

Reference Games as a Testbed for the Alignment of Model Uncertainty and Clarification Requests

Manar Ali^{1,2} Judith Sieker³ Sina Zarriß^{1,3} Hendrik Buschmeier^{1,2}

¹CRC 1646 ‘Linguistic Creativity in Communication’, Bielefeld University, Germany

²Digital Linguistics Lab, Bielefeld University, Germany

³Computational Linguistics Group, Bielefeld University, Germany

{manar.ali|j.sieker|sina.zarriess|hbuschme}@uni-bielefeld.de

Abstract

In human conversation, both interlocutors play an active role in maintaining mutual understanding. When addressees are uncertain about what speakers mean, for example, they can request clarification. It is an open question for language models whether they can assume a similar addressee role, recognizing and expressing their own uncertainty through clarification. We argue that reference games are a good testbed to approach this question as they are controlled, self-contained, and make clarification needs explicit and measurable. To test this, we evaluate three vision-language models comparing a baseline reference resolution task to an experiment where the models are instructed to request clarification when uncertain. The results suggest that even in such simple tasks, models often struggle to recognize internal uncertainty and translate it into adequate clarification behavior. This demonstrates the value of reference games as testbeds for interaction qualities of (vision and) language models.

1 Introduction

Ambiguity and misunderstanding in dialogue is inevitable (Weigand, 1999). Problems in understanding and confusion can arise even for simple utterances, since speakers and addresses process speech subjectively and audience design in language production can never be perfect. For example, a request like “Can you pass me the blue pencil?” may prompt uncertainty in a situation where several blueish pencils are available, as color perception is subjective (Bimler et al., 2004) and context-specific (Mitterer and de Ruiter, 2008). However, if an addressee cannot resolve an ambiguity, communication need not fail. They can initiate *repair*, e.g., by asking a *clarification request* (i.e., “The light blue one?”), which explicitly signals misunderstanding and requests additional information. Repair and clarification are fundamental mechanisms for achieving

intersubjectivity and mutual understanding that humans employ frequently (on average every 84 seconds, as found crosslinguistically by Dingemanse et al., 2015). Often, it is the addressee who takes the lead in signaling an ambiguity, making mutual understanding a shared responsibility (Clark and Wilkes-Gibbs, 1986; Clark, 1996; Dingemanse and Enfield, 2023).

For language models, however, it is still an open question whether they employ similar mechanisms when facing uncertainty in dealing with human input. Although they can generate fluent and contextually appropriate responses, this very fluency can mask underlying comprehension problems, leading humans to overestimate model competence (Sieker et al., 2024; Rathi et al., 2025) and undermining trust in collaboration (Dhuliawala et al., 2023; Si et al., 2024). At the same time, their linguistically communicated confidence is often poorly calibrated to their actual accuracy (Mielke et al., 2022), raising the question of whether they can act adequately given uncertainty, e.g., by requesting clarification – an ability crucial for preventing misunderstandings from being hidden behind confident output.

Addressing the question of whether models can signal their uncertainty through clarification, however, is difficult in open dialogue settings, where clarification needs are hard to define and the space of possible interpretations is unlimited. Here, we propose *reference games* (Dale and Reiter, 1995; Frank and Goodman, 2012) as a testbed that provides an environment that makes clarification needs explicit and measurable. In their basic form, they involve two participants: a speaker and an addressee, who share a set of candidate referents (e.g., images or abstract illustrations). The speaker’s task is to describe a target object, and the listener must identify it among the set candidates. Unlike in broader dialogue research, where clarification often tends to focus on identifying a user’s underlying need or intent, reference games are goal-directed with fixed

alternatives and require no external knowledge: it is immediately clear when a description fails to single out the target and a clarification is needed.

In this paper, we argue that reference games offer a controlled way to test whether language models can convert internal uncertainty into appropriate clarification behavior and investigate whether vision-language models (VLMs), as addresses in a reference game, are able to recognize their own uncertainty and respond with clarification questions.

