Can Vision-Language Models Infer Speaker's Ignorance? The Role of Visual and Linguistic Cues

Ye-eun Cho
English language and literature
Sungkyunkwan University
Seoul, South Korea
joyenn@skku.edu

Yunho Maeng
Ewha Womans University &
LLM Experimental Lab, MODULABS
Seoul, South Korea
yunhomaeng@ewha.ac.kr

Abstract

This study investigates whether vision language models (VLM) can perform pragmatic inference, focusing on ignorance implicatures, utterances that imply the speaker's lack of precise knowledge. To test this, we systematically manipulated contextual cues: the visually depicted situation (visual cue) and QUD-based linguistic prompts (linguistic cue). When only visual cues were provided, three state-of-the-art VLMs (GPT-4o, Gemini 1.5 Pro, and Claude 3.5 sonnet) produced interpretations largely based on the lexical meaning of the modified numerals. When linguistic cues were added to enhance contextual informativeness, Claude exhibited more human-like inference by integrating both types of contextual cues. In contrast, GPT and Gemini favored precise, literal interpretations. Although the influence of contextual cues increased, they treated each contextual cue independently and aligned them with semantic features rather than engaging in context-driven reasoning. These findings suggest that although the models differ in how they handle contextual cues, Claude's ability to combine multiple cues may signal emerging pragmatic reasoning abilities in multimodal models.

1 Introduction

In recent years, many large language models (LLMs) have demonstrated the ability to solve a wide variety of tasks, contributing to their growing popularity. Initially limited to text-based inputs, these models have been extended to incorporate visual inputs, paving the way for vision-language models (VLMs). By bridging vision and language modalities, VLMs have expanded the possibilities for AI applications and become central to the ongoing technological revolution (Radford et al., 2021; Ramesh et al., 2021; Alayrac et al., 2022; Li et al., 2023).

VLMs have enabled various multimodal applications, such as object recognition (Ren et al., 2015; Chen et al., 2020; He et al., 2020), caption generation (Vinyals et al., 2015; Chen et al., 2022; Yu et al., 2022), and visual question answering (Antol et al., 2015). These tasks primarily focus on associations between visual and textual inputs by identifying objects, describing scenes, or responding to straightforward queries. While such capabilities are remarkable, they represent only the surface level of human-like understanding. In fact, real-world communication often requires reasoning about implicit meanings that emerge from the interplay between language and visual information (Sikka et al., 2019). To move toward more human-like multimodal intelligence, VLMs must also be able to engage in this type of context-sensitive and inferential processing (see Kruk et al., 2019). This raises critical questions about VLMs' capacity for context-sensitive reasoning, which underlies the pragmatic reasoning abilities required for real-world communication.

Pragmatics offers an ideal framework for investigating this question. In human communication, pragmatic inference plays a crucial role in understanding intended meanings beyond literal statement (Grice, 1975; Wilson and Sperber, 1995; Levinson, 2000). Contextual cues often provide disambiguating information that influences the interpretation of utterances, making pragmatic reasoning inherently multimodal (Clark, 1996; Kendon, 2004; Martin et al., 2007; McNeill, 2008). While a few studies on pragmatic reasoning have been explored in text-only LLMs (Hu et al., 2022, 2023; Lipkin et al., 2023; Cho and Kim, 2024; Capuano and Kaup, 2024; Tsvilodub et al., 2024), the visual modality enriches meaning construction through interaction with linguistic input. As visual context provides rich, implicit information that influences language interpretation, studying pragmatic phenomena through VLMs presents an intriguing research opportunity. However, how well VLMs can leverage visual information for pragmatic inference

Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025), pages 298–308

remains largely unexplored.

Therefore, we investigate whether VLMs exhibit sensitivity to context, particularly focusing on ignorance implicatures—a pragmatic phenomenon in which a speaker's utterance implies a lack of precise knowledge, and whether this sensitivity can be modulated by a single cue or by the combination of multiple cues. By examining how VLMs handle this phenomenon in comparison to human reasoning, we aim to better understand their strengths and limitations in processing context-dependent pragmatic meaning.

2 Ignorance implicatures

To better understand pragmatic reasoning in real-world language use, we examine the phenomenon of *ignorance implicatures*. Consider the examples in (1).

(1)

a. *(bare numeral)*
Four students passed the exam.

b. *(superlative modified numeral)*
At least four students passed the exam.

c. *(comparative modified numeral)*
More than three students passed the exam.

When it comes to how many students passed the exam, the statement (1a) triggers 'exactly four' interpretation, whereas (1b) and (1c) do not. Both (1b) and (1c) contain modified numerals, suggesting that the speaker may not know the exact number of students who passed the exam. This is known as ignorance implicatures, where the speaker's choice of modifier implies a lack of precise knowledge.

However, not all modifiers give rise to ignorance implicatures to the same extent. Previous studies have shown that superlative modifiers like *at least* tend to trigger ignorance implicatures more consistently than comparative ones like *more than* (Nouwen, 2010; Cummins et al., 2012; Coppock and Brochhagen, 2013b; Mayr and Meyer, 2014; Cremers et al., 2022). In this regard, the likelihood of ignorance inferences typically follows the hierarchy: *superlative modified numerals* > *comparative modified numerals* > *bare numerals*.

