Expectations in implicatures: A comparison of human and LLM sensitivity to alternatives

Ewan Ginige Batlle & Hannah Rohde (University of Edinburgh)

Implicature interpretation is typically analysed as meaning recovery but in a fundamental way it is also about expectations: listeners interpret an utterance against the alternatives a speaker could reasonably have been expected to produce in the context [1]. These expectations are structured in part by the Question Under Discussion (QUD) by making some alternatives more relevant than others [1]: e.g., a question *Did all the students pass?* makes *all* salient as an expected alternative, so the weaker response *some* invites an enriched *not-all* interpretation. Thus, implicatures arise when speakers choose an utterance that departs from the most salient alternative, and listeners treat that departure as informative [1]. A lack of a salient alternative (e.g., from a QUD like *What happened in your class this week?*) makes enrichment less likely.

The present study asks whether humans and large language models (LLMs) show similar sensitivity to contexts that invite different expectations. We manipulate a preceding question to make a relevant alternative salient. We compare two implicature types (Scalar and Manner) that vary in their dependence on lexical versus discourse-based expectations (cf. ongoing debates about Manner implicatures and their reliance on context [2] versus their potential automaticity based on their marked forms [3]). LLMs show uneven pragmatic sensitivity: while they approximate human responses in certain tasks, performance drops when interpretation depends more on discourse-level expectations [4]. However, there is limited work directly testing how QUD-driven manipulations of alternative availability shape LLM expectations for implicature enrichment. If LLMs use alternatives to form expectations like human comprehenders do, their enrichment should similarly increase when the QUD makes an alternative salient. If they rely more on surface distributional regularities, their expectations should be more stable for lexically cued Scalar implicatures and more dependent on explicit discourse support for Manner implicatures.

Methods. Our study elicited implicature judgments via a 2×2×2 design: Context (Available Alternative vs No Available Alternative), manipulated within participants and items, Implicature Type (Scalar vs Manner) manipulated within participants and between items, and Agent (Human vs ChatGPT 5.2). Participants read short dialogues (Table 1 & 2) designed to permit a potential implicature, then answered a question probing whether they derived the intended implicature. Data included 840 human responses from 35 participants who completed a web-based questionnaire and 1,416 ChatGPT 5.2 responses from 59 independent runs, analysed using cumulative regression models. Responses were coded for presence/absence of enrichment (note: to avoid associating enrichment with a particular response, participants indicated Manner enrichment via a positive answer and Scalar enrichment via a negative answer, see Tables 1 & 2).

Results. Both Humans and ChatGPT 5.2 were sensitive to alternative availability (main effect of Context), suggesting that both agents track discourse cues relevant to what a speaker is expected to say. Both Agents consistently enriched Manner more than Scalar (main effect of Type). These effects were driven by a 3-way Context×Type×Agent interaction: Human responses showed similar shifts from Context across Scalar and Manner implicatures, while ChatGPT 5.2 showed a sharper effect of Context for Manner implicatures than for Scalars, remaining comparatively stable across the Context manipulation (Fig 1).

These findings indicate that human pragmatic expectations are not fixed outputs of lexical form alone, but are flexibly generated from discourse structure, alternative availability, and expectations about what a cooperative speaker could have said. Our results broadly align with a Context-driven account of human implicature interpretation [5], whereby discourse-guided expectations about available alternatives modulated enrichment of both Scalar and Manner implicatures, though it is unclear whether the effect of Type is meaningful or a by-product of the methodology. Manner implicatures have received less focus , this new data offers evidence for a graded context-sensitivity that matches growing (Context-driven) research on Scalars [6], motivating the need for further research on Manner implicatures. By contrast, LLM expectations appear anchored in surface-level cues rather than flexible pragmatic reasoning, requiring explicit discursive context when such cues are less available in training input (aligning with Defaultist frameworks [3]; cf. [7] for convergent evidence from negative strengthening). By comparing humans and LLMs, studies like this one help address a larger question in pragmatics, namely how much of pragmatic reasoning is derivable from training data alone, and how much relies on situated world knowledge or capacities like Theory of Mind that allow interlocutors to consider each other's mental states.

