

Probing the Communicative Efficiency of LLMs via Content Selection in Visually Uncommon Reference Games

Simeon Junker[†] Manar Ali[†] Kristina Kobrock^{†,‡} Larissa Koch[†]

Marika Sarzotti^{||} Hendrik Buschmeier[†] Sina Zarrieß[†]

[†]CRC 1646 ‘Linguistic Creativity in Communication’, Bielefeld University, Bielefeld, Germany

[‡]Institute of Cognitive Science, Osnabrück University, Osnabrück, Germany

^{||}Department of Linguistics, University of Potsdam, Potsdam, Germany

Abstract

We study content selection in reference games using “color grids” as a controlled testbed with abstract visual stimuli. We extend an existing human–human dialogue dataset by adding further annotations concerning the grid segments selected by the speakers. We generate referring expressions for the same grids with open-weight LLMs, and evaluate whether these expressions are discriminative without being overly verbose. We find that human content selection is highly efficient and reflects location preferences consistent with known cognitive biases. LLMs above a certain size produce truthful, discriminative descriptions, but frequently over-specify; however, certain models balance effectiveness and effort, especially with “thinking mode” enabled. An analysis of the generated expressions replicates the human location biases, suggesting that certain LLMs reproduce implicit human perceptual biases in content selection.¹

1 Introduction

Human communication is often conceived of as a process in which speakers and listeners aim to be efficient by balancing individual and collaborative processing costs against benefits such as achievement of mutual understanding or communicative goals (Shannon, 1948; Grice, 1975). This trade-off has been studied particularly extensively in the domain of pragmatics and reference production (Clark and Wilkes-Gibbs 1986; Rubio-Fernandez et al. 2021; Degen et al. 2020; Tourtouri et al. 2019; Kobrock et al. 2026, among many others). Recent work in computational linguistics has started to apply theoretical concepts and experimental settings from pragmatics to the analysis of language models and their capabilities (cf. Chang and Bergen, 2024; Ma et al., 2025a). Many existing pragmatic probing

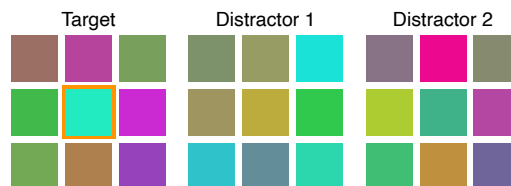


Figure 1: Annotated round from the color-grid reference game (McDowell and Goodman, 2019). A speaker refers to the target grid (left) with the utterance “turquoise looking blue in the middle”. In annotation, the text span was mapped to the middle cell of the target grid (orange highlight).

studies, though, focus on minimal pair-style settings that test models’ abilities in differentiating pragmatically appropriate from inappropriate sentences, cf. Sieker and Zarrieß (2026), rather than studying general concepts such as communicative efficiency. Fried et al. (2023) and Bunzeck et al. (2026) argue that research in this area should move towards more realistic multimodal and interactive tasks.

In this work, we investigate whether and how humans and current open-weight vision-language LLMs achieve communicative efficiency in a visually complex reference game. In this game, an efficient utterance would be a referring expression that is short and effective, i.e., understood by the listener. To investigate how humans and models account for effectiveness and informativeness, we examine the level of *content selection*; specifically how many and which properties of the referential target they include in referring expressions. The problem of content selection has been central to the field since its pioneering work (see Section 2).

In more detail, we investigate content selection in a visually abstract reference game setting with 3×3 color grids (McDowell and Goodman, 2019; see Figure 1), focusing on the question of *which segments* of a target grid are selected for reference in different conditions. For our purposes, this setting offers two main advantages: On the one hand, unlike Referring Expression Generation (REG) set-

¹Data and code for this project are available at: <https://github.com/clause-bielefeld/color-grid-efficiency> and <https://doi.org/10.5281/zenodo.22306409>.

things with natural objects, the feature space that needs to be communicated is comparably small and could, in principle, be exhaustively enumerated, such that informativeness of referring expressions can be quantified by the number of target segments mentioned. On the other hand, the visual domain is unusual and difficult in the sense that discriminative properties that distinguish the target from its distractors need to be determined in effortful visual search. Taken together, this setting allows us to analyze the interplay of communicative effectiveness and informativeness with visual biases in humans’ and models’ referring expression generation.