This observation has prompted researchers to explore how such inferences arise, leading to two main perspectives. One approach suggests that ignorance inference is dependent on the words or phrases themselves (Geurts and Nouwen, 2007; Nouwen, 2010; also see Geurts et al., 2010). Geurts and Nouwen (2007), for example, argued that the semantics of superlative modifiers are inherently more complex. While *more than n* expresses a simple meaning 'larger than *n*', *at least n* can convey both 'possible that there is a set of *n*' and 'certain that there is no smaller set of *n*.' According to Nouwen (2010), when someone has basic knowledge of geometry, (2a) gives the impression that the speaker lacks precise information, as compared to (2b). This attributes the ignorance implicatures to a semantic property specific to *at least*.

(2) a. ? A hexagon has at least five sides.
b. A hexagon has more than four sides.

Under the pragmatic account, on the other hand, ignorance implicatures for both *at least* and *more than* have been primarily explained through Gricean reasoning, particularly the Maxim of Quantity (Grice, 1975), which holds that the speaker's choice to provide a lower-bound statement, rather than a more informative exact number, suggests that the speaker lacks precise knowledge (Büring, 2008; Cummins and Katsos, 2010; Coppock and Brochhagen, 2013b). More recent studies have expanded this account by emphasizing the role of contextual factors (Cummins et al., 2012; Cummins, 2013; Mayr and Meyer, 2014; Westera and Brasoveanu, 2014; Cremers et al., 2022). In these studies, contextual cues, including Question Under Discussion (QUD), preceding discourse, or accompanying visual input, were manipulated to modulate the likelihood of implicature.

For instance, Westera and Brasoveanu (2014) investigated how different types of modified numerals give rise to ignorance implicatures depending on contextual demands and processing cost. In their experiments, participants read short dialogues or utterances containing modified numerals and judged how confident the speaker seemed about the exact quantity, as well as how natural the utterance was. To manipulate the informativeness required by the discourse, the authors introduced different QUDs, such as a 'how many' condition (*How many of the diamonds did you find under the bed?*), which demanded precise answers, and a 'polar' condition (*Did you find {at most | less than} ten of the diamonds under the bed?*), which did not require numerically specific responses, as they could be answered with a simple *yes* or *no*. The results

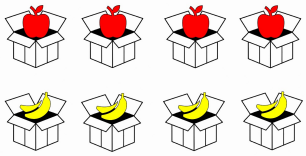
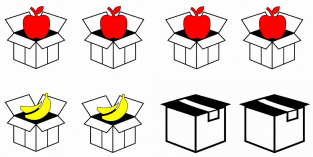
No.	Image	Text	Situation	Modifier
1		*There are four apples in the boxes.*	precise	bare
		There are at least four apples in the boxes.	precise	superlative
		There are more than three apples in the boxes.	precise	comparative
2		*There are four apples in the boxes.*	approximate	bare
		There are at least four apples in the boxes.	approximate	superlative
		There are more than three apples in the boxes.	approximate	comparative

Table 1: A sample set of experimental materials

showed that ignorance inferences occurred significantly more consistently when the QUD demanded precision ('how many' condition), suggesting that contextual expectations about informativeness directly affect how such inferences are drawn.

Likewise, Cremers et al. (2022) systematically manipulated various contextual factors to investigate the conditions under which ignorance implicatures arise. Their experiments involved multiple levels of visual information, QUD types, and textual scenarios. In particular, visual information was used to represent the informativeness of the situation—for example, a precise situation in which all eight cards were face-up, and an approximate situation in which two of the eight cards remained face-down, obscuring the exact quantity. Their findings revealed that ignorance inferences were more likely when the QUD required a precise answer ('howmany' condition) and when the visual context left room for uncertainty ('approximate' condition). These results highlight that ignorance implicatures are largely influenced by both linguistic and non-linguistic contextual cues.

These findings raise the question of whether and how visual and linguistic contextual information can enhance the pragmatic reasoning abilities of VLMs. To address this question, the present study examines whether VLMs exhibit sensitivity to ignorance implicatures across multiple contextual cues.

3 Methods

3.1 Data

As presented in Table 1, experimental materials were designed using two images for contextual precision (henceforth, 'situation') and texts including bare numeral, superlative, and comparative modifiers (henceforth, 'modifier').

In detail, images were used to manipulate the contextual precision, where a picture showing all 8 boxes open and providing the exact number of target objects was labeled as 'precise', and a picture with 2 out of 8 boxes remaining closed and an uncertain number of target objects was labeled as 'approximate'. In both types of situation, the target objects consistently appeared in 4 boxes. For example, in the image for precise situation, all 8 boxes are open and 4 of them contain apples. Since all boxes are open, we can tell that the target objects are exactly 4. On the other hand, in the image for approximate situation, 2 out of the 8 boxes remain closed and 4 of the open boxes contain apples. As what is inside the closed boxes is unknown, the target objects could be 4 or more. The corresponding texts were categorized based on the modifier types, including 'bare' (bare numeral *n*), 'superlative' (*at least n*), and 'comparative' (*more than n*).

