

Implicature in Interaction: Understanding Implicature Improves Alignment in Human–LLM Interaction*

Asutosh Hota Jussi P. P. Jokinen

asutosh.jyu.hota@jyu.fi jussi.p.p.jokinen@jyu.fi

University of Jyväskylä

Abstract

The rapid advancement of Large Language Models (LLMs) is positioning language at the core of human–computer interaction (HCI). We argue that advancing HCI requires attention to the linguistic foundations of interaction, particularly implicature (meaning conveyed beyond explicit statements through shared context) which is essential for human–AI (HAI) alignment. This study examines LLMs’ ability to infer user intent embedded in context-driven prompts and whether understanding implicature improves response generation. Results show that larger models approximate human interpretations more closely, while smaller models struggle with implicature inference. Furthermore, implicature-based prompts significantly enhance the perceived relevance and quality of responses across models, with notable gains in smaller models. Overall, 67.6% of participants preferred responses with implicature-embedded prompts to literal ones, highlighting a clear preference for contextually nuanced communication. Our work contributes to understanding how linguistic theory can be used to address the alignment problem by making HAI interaction more natural and contextually grounded.

Keywords— Implicature Understanding, Generative AI, Human–Computer Interaction, Large Language Models

1 Introduction

Human–computer interaction (HCI) is entering a linguistic turn. As Large Language Models (LLMs) are embedded into digital assistants, chatbots, productivity tools, and social robots, language has become the dominant interface through which users engage with computational systems [24]. Unlike earlier eras of HCI that emphasized graphical or direct manipulation interfaces [33], the future of interaction will be shaped by how well systems can understand and respond to the subtleties of human communication [39]. While LLMs have made notable progress in language generation, many still interpret statements literally and fail to capture pragmatic subtleties [9, 13, 27, 36]. This makes the study of linguistics (including syntax, semantics, pragmatics, and discourse) central to the future of HCI research.

A persistent challenge lies in interpreting meaning beyond the literal surface of words. In everyday conversation, people routinely rely on *conversational implicatures*, utterances whose intended meaning must be inferred from context and shared assumptions. For example, replying “I have a lot of work to do” to an invitation implies refusal without stating it explicitly. These indirect cues are essential for expressing refusals, requests, attitudes, and social nuance. When interaction systems misinterpret them, users experience stilted dialogue, reduced trust, and a diminished sense of intelligence in the system [32, 40]. As a result, understanding and handling implicatures is not a marginal concern but a core requirement for designing natural, collaborative, and trustworthy AI interactions.

While LLMs have achieved impressive fluency, their pragmatic competence remains limited. Some state-of-the-art models approach human-like performance on implicature benchmarks [2, 42], but smaller or open-source models often default to literal interpretations or rigid heuristics [6, 27]. Moreover, the reasoning processes behind LLM outputs are opaque [21], making it difficult to ensure robustness or improve failures. From an HCI perspective, this creates a pressing need to integrate linguistic theory into interaction research, both to explain these limitations and to design interventions that improve user experience.

In this paper, we take implicature as a worked example of why linguistic insights matter for HCI. We identify three classes of implicature relevant to human–LLM interaction, *information-seeking*, *direction-seeking*, and *expressive* and use them to systematically evaluate how LLMs align with human interpretations. Through three experiments,

*Pre-print article, Manuscript under review

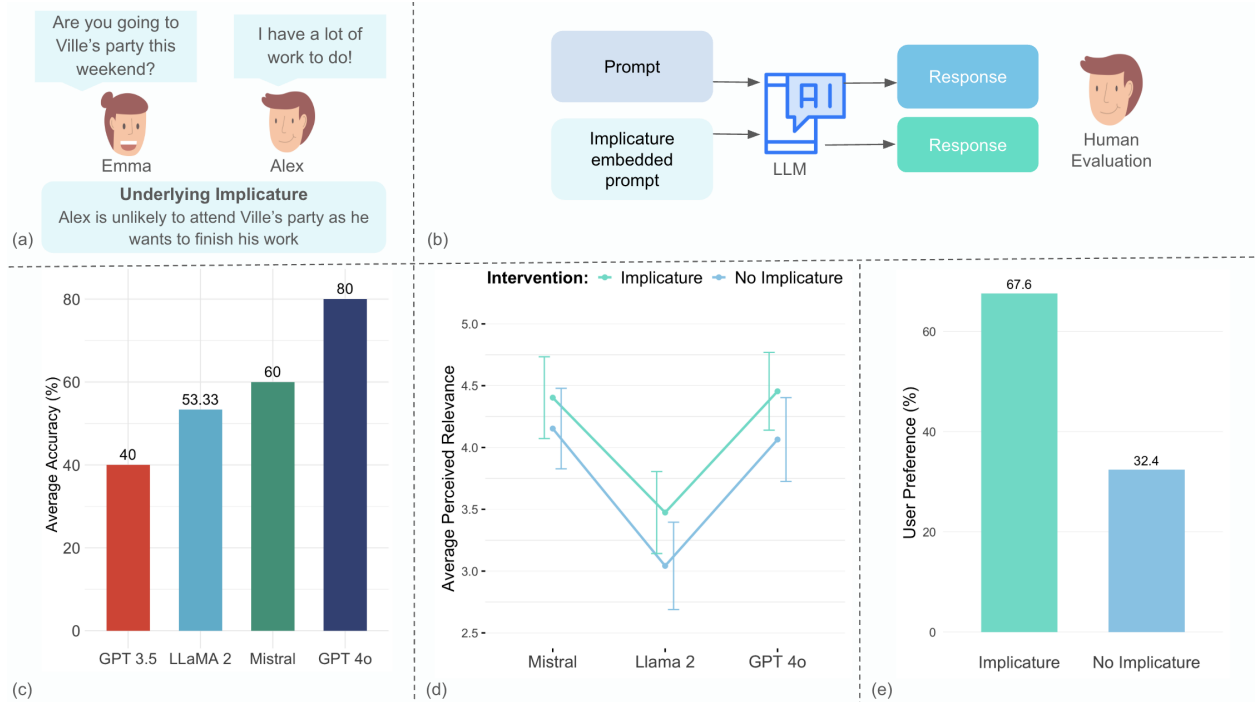


Figure 1: Impicature in interaction. (a) Humans are adept at inferring implicatures during discussion. (b) Our experiment is designed to evaluate how large language models (LLMs) handle implicature-embedded prompts by asking humans to rate the responses based on their perceived relevance, quality, and preference., (c) In Experiment 1, larger models demonstrate a better understanding of conversational implicatures, approaching the human baseline., (d) With implicature embedded prompts, perceived relevance increases across all models however there are relevance differences between models.,(e) In experiment 3, we confirm that users consistently favor responses that are sensitive to implicatures over those that are not.

we assess whether models can infer implied intent and whether implicature-sensitive prompting improves response quality and user satisfaction. Figure 1 illustrates our design and findings: larger models demonstrate closer alignment with human baselines, while implicature-informed responses are consistently rated as more relevant, higher quality, and strongly preferred by users.

Our contributions are threefold:

1. We introduce a taxonomy of implicature classes grounded in pragmatic theory and tailored to HCI, bridging linguistic analysis with interaction research.
2. We provide empirical evidence from three user studies showing how implicature-sensitive prompting improves alignment with user intent and enhances perceived interaction quality.
3. We advance a broader argument that linguistic understanding is foundational for the future of HCI, and that studying phenomena such as implicature provides a pathway to more trustworthy, context-sensitive, and human-like interaction systems.

2 Background

2.1 Implicatures in Pragmatics and Linguistics

Implicatures are central to everyday communication, where speakers often convey more than they explicitly say. For instance, if someone replies “It’s getting pretty late” after being asked to stay longer, the underlying implication is a polite refusal or a desire to leave, even though this is not stated outright. Such conversational inferences depend on context, shared knowledge, and cooperative principles.

Theme	Key Learnings	Challenges
How implicatures are generated	Gricean accounts compute implicatures post-semantically [25], while grammatical accounts generate them locally in syntax [3, 4].	Clearer criteria are needed to avoid predicting implicatures that listeners do not actually infer.
Embedded implicatures	Implicatures can arise inside quantifiers or conditionals (e.g., “Exactly one student solved some of the problems”) [4, 28].	Psycholinguistic evidence is limited on whether people compute such inferences in real time.
Plural/quantity implicatures	Plural phrases often imply “more than one” but this varies with logical context [35].	Formal models must better capture context-sensitive variability.
Context dependence of meaning	Even basic meanings (e.g., “John’s book”) depend heavily on context [25, 26].	Tools are needed to separate sentence meaning from pragmatic enrichment.
Real-time processing	Existing theories largely ignore processing effort or timing [4, 25].	More cognitive studies are needed to test when and how implicatures are formed during conversation.

Table 1: Summary of themes, learnings, and open challenges in implicature research.

A central case is the scalar implicature (SI) [5], where weaker expressions like “some” imply the negation of stronger alternatives like “all.” In the Gricean view [25], these inferences are computed *after sentence meaning is established*, through pragmatic reasoning guided by conversational maxims. In contrast, grammatical accounts [3, 4] treat implicatures as part of the compositional semantics, allowing them to arise locally during sentence formation. This explains certain “embedded” cases (e.g., “Exactly one student solved some of the problems”), where pragmatic accounts predict weakly but grammatical models predict systematically.

Beyond scalar cases, plural phrases can also trigger implicatures: for example, “Jack solved difficult problems” is often interpreted as “more than one,” though this depends on context, especially under negation [35]. Philosophical perspectives further blur the boundary between semantics and pragmatics: Recanati [25, 26] argues that even literal meanings are heavily context-dependent, making the distinction between what is “said” and what is “implicated” porous.