We start by collecting an additional layer of annotations for the data of [McDowell and Goodman \(2019\)](#), specifying which parts of the grids are referred to in which spans of the descriptions (see [Figure 1](#)). We then generate referring expressions for the same grids using a diverse set of state-of-the-art, open-weight LLMs, and conduct a human evaluation assessing whether these expressions are contextually discriminative without being excessively over-specified. Finally, we conduct an LLM-based annotation of the target segments referred to in model-generated expressions, using few-shot examples from our manual annotation.

Our work aims to answer two central questions: (i) to what degree can current LLMs balance effort and task success as observed in human communication, and (ii) what biases do humans and LLMs exhibit when selecting target properties in our visual setting. Our results show that while sufficiently large LLMs generally produce descriptions that effectively discriminate targets from distractors, this is often achieved at the cost of substantial over-specification. Some systems, however, handle the trade-off between referential success and conciseness at near-human levels. In addition, an in-depth analysis of selected target cells reveals systematic location biases in the human data that reflect well-established cognitive biases in visual attention, and our best-performing systems mirror these biases to a certain extent.

2 Background

Referring Expression Generation and Content Selection Research on reference games, in which a speaker describes a target object among distractors to a listener who must identify it, examines how interlocutors reason about situational context, when and to what extent they behave cooperatively or rationally, and how they coordinate behavior over time

([Garrod and Anderson, 1987](#); [Clark and Wilkes-Gibbs, 1986](#); [Dale and Reiter, 1995](#); [Krahmer and van Deemter, 2019](#); [Goodman and Frank, 2016](#)). From a Gricean perspective ([Grice, 1975](#)), the task of a speaker in a reference game can be seen as a balancing act between producing a referring expression that is truthful and achieves communicative success while also being efficient, i.e., not overly costly in production and comprehension. [Dale and Haddock \(1991\)](#) and [Dale \(1992\)](#) argue that Referring Expression Generation, the task of producing the speaker’s description, is guided by Gricean-like conversational maxims, namely adequacy (uniquely identify the referent), efficiency (avoid redundancy), and sensitivity (adapt to listener). These maxims were later operationalized into computational interpretations, most notably the Incremental Algorithm ([Dale and Reiter, 1995](#)), which incrementally adds attributes that eliminate distractors, ensuring adequacy but relaxing efficiency; this mirrors human speakers’ well-documented tendency to *overspecify* referents ([Pechmann, 1989](#)). Such approaches often focus on the problem of *content selection*, i.e., deciding which attributes distinguish the target from distractors (see [Schüz et al. 2023](#); [Ali et al. 2025](#) for recent REG overviews).

Task Success and Efficiency in LLMs More recently, research on LLMs has also adopted paradigms and datasets from pragmatics ([Ma et al., 2025a](#)). [Askari et al. \(2025\)](#), for instance, show that even small language models show some sensitivity to violations of Gricean maxims, but still fall below human and large-scale LLM performance and exhibit the clearest limitations in capturing the Gricean maxims of quantity. In the specific context of REG, [Ma et al. \(2025b\)](#) show that LLMs produce overinformative descriptions of the target and favor different attributes than humans. In a similar vein, other studies found LLMs to not perform well as pragmatic listeners ([Junker et al., 2025](#)) and failing to generate appropriate clarification requests in interactive reference game settings ([Ali et al., 2026](#)). In this work, we contribute to this line of research with a focus on content selection, examining how humans and LLMs balance task success and efficiency when generating referring expressions for visually uncommon targets.

3 Dataset

This section introduces the existing dataset of human reference games on which we base our study



Figure 2: Frequency of mentions of individual target grid segments in human and LLM descriptions for the full sample of 25 games and the individual difficulty conditions (*far*, *split*, *close*). (a) is based on human annotations, (b) on LLM-generated annotations. Further examples are shown in Appendix J.

(Section 3.1). We extend this data with referring expressions generated by LLMs (Section 3.2).

3.1 Human Data

We use data from McDowell and Goodman (2019) (see Figure 1), consisting of 197 reference games with 60 rounds each. In each round, participants were assigned the role of “speaker” or “listener” and were shown three 3×3 grids of color patches, presented in a different order. The speaker was informed of the “target” grid and produced a textual description that the listener used to identify it. Items vary in difficulty depending on the color distance between the target and distractors: In the *far* condition (easy), both distractors are dissimilar to the target; in *split* (medium), one distractor is similar and the other dissimilar; and in *close* (most difficult), both are similar to the target.