Image data was created by combining open and closed boxes generated by GPT-4o (OpenAI, 2024) with standard icons for target objects. In this study, the number of target objects was consistently set to four, as previous work has shown that VLMs often exhibited limited performance on numerical reasoning tasks and experience a marked decline in accuracy when counting more than four items (Paiss et al., 2023). Additionally, since VLMs tend to struggle with counting when objects are presented in unstructured or cluttered spatial arrangements (Liu et al., 2019; Rahmanzadehgervi et al., 2024), the experimental images were carefully constructed with precisely aligned rows and columns. In this manner, each set of materials consisted of 2 images, each paired with 3 corresponding texts. In total, 70 sets of materials were used in the experiment.

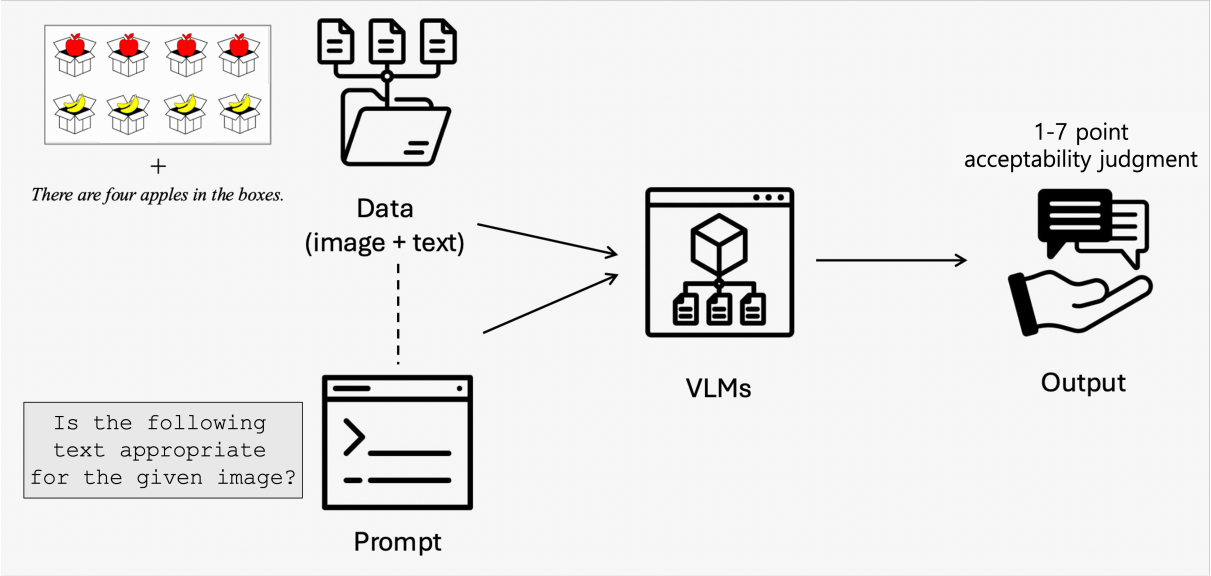


Figure 1: Overview of the experimental procedure

3.2 Models and Procedure

As VLMs for the experiment, we used GPT-4o (OpenAI, 2024), Gemini 1.5 Pro (Team et al., 2024), and Claude 3.5 sonnet (Anthropic, 2024). These models were selected due to their ability to process both image and text inputs simultaneously. They not only allow for the matching of images with text to determine their relationship but also provide the functionality to selectively query specific parts of the text within a broader context. This makes them well-suited for a series of our experiments.

For the experiment, these models were initialized using API keys. The images were then resized to a standard size of 224x224 pixels using the Pillow library (Clark et al., 2015) to ensure consistency in input dimensions and optimize processing efficiency. After resizing, the images were encoded into base64 format to ensure compatibility for input into the model's API.

Each experiment involved presenting the image alongside text prompts, which were specifically tailored for each task. All the materials, code and result of the experiment are publicly available.[1]

4 Experiment 1

4.1 Prompt

Cremers et al. (2022) argued that the disagreement over main findings related to ignorance inference was that the detection depends on the types of tasks participants were asked to perform. Specifically, it varied depending on whether participants were given an acceptability judgment task (Coppock and Brochhagen, 2013a; Westera and Brasoveanu, 2014; Cremers et al., 2022), where they judged the acceptability of the given sentences with respect to the depicted scenarios or images, or an inference task (Geurts et al., 2010), where they judged whether *exactly n* implies *at least n*. Cremers et al. (2022) argued that ignorance inference is more accurately assessed when evaluating the appropriateness of a sentence in relation to the context, rather than through the logical reasoning involved in inference tasks. In this regard, the acceptability judgment task serves as an effective method, guiding participants to evaluate whether a sentence is contextually appropriate. The acceptability judgment task typically involves either a true/false response format, as in truth-value judgment tasks (Coppock and Brochhagen, 2013a), or a numerical scale to capture the degree of ignorance implicatures in a more fine-grained manner (Westera and Brasoveanu, 2014; Cremers et al., 2022). In our experiment, we adopt a 1-7 scale to assess the appropriateness of image-text pairings in a more fine-grained manner.