2 Background

While research in NLP has made significant progress toward modeling dialogue behavior, including clarification and repair, recent analyses show that LLMs still differ in core interactional mechanisms such as turn-taking, feedback, and repair (Pilán et al., 2024; Mayor et al., 2025). Furthermore, LLMs are far less likely than humans to engage in grounding acts, such as rejecting problematic input (Lachenmaier et al., 2025), often presuming common ground instead (Shaikh et al., 2024). Crucially, LLMs rarely initiate repair proactively, typically relying on humans to do so (Pütz and Esposito, 2024), and when they do produce clarification requests, their behavior is often miscalibrated compared to human patterns (Madge et al., 2025). This raises the central question of whether language models can appropriately act on their own uncertainty by formulating clarification requests. Quantifying uncertainty has become an increasingly central topic in NLP and the study on large language models (Vashurin et al., 2025; Shorinwa et al., 2025), yet few studies have connected these phenomena to clarification behavior. Testoni and Fernández (2024) and Zhang and Choi (2025), e.g., propose entropy-based uncertainty measures to guide whether a clarification should be requested.

However, such approaches typically focus on open-ended intent resolution tasks, where the space of possible interpretations is difficult to delimit. Here, reference games offer a controlled alternative. Reference games have long served as controlled settings in (computational) linguistics to study how interlocutors establish shared understanding (Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996). Recent work has shown that even state-of-the-art VLMs struggle in these tasks, particularly when distractors are similar, the input has nested structure, or when descriptions are underspecified (Junker et al., 2025; Testoni et al., 2025). These findings suggest

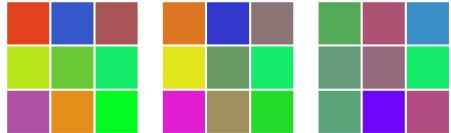


Figure 1: Example item from the dataset. The speaker referred to the second item with the description “Bottom left is bright pink.”

that reference games remain a useful benchmark for probing models’ pragmatic abilities. In this paper, we show that they can not only be used to evaluate reference resolution, but, beyond, to test whether VLMs can express their internal uncertainty through clarification behavior.

3 Methodology

3.1 Models and Data

To show that reference games are a suitable testbed for clarification behavior, we test three models on a simple reference game dataset. The 7B and 72B version of the Qwen2.5-VL family (Qwen Team, 2025; Bai et al., 2023), as its predecessor, Qwen2-VL, achieved the best performance on this task in (Junker et al., 2025), and GPT-5-mini (OpenAI, 2025), as a commercial state-of-the-art model.

We use the color-grid reference game dataset (McDowell and Goodman, 2019), which consists of 197 games each with 60 rounds. In each round, a speaker sees three 3×3 color grids (one target and two distractors) and describes the target to an addressee. The addressee must then identify the target based on this description (see Figure 1 for an example). The dataset distinguishes three conditions with different levels of difficulty based on the color similarity between the target and the distractors: *far* (easiest), *split* (medium), and *close* (hardest). These graded conditions make the task particularly well suited for quantifying how model behavior varies with referential difficulty. While the original setup allows multi-turn dialogue where dyads can interact freely, we focus on the initial speaker description only, in order to examine models’ clarification behavior before any exchange.

While we evaluate the two Qwen2.5-VL models on the complete dataset, we had to limit GPT-5-mini’s evaluation to a subset of 500 data points due to resource constraints. For the Qwen2.5-VL models we report results both for the ‘subset’ as well as for the ‘full dataset’.

3.2 Experiments

Human Data Based on the English data from McDowell and Goodman (2019), we compute **human accuracy** on the color-grid reference game in terms of the ratio of successful trials. Overall human accuracy across all conditions (on the full data) is 0.92; conditions individually increases from *close* (0.91) to *split* (0.93) to *far* (0.97), see Table 2.

Baseline Experiment Our baseline experiment is a conceptual replication of the experiment described in Junker et al. (2025). We prompt the models with a concatenated image of the three color grids and the speaker’s utterance (see Appendix C for the prompt). The model must predict the position of the target, playing the addressee role in the game.

In contrast to Junker et al. (2025), we sample each model five times per example and use the majority vote as the predicted answer. A model **Baseline Accuracy** is computed, in the same way as in the human data, based on this predicted answer. Diversity sampling (Vashurin et al., 2025) provides a more reliable estimate of a model’s performance and additionally enables us to quantify model uncertainty and to compute a simple **Baseline Confidence** measured in terms of proportion of samples matching its majority prediction. This approach yields discrete confidence levels {0.4, 0.6, 0.8, 1.0}, e.g., if 4 out of 5 samples predict the same grid, the confidence is 0.8. For Qwen-72B, we additionally report information-based uncertainty estimates based on the model’s probability distribution restricted to the three answer options (see Appendix E).

