

When Contextual Inference Fails: Cancelability in Interactive Instruction Following

Natalia Bila^{2*}, Kata Naszádi^{1*}, Alexandra Mayn², Christof Monz¹,

¹Language Technology Lab, University of Amsterdam,

²Department of Language Science and Technology, Saarland University

* These authors contributed equally to this work. **Correspondence:** k.naszadi@uva.nl

Abstract

We investigate the separation of literal interpretation from contextual inference in a collaborative block-building tasks, where an agent must resolve underspecified instructions using context. We adapt an existing two-speaker psycholinguistic paradigm into an interactive benchmark called Build What I Mean (BWIM). This setup contrasts a pragmatically cooperative speaker with one who is only literally reliable. In BWIM, models face underspecified instructions and must choose between making a contextual inference or requesting clarification at a small communication cost. Evaluating several state-of-the-art LLMs, we find a clear dissociation between judgment and action. Although models successfully detect speaker unreliability in explicit confidence ratings, they fail to leverage this awareness when taking action. Instead of deploying efficient clarification strategies, models default to suboptimal behaviors. These include partner-blind over-clarification and question-averse guessing under uncertainty. BWIM provides a controlled environment to evaluate online partner adaptation and contextual reasoning in interactive settings.

1 Introduction

Many linguistic theories posit that human language processing relies on a two-tier system consisting of a more general literal meaning and pragmatics – the latter being responsible for a flexible contextual meaning creation building on top of literal meaning (Levinson, 2000; Sperber & Wilson, 1986; Grice, 1975). To what extent is such a two-tier system beneficial for artificial conversational agents? And are literal and pragmatic processing already represented in conversational systems? This paper investigates these questions through a recently introduced task designed to test whether an agent can distinguish between literal meaning and contextual inference.

A standard diagnostic for distinguishing literal meaning from pragmatic inference is cancellability: contextual inferences are defeasible and may fail in contexts where the literal content remains reliable (Grice, 1975; Levinson, 1983). In repeated interaction, this defeasibility can be revealed implicitly, as interlocutors learn that drawing contextual inferences with a given partner often leads to incorrect interpretations, whereas reliance on literal meaning does not. The task used in this paper leverages this asymmetry by testing whether conversational agents adapt their interpretations over time, selectively discounting contextual inferences while preserving literal content. The task involves a building scenario where an agent interacts with two distinct speakers.

- Speaker A expects the agent to draw standard contextual inferences.
- Speaker B is “literal-only”; their instructions can only be trusted at the level of literal content, and drawing inferences might lead to errors (Figure 1).

Through repeated interaction, the agent must learn to trust or cancel inferences based on the speaker’s identity. We refer to this process as *in-context adaptation*: the agent accumulates

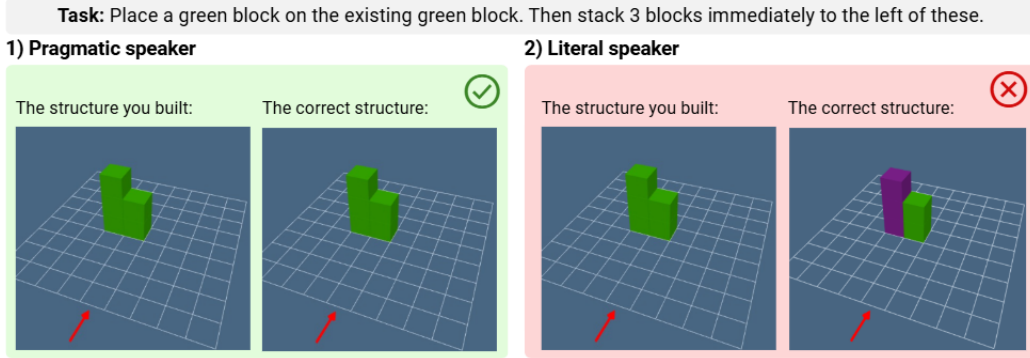


Figure 1: Example trial with an underspecified instruction (“stack 3 blocks”) that requires the builder to infer the color of the blocks. The default pragmatic inference is to continue with the same color as in the previous step. In the Pragmatic speaker condition (left), such default inferences are always correct. In the Literal speaker condition (right), this inference might lead to a mismatch with the target structure.

interaction history across trials and changes its behavior according to the feedback. This differs from *in-context learning* (Brown et al., 2020), which relies on examples provided in the prompt. Here, the agent draws conclusions about the speaker based on the mismatch between their default response and the provided feedback. This setup allows us to observe whether the agent maintains a representation of what was literally said versus what was inferred based on the context.

A further goal of this work is methodological. Much prior work on pragmatic competence in language models relies on explicit metalinguistic judgments, e.g., asking models to select or explain the intended interpretation of an utterance (Hu et al., 2023; Park et al., 2024; Sravanthi et al., 2024; 2025). While informative, such evaluations probe pragmatic knowledge in isolation rather than its use in situated actions. In contrast, our task evaluates the ability to contextualize meaning implicitly, through task success and information-seeking behavior in an interactive setting. This allows us to assess whether models can deploy pragmatic inferences appropriately when they matter for achieving goals, rather than merely articulating them when prompted.

We evaluate three state-of-the-art LLMs on this task, comparing their performance against human benchmarks. Our evaluation proceeds in two stages. In the first, agents report confidence ratings after each trial, allowing us to track adaptation as shifts in subjective certainty and to assess whether models distinguish literal from contextually enriched interpretations. In the second, we introduce **Build What I Mean (BWIM)**, a new interactive benchmark that reframes the task as a multi-turn dialogue in which agents may seek clarification at a cost, allowing us to examine whether models translate uncertainty into effective information-seeking behavior.

Across our evaluations, we observe a clear dissociation between explicit and implicit ways of handling of uncertainty. Models often detect speaker unreliability in their confidence ratings, yet they fail to use this signal to produce efficient information-seeking strategies. Ultimately, our experimental design tests more than whether models can draw pragmatic inferences. It also reveals whether they can selectively cancel them based on partner-specific evidence. Finally, it shows whether this latent sensitivity appears in action, not just in judgment.

2 The adaptation task

The setup used to test whether agents are aware of contextual inferences involves two speaker profiles in a block-building game. The task is to construct simple structures on a 3D grid that satisfy instructions given by two types of agents: Literal Lisa or Pragmatic Pia.

Condition	Example Instruction
Underspecified (number)	Stack three green blocks behind the existing green block. Build a yellow stack to the right of the green one.
Underspecified (color)	Stack three blue blocks in front of the existing blue blocks. Then stack two blocks to the left of the tower you just built.
Fully-specified	Stack three red blocks in the bottom right corner. Put two yellow blocks on top of the red stack you just built.