As presented in Figure 1, each model was prompted to rate whether the texts with bare numerals, superlative, and comparative modifiers are appropriate for the given image on a scale from 1 to 7. The phrases used in the prompt were adapted from Experiment 3 in Cremers et al. (2022). In this manner, 70 sets of experimental items were re-

[1] https://github.com/joyennn/ignorance-implicature

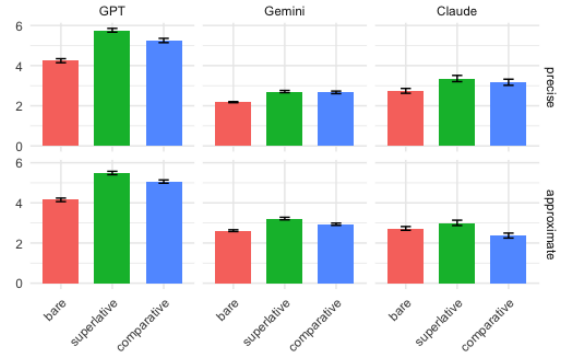


Figure 2: Result of Experiment1 — Mean scores for the appropriateness of image-text pairs based on modifiers, situations and models

peated 5 times to improve the reliability, resulting in a total of 2,100 individual responses from each of three models, as detailed below.

> Is the following text appropriate for the given image?
>
> Please reply with a single integer between 1 and 7, where 1 means "not at all appropriate" and 7 means "completely appropriate."
>
> **Text:** {text}

4.2 Result

Figure 2 shows the mean scores for the appropriateness of image-text pairs based on the types of modifiers, situations, and models. For the statistical analysis, we built mixed-effects logistic regression models (Baayen, 2008; Baayen et al., 2008; Jaeger, 2008; Jaeger et al., 2011) to analyze the results for each model, using the `lme4` package (Bates et al., 2015) in R software (Team, 2023). To examine the fixed effects, modifier and situation were set as independent variables, with appropriate scores as the dependent variable. Image and text were specified as random effects. For independent variables, the bare condition and the precise condition were set as the reference levels for modifier and situation, respectively.

As a result, the scores for appropriateness of image-text pairs followed the order of *superlative* > *comparative* > *bare* in both types of situations across almost all models, except for the approximate condition of Claude. While these predominant results aligned with findings from Cremers et al. (2022), where participants preferred the text with superlative and comparative in the approximate condition. However, the similar pattern in the precise situation was unexpected. In this situation, texts containing bare numerals should have been considered more appropriate than those with the other modifiers, as the number of target objects was explicitly defined.

Statistically, these results were influenced mostly modifier, which showed main effects in GPT ($p < 0.001$), Gemini ($p < 0.001$), and Claude ($p < 0.01$, 0.05). However, there were no significant effects on situation alone in GPT ($p = 0.66$) and Claude ($p = 0.96$), nor in the interaction of situation and modifier in GPT ($p = 0.06$, 0.33) and Gemini ($p = 0.11$, < 0.001).

In summary, superlative and comparative modifiers, which imply uncertainty, consistently led to higher appropriateness ratings even in both types of situations. This suggests that the models' responses were more influenced by the modifiers rather than the situations. The models' reliance on the semantic information inherent in the modifiers, rather than utilizing the contextual cues, indicates that the models are not effectively applying visually presented contextual information to ignorance inference.

5 Experiment 2

5.1 Prompt

Experiment 2 was conducted with the assumption that providing multiple pieces of contextual information would bring about pragmatic interpretation. Thus, another contextual cue, QUD, was added to the previous experimental setup. For QUDs, two types of conditions were designed, such as 'how-many' and 'polar' as below.

In the howmany condition, the question focuses on a specific number of target objects, leading responses containing superlative and comparative modifiers to introduce uncertainty or information gaps, which in turn trigger ignorance inferences. In contrast, the 'polar' condition elicits a simple yes/no response, placing minimal demands on numerical precision. Accordingly, it serves as a baseline for assessing the effects of the howmany QUD.

The appropriateness of the text in response to either of the two questions was measured on a 1-7 scale. For each condition, all experimental sets

were repeated 5 times, resulting in a total of 4,200 individual responses.

> Is the following answer to the question appropriate for the given image?
>
> Please reply with a single integer between 1 and 7, where 1 means "not at all appropriate" and 7 means "completely appropriate."
>
> *(QUD: howmany)*
> **Question:** How many {objects} did you find in the boxes?
> **Answer:** {text}
>
> *(QUD: polar)*
> **Question:** Did you find four {objects} in the boxes?
> **Answer:** {text}

5.2 Result

Figure 3 shows the mean scores for the appropriateness of image-text pairs, when the text was given as a response of howmany and polar questions, based on the types of modifiers, situations, and models. The statistical analysis was the same as in experiment 1, with QUD added as an independent variable in the fixed effects. For QUD, the polar condition was set as the reference level.

In the howmany condition, we observed that the score for bare numerals increased compared to the results from Experiment 1. In most cases observed in GPT and Gemini, bare numerals received the highest score, with the order being *bare* > *superlative* > *comparative*. For the precise condition of Gemini, the order was *superlative* > *bare* > *comparative*, but again, the score for bare numerals increased compared to the previous experiment.