Clarification Experiment In a second experiment, we test whether the models are capable of asking for clarification when they are uncertain about the resolution of the reference. We modified the prompt from the baseline experiment and explicitly instruct the model to ask a clarification question when it is uncertain (see Appendix C for prompt).

We sample each model once and use the model’s answer to evaluate its clarification behavior, yielding (i) **Clarification Request Rate** (CR-Rate; the proportion of examples where a clarification question is generated), (ii) **Accuracy** (for the cases where the model generates a response with a prediction, i.e., not a clarification request, what is the ratio of these being correct), and (iii) **Relaxed accuracy** (rate of responses being either correct or clarification requests).

4 Results

Table 1 shows the main results across all models and conditions. Reported results are for the 500 item subset; for the Qwen2.5-VL models, we additionally report performance on the full dataset in parentheses for comparison. In the clarification experiment, it would be reasonable for the models to align their clarification behavior according to their uncertainty (as quantified in the baseline experiment): when confident about the target, they should respond with their prediction; when uncertain, they should generate a clarification request.¹

In the **Baseline Experiment**, we evaluate how accurately and confidently models identify the target without any opportunity for clarification. Model accuracy and confidence decrease with task difficulty (*far* > *split* > *close*), mirroring human performance (Table 2 in the Appendix) and previous findings (Junker et al., 2025). GPT-5-mini reaches the highest baseline accuracy (91%) and has a very high confidence (99%) in its decisions, followed by Qwen-72B with 77% (71%) accuracy and slightly lower, but still high confidence of 91% (92%). Qwen-7B’s baseline accuracy is significantly lower 53% (52%), again with slightly lower but still high confidence of 88% (0.87%). The high confidence across all models indicates a disposition to be overconfident, particularly for the Qwen models.

In the **Clarification Experiment**, we examine how frequently models generate clarification requests and how this behavior relates to their task performance and confidence. GPT-5-mini generates clarification questions in 13% of items, while Qwen-72B does so in 24% (both for the subset as well as for the full dataset). Qwen-7B, in contrast, almost never generates clarification requests (< 0.1%). Clarification request rates also vary systematically across difficulty conditions for GPT-5-mini, rising from 6% (*far*) to 17% (*split* and *close*). This is not the case for Qwen-72B, where rates do not systematically track task difficulty.

For GPT-5-mini, accuracy of generated predictions (i.e., no clarification requests) exceeds its baseline accuracy (94% vs. 91%) consistent with reasonable behavior of only generating a response when confident. It even matches human performance (see Table 2). Qwen-72B’s accuracy of generated predictions on the subset was slightly lower (73% vs. 77% baseline) with no difference on

¹Figure 2 in the Appendix illustrates how model responses transition between the baseline and clarification experiments.

Model	Cond.	Baseline		Clarification		Comparison CR-Baseline	
		Accuracy	Confidence	CR-Rate	Accuracy	Relaxed	Accuracy
Qwen2.5 VL-7B	close	0.52 (0.47)	0.87 (0.85)	0.0 (0.002)	0.46 (0.40)	0.46 (0.40)	– (0.22)
	split	0.52 (0.51)	0.87 (0.87)	0.0 (0.001)	0.44 (0.41)	0.44 (0.41)	– (0.25)
	far	0.56 (0.57)	0.89 (0.88)	0.0 (0.002)	0.46 (0.44)	0.46 (0.44)	– (0.14)
	ALL	0.53 (0.52)	0.88 (0.87)	0.0 (0.002)	0.46 (0.42)	0.46 (0.42)	– (0.20)
Qwen2.5 VL-72B	close	0.68 (0.65)	0.90 (0.90)	0.24 (0.23)	0.63 (0.64)	0.71 (0.73)	0.70 (0.57)
	split	0.75 (0.71)	0.91 (0.92)	0.23 (0.24)	0.73 (0.69)	0.79 (0.76)	0.76 (0.62)
	far	0.86 (0.78)	0.93 (0.94)	0.26 (0.26)	0.85 (0.79)	0.89 (0.85)	0.86 (0.68)
	ALL	0.77 (0.71)	0.91 (0.92)	0.24 (0.24)	0.73 (0.71)	0.80 (0.78)	0.78 (0.63)
GPT-5 mini	close	0.87	0.97	0.17	0.91	0.92	0.58
	split	0.91	0.99	0.17	0.90	0.93	0.78
	far	0.98	1.00	0.06	0.99	0.98	0.89
	ALL	0.91	0.99	0.13	0.94	0.94	0.71