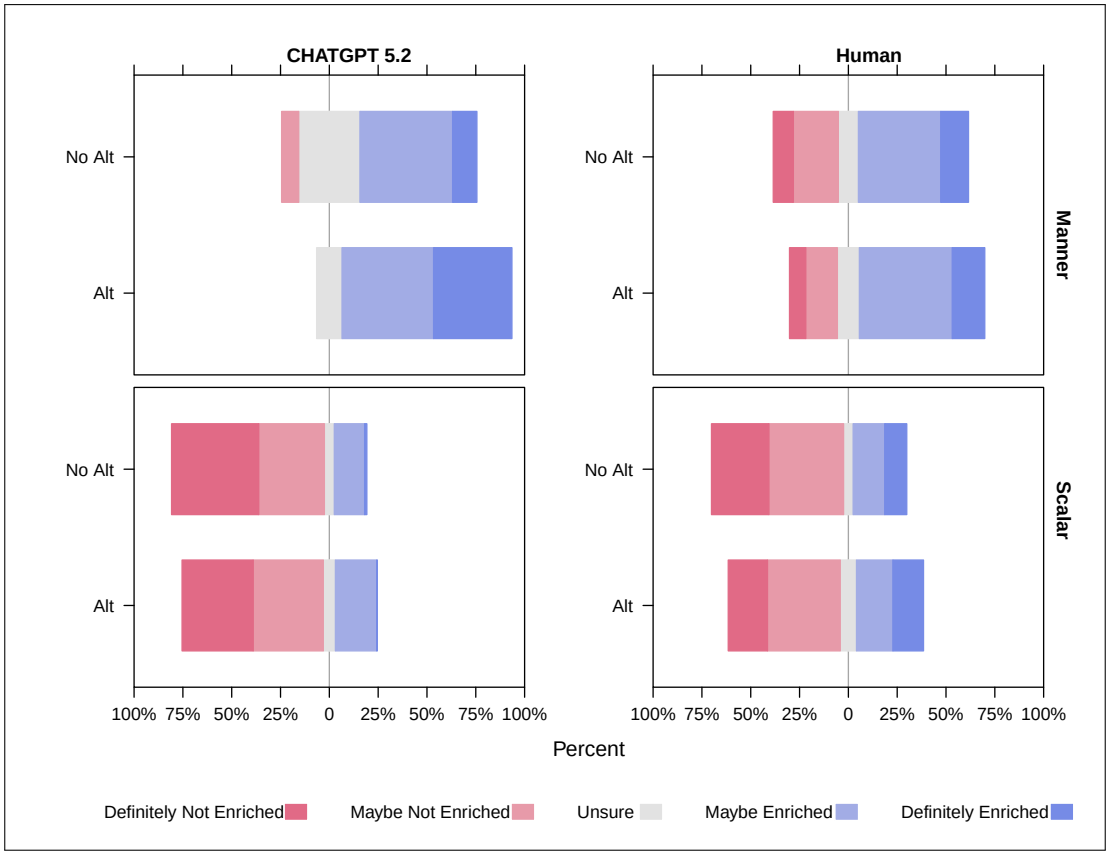

Figure 1: Proportion of Responses across Context, Implicature Type, and Agent

Available Alternative	**No Available Alternative**
Sofia: Is John's haircut any good?	Sofia: Anything new with John?
Ethan: It looks like someone went to his head with electric-powered cutting implements.	Ethan: It looks like someone went to his head with electric-powered cutting implements.
Could Ethan be conveying that John's haircut looks bad?	**Could Ethan be conveying that John's haircut looks bad?**

Table 1: Manner implicature item pair across No Available Alternative and Alternative conditions.

Available Alternative	**No Available Alternative**
Jeanne: Toby's house is tiny, right?	Jeanne: Have you been to Toby's house?
Manu: It's small.	Manu: It's small.
Could Manu be conveying that Toby's house is tiny?	**Could Manu be conveying that Toby's house is tiny?**

Table 2: Scalar implicature item pair across No Available Alternative and Alternative conditions.

[1] Grice, H. P. (1975). Logic and Conversation. In *Syntax and Semantics, Volume 3: Speech Acts* (pp. 41–58). Academic Press. [2] Wilson, E., & Katsos, N. (2016). In a manner of speaking: An empirical investigation of manner implicatures. *Pre-proceedings of Trends in Experimental Pragmatics*, 170–176. [3] Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. The MIT Press. [4] Cong, Y. (2024). Manner implicatures in large language models. *Scientific Reports*, *14*(1), 29113. [5] Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (2. ed., Repr). Blackwell. [6] Khorsheed, A., & Gotzner, N. (2023). A closer look at the sources of variability in scalar implicature derivation: A review. *Frontiers in Communication*, *8*, 1187970. [7] Kohler, J., & Gotzner, N. (2026). Polarity asymmetries in large language models: Human's and GPT's interpretation of indirect language use [Preprint].