Statistical analysis revealed that, for both GPT and Gemini, no significant effects were observed for the modifier (GPT: $p = 0.11, < 0.001$ | Gemini: $p < 0.001, = 0.46$). However, main effects were captured for the two contextual cues, situation (GPT: $p < 0.05$ | Gemini: $p < 0.001$) and QUD (GPT: $p < 0.001$ | Gemini: $p < 0.001$). Additionally, significant interactions were observed between each contextual cue and the modifier, specifically for the interactions of situation and modifier (GPT: $p < 0.05, 0.001$ | Gemini: $p < 0.001$), and QUD and modifier (GPT: $p < 0.001, 0.05$ | Gemini: $p < 0.001$). However, the interaction between the two contextual cues, situation and QUD (GPT: $p = 0.58$ | Gemini: $p = 0.92$), as well as the interaction of modifier, situation, and QUD (GPT: $p = 0.69, 0.92$ | Gemini: $p < 0.05, = 0.92$), were not significant. This finding suggests that while the influence of modifiers remains present, the increased availability of contextual information appears to guide the models toward a more context-driven interpretation strategy.

On the other hand, in case of Claude, the scores followed the order of *bare* > *superlative* > *comparative* in the precise situation, while the order was *superlative* > *bare* > *comparative* in the approximate situation. Although this does not perfectly align with human experimental results, it reflects a pattern similar to our expectations, where bare numerals would be preferred in the precise situation, and either superlative or comparative modifiers would be preferred in the approximate situation. Furthermore, statistical analysis showed a significant effect only in the interaction of both contextual cues, situation and QUD ($p < 0.01$). Considering that in Experiment 1, Claude did not show a similar pattern to human results based on visually encoded context, and that its results were strongly influenced by the modifiers, and the combination of modifier and situation, these findings suggest the possibility that when multiple contextual cues are provided, the model may combine them in a way that aligns more closely with human ignorance inferences, showing a tendency to rely more on contextual cues than on semantic modifiers.

In the polar as a control condition, the appropriateness score for bare numerals was higher compared to the result of Experiment 1, but still lower than in the howmany condition across all the models.

In summary, when the contextual cue QUD was added to the previous experimental setup, GPT and Gemini showed a tendency to prefer bare numerals, which refer to precise knowledge, compared to when only a single contextual cue was provided. In contrast, Claude demonstrated a more integrated approach by utilizing multiple contextual cues together, leading to an interpretation that was closer to pragmatic inference. Despite differences in how these models interpreted the stimuli, all models showed a common pattern of shifting reliance from modifiers to contextual cues when multiple cues were provided.

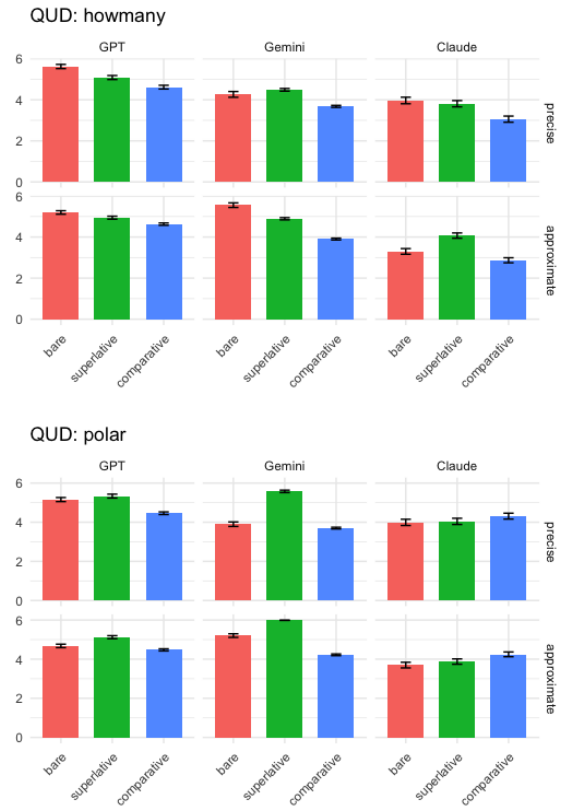


Figure 3: Result of Experiment2 — Mean scores for the appropriateness of image-text pairs based on modifiers, situations, and models across QUDs

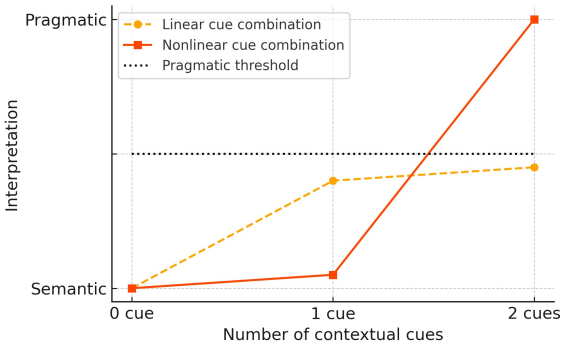


Figure 4: Modeling a threshold effect via linear and nonlinear cue combination as a function of contextual cue number (adapted from Parker, 2019)

6 Discussion

This study aimed to investigate the influence of contextual cues in the interpretation of ignorance inference within VLMs. In Experiment 1, we investigated how visually depicted situation (precise and approximate) and different types of modifiers (bare, superlative, and comparative) influenced appropriateness ratings of image-text pairs. Results revealed that appropriateness ratings consistently followed the order of *superlative* > *comparative* > *bare* across almost all models, regardless of situation types. This pattern suggests that the models primarily relied on the semantic features of modifiers rather than incorporating contextual information into their judgments.