Table 1: Performance and clarification behavior by model and condition (close, split, far, and ALL combined). The table shows the results for the baseline experiment, the clarification experiment, and for the comparison of clarification requests (CR) and baseline. Numbers reported are on the 500 item subset, and, for the Qwen models, also on the full dataset (in parentheses).

the full dataset (71%). Under the relaxed accuracy measure, GPT-5-mini stays at 94%, while Qwen-72B improves to 80% (78% on the full dataset). These increases suggest that clarification requests are generated for items that would otherwise result in wrong responses.

We now **compare** clarification requests to the baseline, examining how clarification behaviour relates to baseline accuracy and confidence. We begin by examining confidence patterns. Across models, items that elicited clarification requests have slightly lower confidence values than baseline confidence, which one could expect from a reasonable model that links clarification behavior to internal uncertainty. However, confidence remains uniformly high, revealing overconfidence especially for the Qwen models. For instance, Qwen-72B’s mean confidence only drops from 91% to 87% items where it generated a clarification request.

Accuracy patterns show clearer differences when comparing clarification requests with the baseline. For GPT-5-mini, items that elicited clarification requests have substantially lower accuracy than the baseline (71% vs. 91%), suggesting that the model tends to generate clarification requests on difficult items. This pattern is not stable for Qwen-72B. For the subset, accuracy is actually moderately higher than baseline (78% vs. 77%), but this reverses on the full dataset (63% vs. 71%), indicating that it tends to generate clarification requests on more difficult items. Qwen-7B, generating almost no clarification requests, in contrast, does not allow for meaningful comparison. On the few items of the full dataset

where it generated one, accuracy is much lower than baseline (20% vs. 52%) though.

5 Discussion and Conclusion

In this paper, we examined whether vision-language models (VLMs), as addressees in reference games, can deal with internal uncertainty by generating clarification requests.

Our results show that VLMs display only a limited ability to do so. GPT-5-mini achieves high accuracy overall and generates clarification requests more often when uncertain. In contrasts, the Qwen2.5-VL family models’ use of clarification requests appears to be largely decoupled from their internal uncertainty and overall task performance. Given the relative simplicity and self-contained nature of reference games, these findings highlight a gap in pragmatic abilities: even in a setting where uncertainty is quantifiable and clarification needs should be apparent, models struggle to translate uncertainty into appropriate interactional responses.

This conclusion is reinforced by our interaction analysis that shows that providing human-in-the-loop responses to these requests rarely improves performance (see Appendix F). Many of the generated clarification requests are uninformative or even inappropriate given the task, rather than genuine attempts to resolve internal uncertainty – let alone adhering to universal principles for sharing responsibility in achieving intersubjectivity in dialogue (Dingemanse et al., 2015).

Limitations

Our experiment has several limitations. In the human data (McDowell and Goodman, 2019), dyads interacted with each other for 60 rounds. This extended interaction could allow participants to build common ground (Clark and Wilkes-Gibbs, 1986; Clark, 1996) and develop conceptual pacts (Brennan and Clark, 1996). In contrast, the VLMs lack access to this iterative grounding process, which may put them at a disadvantage compared to human participants. Additionally, the models' uniformly high confidence, based on diversity sampling, may indicate poor calibration. Information-based uncertainty quantification methods may offer a more accurate reflection of internal uncertainty. However, while for Qwen-72B, such estimates show lower but more graded confidence values, the model remains overconfident and does not better align clarification behavior with uncertainty (Appendix E). Moreover, the accuracy of the baseline models could stem from potentially flawed descriptions of the speaker (mistakes in their utterance, describing the wrong grid), rather than always from the addressee.