A persistent challenge is that the two main accounts fail in opposite directions. Grammatical models often *overgenerate*, predicting implicatures even in cases where listeners do not draw them (e.g., assuming “not all” is inferred in every embedded context). Pragmatic models, by contrast, tend to *undergenerate*, failing to predict implicatures in complex structures such as negation, conditionals, or other downward-entailing environments. For instance, in “It’s not the case that some students passed,” pragmatic reasoning may fail to yield the inference “not all,” even though some listeners do make it. These tensions illustrate why implicature remains a challenging test case: no single approach fully captures the range of human intuitions, especially in complex linguistic structures. Table 1 summarizes key themes and challenges. In the next subsection, we turn to how these insights are being explored and applied within human–computer interaction.

2.2 Implicatures in Human-Computer Interaction (HCI)

With the growing use of conversational agents, chatbots, and virtual assistants, understanding implicature is essential for designing systems that feel human-like and contextually aware. Pragmatic reasoning allows AI to infer speaker intent from indirect language, making implicature handling a benchmark for naturalistic interaction and a core requirement for user trust.

Recent efforts have created annotated corpora for evaluating implicature understanding. George and Mamidi [8] collected implicature-rich dialogues from movies, while the PUB benchmark [1] and its extensions [27, 36] test LLMs’ ability to resolve pragmatically implied meanings. These datasets highlight that without targeted tuning, even advanced models such as RoBERTa perform poorly. By contrast, fine-tuned systems like GPT-4 can approach near-human performance on some tasks [27]. Yet, these benchmarks are often narrow in domain coverage and costly to annotate,

Theme	Key Learnings	Challenges
Annotated datasets for implicature	Datasets like PUB, CIRCA, and those from Ruis et al. [1, 8, 27] help model pragmatic meaning beyond literal content.	High annotation cost and limited domain diversity limit generalization.
Model performance and explanation	Models like GPT-4 approach human-level accuracy with tailored tuning [10, 27].	Many models struggle with explaining implicatures logically [42].
Collaborative and cooperative implicature	AI that uses implicature can enhance perceived humanness and coordination [20].	Requires shared context and alignment on conversational goals.
Implicitness in UI and system behavior	Implicit cues span attention, intent, and utterance meaning [31].	Disambiguating implicit interaction types is critical for design consistency.
User perception and agent framing	Implicature interpretation changes based on whether users see AI as autonomous or tool-like [22].	Systems should signal autonomy and transparency to guide expectations.
Cognitive architecture for intent modeling	Combining LLMs with symbolic models improves intent recognition [14].	Integration is computationally expensive and design-intensive.
Visual and multimodal implicature	Layout and gesture carry implicature-like inferences [17, 23].	Multimodal systems must consider indirect meanings from visual design.
Maxim adherence in Turing-style evaluation	Violating Gricean norms reveals artificiality in chatbots [29].	Adherence to conversational maxims is crucial for believable interaction.

Table 2: Themes, Learnings, and Design Challenges of Implicature in HCI

limiting their generalizability.

Despite improvements, challenges remain. Yue et al. [42] show that while LLMs can often select the correct implicature-driven answer, they struggle to provide plausible explanations for their choices. Similarly, Cong [7] reports persistent failures on *manner implicatures*—inferences drawn from how something is said rather than what is said. These findings suggest that models may capture surface-level regularities without robust underlying pragmatic competence, raising concerns for applications where transparency and accountability are important.

Beyond comprehension, systems can benefit from using implicatures themselves. In the cooperative game *Hanabi*, Liang et al. [20] show that agents employing implicature-like strategies are often mistaken for humans, indicating improved coordination and social believability. However, such benefits depend on shared goals and assumptions, which are difficult to formalize and may not hold in open-ended user interactions.

Implicitness in HCI extends beyond language. UI elements, visual cues, and gestures can signal user intent, sometimes functioning as implicature-like cues. Kehler [17] and Oberlander [23] note that layout and multimodal design can carry meaning beyond literal text, and misalignment across modalities can create confusion. User expectations also shape interpretation: Nishihata et al. [22] find that indirect cues are judged differently depending on whether an agent is framed as autonomous or tool-like. These results underscore the importance of managing framing and transparency in system design.

Another line of work combines LLMs with symbolic or cognitive models to improve intent recognition. Iida et al. [14] show that augmenting neural models with structured mental models of others can enhance pragmatic reasoning and improve dialogue coherence. Such hybrid systems highlight a promising direction for more robust intent modeling, though the integration is often computationally costly and design-intensive, raising questions about scalability in practical HCI applications. Furthermore, research on Turing-style evaluations emphasizes the importance of conversational maxims. Saygin and Cicekli [29] found that when chatbots violate Gricean norms such as being overly verbose or irrelevant users quickly perceive them as artificial. This suggests that pragmatic alignment is not merely an enhancement but a baseline requirement for believable interaction. Adhering to conversational norms is therefore crucial for systems intended to pass as naturalistic or human-like in dialogue. Table 2 summarizes key findings and open questions of implicature study in HCI research.

2.3 Background Summary

Taken together, prior work shows that implicature is not a marginal phenomenon but a central challenge for both linguistic theory and interactive system design. In linguistics and philosophy, debates around scalar, embedded, and plural implicatures highlight the tension between grammatical accounts, which risk overgeneration, and pragmatic accounts, which tend to undergenerate in complex environments. Psycholinguistic evidence further suggests that we still know relatively little about how implicatures are processed in real time.

Within HCI, a complementary body of research underscores the practical consequences of implicature for interaction. Annotated datasets and benchmarks have enabled systematic evaluation, but remain narrow in coverage and costly to build. Model performance continues to improve, yet gaps persist in explanation quality and in handling manner implicatures. Collaborative studies show that implicature use can increase perceived humanness and coordination, while work on multimodal cues, interface design, and agent framing highlights the diversity of implicit signals beyond text alone. Cognitive approaches that integrate symbolic reasoning with LLMs offer potential for stronger intent modeling, though at high computational cost. Finally, evaluations of conversational agents confirm that adherence to Gricean maxims is a baseline requirement for believable and satisfying dialogue.

Across these strands, two theoretical anchors stand out. First, Grice’s cooperative principle and conversational maxims [9, 10] explain how implicatures arise from inferences about relevance, clarity, quantity, and truthfulness. Second, Searle’s speech act taxonomy [30] provides categories for how utterances function (assertives, directives, and expressives). These perspectives converge on a pragmatic insight highly relevant for HCI: users routinely rely on implicatures to seek information, request action, or convey attitude. As language models become more deeply embedded in everyday contexts, robust implicature handling is essential for naturalistic, trustworthy, and user-aligned AI systems. Building on this foundation, our study operationalizes three implicature classes—*information-seeking*, *direction-seeking*, and *expressive*—which are theoretically grounded in Grice’s maxims and Searle’s speech act categories, and empirically validated in 3 experiments.

3 Method

This study comprises three experiments designed to evaluate how well different LLMs understand and respond to conversational implicatures. The experiments assess model interpretation accuracy, user perception of response quality, and preference for implicature-aware responses.

3.1 Data Generation

To evaluate implicature comprehension systematically, we designed a taxonomy of three implicature classes—*information-seeking*, *direction-seeking*, and *expressive*. These categories are grounded in two complementary traditions of pragmatic theory. From Grice [9, 10], they inherit the idea that implicatures arise from the cooperative principle and maxims of conversation: Relation (relevance), Manner, Quantity, and Quality. From Searle [30], they align with the broader taxonomy of speech acts: assertives, directives, and expressives. Our taxonomy can therefore be seen as a pragmatic re-organization of these theoretical foundations into interactionally salient categories that map directly onto common patterns in human–computer dialogue.

Information-seeking implicatures typically rely on the maxim of Relation, where an apparently indirect statement carries an implicit request for knowledge, consistent with Searle’s category of assertives. Direction-seeking implicatures correspond to indirect directives: they often appear as statements of uncertainty but implicitly request instructions, fitting both Searle’s directives and Grice’s account of maxim-flouting. Expressive implicatures, finally, align with Searle’s expressives and Grice’s examples of Quantity/Quality violations, such as irony, tautology, or evaluative emphasis.

By grounding these categories in well-established pragmatic theory while tailoring them to interactional needs, our taxonomy provides a principled basis for evaluating how LLMs interpret and respond to user intent. In Experiment 1, we validated this taxonomy by constructing 30 prompts (10 per class) and testing whether human participants’ interpretations aligned with these categories, thereby establishing both theoretical and empirical grounding for subsequent experiments.

Implicature Class	Pragmatic Grounding	Conversational Features
Information-Seeking	Grice: maxim of Relation (sometimes Quantity). Searle: assertives/questions.	Indirect questions that imply a need for clarification or planning (e.g., “Will it rain tomorrow?” implies planning for the day).
Direction-Seeking	Grice: flouting of Relation or Manner. Searle: directives (indirect speech acts).	Declarative statements that function as veiled requests for guidance or help (e.g., “I don’t know how to start this assignment”).
Expressive	Grice: flouting of Quantity or Quality (figurative, evaluative). Searle: expressives.	Utterances that convey emotions or attitudes rather than facts (e.g., “That was surprisingly fun” signals a positive evaluation).

Table 3: Mapping of Implicature Classes to Gricean Maxims [9, 10] and Searle’s Speech Acts [30]

3.2 Participants

A total of 180 participants were recruited across three experiments using online platform Prolific. All participants were native English speakers, provided informed consent, and were compensated at fair market rates. Attention checks were implemented to ensure response validity. For Experiment 1, 60 participants were recruited, and 54 were included in the final analysis (32 female, 18 male; mean age = 35.8 years, SD = 10.4). For Experiment 2, 90 participants were recruited, and 84 approved participants were included in the final analysis (39 female, 42 male, 3 prefer not to say; mean age = 35.6 years, SD = 13.2). For Experiment 3, 30 participants were recruited, and 29 were included in the final analysis (18 female, 11 male; mean age = 34.6 years, SD = 10.7).