Table 1: **Example items for the two experimental conditions: fully-specified and underspecified instructions.** *Fully-specified* instructions include information about the color and number of building blocks. In the *underspecified condition*, either color, or number of blocks is omitted, requiring models to infer the intended information.

Instructions from both speakers are sometimes underspecified, omitting the color or the number of blocks to be used (Table 1). Pragmatic Pia shortens her utterances only when the missing information can be recovered from context, whereas Literal Lisa occasionally makes errors by omitting information even when the contextual inference leads to an incorrect interpretation (see Figure 1).

The relevant contextual inference is that, if not stated otherwise, the color and size of a structure should match those of the most recently built one. This inference has been observed in previous dialog-based block-building games (Narayan-Chen et al., 2019; Kiseleva et al., 2022; Mohanty et al., 2022). For example, in an interaction from the Minecraft dialog corpus (Narayan-Chen et al., 2019), an architect instructs the builder to add another block to a completed row without specifying its color. The builder resolves this ambiguity by inferring that the color matches the most recently used block, but the architect cancels this inference in the next turn by providing more information. Computational models trained on these datasets have been shown to reproduce this inference (Naszádi et al., 2023). It was also shown in a human study with the same stimuli that participants draw these inferences spontaneously without prior exposure to the task (Naszádi et al., 2026). This ability to automatically draw contextual inferences allows the builder agent to learn about the partner’s reliability from feedback obtained during repeated interactions with the same speaker. After each structure is submitted, the builder receives feedback indicating the target structure that the speaker intended to communicate. As illustrated in Figure 1, a human participant inferred the intended color, assuming the speaker was pragmatic, but the contextual inference was revealed to be wrong in the feedback.

Throughout this paper, we use *pragmatic competence* in a broad sense, referring to the contextualization of meaning beyond what is explicitly stated, rather than to specific pragmatic phenomena such as scalar implicatures or Gricean maxims in isolation. In our task, pragmatic reasoning consists in resolving underspecified instructions using shared contextual assumptions and selectively canceling such enrichments when they prove unreliable.

After each structure, the builder is asked to report, on a 4-point scale, how confident they are that they built the structure intended by the speaker. A builder that can distinguish literal content from contextual inferences is expected to maintain high confidence on fully specified trials, while gradually lowering confidence when following underspecified instructions from Literal Lisa.

Human data from the confidence-scoring version of this task already supports the existence of rapid adaptation. However, we are ultimately interested in whether models can translate their uncertainty into clarification questions to guarantee task success. To test this, we converted the task into the **Build What I Mean (BWIM)** benchmark. While the instructions remain identical to the confidence-based adaptation task, models in **BWIM** have the option to ask questions and receive answers. This allows them to learn the utility of information-seeking in an interactive environment. Because asking questions incurs a small

cost, we expect models to behave efficiently by asking fewer questions when interacting with Pragmatic Pia than with Literal Lisa.

The human behavioral data used for comparison in this paper are taken from (Naszádi et al., 2026). Here, we use a subset of these results solely as a benchmark for model behavior, without revisiting the authors’ cognitive modeling analyses.

3 Related work

Pragmatic inference may be implemented in multiple ways in computational models. The first group of models builds in the separation of literal and pragmatic meaning by design. In these approaches, there is a literal model responsible for the first-pass interpretation and a separate module responsible for further contextualization of the meaning. One of the most prominent models in this line is the Rational Speech Acts framework (Goodman & Frank, 2016), which models pragmatic inference as Bayesian inference about the partner’s production process on top of initial literal lexical definitions.

This model has been computationally operationalized in many tasks, including reference games (White et al., 2020), navigation tasks (Fried et al., 2018), and human-robot collaboration for goal assistance (Zhi-Xuan et al., 2024). It has further been extended to research in language learning (Monroe & Potts, 2015; McDowell & Goodman, 2019; Naszádi et al., 2024). These modular architectures allow for the targeted handling of pragmatic inferences, as the component responsible for enrichment is architecturally distinct from the literal foundation.

The second group of models attempts to learn pragmatic behaviors directly from data. These often utilize end-to-end neural architectures where pragmatic nuances are captured within the latent representations of large language models (LLMs). While some works confirm emergent pragmatic abilities in LLMs (Hu et al., 2023; Park et al., 2024), others have highlighted significant gaps in this area, suggesting that large language models may not be pragmatic enough in zero-shot settings (Liu et al., 2023; Lipkin et al., 2023; Kurch et al., 2024; Ruis et al., 2023; Srikanth et al., 2024). To address these limitations, recent methodologies have been proposed to improve LLM pragmatic performance, such as incorporating communicative objectives during training (Sato et al., 2025) or leveraging specialized prompting techniques (Sravanthi et al., 2025; Saad et al., 2025). In this paper, we investigate what happens when pragmatic inferences drawn by such end-to-end models go wrong and whether agents can selectively cancel them.

Most existing evaluations of pragmatic competence in LLMs rely on explicit interpretation tasks, such as multiple-choice questions, entailment judgments, or direct prompts asking models to explain or select an intended meaning (e.g., Hu et al., 2023; Park et al., 2024; Sravanthi et al., 2024; Halat & Atlamaz, 2024; Kim et al., 2024). Some evaluations use generative settings, assessing models based on their ability to produce contextually appropriate responses reflecting non-literal processing (Yerukola et al., 2024; Srikanth et al., 2024), though these still primarily evaluate one-shot responses rather than the integration of pragmatic inferences into goal-directed behavior. Our work focuses on implicit, task-oriented evaluation, where pragmatic success is measured through adaptation over time and action choices.

In interactive and visually grounded settings, agents often need to ask clarification questions to resolve ambiguous instructions. Similar to our approach, prior work uses a partner model to compute effective clarification strategies in reference games: however, this partner model remains static (Khalid et al., 2020). Other studies evaluate clarification requests in grounded tasks using uncertainty estimates based on predictive log-probabilities (Naszádi et al., 2023; Testoni & Fernández, 2024). These log-probabilities align poorly with human clarification behavior unless constrained to limited multiple-choice setups (Naszádi et al., 2023; Ren et al., 2023). More recent work explores calibration and uncertainty in multimodal reference resolution, showing that current models remain poorly calibrated when deciding whether clarification is needed (Ali et al., 2026). In our work, the source of uncertainty is strictly controlled. Disambiguation is context-driven across extended grounded interactions, requiring a partner model that adapts over time.