We manually annotated a subset of this dataset to align the colors and cells mentioned in the referring expressions with visual representation of the grid. We randomly sample 25 games, totaling 1500 rounds, after excluding games with success rates below 80 % or any missing rounds (18.7 % of all games excluded). The annotation procedure is described below. We use the same sample for our generation experiments.

3.2 LLM-generated Descriptions

We generate referring expressions for item triplets in our dataset, each consisting of a target grid and two distractor grids, using a diverse set of instruction-tuned vision-language LLMs of varying sizes, both with and without “thinking” mode, i.e., their inter-

nal chain-of-thought reasoning mechanisms (Wei et al., 2022; OpenAI, 2024)². Implementation details are provided in Appendix D. We evaluate several variants of Qwen: 3.5-4B, 3.5-9B, 3.5-27B (Qwen Team, 2026a), and 3.6-27B (Qwen Team, 2026b). We also evaluate Gemma-4 (Gemma Team, 2026): the dense models E4B-it and 12B-it, and the mixture of experts (MoE) model 26B-A4B-it.

The LLM prompts consist of rendered images of the color grids for a given round, with the target highlighted by a green border (see Figure 3). We prompt the models to generate descriptions of the target grids that are concise but sufficiently informative (full prompts in Appendix E). We run inference with greedy decoding.

4 Annotation and Evaluation

The aim of our study is to compare LLM-generated to human-produced referring expressions in terms of their communicative efficiency. First, we carry out a human evaluation of human and LLM-generated expressions, assessing accuracy and informativeness (Section 4.1). To be able to analyze and quantify which parts of the grids humans and LLMs refer to, we manually annotate referred-to segments in the human referring expressions (Section 4.2). Based on this, we collect and validate model-based segment annotations (Section 4.3).

²We note that the term “thinking” is not ideal, as it conflates the technical operations of LLMs with human cognitive abilities (Bender and Inie, 2026). Nevertheless, we use the term in the following, as it is widely used in the related literature, as well as in the model reports and frameworks used in this work.

4.1 Human Evaluation

We conduct a human evaluation to assess the LLMs’ ability to produce referring expressions that are communicatively efficient, i.e., discriminative in context but also concise. We report target identification accuracy, average informativeness of the descriptions, and the frequency of cases in which descriptions are ambiguous or fail to accurately describe any of the presented grids.

We sample 150 item IDs from our annotated subset of the original dataset for our human evaluation, with 50 items drawn from each of the *far*, *split*, and *close* conditions. To ensure comparability between human and LLM descriptions, we restrict the sample to items whose referential dialogues consist of a single speaker utterance, after which the listener successfully decided on a target grid. We select the speaker-produced descriptions from McDowell and Goodman (2019) along with the descriptions generated by each system for these items, and distribute them across three expert evaluators (including one of the authors). Each evaluator annotates a distinct subset of items, with three items per system (one per condition) shared across annotators to estimate inter-annotator agreement.

We conducted the evaluation using oTree (Chen et al., 2016). For each item, evaluators were shown the target and distractor grids in a web-based interface and asked to select the grid corresponding to the displayed description. Alternatively, they could indicate that no grid was truthfully described (*none fits*) or that the description was too ambiguous to identify a unique target (*multiple fit*). In addition, evaluators rated the informativeness of each description on a 5-point scale, where 1 indicates a severe lack of information, 5 excessive information, and 3 an appropriate level of detail. For the analysis, we center these ratings to the interval $(-2, 2)$, such that scores < 0 indicate insufficient information, scores > 0 indicate overly detailed descriptions, and 0 corresponds to the optimal midpoint.

We estimate inter-annotator agreement using Fleiss’ kappa (Fleiss, 1971). For target selection, including the *N/A* alternatives, we obtain $\kappa = 0.77$ (*substantial* agreement). For the informativeness ratings, again treating *none fits* as an *N/A* category, we obtain $\kappa = 0.41$ (*moderate* agreement).

4.2 Manual Segment Annotations

We add manual annotation to our sample of McDowell and Goodman (2019)’s dataset, consisting

of mappings between spans in the referential dialogues and segments of the target grids that are described by the respective text sections. We use the annotations to analyze content selection behavior and visual biases in the human descriptions, and to sample few-shot examples for our LLM-based annotation procedure (see below).

