, drawing on its extensive exposure to language data to compute subtle distinctions between alternatives, rather than articulating or applying abstract pragmatic principles such as the Q-, the M-, or the I-principle.

Despite its effectiveness in deriving implicated meaning, GPT-4.1 demonstrates little evidence of explicit rule-based reasoning or systematic application of pragmatic

principles. Instead, its success appears to stem from its capacity to evaluate linguistic alternatives within layered contexts, predicting meaning based on statistical regularities and expectedness in discourse. Occasional inconsistencies in logical priority, particularly regarding the priority of pragmatic principles, further suggest limitations in the model's ability to handle abstract theoretical rules.

Our case analysis raises questions about the depth of its pragmatic understanding and the mechanisms driving its performance. Future research can further explore the extent to which large language models can internalise and apply abstract pragmatic rules.

References

- Acton, Eric K. and Matthew Hunt. 2025. Pragmatic alternatives and social meaning. *Journal of Pragmatics* 243: 42-52. doi: 10.1016/j.pragma.2025.05.001.
- Bojić, Ljubiša, Predrag Kovačević, and Milan Čabarkapa. 2025. Does GPT-4 surpass human performance in linguistic pragmatics? *Humanities & Social Sciences Communications* 12(1): 794. doi: doi.org/10.1057/s41599-025-04912-x.
- Cho, Hyesun, Sunwoo Park, Sanghoun Song, and Eunjin Oh. 2025. Investigating ChatGPT's phonology problem-solving abilities through reasoning with varying custom instructions. *Linguistic Research* 42(1): 53-93. doi: 10.17250/khisli.42.1.202503.003.
- Cho, Ye-eun Kim and Seong Mook. 2024. Pragmatic inference of scalar implicature by LLMs. Presented at *62nd Annual Meeting of the Association-for-Computational-Linguistics (ACL) / Student Research Workshop (SRW)*, 10-20. Bangkok, Thailand, August 11-16.
- Cong, Yan. 2024. Manner implicatures in large language models. *Scientific Reports (Nature Publisher Group)* 14(1): 29113. doi: doi.org/10.1038/s41598-024-80571-3.
- Grice, Paul. 1975. Logic and conversation. In Peter Cole and Jerry Morgan (eds.), *Syntax and Semantics, vol. 3: Speech acts*, 41-58. New York: Academic Press.
- Hu, Jennifer, Roger Levy, Judith Degen, and Sebastian Schuster. 2023. Expectations over unspoken alternatives predict pragmatic inferences. *Transactions of the Association for Computational Linguistics* 11: 885-901. doi: 10.1162/tacl_a_00579.
- Hu, Jennifer, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.
- Hu, Jingwen, Yuanxing Liu, Longxuan Ma, Wei-Nan Zhang, and Ting Liu. 2025. Pragmatic

- inference and mapping for conversational implicature. *International Journal of Machine Learning and Cybernetics*. doi: 10.1007/s13042-025-02741-7.
- Huang, Yan. 2014. *Pragmatics (2nd Edition)*. Oxford: Oxford University Press.
- Levinson, Stephen C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.
- Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.
- Molares-Cardoso, Julinda, Vicente Badenes-Plá, and Carmen Maiz-Bar. 2024. Human creativity vs. artificial creativity: Comparative study between university students and chatbots in idea generation. *Palabra Clave* 27(1): 1-27. doi: 10.5294/pacla.2024.27.1.10.
- Qiu, Zhuang, Xufeng Duan, and Zhenguang Cai. 2023. Does ChatGPT resemble humans in processing implicatures? *Proceedings of the 4th natural logic meets machine learning workshop (NALOMA23)*, 25-34. Nancy, France: Association for Computational Linguistics.
- Strachan, James W. A., Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing Theory of Mind in large language models and humans. *Nature Human Behaviour* 8(7): 1285–1295. doi: 10.1038/s41562-024-01882-z.
- Wu, Shengguang, Shusheng Yang, Zhenglun Chen, and Qi Su. 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 (EMNLP 2024)*, 22583-22599. Miami, FL: Association for Computational Linguistics.
- Yue, Shisen, Siyuan Song, Xinyuan Cheng, and Hai Hu. 2024. Do large language models understand conversational implicature – A case study with a Chinese sitcom. arxiv.org/abs/2404.19509.
- Zhao, Yanjie, Xinyi Hou, Shenao Wang, and Haoyu Wang. 2025. LLM app store analysis: A vision and roadmap. *ACM Transactions on Software Engineering and Methodology* 34(5): 1-25. doi: 10.1145/3708530.

Vincent Xian Wang

Associate Professor

Department of English

University of Macau

Avenida da Universidade, Taipa,

Macao, China

E-mail: vxwang@um.edu.mo

Lily Lim

Associate Professor

MPU-Bell Centre of English

Macao Polytechnic University

R. de Luís Gonzaga Gomes,

Macao, China

E-mail: llim@mpu.edu.mo

Received: 2025. 07. 28.

Revised: 2025. 08. 15.

Accepted: 2025. 08. 15.