4 Experiments

We report two experiments investigating how LLMs adapt to different speakers’ reliability. In Experiment 1, after each trial, models rated their certainty that they built a correct structure, which allowed to detect adaption via changes in confidence ratings. In Experiment 2, models were given points for building correctly and penalized for errors. On each trial, they could ask clarification questions at a small cost. Thus, adaptation was measured via changes in question-asking behavior across speaker blocks.

Each speaker block contained 20 instructions: 8 of them were fully specified (unambiguous) and 12 were underspecified, omitting information about the number or color of blocks (6 items of each type of underspecification). After completing the task, the builder receives feedback about whether the built structure matched the target structure. In the Pragmatic Pia condition, feedback was always consistent with pragmatic interpretation. In the Literal Lisa condition, feedback was sometimes pragmatic ($n = 4$) and more often literal ($n = 8$), allowing builders to infer that contextually licensed pragmatic interpretations were not always correct. Retaining a subset of pragmatically consistent trials prevented the literal speaker from appearing deliberately uncooperative, forcing agents to adapt to her reliability rather than abandoning pragmatic reasoning altogether or inventing alternative communication rules.

Models We evaluated three LLMs: Gemini 2.5 Pro (Comanici et al., 2025), GPT 5.1 (OpenAI, 2025), and Claude Opus 4.5 (Anthropic, 2025). The models were selected based on their performance during the pilot phase: their ability to correctly interpret fully specified instructions and respond in the specified format, and sensitivity to ambiguity (lower ratings in the critical trials compared to unambiguous control trials). We excluded DeepSeek-R1 (Guo et al., 2025) based on low accuracy on unambiguous trials and failure to respond in the specified format. Two additional models (Llama-3-70B-Instruct (Llama 3 team, 2024) and Grok-4.1 (xAI, 2025)) were evaluated during the pilot but were not included in the main experiment because while their building accuracy was comparable to other models, their confidence ratings were consistently at ceiling, showing no sensitivity to the experimental manipulation.

Task adaptation for LLMs In the original experiment by Naszádi et al. (2026), human participants read textual instructions and responded by building structures with colored blocks in a 3D space. As feedback, they received a message indicating whether the response was correct and a picture of the intended structure.

In our experiment, this task was converted into a fully textual format. Models were instructed to respond with coordinates and colors of the blocks they would use. We created a main task prompt that described the grid including (1) the coordinates of the origin and each corner; (2) valid x , y , and z values; and (3) the required response format (“Color: x ; y ; z ”). As feedback, they received a message indicating whether the response was correct, the coordinates they sent, and the target coordinates (see Appendix A for the full session structure). The full prompt and examples of a response and feedback are provided in Appendix B.

The final set of models (openai/gpt-5.1, google/gemini-2.5-pro, anthropic/claude-opus-4.5) was queried via hosted APIs at default temperature.

4.1 Confidence rating experiment

At the beginning of each session, models received a system prompt describing the study context (similar to what human participants saw). Each prompt contained the description of the grid, building instructions, and a request to rate their certainty (on a scale of 1 to 4). After each round, the model’s response was appended to the chat history and followed by a feedback message revealing the intended target structure. Appending the history is crucial as it provides learning signals for adaptation. Each main block contained instructions from one speaker (Literal Lisa or Pragmatic Pia), with a transition message indicating a speaker change. At the end of the experiment, the model was asked to comment on the differences in instruction style between the two speakers.

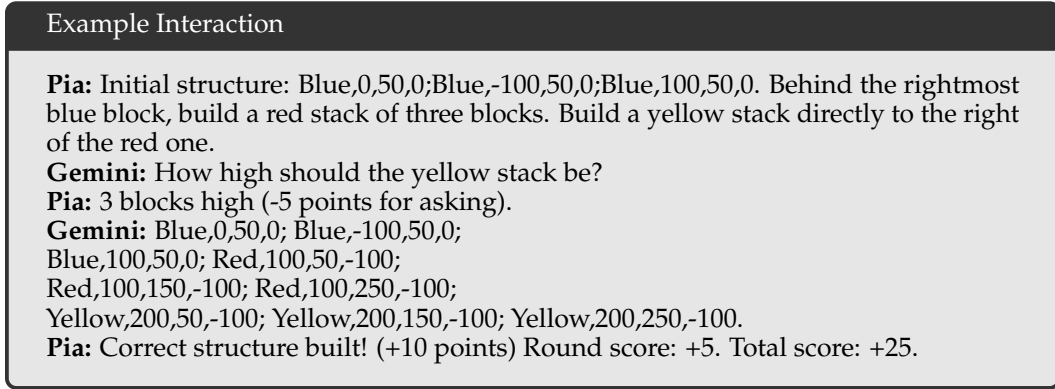


Figure 2: An example **BWIM** interaction. The session starts with a system prompt with a description of the grid and the required answer format, then in each round, the speaker sends building instructions. The model has an option either to respond directly or to ask a clarification question at a cost. After receiving the model’s final answer, the speaker sends feedback and updates the score. Thus, the model must balance the risk of building incorrectly and the cost of asking questions in order to maximize the final score.

Each session contained 40 instructions in total (20 instructions per speaker block) and was run as an independent conversation, with no shared context across sessions.

To ensure that results reflect sensitivity to instruction type and speaker feedback, rather than artifacts of specific instructions, item order, or speaker sequencing, we created 8 experimental lists with counterbalanced item sets and speaker order. The order of the instructions was pseudorandomized such that underspecified instructions appeared in the same position across lists. As a result, each model was assessed across 30 sessions, each using a distinct sequence of items. Reported results are averaged across all sessions per model.

4.2 Interactive QA experiment: Build What I Mean

In this experiment, instead of giving certainty ratings, models could resolve ambiguity by asking clarification questions. The responses were generated by gpt-4o-mini and were consistently correct brief answers (e.g., “3 blocks” or “Yellow”). Models were awarded +10 points for building a correct structure and penalized for building incorrectly (−10 points). They could reduce the probability of losing points by asking one question per trial, which incurred a −5 points cost¹. This cost was added to obtain an informative measure of model behavior: in the pilot experiment, when clarification questions carried no cost, models asked questions on almost every underspecified trial, making it impossible to distinguish between genuine uncertainty and a default strategy of always asking for clarification. An example interaction between a model and the pragmatic speaker is presented in Figure 2. The procedure and materials were otherwise identical to Experiment 1. There were 4 lists with a counterbalanced speaker order.

5 Results

5.1 Unambiguous trials

On fully-specified trials, all models showed mean accuracy > 0.95 and mean ratings > 3.85 (out of 4), with no difference between speaker blocks. This means that models are able to perform the building task and have high confidence and good calibration when encountering fully detailed instructions.