We set up a custom annotation interface with Label Studio (Tkachenko et al., 2025) where color grid triplets are displayed jointly with the speaker and listener utterances. Annotators mark text spans that refer to parts of the grids, and link those spans to associated grid cells by selecting from 27 unique labels, one for each cell of the target and distractor grids. The annotation guidelines were tested and refined in two pilot annotation phases with a small subset of the data; the main annotation was performed independently by three trained annotators (including two of the authors). The final spans and cell labels are determined by majority voting (details in Appendix C).

For the start and end positions of the annotated spans, all annotators agree in 59.6 % of cases, and a consensus can be determined by majority voting in a total of 82.9 % of cases. When considering both span boundaries and assigned labels, we find perfect agreement in 48.5 % of cases and a majority vote for both spans and labels in 78.3 % of cases. Our final annotations include an average of 1.36 spans per round and 1.45 labels per annotated span, with 1.35 being newly introduced labels and 0.1 being repeated labels that had already been mentioned in an earlier span for the given item.

4.3 Model-based Segment Annotations

As with our sample of human referential dialogues, we collect annotations of mappings between cells mentioned in the LLM-generated referring expressions and the grid, to analyze content selection in LLM-generated descriptions. Due to time and financial constraints, obtaining human annotations for all LLM descriptions is not feasible. We therefore perform an additional LLM-based annotation following the procedure described above, using the vision-language MoE model Qwen36-35B-A3B with FP8 quantization as a proxy for human annotators (Qwen Team, 2026c).

The annotation prompt is based on the instructions provided to human annotators (see Appendix E). Given the same images used for the human annotation, the model produces a JSON object for each item, identifying the referential spans in the

description, along with the target and distractor cells referenced by each span. We use an in-context learning setup in which the model is provided with annotations for five distinct items sampled from the human-annotated data, selected to ensure sufficient diversity among the examples (see Appendix F).

To validate this procedure, we re-annotate the human descriptions using our automatic method and compare the resulting annotations against the human labels obtained through majority voting. When considering unique cells referenced in each description and excluding references to distractor cells, the human and LLM annotations are identical for 90.6 % of items.

4.4 Segment-level Evaluation

The segment-level annotation described in Section 4.3 allows us to analyze how many and which cells from a color grid humans and LLMs select when referring. In the following sections, we report: (i) the average number of referred-to cells ranging between 1 and 9, (ii) location-wise distributions of referred-to cells, and (iii) the Intersection-over-Union (IoU) between cells in human-produced and LLM-generated descriptions for corresponding items.

5 Results: Human Descriptions

Task Success Based on the listener data in [McDowell and Goodman \(2019\)](#), the macro-averaged success rate across all games in our annotated subset is 95.0 % (SD = 4.1; range: 84.5–100.0 %), indicating that in the vast majority of rounds, listeners are able to identify the correct target based on speaker descriptions. The accuracy score in our human (listener) evaluation with 150 items is lower but remains high: 88.0 % of the human descriptions are correctly resolved by our evaluators (Table 1), with *none fits* selected in 2.7 % of cases and *multiple fit* slightly more common at 5.3 %. Importantly, the original data collection was carried out in an interactive dialogue setting, where listeners could ask for additional information or clarification if they could not identify the target based on the speaker’s initial description.

Informativeness On average, the human speaker-listener dyads in the original dataset produced 1.33 utterances per round (SD = 0.50; range: 1.00–3.53), suggesting that a high proportion of rounds were resolved within a single conversational turn. Our results confirm this brevity tendency on the level of content selection within the referring ex-

pressions: Based on our human cell annotations, the number of target cells described in referential dialogues was, on average, 1.56. The results of our LLM-based cell annotation with 150 items support these results, with a slightly higher average of mentioned cells (1.63 unique cells per description, see Table 1). Thus, most speaker predominantly rely on minimal, single-cell descriptions while others tend to select a higher number of target segments (SD = 0.51; range: 1.07–3.13). When broken down by condition, a slightly higher number of cells are selected as contextual pressures increase: speakers in the far condition describe the fewest target cells on average (1.39, SD = 0.47, range: 1.00–3.00), followed by the split condition (1.63, SD = 0.58, range: 1.00–3.20). The close condition shows the highest mean number of target cells (1.67, SD = 0.55, range: 1.05–3.25).

According to the human evaluation of informativeness (scale: –2 to 2), human descriptions sampled from the original dataset were rated with an average of –0.16, i.e., slightly under-informative, but received optimal ratings in most cases (see Appendix I). We note, however, that the descriptions were produced in a dialogue setting where speakers and listeners can generally adapt the ways in which they produce and understand linguistic utterances, possibly requiring less overt information for identification than if the descriptions were presented in isolation.

