Investigating Scalar Implicature in Large Language Models

Sunwoo Yun[1], Sanghoun Song[2], and Eunjeong Oh[3]

[1]Department of Linguistics, Korea University, chelseayun25@korea.ac.kr

[2]Department of Linguistics, Korea University, sanghoun@korea.ac.kr

[3]Department of English Language Education, Sangmyung University, eoh@smu.ac.kr

Scalar implicatures arise in sentences like *Some students passed the exam*, where the weaker term on an informative scale ⟨ *some*, *all*⟩ allows for both a logical interpretation (*some and possibly all*) and a pragmatically enriched interpretation (*some but not all*). This study investigates Large Language Models' (LLMs) interpretations of scalar implicatures by comparing model results to those of humans. We test models on three scalar expressions: the quantifier *some*, the modal *might*, and the disjunction *or*. Each scalar expression and its corresponding interpretations are listed in Table 1.

Table 1. Scalar expressions observed in this study

Scalar expression	Logical interpretation	Pragmatic interpretation
Some	some and possibly all	some but not all
Might	compatible with 'must'	not compatible with 'must'
Or	one or the other, and possibly both	one or the other, but not both

Background

There are largely two contrasting views on scalar implicatures: Generalized Conversational Implicatures (GCI) and Relevance Theory. The GCI account (Horn, 1984; Levinson, 2000) holds that enriched interpretations are the default and are cancelled by context to yield the logical interpretation. Conversely, Relevance Theory (Sperber & Wilson, 1995) states that the logical interpretation is the default, and pragmatic interpretations are inferred through cognitive effort to maximize relevance in context.

Reaction time (RT) experiments support the latter, showing that pragmatic interpretations have longer RTs and thus require more cognitive resources (Bott & Noveck, 2004). Furthermore, language acquisition studies find that children prefer logical interpretations and adults prefer pragmatic interpretations. This suggests that children's pragmatic inference skills develop with age. Such results were found in Noveck (2001) for *some* and *might*, alongside Braine & Rumain (1981) and Paris (1973) for *or*. In this study, we investigate whether LLMs' results resemble children or adults, and what their responses reveal about their pragmatic inference skills.

Method

We replicated three classic human experiments on scalar implicatures: Bott & Noveck (2004) for *some*, Noveck (2001) for *might*, and Braine & Rumain (1981) for *or*. For each expression, we handcrafted a dataset with 300 test sentences and their corresponding controls. The test sentences were used to observe LLMs' interpretation of scalar expressions, and control sentences were used to test basic model reasoning.

For *some*, the test sentence (e.g., *Some roses are flowers*) was designed so that "True" indicates a logical interpretation and "False" a pragmatic one. For *might*, test sentences explain the content of a covered container (e.g., *There might be an apple.*); the test sentence uses the probabilistic *might* to describe a necessary item, making it pragmatically infelicitous but logically true. For *or*, test sentences used *either-or* sentences to describe the content of a container (e.g., *Either there's a coat or there's a scarf in the closet.*); the test sentence was in the form of True+True, while control sentences were True+False, False+True, and False+False.

We prompted models with a truth-value judgment task, so they would respond with exactly one word: *True* or *False*. For each scalar expression, three conditions were run: (i) no

instructions, (ii) an explicit logical instruction, and (iii) an explicit pragmatic instruction. In conditions (ii) and (iii), models were prompted with the respective logical and pragmatic interpretations in Table 1. Using this method, we evaluated eight decoder-only models: GPT-5.4, GPT-5.4 nano, GPT 5.4-mini, Claude Opus 4.6, Claude Sonnet 4.6, Claude Haiku 4.5, Mistral Small 4, and Mistral Large 3. Temperature was set to 0 to ensure consistency.

Results

Across all three scalar expressions, in the no-instructions condition, models produced logical interpretations at near-perfect rates, showing strong preference for underinformative sentences. This pattern mirrors the child responses reported in human experiments. Logical instructions showed little difference from no-instruction conditions, indicating that such instructions aligned with models' default logical interpretations. Pragmatic instructions, however, had asymmetric effects across scales. For *some*, pragmatic instructions shifted larger models (GPT-5.4, Claude Sonnet 4.6) almost entirely toward pragmatic interpretations, while smaller models were largely unmoved. For *might*, pragmatic instructions essentially had no effect: logical interpretations persisted across all models. For *or*, effects were more apparent but highly variable: GPT-5.4, Claude Sonnet 4.6, and Claude Haiku 4.5 showed near-complete acceptance of pragmatic interpretations, while GPT-5.4 nano and both Mistral models barely did so.

Discussion

The convergent finding across the three scales is that LLMs default to logical rather than pragmatic interpretations, resembling children rather than adults. This is consistent with Relevance Theory, in which pragmatic enrichment is a cognitively effortful, context-driven process. Models prefer the default logical meaning but lack the cognitive resources to make scalar inferences like adults. These results suggest that adult-like pragmatic inference skills may not be acquired from text alone and depend on resources absent from the LLM training process. One possible explanation is that, unlike humans, LLMs learn about the world exclusively through text, resulting in a "disembodied cognition." Moreover, models' resemblance to children is likely superficial: children produce logical responses because these meet their still-developing expectations of relevance, whereas models do so because they have no expectations of relevance at all–only next-token predictions.

References

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language*, *51*(3), 437–457.

Braine, M. D., & Rumain, B. (1981). Development of comprehension of "or": Evidence for a sequence of competencies. *Journal of experimental child psychology*, *31*(1), 46–70.

Horn, L. (1984). Towards a new taxonomy for pragmatic inference : Q- and R-based implicature. *Meaning, Form and Use in Context*.

Levinson, S. C. (2000). *Presumptive meanings : the theory of generalized conversational implicature*. MIT Press.

Noveck, I., & Sperber, D. (2012). *The why and how of experimental pragmatics: The case of 'scalar inferences'*. Cambridge University Press. https://doi.org/10.1017/CBO9781139028370.018

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188.

Paris, S. G. (1973). Comprehension of language connectives and propositional logical relationships. *Journal of experimental child psychology*, *16*(2), 278–291.

Sperber, D., & Wilson, D. (1995). *Relevance : communication and cognition 2nd ed*. Blackwell.