Building upon these findings, Experiment 2 introduced an additional contextual cue, QUD, with two conditions (howmany and polar). This experiment aimed to determine whether multiple contextual cues would facilitate more sophisticated pragmatic inference. Interestingly, when presented with the howmany QUD, both GPT and Gemini models gave higher appropriateness ratings to bare numerals, which aligned with expectations for precise situation, but not for approximate situation. In our analysis, these models showed greater improvement in providing precise information than in engaging in pragmatic reasoning. While the influence of modifiers remained, there was a modest increase in sensitivity to contextual cues.

In contrast, Claude demonstrated a more integrated approach to contextual reasoning, using both situation and QUD simultaneously. This integration pattern suggests Claude may be moving closer to human-like pragmatic reasoning, which typically involves holistic consideration of multiple contextual factors. This leads to the assumption that Claude may have benefited from cue combination, where the presence of two contextual cues, rather than a single cue, led to pragmatic interpretation.

This pattern resonates with Parker (2019)'s cue combination scheme, which posits the processing benefit of retrieval cues for anaphora in memory is not merely additive but emerges nonlinearly when multiple cues are jointly available. Extending this idea to our findings, contextual cue combination for ignorance inference in VLMs may similarly follow the nonlinear cue combination method: one contextual cue alone may not significantly affect the context-sensitive reasoning, but the addition of the second cue increases the "cue weight" enough to reach the threshold for pragmatic interpretation. This threshold effect is visualized in Figure 4, which contrasts linear and nonlinear cue integration patterns as a function of contextual cue number.

Then, do GPT and Gemini follow the linear cue combination? It would be insufficient to characterize these models' behavior merely as examples of linear cue combination. Rather, the observed

pattern suggests a difference in how these models represent and utilize contextual information. While Claude appears to engage in combining two contextual cues into a unified pragmatic representation, GPT and Gemini exhibit a pattern of local alignment, in which modifiers are evaluated separately with each cue, but the cues themselves remain structurally unbound. In this sense, their responses are not limited because they combine cues linearly, but because their internal processing architecture does not support contextual cue combination in the first place. Consequently, their outputs reflect a tendency to prioritize informational precision over pragmatic reasoning. As more cues become available, the models tend to converge on more semantically determinate interpretations, aiming to reduce uncertainty in a localized manner rather than resolving it through holistic, context-sensitive inference.

Taken together, these findings offer new insights into how current VLMs differ in their capacity for contextual cue combination in pragmatic inference. By introducing multiple types of cues—both visual and linguistic—within a controlled experimental setting, this study provides empirical evidence that not all models process contextual information in the same way, and that the ability to integrate multiple cues holistically may serve as a crucial indicator of emerging pragmatic reasoning abilities in VLMs. In doing so, this research contributes to the growing body of work on multimodal language processing by highlighting the need to evaluate not only what models generate and understand, but also how they integrate diverse contextual cues to infer meaning.

7 Conclusion

This study examined whether and how current VLMs engage in pragmatic inference, particularly focusing on ignorance implicatures, when provided with visual and linguistic contextual cues. Through two experiments manipulating modifier types and contextual cues—including situation and QUDs—we found that not all VLMs process such information in the same way. Claude demonstrated the ability to integrate multiple contextual cues into a unified interpretation, exhibiting a threshold effect in pragmatic reasoning when both contextual cues were available. In contrast, GPT and Gemini tended to treat these cues independently, prioritizing precision over context-sensitive inference. This suggests a fundamental difference not only in cue weighting tendencies but also in how models internally represent and combine contextual information.

By systematically evaluating ignorance implicatures across VLMs, this study contributes to our understanding of the mechanisms underlying pragmatic behavior in VLMs. Importantly, it highlights that the capacity for contextual cue combination may serve as one of the key indicators of emerging pragmatic reasoning abilities in VLMs. These findings open new directions for evaluating and developing VLMs that move beyond literal interpretation toward more human-like pragmatic inference.

Acknowledgments

This research was supported by Brian Impact Foundation, a non-profit organization dedicated to the advancement of science and technology for all.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Anthropic. 2024. *Claude 3.5 Sonnet*. Anthropic.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

R Harald Baayen. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge university press.

R Harald Baayen, Douglas J Davidson, and Douglas M Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Gabor Grothendieck, Peter Green, and M Ben Bolker. 2015. Package 'lme4'. *convergence*, 12(1):2.

Daniel Büring. 2008. The least at least can do. In *West Coast Conference on Formal Linguistics (WCCFL)*, volume 26, pages 114–120. Citeseer.