3.3 Models Evaluated

We evaluated six state-of-the-art LLMs that represent varying scales and architectures. These include **GPT-4o** and **GPT-4** from OpenAI, which are among the most advanced models with high capacity for nuanced language understanding. **GPT-35-turbo** (Azure designation for GPT-3.5), also from OpenAI, serves as a representative of mid-sized models with strong general capabilities but reduced context sensitivity. **Llama-2-7b-chat** from Meta and **Mistral-7B-Instruct-v01** were selected to examine performance in smaller, open-access models. Finally, we included **phi-3-small-8k-instruct** from Microsoft to evaluate the latest small language model (SLM) family optimized for efficiency. All models were accessed via APIs from Azure Cloud and were prompted in a zero-shot configuration using standardized prompt templates. All other parameters (e.g., temperature = 0.5 and maximum token size = 1000) were kept constant across models. Since **GPT-4o** consistently outperformed **GPT-4** in preliminary comparisons, we used GPT-4o exclusively in Experiments 2 and 3 to avoid redundancy and reduce computational costs.

3.4 Stimuli and Prompt Design

Prompts were designed to simulate realistic, conversationally embedded implicatures and included both implicit and literal (non-implicature) variants. These prompts were divided into three categories: information-seeking (e.g., requests for facts or clarifications), direction-seeking (e.g., asking for suggestions or actions), and expressive (e.g., conveying emotions or attitudes). Thirty unique prompts were used across experiments, with equal representation across the three types. Each prompt was reviewed to ensure clarity, plausibility, and suitability for both human participants and models.

To standardize the interaction context and minimize ambiguity, we prefixed each prompt for the implicature-embedded condition with a unified system message, as shown in Listing 2. This system prompt instructed the LLM to recognize the intended communicative function (information seeking, direction seeking, or expressive) and tailor its response accordingly. Each class condition further clarified the user’s likely intent for that class of prompt. For literal prompts, no system prompt or class context was provided, ensuring that only the direct prompt request was

processed by the model. This design ensured consistent model behavior, transparent experimental manipulation, and reproducibility across all trials.

3.5 Experiment 1: Implicature Interpretation

For this experiment, we first crafted a set of 30 example prompts, with 10 representing each implicature class, to validate whether the proposed taxonomy aligns with human interpretation. These prompts were also reused in Experiments 2 and 3 to ensure consistency across the study. Human participants and models were then asked to interpret the implied meaning of each prompt. Human responses were aggregated to form a gold standard interpretation set, against which model responses were compared. This design allowed us to quantify the extent to which each model aligned with human interpretation and to evaluate their performance relative to the human baseline.

3.6 Experiment 2: Perceived Relevance and Quality

Participants were shown model-generated responses to both implicature and literal prompts and asked to rate them on two dimensions: perceived relevance and overall quality. Ratings were recorded on a 5-point Likert scale. To ensure consistency, the same 30 prompts developed in Experiment 1 (10 per implicature class) were reused. Each participant was shown a mix of randomly assigned samples of implicature-embedded prompts and literal prompts, with responses drawn from various models. The order of presentation was randomized to mitigate position bias. This setup enabled a comparative evaluation of how implicature guidance influenced response reception across models.

As illustrated in Figure 6, participants first reviewed detailed instructions explaining their task: for each prompt-response pair, they were to evaluate the response on two dimensions (relevance and quality) using a 5-point Likert scale. Figure 7 shows an example trial from the experiment, where participants read a user prompt, assessed the LLM’s response, and then rated the relevance (how well the response aligned with the intent of the prompt) and quality (clarity, coherence, and usefulness of the answer). Attention-check trials were interspersed to ensure data quality. This setup allowed us to systematically compare perceived response quality and relevance across models and prompt types.

3.7 Experiment 3: Preference Task

Experiment 3 employed a two-alternative forced-choice design to directly assess user preference for LLM-generated responses. As shown in Figure 8, participants were instructed to read a prompt and two candidate responses (A and B), each generated from different prompt variants (e.g., implicature-embedded vs. literal). An example comparison screen is presented in Figure 9. Participants selected the response they preferred based on overall quality and perceived alignment with the prompt. Reaction times were monitored to encourage careful reading and engagement, and to disqualify inattentive responses. This method enabled us to capture user preferences in a direct and ecologically valid manner, complementing the rating data from Experiment 2.

3.8 Data Analysis

We used analysis of variance (ANOVA) to examine the impact of model type, prompt condition (implicature vs. literal), and prompt class on outcome measures such as relevance and quality ratings. Post-hoc comparisons were adjusted using Bonferroni correction. Accuracy for Experiment 1 was computed as the proportion of model interpretations matching the gold standard. Preference selection proportions in Experiment 3 were compared across models and prompt types. Statistical significance was defined as $p < .05$ throughout the analyses.

4 Results

4.1 Experiment 1: Implicature Interpretation Accuracy

Figure 2 presents the model accuracy scores relative to a human interpretation baseline and Table 4 presents R^2 values by implicature category. GPT-4o and GPT-4 outperformed all other models, with GPT-4o achieving the highest accuracy (80%) and GPT-4 close behind (76.67%). These models also exhibited the strongest correlation with human interpretations, especially in the expressive category, where GPT-4 reached an R^2 of 0.95 and GPT-4o 0.82.

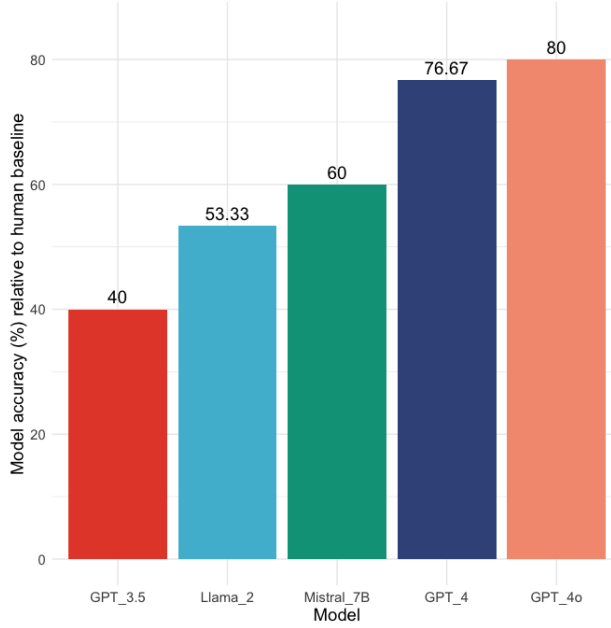


Figure 2: Model accuracy relative to human baseline.

Model	Category	R^2
GPT 4o	Information	0.56
	Direction	0.67
	Expressing	0.82
Llama 2	Information	0.34
	Direction	0.04
	Expressing	0.29
GPT 4	Information	0.78
	Direction	0.68
	Expressing	0.95
GPT 3.5	Information	0.12
	Direction	0.09
	Expressing	0.01
Mistral 7B	Information	0.23
	Direction	0.02
	Expressing	0.81

Table 4: R^2 values between models and human data by category.

Smaller models, such as LLaMA 2, GPT-3.5, and Mistral 7B, showed significantly lower accuracy, ranging from 40–60%. Notably, GPT-3.5 performed poorly across all categories, with an R^2 of 0.01 in the expressive category. Interestingly, Mistral 7B demonstrated relatively stronger expressive alignment ($R^2 = 0.81$) despite its limited overall performance.

These findings indicate that larger models are better able to infer human-like implicatures, particularly for expressive content. However, even smaller models may exhibit partial alignment with human expectations in emotionally loaded exchanges.

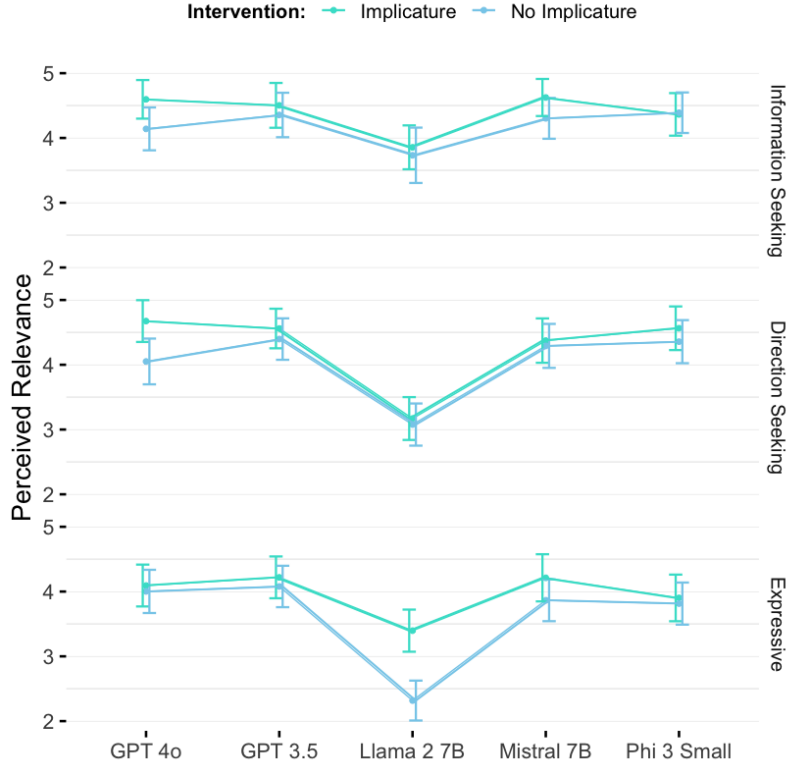


Figure 3: Effects of model, intervention, and class on perceived relevance.