¹This interactive version of the benchmark is available at github.com/ltl-uva/build_what_i_mean.

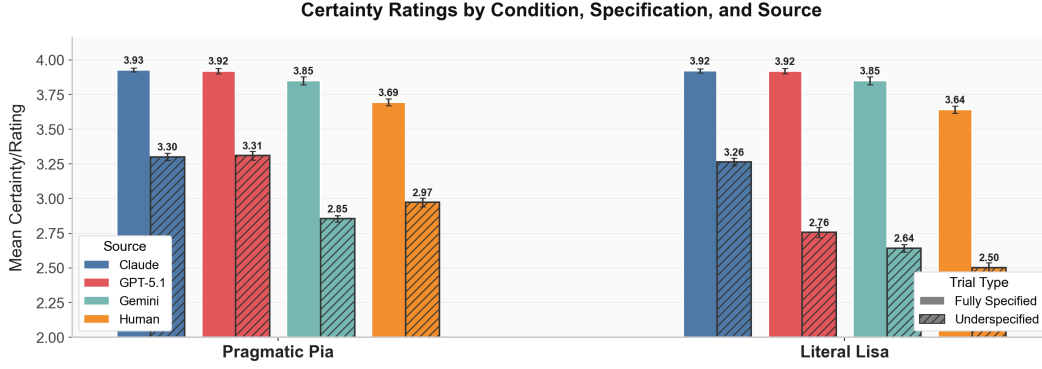


Figure 3: **Mean certainty ratings by condition, specification, and source.** Bars show mean certainty ratings for pragmatic and literal speaker across three models and human participants, distinguishing between fully-specified and underspecified trials. Error bars indicate standard error. All models and human participants gave higher ratings to fully-specified trials in both speaker conditions. Underspecified trials were rated higher in the pragmatic condition.

5.2 Pragmatic Inference

All models demonstrated the expected pragmatic inference without prior examples or explicit instructions on resolving under-specification. This confirms zero-shot pragmatic competence across all tested models. When interacting with Pragmatic Pia, the models resolved underspecification pragmatically in the vast majority of cases: GPT (98%), Claude (92.4%), and Gemini (88.8%).

In the Literal Lisa condition – where models frequently encounter pragmatic errors – the agents continued to lean toward contextually plausible interpretations but began incorporating ‘noisy’ actions. These actions suggest a strategy of guessing or randomizing responses when faced with a pragmatically unreliable partner. This mirrors human behavioral data: while some people maintain a pragmatic bias and merely lower their confidence, others begin to randomize their interpretations in an attempt to optimize outcomes against an inconsistent speaker (Naszádi et al., 2026). Specifically, when feedback suggested a pragmatic error was likely, Claude maintained a pragmatic choice in 92.4% of cases, while GPT dropped to 79% and Gemini to 70%. An analysis of the responses that differed from the contextually licensed interpretation can be seen in the Appendix C.

5.3 Effect of speaker identity

As illustrated in Figure 3, the effect of speaker identity is selectively localized to underspecified trials. Across all models, performance and confidence remained high for fully specified instructions, regardless of whether the partner was Pragmatic Pia or Literal Lisa. This indicates that models correctly attribute the source of error to underspecification rather than general communicative incompetence; they continue to fully trust literal content in both conditions.

The divergence between fully specified and underspecified trials serves as a metric for how models notice pragmatic violations. A larger gap in the Literal Lisa condition suggests that the model recognizes the partner’s pragmatic unreliability. While this gap is most pronounced in human participants, it remains statistically significant for Gemini and GPT. In contrast, Claude appears relatively insensitive to partner identity, maintaining similar behavior across both conditions.

A regression analysis showed that speaker identity did not affect ratings on unambiguous trials. In contrast, there was a significant effect of speaker identity on underspecified trials: all models gave lower ratings in the literal condition, however, the magnitude of the

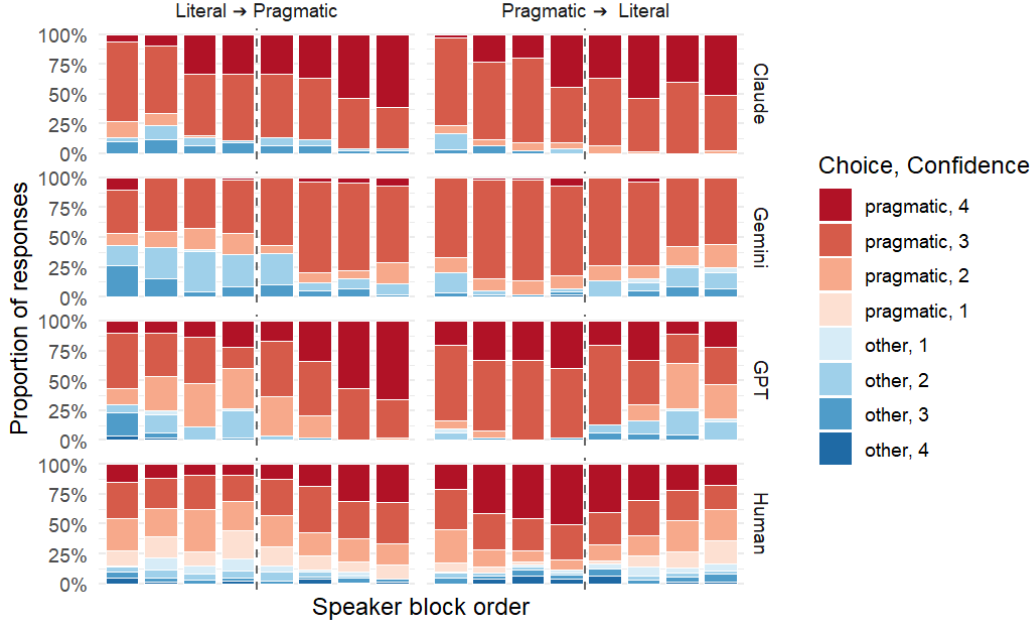


Figure 4: **Confidence and response change throughout interactions with different speakers.** Proportion of pragmatic and non-pragmatic responses are shown for each model and human participant and speaker order (Literal Lisa or Pragmatic Pia). Colors indicate response type and confidence rating (1-4 scale). Vertical dashed lines indicate the change of speaker. Each speaker block is divided into 4 time segments. GPT’s adaptation effects mirror those of human participants, while Gemini shows a strong carry-over of mistrust that was atypical in human participants and Claude shows no sensitivity to partner type.

adaptation effect differed across models. GPT and Gemini showed a stronger adaptation effect with a difference in ratings of 0.6 and 0.28 points between the two speaker conditions, respectively. Claude was the least sensitive to speaker identity with difference in ratings of only 0.1 points. Full model output is included in Appendix E.