Visual Biases Figure 2 illustrates the human location preferences with regard to the grid segments selected for referring to the items, both with respect to the entire data as well as subsets restricted to items from the *far*, *split*, and *close* splits. The heatmaps represent the relative frequency of references to the respective segments of the target grids, regardless of whether other target or distractor segments are also described in the same spans.

For the full data, we find global preferences for certain parts of the target grids: Averaged over all dyads and rounds, the top row (most notably the top left cell) and the center position are more often referred to than the remaining parts of the grids. This reflects more general findings in work on asymmetries in visual perception (e.g., [Spotorno and Tatler 2025](#)).

The heatmaps restricted to items for the far, split, and close conditions show interesting differences: The preferences found in the full data are most pronounced in the far condition, whereas the distri-

	model	think.	Human Evaluation				Cell Annotation			
			info.	acc.	none	multi	err.	no ref.	# cells	IoU
	human	—	−0.16	88.00	2.70	5.30	0.00	0.00	1.63	—
Qwen	3.5-4B	✓	1.02	78.67	12.67	4.00	0.87	6.25	8.45	0.20
		✗	1.08	47.33	48.00	1.33	0.20	27.72	8.74	0.19
	3.5-9B	✓	0.71	88.67	2.00	6.67	2.20	0.14	4.97	0.29
		✗	−0.20	56.67	8.00	16.67	0.33	0.80	3.67	0.29
	3.5-27B	✓	0.02	93.33	0.67	4.67	1.40	0.00	2.22	0.36
		✗	−0.15	81.33	4.00	9.33	0.20	0.80	1.57	0.40
	3.6-27B	✓	−0.01	92.67	1.33	3.33	1.00	0.20	2.21	0.37
		✗	−0.12	81.33	4.00	8.67	0.20	2.14	1.88	0.39
Gemma-4- <i>*-it</i>	E4B	✓	−0.13	81.33	4.67	10.00	0.20	1.67	4.19	0.27
		✗	−0.61	46.00	10.00	30.00	0.80	1.14	7.27	0.19
	12B	✓	1.25	91.33	2.67	3.33	2.73	0.07	7.88	0.22
		✗	1.35	94.00	0.00	3.33	0.20	0.00	8.84	0.19
	26B-A4B	✓	1.26	90.67	4.67	3.33	4.20	0.07	8.48	0.20
		✗	1.03	90.67	1.33	4.67	0.20	0.67	7.82	0.21

Table 1: Results of the human evaluation and model-based cell annotation for human- and LLM-generated descriptions. For the human annotation, *info.* indicates mean informativeness, *acc.* the % of correctly identified targets, *none* the % of descriptions matching none of the presented grids, and *multi* the % of ambiguous descriptions. For the cell annotations, *err.* reports the % of invalid descriptions, *no ref.* the % of descriptions with no information mappable to target cells, *# cells* the average number of target cells described in the remaining descriptions, and *IoU* the Intersection-over-Union with human-selected cells.

bution over grid positions is much more balanced in the close condition, where target and distractor grids are highly similar, requiring more careful visual search procedures. The patterns for the split condition fall between those extremes. This indicates that participants prefer certain parts of the grids when they can decide more freely between alternatives, but adjust to increased task demands by deviating from these preferences.

Summary Our results show that in this dataset, humans achieve a high level of communicative efficiency. The reported success rates in McDowell and Goodman (2019) and the accuracies in our human evaluation show that human descriptions allow for successful target identification in the vast majority of cases. At the same time, dialogue lengths, informativeness ratings and average counts of cells included in the descriptions indicate that humans aim to keep individual as well as collaborative efforts minimal. In addition, our cell annotations point to visual biases that lead to the preference of particular (top and left) parts of the grids.

6 Results: LLM Descriptions

Task Success The results of our human evaluation of LLM descriptions (Table 1) show that, across architectures, models with a sufficiently high parameter count consistently achieve satisfactory

accuracy ratings. All variants of gemma-4-12B-it and gemma-4-26B-A4B-it, but also Qwen3.5-27B and Qwen3.6-27B with thinking achieve better accuracy than human descriptions. For Qwen models, the accuracy consistently increases if model thinking is activated. This is not the case for Gemma, most notably gemma-4-12B-it without thinking, which achieves the highest accuracy of all systems and surpasses the human result by 6 points. The rates of false or ambiguous descriptions (*none fits* and *multiple fit*) are generally low for sufficiently large models across both model families. In this respect, Qwen3.5-9B appears to mark a lower size threshold, yielding satisfactory performance in thinking mode, but not in non-thinking mode.