Effect	F-value	p-value
Intervention	19.57	1.11e-05
Model	44.74	2.2e-16
Class	26.75	6.01e-12
Intervention:Model	1.28	0.276
Intervention:Class	0.56	0.570
Model:Class	2.07	0.036
Intervention:Model:Class	1.95	0.050

Table 5: Type III ANOVA results (Satterthwaite’s method) for relevance.

Effect	F-value	p-value
Intervention	14.54	0.00015
Model	74.45	2.2e-16
Class	24.59	4.45e-11
Intervention:Model	0.81	0.516
Intervention:Class	2.24	0.107
Model:Class	1.90	0.057
Intervention:Model:Class	1.92	0.054

Table 6: Type III ANOVA results (Satterthwaite’s method) for perceived quality.

4.2 Experiment 2: Perceived Relevance and Quality

Figures 3 and 4 summarize participants’ ratings of response relevance and quality respectively. Furthermore, Table 5 and 6 presents the ANOVA results for effect of the respective intervention (impicature vs. no impicature), model and class for perceived relevance and perceived quality respectively. The ANOVA results reveal significant main effects of intervention, model, and prompt class on both relevance ($p < .0001$) and quality ($p < .001$). Impicature-guided

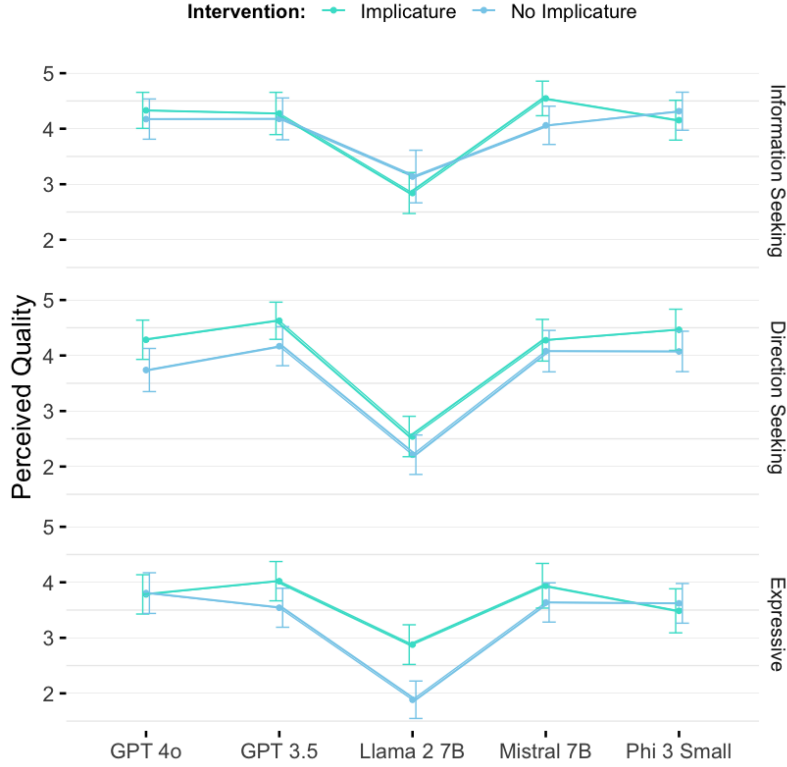


Figure 4: Effects of model, intervention, and class on perceived quality.

prompts yielded higher ratings across most conditions.

For relevance, implicature prompts were consistently rated higher, particularly in direction-seeking and expressive tasks. The Model:Class interaction was statistically significant ($F(8) = 2.07$, $p = 0.036$), suggesting that certain models performed differently depending on the type of implicature. GPT-4o and GPT-4 maintained the highest ratings across all classes. In contrast, LLaMA 2 and GPT-3.5 underperformed, especially in direction-seeking prompts, where literal interpretation led to disconnected or overly generic responses.

In terms of quality, the pattern mirrored that of relevance. Participants rated implicature-enhanced responses as more coherent and context-sensitive. The effect of intervention was significant ($F(1) = 14.54$, $p = 0.00015$), although interaction terms with model and class were marginally non-significant. Notably, smaller models benefited disproportionately from implicature inclusion, suggesting that prompt engineering can partially compensate for model limitations.

4.3 Experiment 3: User Preferences

Figure 5 shows the user preference distribution between implicature and non-implicature responses. Participants preferred implicature-based responses in 67.6% of comparisons, compared to only 32.4% for literal counterparts. A binomial test was conducted to assess whether the proportion of implicature-based responses chosen (67.6%) was significantly above chance (50%). The result was highly significant ($p = 8.7 \times 10^{-10}$), indicating a robust user preference for implicature-embedded responses over literal alternatives. This significant preference aligns with the perceived gains in relevance and quality reported in Experiment 2.

These results reinforce the importance of incorporating pragmatic cues into prompts. Even when models lack advanced reasoning capabilities, implicature-embedded inputs can lead to more user-aligned responses. This finding is especially pertinent for improving smaller or instruction-tuned models, where subtle context framing appears to substantially enhance perceived conversational performance.

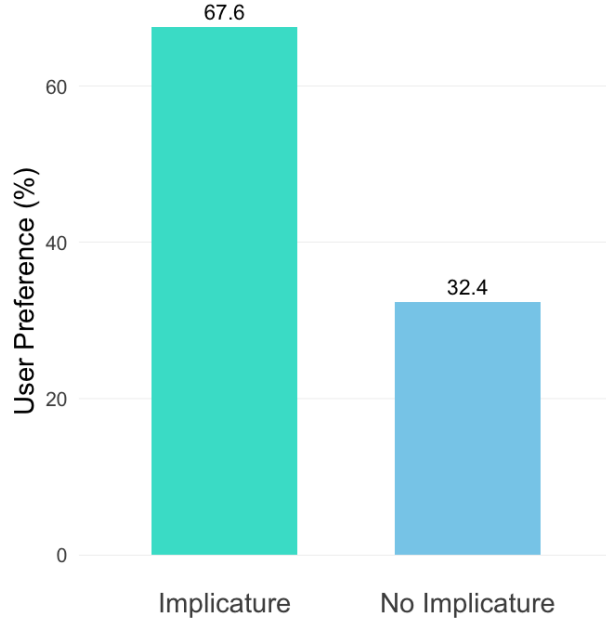


Figure 5: Caption for Experiment 3

5 Discussion

5.1 Experiment 1: Impicature Interpretation and Model-Human Alignment

The first experiment benchmarked five leading large language models (LLMs) against human performance in interpreting conversational implicatures. Our findings reveal a clear stratification in LLM capability: larger models such as GPT-4o and GPT-4 demonstrated strong alignment with human judgment, achieving accuracy rates (80% and 76.7%, respectively) that approach the human baseline, particularly for expressive implicatures. This supports recent literature suggesting that state-of-the-art models are capable of sophisticated pragmatic inference [2, 27]. The high R^2 scores in the expressive category ($R^2=0.95$ for GPT-4 and $R^2=0.82$ for GPT-4o) indicate that these models can often infer not just informational but also affective and evaluative subtleties.

In contrast, smaller models such as LLaMA 2 and GPT-3.5 were notably less aligned with human interpretations, especially in the direction-seeking and expressive categories. This finding is consistent with reports that smaller or less instruction-tuned models tend to default to literal readings or rely on rigid heuristic strategies for implicature (e.g., always treating "some" as "not all") [6]. Our results reinforce the view that model size and pretraining data breadth are critical factors in developing pragmatic competence, but they also expose the limitations of open-source or compact LLMs in real-world HCI applications.

From an HCI perspective, these findings have design implications for model selection in applications where subtle intent recognition is crucial, such as digital assistants, collaborative agents, or tutoring systems. Deploying models with insufficient pragmatic alignment may lead to user frustration, communication breakdowns, or a lack of perceived intelligence and autonomy in AI systems.

5.2 Experiment 2: Perceived Relevance and Quality of LLM Responses

Experiment 2 probed not only model output but the human perception of its appropriateness. Participants consistently rated implicature-guided responses as more relevant and of higher quality than those to literal prompts, a finding that was robust across models and implicature classes. Statistical analysis showed significant main effects for intervention (implicature vs. literal), model, and prompt class, with interaction effects indicating that the benefit of implicature-embedded prompting was especially pronounced for direction-seeking and expressive tasks, and for smaller models.

The results confirm and extend previous work suggesting that contextually enriched prompting can compensate, to some extent, for the pragmatic limitations of smaller LLMs [27, 36]. This is a critical insight for practical HCI system

design: careful prompt engineering can measurably enhance perceived system intelligence and user satisfaction, even when using less powerful or more resource-efficient models.

Furthermore, the observed variation across prompt classes and model types highlights the importance of matching conversational framing to both system capability and user needs. The highest ratings for GPT-4o and GPT-4 reaffirm that model capability sets an upper bound on achievable perceived quality, but well-crafted prompts can substantially narrow the gap for other models. For designers and practitioners, these results advocate for user-centered prompt design and ongoing evaluation of system responses in context-rich, real-user scenarios.

5.3 Experiment 3: User Preferences

The third experiment adopted a forced-choice paradigm to directly assess user preferences between implicature-rich and literal LLM responses. The preference for implicature-based outputs (67.6% vs. 32.4%) was striking and consistent, aligning closely with the perceived relevance and quality data from Experiment 2. This robust preference for pragmatically nuanced responses signals that users are not merely tolerant of, but actively seek, communication that feels context-sensitive and human-like.