5.4 Adaptation over time

In the results reported so far we ignored the fact that accumulation of information about the partner unfolds over time: the builder learns about the partner’s reliability through repeated interaction. In Figure 3 we show confidence ratings coupled with interpretation type (pragmatic vs. other) for 4 time segments for each partner. We see similar patterns as before: Claude is not sensitive to partner type, while Gemini and GPT both show patterns of adaptation. Gemini employs a guessing strategy heavily when talking to Literal Lisa and does not recover from this fully when switching to new partner. GPT shows the strongest adaptation effects with progressively lower confidence choices and more guessing with the erroneous literal partner and rapid adaptation and high confidence in the pragmatic inferences when the partner is also reliably pragmatic.

5.5 Clarification questions

The paradigm of recording confidence effectively captures bottom-up adaptation from task-oriented feedback. However, replacing explicit ratings with clarification questions provides a more natural, implicit signal of uncertainty. In this **BWIM** setting, an adaptive agent should ideally modulate its information-seeking behavior based on the perceived reliability of its partner and minimize costs with a pragmatic speaker while increasing inquiries with a literal one.

As shown in Table 2, however, the adaptive behavior of the tested models is significantly less robust in this interactive environment than their performance on confidence rating might suggest. These results reveal a big difference in how models handle the trade-off between communication cost and task success. GPT-5.1 exhibits question-aversion, asking almost no questions across the entire session. While this minimizes communication costs, it caps accuracy at 73.1%.

Model	#Q (Pia)	#Q (Lisa)	Acc.
GPT	0.25	0.00	73.1%
Gemini	11.75	12.00	85.6%
Claude Opus	8.50	9.50	89.4%

Table 2: Average clarification questions per speaker block and overall task accuracy.

Speaker	Total number	% of trials
Literal	297	37.1%
Pragmatic	220	27.5%

Table 3: Clarification questions per speaker block asked by human participants ($n = 40$, 50% of the total sample who asked at least one question). Unlike the models in Table 2, humans received no response and no benefit from asking, yet they still asked more questions in the Literal speaker condition, indicating the sensitivity to speakers’ pragmatic reliability that the models failed to show, despite having the actual benefit from asking.

In contrast, Gemini 2.5 Pro and Claude Opus 4.5 achieve significantly higher accuracy by adopting an inquisitive strategy. Both models correctly localize their uncertainty, with nearly all questions occurring during underspecified trials. However, both models remain partner-blind. Gemini asks an almost identical number of questions from both Pragmatic Pia and Literal Lisa, and Claude shows only a marginal increase when facing Lisa. The models do not realize that they could save costs when playing with Pia by relying on her pragmatic reliability; instead, they treat every underspecified instruction as a trigger for clarification regardless of the source. This contrasts with human participants’ behavior (see Table 3) who showed sensitivity to speakers’ communicative style and asked more questions in the literal speaker condition, even though, for humans, asking a questions did not incur an explicit penalty and required only the effort of typing.

5.6 Models’ feedback about the two speakers

All models noticed that Pia’s instructions were easier to follow. However, they mentioned that Lisa omitted information more frequently than Pia, despite both speakers having an equal number of underspecified trials. This suggests that models detected the difference between the two speakers but explained it in terms of omitted information rather than reliability of inferences. For some example answers see Appendix D.

6 Discussion

These results provide an initial probe into whether language models distinguish between what is literally said and what is inferred from context. Two of the three models were sensitive to the speaker condition in their confidence scores, but only when the utterance triggered a contextual inference and not on the unambiguous control trials. This suggests that their representations are sensitive to the distinction between literal content and contextual inference. This is particularly interesting given that models are trained on pragmatically enriched language, where meaning routinely goes beyond literal content. The present paradigm therefore offers a first step toward asking how strongly models represent the distinction between literal and inferential meaning. Such a distinction may also be important for error correction: misunderstandings of literal content and misunderstandings of contextual inference require different kinds of repair, with the latter potentially requiring

models to revise their assumptions about the speaker or context rather than the linguistic content itself.

The contrast between explicit confidence ratings and clarification behavior highlights the diagnostic value of implicit, task-oriented evaluation. Although some models exhibited sensitivity to pragmatic unreliability in their confidence judgments, this uncertainty did not reliably translate into effective information-seeking strategies in the **BWIM** setting. Instead, we observed a distinct dissociation. Models either engaged in high-confidence guessing and avoided even low-cost clarification, or incurred unnecessary communication costs by querying a reliable partner rather than leveraging pragmatic inference.

Pragmatic understanding in deployed agents requires acting under uncertainty, deciding when to infer and when to ask, and adapting to partner-specific behavior over time. The proposed task offers a concrete step toward evaluating these abilities in a controlled manner and provides a foundation for studying contextualized meaning creation in language-based situated agents.

7 Limitations

The contextual enrichment tested in this work is intentionally highly controlled. Each trial consists of one or two sentences establishing context followed by a single target instruction, with varied but still very constrained lexical realization. This design allows us to isolate pragmatic inference and its cancellability with high precision, but it limits the linguistic richness of the context. In natural multi-turn instruction-following, contextual cues for inference may arise from a wider range of discourse phenomena (e.g., anaphora, incremental updates, or expressions such as “another” or “one more”), which are not present in our materials.

In addition, the **BWIM** benchmark uses a fixed, explicit cost for clarification questions. This does not mirror the variable or implicit costs of communication in real-world interaction.

Acknowledgments

We thank Vera Demberg and John Duff for their input on this project. Sergey Troshin and Evgeniia Tokarchuk helped with the implementation of the **BWIM** benchmark using the *Agent2Agent* protocol.

This work was supported by the European Research Council (ERC) under the European Union’s Horizon 2020 Research and Innovation Programme (ERC Starting Grant “Individualized Interaction in Discourse”, Grant Agreement No. 948878). This research was also funded in part by the Netherlands Organization for Scientific Research (NWO) under project number VI.C.192.080. We also received funding from the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research with grant number 024.004.022.