Francesca Capuano and Barbara Kaup. 2024. Pragmatic reasoning in gpt models: Replication of a subtle negation effect. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Ye-eun Cho and Seong mook Kim. 2024. Pragmatic inference of scalar implicature by llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 10–20.

Alex Clark et al. 2015. Pillow (pil fork) documentation. *readthedocs*.

Herbert H Clark. 1996. *Using language*. Cambridge University Press.

Elizabeth Coppock and Thomas Brochhagen. 2013a. Diagnosing truth, interactive sincerity, and depictive sincerity. In *Semantics and Linguistic Theory*, pages 358–375.

Elizabeth Coppock and Thomas Brochhagen. 2013b. Raising and resolving issues with scalar modifiers. *Semantics and Pragmatics*, 6:3–1.

Alexandre Cremers, Liz Coppock, Jakub Dotlačil, and Floris Roelofsen. 2022. Ignorance implicatures of modified numerals. *Linguistics and Philosophy*, 45(3):683–740.

Chris Cummins. 2013. Modelling implicatures from modified numerals. *Lingua*, 132:103–114.

Chris Cummins and Napoleon Katsos. 2010. Comparative and superlative quantifiers: Pragmatic effects of comparison type. *Journal of Semantics*, 27(3):271–305.

Chris Cummins, Uli Sauerland, and Stephanie Solt. 2012. Granularity and scalar implicature in numerical expressions. *Linguistics and Philosophy*, 35:135–169.

Bart Geurts, Napoleon Katsos, Chris Cummins, Jonas Moons, and Leo Noordman. 2010. Scalar quantifiers: Logic, acquisition, and processing. *Language and cognitive processes*, 25(1):130–148.

Bart Geurts and Rick Nouwen. 2007. 'at least'et al.: the semantics of scalar modifiers. *Language*, pages 533–559.

H Paul Grice. 1975. *Logic and Conversation*. Brill.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Jennifer Hu, Roger Levy, Judith Degen, and Sebastian Schuster. 2023. Expectations over unspoken alternatives predict pragmatic inferences. *Transactions of the Association for Computational Linguistics*, 11:885–901.

Jennifer Hu, Roger Levy, and Sebastian Schuster. 2022. Predicting scalar diversity with context-driven uncertainty over alternatives. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 68–74.

T Florian Jaeger. 2008. Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of memory and language*, 59(4):434–446.

T Florian Jaeger, Emily M Bender, and Jennifer E Arnold. 2011. Corpus-based research on language production: Information density and reducible subject relatives. *Language from a cognitive perspective: grammar, usage and processing. Studies in honor of Tom Wasow*, pages 161–198.

Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. *arXiv preprint arXiv:1904.09073*.

Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Benjamin Lipkin, Lionel Wong, Gabriel Grand, and Joshua B Tenenbaum. 2023. Evaluating statistical language models as pragmatic reasoners. *arXiv preprint arXiv:2305.01020*.

Weizhe Liu, Mathieu Salzmann, and Pascal Fua. 2019. Context-aware crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5099–5108.

Jean-Claude Martin, Patrizia Paggio, P Kuenlein, Rainer Stiefelhagen, and Fabio Pianesi. 2007. Multimodal corpora for modelling human multimodal behaviour. *Special issue of the International Journal of Language Resources and Evaluation*, 41(3-4).

Clemens Mayr and Marie-Christine Meyer. 2014. More than at least. In *Slides presented at the Two days at least workshop, Utrecht*.

David McNeill. 2008. Gesture and thought. In *Gesture and thought*. University of Chicago press.

Rick Nouwen. 2010. Two kinds of modified numerals. *Semantics and Pragmatics*, 3:3–1.

OpenAI. 2024. *Hello GPT-4o*. OpenAI.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3170–3180.

Dan Parker. 2019. Cue combinatorics in memory retrieval for anaphora. *Cognitive science*, 43(3):e12715.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2024. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Karan Sikka, Lucas Van Bramer, and Ajay Divakaran. 2019. Deep unified multimodal embeddings for understanding both content and users in social media networks. *arXiv preprint arXiv:1905.07075*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Polina Tsvilodub, Paul Marty, Sonia Ramotowska, Jacopo Romoli, and Michael Franke. 2024. Experimental pragmatics with machines: Testing llm predictions for the inferences of plain and embedded disjunctions. *arXiv preprint arXiv:2405.05776*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Matthijs Westera and Adrian Brasoveanu. 2014. Ignorance in context: The interaction of modified numerals and quds. In *Semantics and Linguistic Theory*, pages 414–431.

Deirdre Wilson and Dan Sperber. 1995. *Relevance theory*. Oxford:Blackwell.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Limitations

This study has several limitations that offer directions for future research. First, our experiments focused on a specific pragmatic phenomenon involving modified numerals, which allowed for a controlled testbed but may limit the generalizability of the findings. Extending the investigation to other types of pragmatic inferences would provide a broader understanding of VLMs' pragmatic reasoning abilities. Second, although we tested three state-of-the-art models—GPT-4o, Gemini 1.5 Pro, and Claude 3.5—the results may not fully generalize to other architectures, including open-source models with different training paradigms. Expanding the model pool would help assess the robustness of cue integration effects. Lastly, while our study builds on prior human experiments, it does not include a direct comparison with human performance under identical conditions. Such empirical comparisons would clarify whether model behavior reflects genuine pragmatic reasoning or merely statistical alignment with training data.