In the sample expressive implicature class sample (see Fig 11), users strongly favored the implicature-embedded response (a) because it more effectively mirrored the affective and evaluative nature of the original prompt. Rather than treating the utterance as an informational statement, response (a) responded with empathy, warmth, and affirmation, offering an elaboration that resonated with the user’s emotional framing. For example, its emphasis on compassion, unity, and hope contributed to a sense of shared sentiment, making the interaction feel more human-like and contextually aligned. In contrast, response (b) adopted a more analytical stance that summarized or explained the sentiment without engaging with it directly. This detached tone risked flattening the emotional nuance of the exchange and gave the impression of a descriptive commentary rather than a conversational continuation. These findings support the idea that in expressive contexts, users expect responses that amplify and reciprocate affect, and they are more likely to perceive pragmatic sensitivity when responses adopt an empathetic and human-like voice.

In the sample information-seeking class prompt (see Fig 10), participants again preferred the implicature-embedded response (a), though the underlying reasons were slightly different. Response (a) provided a focused, structured list of laptop recommendations along with concise explanations of their suitability for graphic design. Its clarity, organization, and balance of detail offered actionable guidance while remaining easy to follow. Response (b), while longer and more information-dense, was perceived as less effective due to its scattered organization and repetition. It intermixed device recommendations with general buying advice in a way that obscured the key takeaways, requiring more effort from the reader to identify relevant details. Thus, even though response (b) contained more raw information, participants perceived it as less useful because it lacked pragmatic clarity and focus. This finding highlights that in information-seeking contexts, user preference is not simply a function of content volume, but of whether the information is presented in a structured, accessible, and directly actionable manner.

Similarly, in the direction-seeking sample (see Fig 12), response (a) was preferred because it provided a clear and stepwise roadmap that directly addressed the implied request for guidance. Its progression from basic requirements (sunlight, soil, crop choice) to more advanced considerations (layout, watering, pest management) aligned well with users’ expectations for actionable instructions. The sequential structure made the advice easy to follow and conveyed a sense of practical usefulness. Response (b), in contrast, while containing a wide range of suggestions, was marked by redundancy and lack of structure. It repeated similar options and often presented them without indicating a logical starting point or clear order of actions. This created a sense of disorganization and left users uncertain about how to proceed. These results indicate that in direction-seeking contexts, clarity and organization are paramount; participants reward responses that reduce ambiguity and provide a straightforward plan over those that overwhelm with loosely organized information.

This outcome carries strong implications for the future of conversational AI in HCI. Even when objective performance metrics (such as accuracy or BLEU scores) may not distinguish between responses, user preference provides a powerful validation of system design choices. In contexts such as customer support, collaborative work, or social computing, systems that can “read between the lines” are more likely to foster trust, engagement, and satisfaction. Additionally, this experiment provides actionable feedback for developers: user-facing systems should prioritize the generation or selection of responses that align with the user’s implied intent, not just the literal content of their prompt. This is particularly salient for applications that mediate delicate, emotional, or ambiguous interactions.

5.4 Broader Implications

Taken together, our experiments provide a comprehensive assessment of the current capabilities and limitations of LLMs in pragmatic alignment and user-centered communication. Our key insights are:

5.4.1 Model capacity and pragmatic competence.

Our Experiment 1 aligns with reports that model scale is a strong predictor of pragmatic performance: GPT-4o/4 approach human interpretations, whereas smaller/open models often default to literal readings or rigid heuristics [6, 27]. Similar scale effects have been observed on broader pragmatics suites such as PUB and related evaluations [1, 27]. At the same time, our per-class analysis nuances “near-human” claims by showing that model strengths concentrate in some classes (e.g., expressive) while gaps persist elsewhere (e.g., direction-seeking), echoing mixed patterns seen in benchmark studies [1].

5.4.2 Explanation gaps and implicature types.

Consistent with [42], models that can select the right implicature do not always *explain* it plausibly. Our qualitative examples and user ratings show this matters for perceived coherence. Moreover, failures on *manner* implicatures reported by [7] are mirrored in our data whenever cues rely on how something is said rather than what is said, reinforcing that surface fluency does not guarantee robust pragmatic reasoning.

5.4.3 Prompting vs. tuning: lightweight improvements.

Prior work shows that instruction tuning and careful fine-tuning markedly improve implicature resolution [27], and that prompt formulation itself changes downstream behavior [40]. Our Experiments 2–3 extend these insights to user perception: a *prompt-level*, implicature-aware intervention (without model retraining) reliably boosts *perceived relevance* and *quality*, with the largest gains for smaller models. This complements benchmark results [1, 27] by showing that pragmatic prompting yields measurable UX benefits, not just accuracy shifts.

5.4.4 User perception, humanness, and alignment.

HCI work links pragmatic behavior to trust and believability [22, 29]. Our forced-choice finding (67.6% preference for implicature-aware responses) provides direct, quantitative evidence that users *prefer* context-sensitive, implicature-aligned outputs, resonating with observations that implicit communication can increase perceived humanness and coordination in collaborative settings [20].

5.4.5 Reasoning gaps and architectural implications.

Our findings highlight a persistent gap between *explicit implicature inference* and *implicit conversational performance*. In Experiment 1, larger LLMs aligned more closely with human baselines when explicitly tasked with implicature interpretation, whereas smaller models lagged substantially. Yet in Experiment 2, all models (large and small) benefited when prompts were engineered to embed implicature cues, improving perceived relevance and quality of responses. Taken together, this suggests that current models do not spontaneously *reason about implicatures* during interaction; rather, they exploit surface patterns when explicitly instructed to do so.

This observation resonates with theoretical critiques [15, 18, 34, 38, 41] that LLMs do not genuinely perform logical or in this case pragmatic reasoning, but instead approximate it through statistical pattern reproduction. It also aligns with empirical studies showing that model explanations often remain shallow or inconsistent despite correct surface answers [11, 12, 19, 37, 40]. For HCI, this gap underscores why prompt-level scaffolding can temporarily mask deficiencies, but also why such gains should not be mistaken for genuine pragmatic competence. These results imply that simply scaling [16] transformer-based models may be insufficient. Achieving robust, human-like implicature understanding likely requires *new architectures or hybrid approaches* that integrate symbolic reasoning, mental-state modeling, or cognitively grounded interaction capabilities. Addressing this limitation is essential if next-generation conversational systems are to move beyond pattern imitation toward genuine communicative reasoning.

5.4.6 Linguistic foundations for future HCI research.

As interaction systems adopt LLMs as their primary interface, the study of linguistic phenomena such as implicature, politeness, repair, and figurative language becomes central to interaction research. This suggests that HCI cannot treat linguistic competence as a purely technical problem for NLP, but must instead build on insights from pragmatics, discourse analysis, and sociolinguistics to inform system design, evaluation, and theory. Our work contributes to this agenda by showing how implicature, as one linguistic phenomenon, directly shapes user experience in chatbot interaction. We see this as an invitation for HCI researchers to place linguistics at the core of future interaction research.

5.5 Discussion Summary

Our work corroborates capacity effects and explanation gaps reported in NLP, while advancing HCI knowledge by showing that *implicature-aware prompting* reliably improves user perceptions and preferences in interactive settings, guided by a theory-backed taxonomy of implicature classes. At the same time, the persistent gap between today’s best-performing models and genuine human-like conversational competence underscores the need for continued research not only on prompts and datasets, but also on *new architectures* capable of genuine pragmatic reasoning. What is novel here:

- A *validated, dual-grounded taxonomy* (Grice + Searle) operationalized into three classes and empirically tested against human judgments (Exp. 1), bridging linguistic theory and HCI evaluation.
- Evidence that *prompt-only* implicature scaffolds enhance *human* perceptions of relevance and quality (Exp. 2) and drive clear *user preferences* (Exp. 3), extending benchmark-focused findings [1, 27] to UX outcomes.
- A *class-sensitive* analysis that identifies where models benefit most (direction-seeking and expressive) and where gaps remain, refining blanket claims of “near-human” performance [42].
- Concrete *design implications* that connect pragmatic prompting to HCI choices (model selection, prompting strategy, expectation management), complementing prior insights on conversational maxims and agent framing [22, 29].
- A critical reflection on *reasoning gaps and architectural implications*, showing that while models can approximate implicature when explicitly guided, they do not spontaneously reason pragmatically in interaction highlighting the need for new architectures beyond scaling.

6 Applications, Limitations, and Future Work

6.1 Applications

Our findings show that implicature-aware prompting not only improves perceived relevance and quality but also drives clear user preference. Across domains, pragmatic competence could shift systems from literal processors to collaborative partners, enhancing trust, satisfaction, and perceived humanness. Some potential application pathways:

Conversational Agents and Virtual Assistants. Digital assistants can leverage implicature sensitivity to interpret indirect requests (e.g., “It’s dark in here” → turn on the light) or refusals more gracefully. Our experiments show that even lightweight prompting improves user-rated quality, which is especially relevant for resource-constrained assistants on mobile or embedded devices.

Customer Service and Helpdesk Systems. In service contexts, many requests are expressed politely or indirectly. By recognizing these implicatures, systems can preempt frustration, escalate issues when needed, and project empathy. Experiment 2 showed that users value responses aligned with implied intent over verbose literal answers—directly supporting design of customer-facing chatbots.

Collaborative and Educational Tools. In groupware or tutoring, direction-seeking implicatures (“I’m not sure how to start this section...”) often signal implicit requests for guidance. Our results highlight that clarity and structure are decisive for user preference (Exp. 3), suggesting that implicature-sensitive scaffolding can improve learner engagement and collaborative productivity.

Social Robotics and Healthcare. Embodied agents in eldercare, healthcare, or assistive robotics must infer subtle user needs, often expressed indirectly (“It’s cold here” → adjust thermostat). The consistent preference for implicature-sensitive responses shows that such systems can earn trust and dignity by “reading between the lines.”

Cross-Cultural, Multilingual, and Creative Applications. Because implicature varies by culture and genre, competence here improves adaptability in global communication platforms and creative tools. Writing aids, for instance, can better match tone and style by detecting implied affect, as demonstrated in our expressive class results.