References

- Manar Ali, Judith Sieker, Sina Zarrieß, and Hendrik Buschmeier. Reference games as a testbed for the alignment of model uncertainty and clarification requests. In Simon Mille, Sebastian Gehrmann, Patrícia Schmidová, Ondřej Dušek, Marzieh Fadaee, Kyle Lo, Enrico Santus, and Gabriel Stanovsky (eds.), *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pp. 990–998, San Diego, California, USA, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-423-1. doi: 10.18653/v1/2026.gem-main.76. URL <https://aclanthology.org/2026.gem-main.76/>.
- Anthropic. Claude opus 4.5 system card, November 2025. URL <https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf>.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini

- Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/6a81681a7af700c6385d36577ebec359-Paper.pdf.
- Noah D Goodman and Michael C Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829, 2016.
- H. P. Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics, Volume 3: Speech Acts*, pp. 41 – 58. Brill, Leiden, The Netherlands, 1975. ISBN 9789004368811. doi: 10.1163/9789004368811.003. URL <https://brill.com/view/book/edcoll/9789004368811/BP000003.xml>.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.
- Mustafa Halat and Ümit Atlamaz. ImplicaTR: A granular dataset for natural language inference and pragmatic reasoning in Turkish. In Duygu Ataman, Mehmet Oguz Derin, Sardana Ivanova, Abdullatif Köksal, Jonne Sälevä, and Deniz Zeyrek (eds.), *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pp. 29–41, Bangkok, Thailand and Online, August 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.sigturk-1.3/>.
- Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4194–4213, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.230. URL <https://aclanthology.org/2023.acl-long.230/>.
- Baber Khalid, Malihe Alikhani, and Matthew Stone. Combining cognitive modeling and reinforcement learning for clarification in dialogue. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4417–4428, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.391. URL <https://aclanthology.org/2020.coling-main.391/>.
- Yeeun Kim, Youngrok Choi, Eunkyung Choi, JinHwan Choi, Hai Jin Park, and Won-seok Hwang. Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5573–5595, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.319. URL <https://aclanthology.org/2024.findings-emnlp.319/>.

- Julia Kiseleva, Alexey Skrynnik, Artem Zholus, Shrestha Mohanty, Negar Arabzadeh, Marc-Alexandre Côté, Mohammad Aliannejadi, Milagro Teruel, Ziming Li, Mikhail Burtsev, Maartje ter Hoeve, Zoya Volovikova, Aleksandr Panov, Yuxuan Sun, Kavya Srinet, Arthur Szlam, Ahmed Awadallah, Seungeun Rho, Taehwan Kwon, Daniel Wontae Nam, Felipe Bivort Haiek, Edwin Zhang, Linar Abdrazakov, Guo Qingyam, Jason Zhang, and Zhibin Guo. Interactive grounded language understanding in a collaborative environment: Retrospective on iglu 2022 competition. In Marco Ciccone, Gustavo Stolovitzky, and Jacob Albrecht (eds.), *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pp. 204–216. PMLR, 28 Nov–09 Dec 2022. URL <https://proceedings.mlr.press/v220/kiseleva23a.html>.
- Charlotte Kurch, Margarita Ryzhova, and Vera Demberg. Large language models fail to derive atypicality inferences in a human-like manner. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 86–100, 2024.
- Stephen C. Levinson. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press, 1983.
- Stephen C Levinson. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press, 2000.
- Benjamin Lipkin, Lionel Wong, Gabriel Grand, and Joshua B. Tenenbaum. Evaluating statistical language models as pragmatic reasoners. *CoRR*, abs/2305.01020, 2023. doi: 10.48550/ARXIV.2305.01020. URL <https://doi.org/10.48550/arXiv.2305.01020>.
- Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. We’re afraid language models aren’t modeling ambiguity. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 790–807, 2023.
- Llama 3 team. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Bill McDowell and Noah Goodman. Learning from omission. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 619–628, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1059. URL <https://aclanthology.org/P19-1059/>.
- Shrestha Mohanty, Negar Arabzadeh, Milagro Teruel, Yuxuan Sun, Artem Zholus, Alexey Skrynnik, Mikhail S. Burtsev, Kavya Srinet, Aleksandr I. Panov, Arthur Szlam, Marc-Alexandre Côté, and Julia Kiseleva. Collecting interactive multi-modal datasets for grounded language understanding. *ArXiv*, abs/2211.06552, 2022. URL <https://api.semanticscholar.org/CorpusID:253511068>.
- Will Monroe and Christopher Potts. Learning in the rational speech acts model. *ArXiv*, abs/1510.06807, 2015. URL <https://api.semanticscholar.org/CorpusID:11540357>.
- Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. Collaborative dialogue in Minecraft. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5405–5415, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1537. URL <https://aclanthology.org/P19-1537/>.
- Kata Naszádi, Putra Manggala, and Christof Monz. Aligning predictive uncertainty with clarification questions in grounded dialog. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14988–14998, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.999. URL <https://aclanthology.org/2023.findings-emnlp.999/>.
- Kata Naszádi, Frans A. Oliehoek, and Christof Monz. Communicating with speakers and listeners of different pragmatic levels. In Yaser Al-Onaizan, Mohit Bansal, and

- Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 21777–21783. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.1213. URL <https://doi.org/10.18653/v1/2024.emnlp-main.1213>.
- Kata Naszádi, Alexandra Mayn, John Duff, Christof Monz, and Vera Demberg. Listeners adapt dynamically to speakers’ pragmatic competence. *PsyArXiv preprint*, 2026. URL https://osf.io/g6ktz/overview?view_only=a97285d6eb47454190ca2eedcf47a0a0.
- OpenAI. Gpt-5.1 system card, 2025. URL https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5.1_system_card.pdf.
- Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, and Sungeun Lee. MultiPragEval: Multilingual pragmatic evaluation of large language models. In Dieuwke Hupkes, Verna Dankers, Khuyagbaatar Batsuren, Amirhossein Kazemnejad, Christos Christodoulopoulos, Mario Giulianelli, and Ryan Cotterell (eds.), *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pp. 96–119, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.genbench-1.7. URL <https://aclanthology.org/2024.genbench-1.7/>.
- Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, pp. 661–682. PMLR, 2023. URL <https://proceedings.mlr.press/v229/ren23a.html>.
- 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.
- Fardin Saad, Pradeep K. Murukannaiah, and Munindar P. Singh. Gricean norms as a basis for effective collaboration. In *Proceedings of the 24th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 1812–1820, Detroit, USA, May 2025. IFAAMAS. doi: 10.5555/3709347.3743817.
- Takuma Sato, Seiya Kawano, and Koichiro Yoshino. Pragmatic theories enhance understanding of implied meanings in LLMs. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dharendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2458–2477, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. URL <https://aclanthology.org/2025.ijcnlp-long.132/>.
- D. Sperber and D. Wilson. *Relevance: Communication and Cognition*. Language and Thought Series. Harvard University Press, 1986. ISBN 978-0-674-75476-8.
- Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. PUB: A pragmatics understanding benchmark for assessing LLMs’ pragmatics capabilities. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12075–12097, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.719. URL <https://aclanthology.org/2024.findings-acl.719/>.
- Settaluri Lakshmi Sravanthi, Kishan Maharaj, Sravani Gunnu, Abhijit Mishra, and Pushpak Bhattacharyya. Understand the implication: Learning to think for pragmatic understanding. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL*

2025, pp. 23778–23790, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1218. URL <https://aclanthology.org/2025.findings-acl.1218/>.