A Appendix

	Estimate	**Std**	***t***	***p*-value**
(Intercept)	4.24	0.17	24.31	**<0.001**
Situation	-0.09	0.21	-0.43	0.66
Modifier - Superlative	1.51	0.13	11.1	**<0.001**
Modifier - Comparative	1.01	0.13	7.37	**<0.001**
Situation:Modifier - Superlative	-0.17	0.09	-1.81	0.06
Situation:Modifier - Comparative	-0.09	0.09	-0.96	0.33

Table 2: Summary of fixed effects from mixed-effects logistic regression models by GPT-4o in Experiment 1

	Estimate	Std	*t*	*p*-value
(Intercept)	2.18	0.11	20.51	**<0.001**
Situation	0.43	0.13	3.24	**<0.01**
Modifier - Superlative	0.52	0.07	7.12	**<0.001**
Modifier - Comparative	0.47	0.07	6.32	**<0.001**
Situation:Modifier - Superlative	0.07	0.04	1.61	0.11
Situation:Modifier - Comparative	-0.17	0.04	-3.67	**<0.001**

Table 3: Summary of fixed effects from mixed-effects logistic regression models by Gemini 1.5 Pro in Experiment 1

	Estimate	Std	*t*	*p*-value
(Intercept)	2.74	0.26	10.21	**<0.001**
Situation	-0.01	0.33	-0.04	0.96
Modifier - Superlative	0.61	0.19	3.19	**<0.01**
Modifier - Comparative	0.42	0.19	2.22	**<0.05**
Situation:Modifier - Superlative	-0.33	0.09	-3.71	**<0.001**
Situation:Modifier - Comparative	-0.78	0.09	-8.64	**<0.001**

Table 4: Summary of fixed effects from mixed-effects logistic regression models by Claude 3.5 in Experiment 1

	Estimate	Std	*t*	*p*-value
(Intercept)	5.15	0.15	33.56	**<0.001**
Situation	-0.48	0.20	-2.39	**<0.05**
QUD	0.46	0.07	6.24	**<0.001**
Modifier - Superlative	0.17	0.11	1.59	0.11
Modifier - Comparative	-0.69	0.11	-6.26	**<0.001**
Situation:Modifier - Superlative	0.26	0.10	2.56	**<0.05**
Situation:Modifier - Comparative	0.49	0.10	4.71	**<0.001**
QUD:Modifier - Superlative	-0.71	0.10	-6.81	**<0.001**
QUD:Modifier - Comparative	0.30	0.10	-2.94	**<0.05**
Situation:QUD	0.05	0.10	0.54	0.58
Situation:QUD:Modifier - Superlative	0.01	0.14	-0.38	0.69
Situation:QUD:Modifier - Comparative	-0.05	0.14	0.09	0.92

Table 5: Summary of fixed effects from mixed-effects logistic regression models by GPT-4o in Experiment 2

	Estimate	Std	*t*	*p*-value
(Intercept)	3.83	1.45	26.27	**<0.001**
Situation	1.32	1.19	11.12	**<0.001**
QUD	0.35	6.22	5.72	**<0.001**
Modifier - Superlative	1.74	1.78	9.75	**<0.001**
Modifier - Comparative	0.13	1.78	-0.73	0.46
Situation:Modifier - Superlative	-0.92	8.66	-9.29	**<0.001**
Situation:Modifier - Comparative	-0.81	8.66	-10.61	**<0.001**
QUD:Modifier - Superlative	-1.45	8.69	-16.66	**<0.001**
QUD:Modifier - Comparative	-0.39	8.69	-4.44	**<0.001**
Situation:QUD	0.01	0.09	0.10	0.92
Situation:QUD:Modifier - Superlative	-0.30	0.12	-2.43	**<0.05**
Situation:QUD:Modifier - Comparative	-0.01	0.12	-0.09	0.92

Table 6: Summary of fixed effects from mixed-effects logistic regression models by Gemini 1.5 Pro in Experiment 2

	Estimate	Std	*t*	*p*-value
(Intercept)	3.99	0.29	13.55	**<0.001**
Situation	-0.29	0.40	-0.73	0.46
QUD	-0.02	0.08	-0.29	0.76
Modifier - Superlative	0.05	0.13	0.37	0.70
Modifier - Comparative	0.32	0.13	2.34	**<0.05**
Situation:Modifier - Superlative	0.12	0.12	1.03	0.29
Situation:Modifier - Comparative	0.22	0.12	1.84	0.06
QUD:Modifier - Superlative	-0.20	0.12	-1.68	0.09
QUD:Modifier - Comparative	-1.22	0.12	-9.92	**<0.001**
Situation:QUD	-0.37	0.12	-2.99	**<0.01**
Situation:QUD:Modifier - Superlative	0.80	0.17	1.41	0.15
Situation:QUD:Modifier - Comparative	0.24	0.17	4.56	**<0.001**

Table 7: Summary of fixed effects from mixed-effects logistic regression models by Claude 3.5 in Experiment 2