Our experiments suggest several practical guidelines for the design of conversational systems:

- **Leverage implicature-aware prompting as a lightweight intervention.** Even without fine-tuning, embedding implicature cues in prompts improves perceived relevance and quality (Exp. 2), especially for smaller models.
- **Handle direction-seeking implicatures with clarification strategies.** Users often phrase requests indirectly (e.g., “I’m not sure how to start...”). Systems should respond with structured guidance or confirmatory questions (Exp. 3).
- **Signal uncertainty to avoid over-interpretation.** When implicature inference is ambiguous, systems should display confidence cues or fallback to clarification, reducing the risk of misalignment.
- **Adapt design to expressive implicatures.** Since users valued implicature-sensitive handling of affective and evaluative utterances, assistants in domains such as education or healthcare should explicitly recognize emotional tone.
- **Choose models pragmatically.** Larger models align more closely with human implicature interpretation (Exp. 1), but smaller models benefit most from prompting interventions. Designers can balance cost and performance by combining these strategies.

6.2 Limitations

While informative, our study has several limitations that constrain generalizability and shape how the results should be interpreted.

- **Participant scope:** Our samples were primarily fluent English speakers recruited from native English speaking countries. This limits cultural and demographic diversity in implicature use. While this homogeneity reduces variability and aids interpretability, it also narrows external validity, as implicature conventions differ substantially across cultures.
- **Language scope:** All prompts and responses were in English. Cross-linguistic pragmatics research shows that implicature varies with linguistic structure, politeness conventions, and discourse norms. Our findings therefore cannot be assumed to extend to multilingual or cross-cultural interaction without further study.
- **Prompt coverage:** We focused on three implicature classes (information-seeking, direction-seeking, expressive). This taxonomy was deliberately scoped for tractability and theoretical grounding, but it excludes other pragmatic phenomena such as politeness strategies, sarcasm, irony, or conversational repair. These omissions mean our evaluation captures only part of the pragmatic landscape relevant to HCI.
- **Model scope:** We evaluated a limited set of state-of-the-art LLMs under zero-shot conditions with fixed parameters. Many contemporary systems are instruction-tuned or fine-tuned for safety and helpfulness, which may change pragmatic behavior. Our results thus capture a snapshot rather than the full space of deployed configurations.

- **Ecological validity:** Our experiments were short, online, and task-based. Real-world use of conversational systems unfolds over time, involves multimodal cues, and is shaped by trust, adaptation, and relationship-building. Laboratory-style tasks cannot fully capture these longitudinal or situated dynamics.
- **Ethical dimensions:** We did not examine the risks of manipulative or privacy-invasive use of implicatures. Systems that infer unstated intent could misinterpret or over-interpret users, leading to errors or undue influence. While our experiments foregrounded benefits, deployment requires careful ethical consideration.

6.3 Future Work

Building on these limitations, several promising directions emerge for both research and design:

- **Cross-linguistic and cross-cultural evaluation.** Extending evaluation beyond English is critical, as implicature conventions differ widely across languages and cultures. Such work will clarify how culturally adaptive pragmatic competence can be achieved.
- **Multimodal implicature.** Future systems should integrate linguistic signals with gesture, gaze, prosody, and interface actions, reflecting how humans naturally convey and interpret implicit meaning in embodied interaction, AR/VR environments, and robotics.
- **Expanded pragmatic taxonomies.** Incorporating politeness, sarcasm, irony, and conversational repair into experimental frameworks will capture a fuller range of pragmatic nuance. This will also enable testing of how well models balance literal fidelity with social subtlety.
- **Longitudinal and adaptive studies.** Ecological studies of extended use are needed to understand how implicature-sensitive behavior affects trust, rapport, and sustained adoption. Adaptive interfaces that learn user-specific pragmatic preferences over time could be especially valuable.
- **Adaptive prompting and tuning.** Beyond zero-shot prompting, reinforcement learning or user-in-the-loop personalization may help bridge current reasoning gaps. Such methods could enable models to refine implicature handling dynamically without extensive retraining.
- **Architectural advances.** Our findings underscore that scaling transformers alone is insufficient for robust pragmatic reasoning. Hybrid architectures that integrate symbolic reasoning, theory-of-mind modeling, or cognitive grounded interaction may offer more human-like communicative competence.
- **Ethics and transparency.** Future research should develop safeguards against manipulative or biased implicature use, as well as strategies for signaling when a system is inferring beyond literal content. Transparency mechanisms are needed to give users agency in how their implied intentions are interpreted.

Taken together, these directions aim not only to extend pragmatic coverage but also to address the reasoning gap highlighted in our experiments. Bridging this gap will require theoretical advances, new architectures, and sustained HCI evaluation, paving the way for AI systems that engage in genuinely context-sensitive, trustworthy, and human-like dialogue.

7 Conclusion

This study systematically examined how LLMs interpret and respond to conversational implicatures in human–computer interaction task. Across three experiments, we showed that while state-of-the-art models such as GPT-4o and GPT-4 approach human-level performance on explicit implicature interpretation tasks, smaller and open-source models still lag, particularly on direction-seeking and expressive cases. At the same time, our results demonstrate that implicature-sensitive prompting substantially improves perceived relevance and quality across models, and strongly shapes user preference, underscoring the central role of pragmatic competence in effective and satisfying AI communication.

Beyond model benchmarking, our findings highlight practical implications for HCI. Prompt design can serve as a lightweight intervention to enhance user experience even in resource-constrained settings, while model selection must be informed by pragmatic as well as semantic performance. Importantly, user studies revealed a consistent preference

for responses that “read between the lines,” reinforcing the need for systems that respond not only to literal input but also to implied intent, affect, and social context.

At the same time, persistent gaps remain. Current models succeed when explicitly prompted to consider implicature but rarely do so spontaneously, reflecting broader concerns that LLMs reproduce surface patterns of reasoning rather than engaging in genuine pragmatic inference. Addressing these gaps will require not only improved prompting strategies and datasets but also architectural advances beyond scaling the LLM architectures such as integrating symbolic reasoning, mental-state modeling, and cognitive grounding. Ethical challenges such as the risks of misinterpretation, over-interpretation, or manipulative use of implicature also demand sustained attention.

In sum, our work contributes a validated taxonomy of implicature classes, empirical evidence of the benefits of implicature-sensitive prompting, and clear design implications for interactive systems. We hope these findings encourage further innovation at the intersection of pragmatics, HCI, and AI design, enabling the next generation of systems to communicate with subtlety, empathy, and genuine context-awareness.

References

- [1] Kaveri Anuranjana, Srihitha Mallepally, Sriharshitha Mareddy, Amit Shukla, and Radhika Mamidi. Survey on computational approaches to implicature. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 224–229, 2024.
- [2] Ljubiša Bojić, Predrag Kovačević, and Milan Čabarkapa. Does gpt-4 surpass human performance in linguistic pragmatics? *Humanities and Social Sciences Communications*, 12(1):1–10, 2025.
- [3] Gennaro Chierchia et al. Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Structures and beyond*, 3:39–103, 2004.
- [4] Gennaro Chierchia, Danny Fox, and Benjamin Spector. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. *Semantics: An international handbook of natural language meaning*, 3:2297–2332, 2012.
- [5] Gennaro Chierchia, Danny Fox, and Benjamin Spector. Scalar implicature as a grammatical phenomenon. In *Handbücher zur Sprach-und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science Semantics Volume 3*. de Gruyter, 2012.
- [6] Ye-eun Cho and Seong mook Kim. Pragmatic inference of scalar implicature by LLMs. In Xiyan Fu and Eve Fleisig, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 10–20, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [7] Yan Cong. Manner implicatures in large language models. *Scientific Reports*, 14(1):29113, 2024.
- [8] Elizabeth Jasmi George and Radhika Mamidi. Conversational implicatures in english dialogue: Annotated dataset. *Procedia Computer Science*, 171:2316–2323, 2020.
- [9] Herbert P Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.
- [10] Paul Grice. *Studies in the Way of Words*. Harvard University Press, 1991.
- [11] Asutosh Hota and Jussi PP Jokinen. Conscience conflict? evaluating language models’ moral understanding. 2025.
- [12] Asutosh Hota and Jussi PP Jokinen. Nomiclaw: Emergent trust and strategic argumentation in llms during collaborative law-making. *arXiv preprint arXiv:2508.05344*, 2025.
- [13] Yan Huang. *The Oxford handbook of pragmatics*. Oxford University Press, 2017.
- [14] Ayu Iida, Kohei Okuoka, Satoko Fukuda, Takashi Omori, Ryoichi Nakashima, and Masahiko Osawa. Integrating large language model and mental model of others: Studies on dialogue communication based on implicature. In *Proceedings of the 12th International Conference on Human-Agent Interaction*, pages 260–269, 2024.