Acknowledgments

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project no. 512393437, project B02. We also acknowledge support from SAIL: “SustAInable Life-cycle of Intelligent Socio-Technical Systems” (Grant ID NW21-059A), an initiative of the Ministry of Culture and Science of the German state of North Rhine-Westphalia.

Ethics Statement

We do not consider this work to pose ethical concerns, and, as such, it did not require an ethics review. The dataset we use is publicly available.

References

- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. *Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond*. Preprint, arXiv:2308.12966.
- David L. Bimler, John Kirkland, and Kimberly A. Jameson. 2004. *Quantifying variations in personal color spaces: Are there sex differences in color vision?* *Color Research & Application*, 29(2):128–134.
- Susan E. Brennan and Herbert H. Clark. 1996. *Conceptual pacts and lexical choice in conversation*. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.
- Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.
- Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. *Referring as a collaborative process*. *Cognition*, 22(1):1–39.
- Robert Dale and Ehud Reiter. 1995. *Computational interpretations of the Gricean maxims in the generation of referring expressions*. *Cognitive Science*, 19(2):233–263.
- Shehzaad Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. 2023. *A diachronic perspective on user trust in AI under uncertainty*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5567–5580, Singapore. ACL.
- Mark Dingemanse and N.J. Enfield. 2023. *Interactive repair and the foundations of language*. *Trends in Cognitive Sciences*, 28:30–42.
- Mark Dingemanse, Seán G. Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S. Gisladottir, Kobin H. Kendrick, Stephen C. Levinson, Elizabeth Manrique, Giovanni Rossi, and Nick J. Enfield. 2015. *Universal principles in the repair of communication problems*. *PLoS ONE*, 10:e0136100.
- Michael C. Frank and Noah D. Goodman. 2012. *Predicting pragmatic reasoning in language games*. *Science*, 336(6084):998–998.
- Simeon Junker, Manar Ali, Larissa Koch, Sina Zarrieß, and Hendrik Buschmeier. 2025. *Are multimodal large language models pragmatically competent listeners in simple reference resolution tasks?* In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24101–24109, Vienna, Austria. ACL.
- Clara Lachenmaier, Judith Sieker, and Sina Zarrieß. 2025. *Can LLMs ground when they (don't) know: A study on direct and loaded political questions*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 14956–14975, Vienna, Austria. ACL.
- Chris Madge, Matthew Purver, and Massimo Poevio. 2025. *Referential ambiguity and clarification requests: Comparing human and LLM behaviour*. Preprint, arXiv:2507.10445.
- Eric Mayor, Lucas M. Bietti, and Adrian Bangerter. 2025. *Can large language models simulate spoken human conversations?* *Cognitive Science*, 49(9):e70106.
- Bill McDowell and Noah Goodman. 2019. *Learning from omission*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 619–628, Florence, Italy. ACL.

- Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents’ overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.
- Holger Mitterer and Jan Peter de Ruiter. 2008. Re-calibrating color categories using world knowledge. *Psychological Science*, 19(7):629–634.
- OpenAI. 2025. GPT-5 System Card.
- Ildikó Pilán, Laurent Prévot, Hendrik Buschmeier, and Pierre Lison. 2024. Conversational feedback in scripted versus spontaneous dialogues: A comparative analysis. In *Proceedings of the 25th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 440–457, Kyoto, Japan. ACL.
- Ole Pütz and Elena Esposito. 2024. Performance without understanding: How ChatGPT relies on humans to repair conversational trouble. *Discourse & Communication*, 18(6):859–868.
- Qwen Team. 2025. Qwen2.5-VL technical report. *Preprint*, arXiv:2502.13923.
- Neil Rathi, Dan Jurafsky, and Kaitlyn Zhou. 2025. Humans overrely on overconfident language models, across languages. In *Second Conference on Language Modeling*, Montreal, Canada.
- Omar Shaikh, Kristina Gligoric, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6279–6296, Mexico City, Mexico. ACL.
- Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Comput. Surv.*, 58(3).
- Chenglei Si, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. 2024. Large language models help humans verify truthfulness – Except When They Are Convincingly Wrong. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1459–1474, Mexico City, Mexico. ACL.
- Judith Sieker, Simeon Junker, Ronja Utescher, Nazia Attari, Heiko Wersing, Hendrik Buschmeier, and Sina Zarrieß. 2024. The illusion of competence: Evaluating the effect of explanations on users’ mental models of visual question answering systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19459–19475, Miami, FL, USA. ACL.
- Alberto Testoni and Raquel Fernández. 2024. Asking the right question at the right time: Human and model uncertainty guidance to ask clarification questions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 258–275, St. Julian’s, Malta. ACL.
- Alberto Testoni, Barbara Plank, and Raquel Fernández. 2025. Racquet: Unveiling the dangers of overlooked referential ambiguity in visual LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China. ACL.
- Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.
- Edda Weigand. 1999. Misunderstanding: The standard case. *Journal of Pragmatics*, 31:763–785.
- Michael J. Q. Zhang and Eunsol Choi. 2025. Clarify when necessary: Resolving ambiguity through interaction with LMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5526–5543, Albuquerque, NM, USA. ACL.