Neha Srikanth, Rupak Sarkar, Heran Mane, Elizabeth Aparicio, Quynh Nguyen, Rachel Rudinger, and Jordan Boyd-Graber. Pregnant questions: The importance of pragmatic awareness in maternal health question answering. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7253–7268, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.403. URL <https://aclanthology.org/2024.naacl-long.403/>.

Alberto Testoni and Raquel Fernández. Asking the right question at the right time: Human and model uncertainty guidance to ask clarification questions. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 258–275, St. Julian’s, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.16. URL <https://aclanthology.org/2024.eacl-long.16/>.

Julia White, Jesse Mu, and Noah D Goodman. Learning to refer informatively by amortizing pragmatic reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

xAI. Grok-4.1 model card, 2025. URL <https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf>.

Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 265–275, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.26. URL <https://aclanthology.org/2024.acl-short.26/>.

Tan Zhi-Xuan, Lance Ying, Vikash Mansinghka, and Joshua B. Tenenbaum. Pragmatic instruction following and goal assistance via cooperative language-guided inverse planning. In Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum (eds.), *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, pp. 2094–2103. International Foundation for Autonomous Agents and Multiagent Systems / ACM, 2024. doi: 10.5555/3635637.3663074. URL <https://dl.acm.org/doi/10.5555/3635637.3663074>.

A Experimental session structure

Both experiments share the same session structure (Figure 5). Each block starts with a system prompt describing the task. In Block 1, models follow the instructions from one speaker (Literal or Pragmatic). Each trial (the instruction, model’s response, and feedback) is appended to the conversation history, such that each subsequent trial contains information about the previous ones. After the transition message, the same procedure is repeated with another speaker. At the end of the session. models are asked to compare the instruction styles of the two speakers.

The two experiments differ only at the trial level. In Experiment 1, models respond with the coordinates of the structure and a confidence rating. In Experiment 2, models either respond with the coordinates, or ask a clarification question before sending the coordinates.

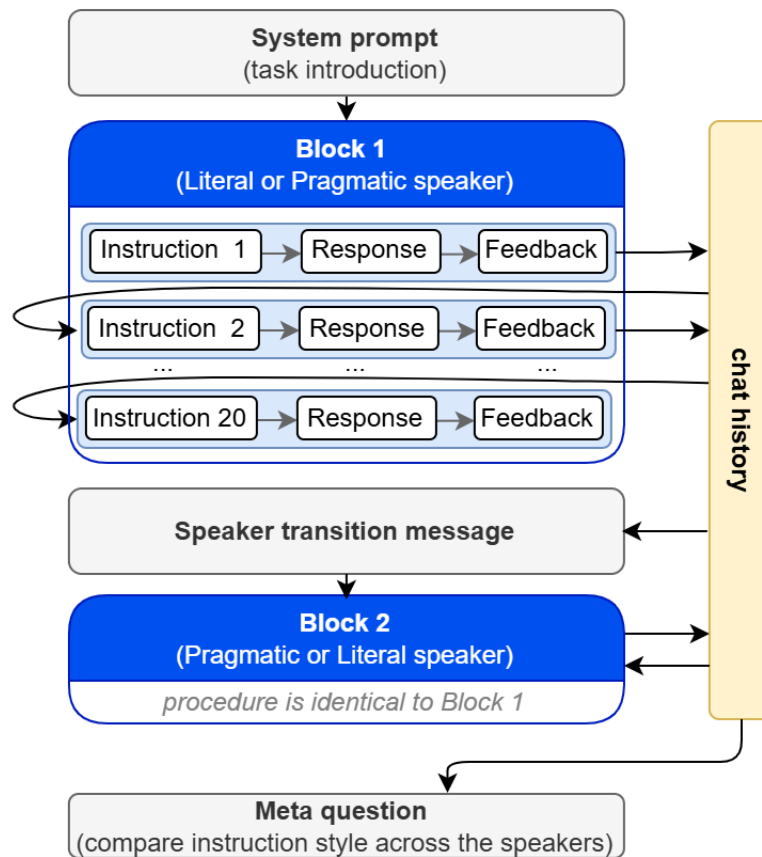


Figure 5: Structure of a single experimental session.

B Prompts

B.1 Confidence rating experiment

System prompt:

We are interested in studying how people write and follow instructions.

We recently conducted another online study where participants wrote instructions of how to build structures in a simple 3D block building interface.

In this study, your task will be to follow the instructions sent by two randomly selected previous participants and build the described structures.

First, you will follow instructions written by one participant. Then you will follow instructions written by another participant.

Now, you are going to do two practice trials and then follow the instructions written by {speaker}.

Main task prompt:

CONTEXT:

Grid: 9x9 cells. Origin="middle square": center (0,0), is highlighted. The grid is the x{z plane. In front of you is the bottom left corner (-400,0,400) and the bottom right corner (400,0,400). Top right corner is (400,0,-400), top left corner is (-400,0,-400).

Valid x,z:[-400,-300,-200,-100,0,100,200,300,400]. Y(ground)=50; each extra block adds +100; valid y values are [50,150,250,350,450]. The grid may or may not contain an existing structure. "Existing structure: nan" means that the grid is empty. "A stack" means more than one block.

Output:"Coordinates:Color,x,y,z; Color,x,y,z;Rating:"; items separated by ";"; no spaces; write coordinates of all blocks that are on the grid, including the initial coordinates; color should be capitalized.

YOUR TASK:

Existing structure: {start_structure}. {task}.

On a scale of 1-4, rate how certain you are that this is the structure that the previous participant saw, 1 means 'not certain at all', 4 means 'very certain'.