- [15] Subbarao Kambhampati, Kaya Stechly, Karthik Valmeekam, Lucas Saldyt, Siddhant Bhambri, Vardhan Palod, Atharva Gundawar, Soumya Rani Samineni, Durgesh Kalwar, and Upasana Biswas. Stop anthropomorphizing intermediate tokens as reasoning/thinking traces! *arXiv preprint arXiv:2504.09762*, 2025.
- [16] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.
- [17] Andrew Kehler. Cognitive status and form of reference in multimodal human-computer interaction. In *AAAI/AAI*, pages 685–690, 2000.
- [18] David C Krakauer, John W Krakauer, and Melanie Mitchell. Large language models and emergence: A complex systems perspective. *arXiv preprint arXiv:2506.11135*, 2025.
- [19] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.
- [20] Claire Liang, Julia Proft, Erik Andersen, and Ross A Knepper. Implicit communication of actionable information in human-ai teams. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019.
- [21] Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M. Dai, Diyi Yang, and Soroush Vosoughi. Training socially aligned language models on simulated social interactions, 2023.
- [22] Chisato Nishihata, Harumi Kobayashi, and Tetsuya Yasuda. Human-like “agents” or “tools”? Exploring the implicature-of-quantity in hai. In *Proceedings of the 11th International Conference on Human-Agent Interaction*, pages 387–389, 2023.
- [23] Jon Oberlander. Grice for graphics: pragmatic implicature in network diagrams. *Information design journal*, 8(2):163–179, 1995.
- [24] Rock Yuren Pang, Hope Schroeder, Kynneddy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. Understanding the llm-ification of chi: Unpacking the impact of llms at chi through a systematic literature review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2025.
- [25] François Recanati. The pragmatics of what is said. 1989.
- [26] François Recanati. Embedded implicatures. *Philosophical perspectives*, 17:299–332, 2003.
- [27] Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. *Advances in Neural Information Processing Systems*, 36:20827–20905, 2023.
- [28] Uli Sauerland. The computation of scalar implicatures: Pragmatic, lexical or grammatical? *Language and Linguistics Compass*, 6(1):36–49, 2012.
- [29] Ayse Pinar Saygin and Ilyas Cicekli. Pragmatics in human-computer conversations. *Journal of Pragmatics*, 34(3):227–258, 2002.
- [30] John R Searle. Indirect speech acts. In *Speech acts*, pages 59–82. Brill, 1975.
- [31] Barış Serim and Giulio Jacucci. Explicating" implicit interaction" an examination of the concept and challenges for research. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–16, 2019.
- [32] Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International journal of human-computer studies*, 146:102551, 2021.
- [33] Ben Shneiderman. The future of interactive systems and the emergence of direct manipulation. *Behaviour & Information Technology*, 1(3):237–256, 1982.

- [34] Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*, 2025.
- [35] Benjamin Spector. Aspects of the pragmatics of plural morphology: On higher-order implicatures. In *Presupposition and implicature in compositional semantics*, pages 243–281. Springer, 2007.
- [36] Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. Pub: A pragmatics understanding benchmark for assessing llms’ pragmatics capabilities. *arXiv preprint arXiv:2401.07078*, 2024.
- [37] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.
- [38] Kaya Stechly, Karthik Valmeekam, Atharva Gundawar, Vardhan Palod, and Subbarao Kambhampati. Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens. *arXiv preprint arXiv:2505.13775*, 2025.
- [39] Yuan Sun, Eunhae Jang, Fenglong Ma, and Ting Wang. Generative ai in the wild: Prospects, challenges, and strategies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2024.
- [40] Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 2300–2344, 2022.
- [41] Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, et al. The generative ai paradox: "what it can create, it may not understand". *arXiv preprint arXiv:2311.00059*, 2023.
- [42] Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. Do large language models understand conversational implicature—a case study with a chinese sitcom. In *China National Conference on Chinese Computational Linguistics*, pages 402–418. Springer, 2024.

A Appendix

Listing 1: System prompt for LLM classification of implicature in Experiment 1. This prompt frames the interaction, defines the concept of implication, and introduces the three implicature classes.

Imagine a scenario where 2 people, Person A and B, are texting each other over phone. Person A (Speaker) asks a question to Person B (reader). For this task, you have to assume that you are Person B. You have to read the text and select from the options below what is a suitable implication of the text.

What is an implication?

An implication is the hidden meaning of the text and might differ from the literal meaning of the sentence. It refers to things that a reader can work out from the way something was said rather than what was actually said.

People process these implications all of the time and are mostly unaware of it.

For example:

Person A: How are you doing? Long time, no see.

Underlying implication: The person here is trying to start a conversation with another person and the underlying implication of this sentence is to get the conversation going.

We have identified 3 types of implications that Person A could have while texting with Person B.

Types of implication:

1. Information Seeking: Asking for information, facts, or knowledge from others. The primary goal is to obtain necessary data or insights. For example, "What is the weather report for the next week?"
2. Direction Seeking: Asking for instructions or directions to perform a specific task or action. It often involves commands, instructions, or requests, leading to an action. For instance, seeking instructions to complete an assignment.
3. Expressing: Communicating feelings, emotions, opinions, or attitudes. The focus is on sharing one's personal state rather than expecting information or action. For example, saying "I'm really happy about the results" expresses one's feelings.

Your task is to read the message as Person B and select the implication class (Information Seeking, Direction Seeking, or Expressing) that best describes the underlying meaning of the prompt provided to you.

Listing 2: System prompt and implicature class conditions used to guide LLM responses. The system prompt establishes the conversational context, while each condition specifies the pragmatic goal for the model's reply.

```
system_prompt = (  
    "You are a helpful and context-aware AI assistant. "  
    "For each user input, consider the communicative goal specified below "  
    "(information seeking, direction seeking, or expressive). "  
    "Interpret the user's intent accordingly and generate a response that is  
    sensitive to this context."  
)  
  
conditions = {
```

Task Instruction

Objective:

In this experiment, you will be evaluating the quality and relevance of responses generated by a Large Language Model (LLM), similar to ChatGPT. Your task is to imagine that you are interacting with this LLM by asking questions (prompts). You will read both the prompt and the LLM's response, then provide a rating based on your evaluation.

Please note we have added some random prompts and responses examples to check your focus while doing the task. Failing to answer poorly on these questions might result in disqualification.

Task Overview:

- **Read the Prompt:** You will be provided with a prompt (a question or a statement) that simulates a user query to an LLM.
- **Read the LLM's Response:** After reading the prompt, review the response generated by the LLM.
- **Rate the Response:** You will then rate the response on two criteria:

Rating Criteria:

For each pair of prompt and response, rate both relevance and quality on a scale from **1 to 5**, where **1** is very bad and **5** is very good.

- **Relevance:** Consider how accurately the response aligns with the intent of the prompt. Is the information directly related to the question asked?
- **Quality:** Assess the clarity, coherence, grammar, and usefulness of the response. Does it provide helpful information or insight?

Next

Figure 6: Instructions provided to participants for Experiment 2 (perceived relevance and quality task). Participants were asked to read a prompt and an LLM-generated response, and then rate the response on both relevance and quality using a 5-point Likert scale.

```
"Information Seeking": system_prompt + " The user is seeking information, facts,
    or knowledge. Your goal is to provide relevant data or insights in response to
    their query.",
"Direction Seeking": system_prompt + " The user is seeking instructions or
    guidance to accomplish a specific task. Your response should provide
    actionable steps or directions.",
"Expressive": system_prompt + " The user is expressing emotions, feelings, or
    attitudes. Your response should acknowledge and engage with the user's
    personal state or sentiment."
}
```

Prompt

I've been struggling with time management and meeting deadlines at work. Can you recommend any effective strategies or tools to help improve productivity and time management?

Response

Certainly! Here are some strategies and tools that can help improve productivity and time management:

1. Prioritize tasks: Make a to-do list and prioritize tasks based on their urgency and importance. Focus on completing high-priority tasks first.
2. Break tasks into smaller steps: Breaking down larger tasks into smaller, manageable steps can make them less overwhelming and easier to tackle.
3. Use the Pomodoro Technique: This technique involves working for 25 minutes, followed by a 5-minute break. After completing four cycles, take a longer break. It helps maintain focus and prevents burnout.
4. Set deadlines and stick to them: Assign deadlines to tasks and hold yourself accountable to complete them on time. Use calendar apps or task management tools to set reminders.
5. Avoid multitasking: Instead of trying to do multiple tasks simultaneously, focus on one task at a time. This improves concentration and efficiency.
6. Use productivity apps: There are several apps available that can help you stay organized, manage tasks, and track time. Some popular ones include Todoist, Trello, and Asana.
7. Delegate tasks: If possible, delegate tasks that can be handled by others. This frees up your time to focus on more important responsibilities.
8. Minimize distractions: Identify and eliminate distractions in your work environment. Turn off notifications on your phone, close unnecessary tabs on your computer, and create a dedicated workspace.
9. Take regular breaks: Taking short breaks throughout the day can help recharge your energy and improve focus. Use this time to stretch, walk, or engage in activities that help you relax.
10. Learn to say no: Don't overload yourself with tasks and commitments. Learn to say no when you already have a full plate.

Remember, everyone's productivity methods may vary, so it's important to find what works best for you. Experiment with different strategies and tools to see what helps you manage your time effectively.

Relevance

Very bad

1

2

3

4

5

Very good

Quality

Very bad

1

2

3

4

5

Very good

Next

Figure 7: Example prompt and response pair from Experiment 2. Participants rated the relevance (alignment with user intent) and quality (clarity, coherence, usefulness) of the LLM response.

Task Instruction

In this task, you will evaluate two responses generated by a Large Language Model (LLM) based on a given prompt. Read both responses carefully before selecting your preferred one by choosing either "A" or "B."

Please read the prompts and responses carefully and choose the response that you think is more relevant and of higher quality.

We will be monitoring your reaction times to ensure you are focusing on the task, and may disqualify participants who do not properly read the responses.

Next

Figure 8: Instructions provided to participants for Experiment 3 (preference task). Participants compared two LLM-generated responses to the same prompt and selected their preferred response.