A Implementation Details

We used the following models for our experiment:

- Qwen2.5-VL-7B-Instruct: temperature 0.7
- Qwen2.5-VL-72B-Instruct (quantized): temperature 0.7
- GPT-5-mini: No temperature control

We used a single NVIDIA RTX A6000 GPU for inference with the Qwen models. Depending on model size, generating responses took between 4 h and 96 h for the data. Budget for cloud infrastructure was around 5 USD (GPT-5-mini inference).

B Scientific Artifacts

In our work, we mainly used scientific artifacts in the form of publicly available datasets and model implementations (MIT, Apache 2.0). The color grid dataset is available on [GitHub](#) (MIT License). We are confident that our work is consistent with their intended use.

Data and code will be made available as a data publication on Zenodo.

Condition	Subset	Full Dataset
close	0.91	0.90
split	0.93	0.92
far	0.97	0.96
ALL	0.94	0.92

Table 2: Human accuracy by condition on the subset and full dataset.

C Prompts

Baseline Experiment Prompt

You see three color grids arranged from left to right.
A speaker describes one of these grids:
{utterance}

Which grid is the speaker referring to? Answer with exactly one word: first, second, or third.

Clarification Experiment Prompt

You see three color grids arranged from left to right.
A speaker describes one of these grids:
{utterance}

INSTRUCTIONS:

- If you can clearly identify which grid (first, second, or third) the speaker means, respond with exactly one word: first, second, or third.
- If you are uncertain, unclear, or cannot confidently determine which grid is meant, you MUST ask a clarifying question starting with "QUESTION:"

Response:

Interaction Prompt

You see three color grids arranged from left to right.
A speaker describes one of these grids:
{utterance}
The listener asked for clarification:
{question}
The speaker clarified: {answer}"

Which grid is the speaker referring to? Answer with exactly one word: first, second, or third.""
Response:

D Additional Material

Table 2 reports the human performance on color-grid data. Figure 2 visualizes for which baseline items clarification requests were generated and how consistent responses were.

Figures 3 and 4 show examples of clarification

requests generated by GPT-5-mini and Qwen2.5-VL-72B.

E Information-Based Uncertainty Estimates

In our experiment, we used consistency-based uncertainty estimations as it is the only uncertainty signal uniformly available across all models we evaluate, making it the fairest comparison. There are other, model-internal ways of quantifying uncertainty, e.g., using information theory, but these measures are not readily applicable when working with commercial models (such as GPT-5-mini).

In order to investigate, whether such information-based metrics would change our results, we explored them for the open-weights model Qwen-72B. Specifically, we re-computed confidence values using MSP (maximum softmax probability) based on the model’s probability distribution restricted to the three answer options.

Table 3 shows that, overall, the resulting confidence values are lower (compared to consistency-based confidence; Table 1) and vary more gradually with task difficulty. Similar to the consistency-based measure, MSP confidence decreases on items where the model requests clarification, but this reduction is relatively small. Despite the small improvement in confidence estimation, the model nonetheless remained overconfident when using information-based metrics.

F Interaction Experiment

To investigate whether model’s generated clarifications are effective, i.e., seek information that would help the model to correctly resolve the reference to the target grid, we conducted a follow-up interaction experiment in which a human-in-the-loop answered the model’s clarification request, giving the model another, final, chance to succeed.