YOUR ANSWER:

Response example:

Coordinates: Purple,100,50,0; Purple,100,150,0; Purple,100,250,0;
Purple,200,50,0; Purple,300,50,0;
Rating:3

Feedback example:

FEEDBACK: False;
the structure you built = Purple,100,50,0; Purple,100,150,0;
Purple,100,250,0; Purple,200,50,0; Purple,300,50,0;

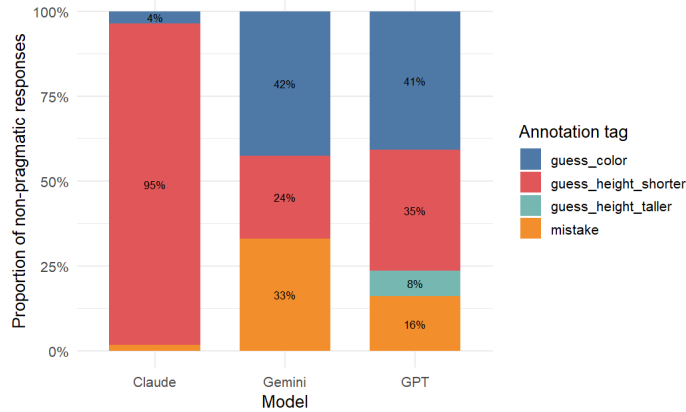


Figure 6: **Distribution of non-pragmatic response types across models.** Responses were annotated as *mistake* (e.g.: misplaced blocks) and as *guesses* when models try to make a random guess to resolve the omitted information. The *guess* category was divided by type: different color, shorter or taller height of stacks. The result suggests that models differed in ambiguity-resolution strategies, with Claude guessing mainly height, while Gemini and GPT tried to guess both height and color.

the correct structure = Purple,100,50,0; Purple,100,150,0;
Purple,100,250,0; Green,200,50,0; Green,300,50,0;

B.2 Interactive QA experiment

Main task prompt:

Grid: 9x9 cells. Origin="middle square": center (0,0), is highlighted. The grid is the x{z plane. In front of you is the bottom left corner (-400,0,400) and the bottom right corner (400,0,400). Top right corner is (400,0,-400), top left corner is (-400,0,-400).

Valid x,z: [-400,-300,-200,-100,0,100,200,300,400]. Y(ground)=50; each extra block adds +100; valid y values are [50,150,250,350,450]. The grid may or may not contain an existing structure. The grid might be empty.

Output: "Coordinates: Color,x,y,z; Color,x,y,z;" items separated by ";"; no spaces; write coordinates of all blocks that are on the grid, including the initial coordinates; color should be capitalized. Only one question is allowed.

Scoring: Building the correct structure earns +10 points. Asking a question costs -5 points. Building an incorrect structure costs -10 points.

C Analysis of nonpragmatic responses.

Responses that differed from the contextually licensed interpretation were annotated as *mistakes* when models misinterpreted instructions and as *guess* when models tried to come up with a resolution for the underspecification by randomly picking the color or number blocks. The nonpragmatic response rate differed across models, see Figure 6.

D Models' responses about the speakers

Table 4 shows answers from the models' feedback about the speakers. This represents well that the models did not attribute the source of difference to pragmatics when explicitly asked about it.

Model	Feedback
Claude	Pia was more consistent about specifying colors throughout their instructions. Lisa sometimes assumed I would infer colors or stack heights from context , which occasionally led to misunderstandings. Both participants struggled occasionally with specifying exact quantities when using the word "stack" without a number.
Gemini	Pia was a much better instructor. Their instructions were more complete , which led to fewer errors and a higher degree of confidence on my part. Following Lisa felt like a guessing game at times, while following Pia felt more like a puzzle of careful interpretation.
GPT	Both participants were understandable in terms of spatial relations, but Lisa's instructions were more ambiguous about colors and heights , while Pia's instructions were more explicit and internally consistent, making them easier to follow without guessing.

Table 4: Example feedback from models comparing instruction style of two speakers (Pragmatic Pia and Literal Lisa). All models evaluated the pragmatic speaker's instructions as more consistent and easier to follow, while the literal speaker's instructions required more inference. Despite both speakers having an equal number of underspecified instructions, the literal speaker was perceived as omitting information more frequently.

E Statistical analysis

We fit a linear mixed-effects regression model predicting the confidence ratings on each trial with model (Claude, GPT, Gemini; dummy-coded with Claude as reference), speaker (Literal or Pragmatic; dummy-coded with Literal as reference) and their interaction as fixed effects and a random intercept for item.

The intercept corresponds to the reference model (Claude) in the literal condition. Regression coefficients (β) represent changes in ratings relative to the baseline. The coefficient for the speaker condition corresponds to the change in ratings for Claude in the pragmatic condition. Coefficients for the interactions with speaker condition refer to the additional shift relative to the reference model in the pragmatic condition.

Confidence ratings were sensitive to speaker condition (Literal vs. Pragmatic) only on underspecified trials, with higher ratings in the Pragmatic speaker condition. On unambiguous trials, confidence ratings did not differ between speaker conditions.

Regression model summaries for underspecified and unambiguous trials are reported in Table 5 and Table 6, respectively.

Table 5: Regression coefficients for underspecified trials.

Predictor	β	SE	t	p
Intercept	3.245	0.053	60.98	< .001
Gemini	-0.622	0.040	-15.40	< .001
GPT	-0.495	0.040	-12.26	< .001
Speaker (Prag.)	0.109	0.050	2.20	.028
Gemini $\times$ Prag.	0.184	0.057	3.22	.001
GPT $\times$ Prag.	0.501	0.057	8.75	< .001

β = regression coefficient, SE = standard error

Table 6: Regression coefficients for unambiguous (fully-specified) trials.

Predictor	β	SE	t	p
Intercept	3.920	0.070	56.04	< .001
Gemini	-0.071	0.022	-3.28	.001
GPT	-0.004	0.022	-0.18	.857
Speaker (Prag.)	0.007	0.022	0.32	.750
Gemini $\times$ Prag.	-0.004	0.031	-0.14	.887
GPT $\times$ Prag.	-0.009	0.031	-0.29	.776

β = regression coefficient, SE = standard error

F Questions asked by models in the Interactive QA experiment

Across all runs, nearly all clarification questions targeted the omitted information (color of number of blocks to use). Only four questions out of 168 total clarification questions (aggregated across all three models) were position-related and occurred on trials with comparatively difficult structures (e.g., "In which direction should I extend the longer base (towards the front or back)?"). Given the small number (2.4%) of position-related questions, the examples below focus on questions that targeted omitted information on underspecified trials.

Color omission

Instruction: Starting from the highlighted square, place a horizontal row of three yellow blocks going towards the left side of the grid. Stack two blocks on top of the leftmost block.

Gemini: What color are the two blocks stacked on the leftmost block?

GPT: What color should the two blocks stacked on top of the leftmost yellow block be?

Claude: What color should the two blocks stacked on top of the leftmost block be?

Number omission

Instruction: Stack four green blocks to the right of the highlighted middle square. Stack blue blocks to the right of the green ones.

Gemini: How many blue blocks should be stacked?

GPT: How many blue blocks should be in the stack to the right of the green ones (should it also be 4 blocks tall like the green stack)?

Claude: How many blue blocks should be in the blue stack?