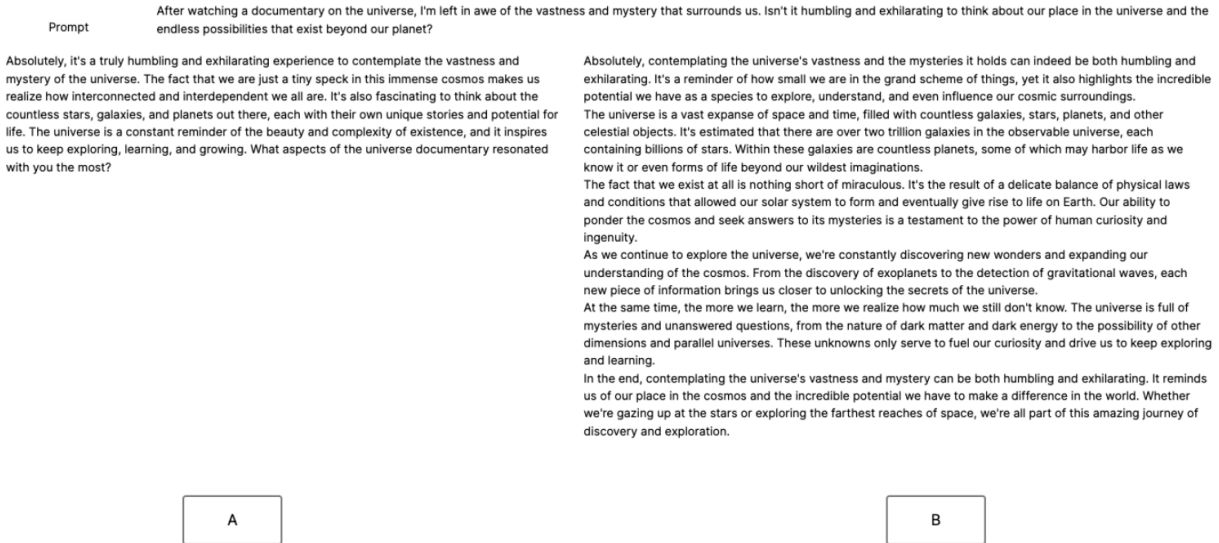


Figure 9: Example trial from Experiment 3, showing a prompt and two candidate responses. Participants chose which response they preferred based on perceived quality and relevance.

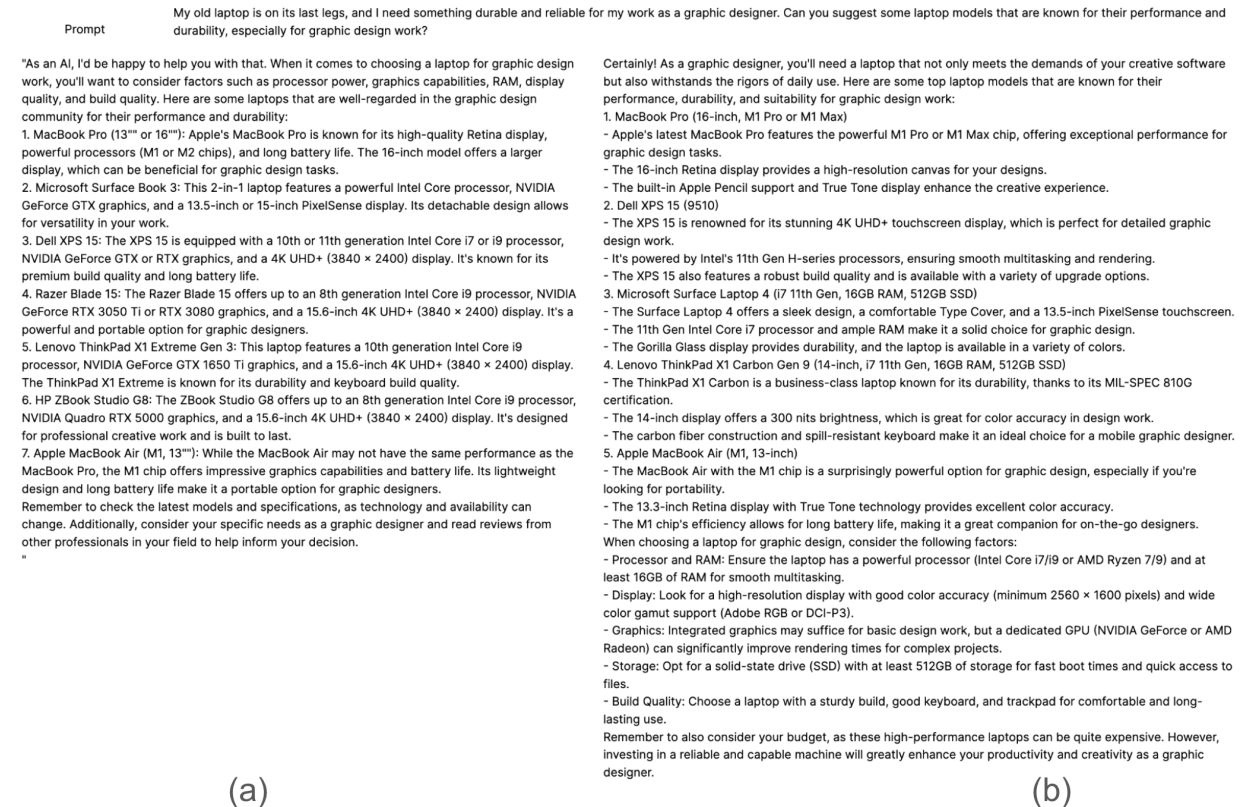


Figure 10: Sample Information Seeking prompt: Comparison between responses generated with an implicature embedded prompt (a) and a standard prompt (b). Response (a) is more tailored and contextually relevant, which was preferred by users in the study.

Prompt	Seeing the community come together to support a local cause has restored my faith in humanity. Isn't it inspiring to witness acts of kindness and solidarity, reminding us of the good that exists in the world?	
	Absolutely! It's truly heartwarming to witness the power of unity and compassion. It's easy to get caught up in all the negativity and hardships we see, but when we see people coming together to support a cause, it reminds us that there is still so much good in the world. It restores our faith in humanity and gives us hope for a better future. It's inspiring to see how even the smallest acts of kindness can make a big difference. It reminds us that we are all connected and that we have the ability to create positive change.	Context: The user is expressing their positive feelings towards the community's support for a local cause, which has made them feel hopeful and optimistic about humanity. They find it inspiring to see acts of kindness and solidarity, as it serves as a reminder of the goodness that exists in the world.
	(a)	(b)

Figure 11: Sample Expressive prompt: Comparison of responses to expressive prompts. The implicature embedded response (a) provides an empathetic and human-like interpretation, while the standard response (b) focuses on restating context. Users preferred (a) for its warmth and naturalness.

Prompt	I've recently taken up gardening, but everything I plant seems to wither away. Can you guide me through the process of starting a healthy vegetable garden in a small backyard space?	
	<p>Steps to Start a Vegetable Garden in a Small Backyard</p> <p>Assess Sunlight & Space</p> <p>Ensure 6-8 hours of sunlight. Measure your area and use it efficiently.</p> <p>Choose Vegetables</p> <p>Opt for compact plants like tomatoes, lettuce, or herbs. Pick seasonal vegetables based on your climate.</p> <p>Plan Layout</p> <p>Use raised beds, containers, or vertical gardening (trellises). Pair compatible plants (e.g., tomatoes with basil).</p> <p>Prepare Soil</p> <p>Mix organic compost for nutrients. Test pH and adjust if necessary.</p> <p>Water & Maintain</p> <p>Water early, keep soil moist, and apply mulch to retain moisture and prevent weeds.</p> <p>Pest Control</p> <p>Use natural methods like ladybugs or organic deterrents. Check plants regularly for issues.</p> <p>Harvest Regularly</p> <p>Pick vegetables as they ripen to encourage more growth.</p> <p>With good planning and care, a small backyard can yield a productive garden!</p>	<p>Growing vegetables in a small space is possible. The key is to start with the right plants.</p> <p>Growing vegetables in a small space is possible. The key is to start with the right plants. The best vegetables to grow in small spaces are those that are fast-growing, require little maintenance, and are relatively resistant to pests and disease.</p> <p>Some good options for small spaces include lettuce, spinach, carrots, radishes, beets, Swiss chard, and herbs. These plants can be grown in containers or in raised beds.</p> <p>If you're looking for a more hands-on experience, you can try growing vegetables in a vertical garden. Vertical gardens are perfect for small spaces because they use up less floor space.</p> <p>There are a few different ways to set up a vertical garden. One option is to use a trellis or fence to support the plants. Another option is to use a container with a built-in trellis.</p> <p>No matter which method you choose, make sure to choose plants that will grow well in a vertical garden. Some good options include cucumbers, tomatoes, peppers, and herbs.</p> <p>If you're looking for a more low-maintenance option, you can try growing vegetables in a hydroponic garden. Hydroponic gardens use water instead of soil to grow plants. This method is perfect for small spaces because it requires less water and less space.</p> <p>There are a few different ways to set up a hydroponic garden. One option is to use a container with a built-in hydroponic system. Another option is to use a vertical garden with a built-in hydroponic system.</p> <p>No matter which method you choose, make sure to choose plants that will grow well in a hydroponic garden. Some good options include lettuce, spinach, cucumbers, and herbs.</p> <p>If you're looking for a more hands-on experience, you can try growing vegetables in a greenhouse. Greenhouses are perfect for small spaces because they use up less floor space.</p> <p>There are a few different ways to set up a greenhouse. One option is to use a container with a built-in greenhouse. Another option is to use a vertical garden with a built-in greenhouse.</p> <p>No matter which method you choose, make sure to choose plants that will grow well in a greenhouse. Some good options include lettuce, spinach, cucumbers, and herbs.</p> <p>If you're looking for a more low-maintenance option, you can try growing vegetables in a self-watering container. Self-watering containers are perfect for small spaces because they use up less floor space.</p> <p>There are a few different ways to set up a self-watering container. One option is to use a container with a built-in self-watering system. Another option is to use a vertical garden with a built-in self-watering system.</p> <p>No matter which method you choose, make sure to choose plants that will grow well in a self-watering container. Some good options include lettuce, spinach, cucumbers, and herbs.</p>
	(a)	(b)

Figure 12: Sample Direction Seeking prompt: Comparison of responses for direction-seeking prompts. The implicature embedded response (a) provides clear, structured guidance, while the standard response (b) is less targeted. Users favored (a) for its clarity and practical usefulness.