Annotation and Human-in-the-Loop Response

To investigate the quality of the model’s clarification requests and their impact on model confidence and accuracy, we manually inspect all model-generated questions for Qwen-72B (116 items) and annotated whether each question is to be considered *task-relevant* or *not task-relevant*, given the color grid and original human description. We find that only 42% of the questions provided by Qwen-72B are task-relevant.

For each clarification request, a human-in-the-loop then responded to the model as a cooperative

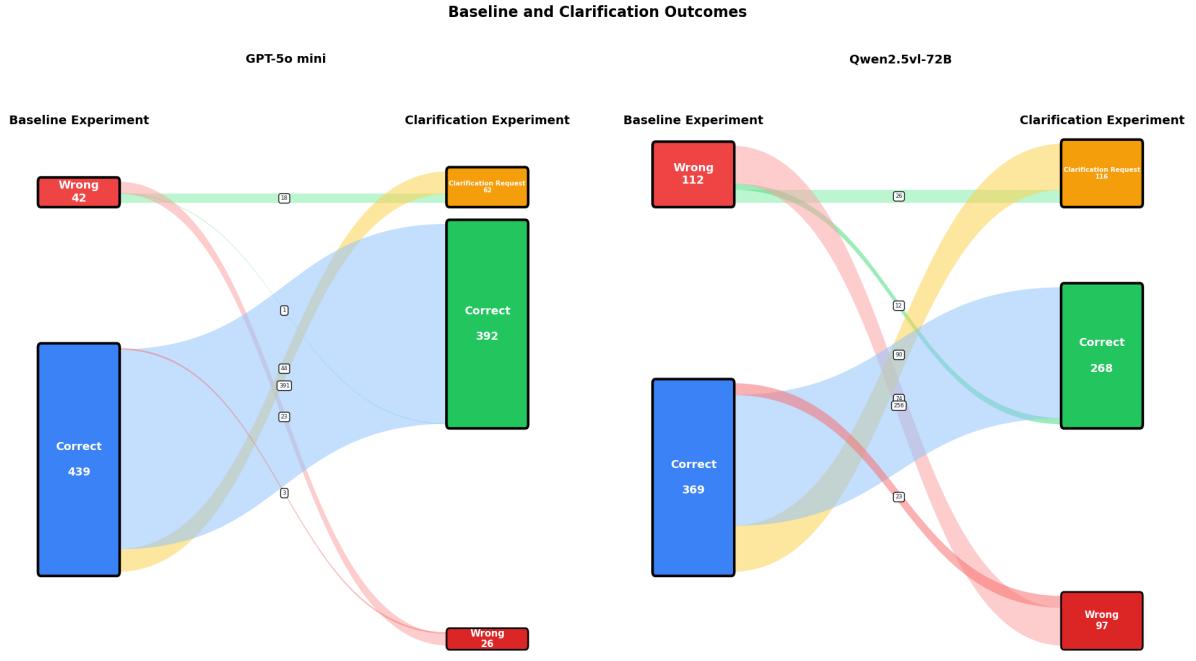
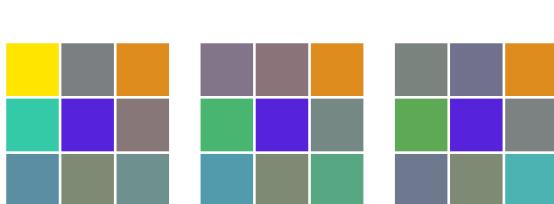


Figure 2: Sankey diagrams for GPT-5-mini (left) and Qwen2.5 VL-72B (right) showing each model’s outcomes in the baseline and clarification experiments. The flow indicates for which baseline items clarification requests were generated and how consistent responses were.

Cond.	Baseline		Comparison CR-Baseline	
	Accuracy	MSP Confidence	Accuracy	MSP Confidence
close	0.66	0.80	0.70	0.73
split	0.75	0.83	0.68	0.77
far	0.88	0.88	0.88	0.80
ALL	0.76	0.84	0.76	0.77

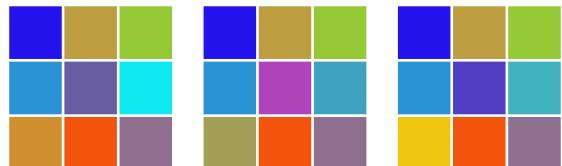
Table 3: Qwen-72B performance and *information-based confidence estimate* by condition (CLOSE, SPLIT, FAR, and ALL combined). The table shows the results for the baseline experiment with MSP (maximum softmax probability) confidence results and the comparison of clarification requests (CR) and baseline. Numbers reported are on the 500-item subset. We use the maximum MSP as a confidence score and select the highest-probability option to predict the VLM answer.



Human speaker: “bottom right is grey not the bluish not the greenish” (referring to the first item).

GPT-5-mini: “QUESTION: Which of the three grids do you mean — the left, middle, or right one? None of their bottom-right squares looks plainly grey, so can you confirm which grid or clarify the shade you mean?”

Figure 3: Example of GPT-5-mini behavior in the clarification experiment.



Human speaker: “Middle square is dark purple.” (referring to the third item).

Qwen2.5-VL-72B Baseline: “First” (incorrect).

Qwen2.5-VL-72B Clarification: “Could you please clarify which grid you are referring to?”

Annotation: not task-relevant (generic, fails to identify specific ambiguity)

Human-in-the-loop: “The grid with the darkest purple in the middle.”

Qwen2.5-VL-72B: “Third” (correct).

Figure 4: Example of Qwen2.5-VL-72B behavior with human-in-the-loop interaction.

Items	Condition	Before		After		Δ	
		Accuracy	Confidence	Accuracy	Confidence	Accuracy	Confidence
CR-only	close	0.703	0.859	0.784	0.859	+0.081	+0.000
	split	0.757	0.886	0.649	0.876	-0.108	-0.010
	far	0.857	0.867	0.786	0.962	-0.071	+0.095
	ALL	0.776	0.871	0.741	0.902	-0.035	+0.031
full	close	0.682	0.898	0.701	0.898	+0.019	+0.000
	split	0.753	0.912	0.728	0.910	-0.025	-0.002
	far	0.864	0.930	0.846	0.954	-0.018	+0.024
	ALL	0.767	0.914	0.759	0.921	-0.008	+0.007

Table 4: Accuracy and confidence of the Qwen2.5-VL-72B model in the interaction experiment by condition (close, split, far, and ALL combined). **Before** corresponds to the original speaker utterance-only baseline prompt, whereas **After** replaces the utterance with a dialogue-conditioned prompt containing the human-in-the-loop response. Δ is the difference between Before and After. The upper part of the table (CR-only) reports on the subset of 116 items where the model requested clarification. The lower part (full) reports on all items. Here, after keeps the 365 items without clarification requests unchanged and replaces the 116 items on which the model previously asked a clarification request with the dialogue-conditioned prompts using human-provided clarification answers. We report majority-vote accuracy and mean consistency confidence.

speaker. When the model’s clarification request was labeled task-relevant, the human answered it directly in order to resolve the model’s difficulty (e.g., a perceived ambiguity in the original description). When the clarification request was not considered task-relevant (see Figure 4 for an example), the human instead reformulated the original utterance to provide a clearer or more detailed description of the target grid. Given the constrained nature of the reference game, we employed a single expert as annotator and human-in-the-loop. This person (one of the authors) was familiar with the task to provide clarification responses and had experience with the color-grid dataset.

Interactivity Effect To investigate interactivity effects on uncertainty and end-to-end accuracy, we re-run the Qwen-72B model on the human-in-the-loop clarified dialogues (see Appendix C for the prompt) and recomputed accuracy and uncertainty estimates, expecting both accuracy and

confidence to increase after clarification. However, Table 4 shows that providing human answers to Qwen-72B’s clarification requests does not lead to performance improvements. While confidence increases after clarification, end-to-end accuracy decreases slightly (except for the close condition), suggesting that the provided answers to the clarification request are either not utilized well by the model, or that the model’s clarification request was not useful in the first place (either not task-relevant or not seeking information that the model actually needed).

G Use of AI Assistants

AI assistants were used during manuscript preparation solely for specific linguistic reformulation to refine clarity and style, and to assist with code writing, primarily for the visualization (Figure 2). No AI-generated content was used to interpret results, and all conclusions were drawn by the authors.