In the LLM-based cell annotations, we observe a strikingly high proportion of *no reference* annotations for Qwen3.5-4B without thinking (Table 1). A closer inspection reveals that this can be mainly attributed to cases in which the description model refers to the target by means of its target marker (e.g., “the grid with the green rectangle”), which was determined by our annotation LLM as lacking referential information that can be mapped to particular target cells (see Figure 3), also see Appendix F. This suggests that, without thinking, this model struggles with fundamental aspects of the task at hand, namely, describing the target grid based on inherent properties. The rates of errors, where the

annotation LLM was unable to process or parse the input descriptions, generally increase with activated thinking modes, which can be attributed to e.g., timeout errors due to increased computational demands.

Informativeness The accuracy results of our human evaluation show that several LLMs outperform the human baseline with regard to resolution accuracy. However, in many cases this advantage appears to be driven by severe overspecification: For example, all Gemma systems with at least 12B size deviate from the optimal midpoint by more than one rating step, indicating excessive information provided in the descriptions. By contrast, Qwen3.5/3.6-27B in thinking mode produce descriptions that are more effective than human-generated descriptions while also remaining closer to the optimal level of informativeness.

This is also supported by the automatic cell annotations in Table 1. There is high variance regarding the average numbers of target cells that are mentioned in descriptions: Qwen3.5/3.6-27B closely resemble human brevity when thinking modes are disabled, and remain only slightly more verbose with explicit reasoning. In contrast, other systems resort to near-exhaustive descriptions, including Qwen3.5-4B and all Gemma variants except gemma-4-E4B-it in its non-thinking mode. Systems judged to overspecify in the human evaluation generally also show high cell counts in the automatic annotations. However, some systems judged to *underspecify* (e.g., gemma-4-E4B-it) show similarly high cell counts, raising questions about their capacity to produce accurate or sufficiently distinctive descriptions. The tendency of many systems to provide excessive information may reflect known biases toward verbose LLM responses (Saito et al., 2023), but also the absence of pressures that shaped human behavior in the original task: Human speakers operated in real-time chat with a listener awaiting a timely response, and could add information in later turns if an initial description proved insufficient. Interestingly, within the Qwen model family, the number of mentioned target cells generally decreases as model size increases, mirroring our earlier finding of an increasingly favorable balance between effectiveness and efficiency in these systems. This suggests that Qwen3.5/3.6-27B in particular exhibit signs of strong contextual reasoning, being able to reliably identify a small number of target segments that are informative for listeners.



human	brightest pink middle → [5]
Qwen3.5-4B	A 3x3 grid of colored squares with a thick white border surrounding it → []
+ thinking	A 3x3 grid with a thick green border. Top row: green, grey-blue, teal. Middle row: blue, magenta, grey-green. Bottom row: dark blue, teal, olive green → [1,2,3,4,5,6,7,8,9]
Qwen3.5-27B	The grid with a bright pink center square and a dark blue square in the bottom left corner → [5,7]
+ thinking	The grid with a green square in the top left corner and a dark blue square in the bottom left corner → [1,7]

Figure 3: Descriptions for the target grid (marked green) generated with Qwen3.5-4B and Qwen3.5-27B, LLM-based cell annotations presented in → [square brackets]. Qwen3.5-4B fails to produce useful references without and generates exhaustive descriptions with thinking mode. Qwen3.5-27b produces concise descriptions that are slightly more discriminative with thinking.

We used a Bayesian ordinal regression model in R with brms to quantify how humans and LLMs differ in informativeness across the three different context conditions (Bürkner and Vuorre, 2019; Nicenboim, 2026; Bürkner, 2017). For this model, we excluded data from smaller LLMs ($\leq 4B$), as the above analyses have shown their outputs to be not reliable. We define a variable `human_vs_llm`, that encodes whether the utterance stems from a human or LLM, to facilitate comparison between human informativeness and the LLMs’ informativeness. Our statistical model predicts the informativeness of utterances on a 5-point-ordinal scale, ranging from -2 (underinformative) to 2 (overinformative), by context condition (*split*, *far*, *close*), `human_vs_llm` (*human*, *LLM*), and the interaction of these two variables. We used the cumulative model family with link ‘probit’ and default priors. The regression model predicts a substantial difference between humans’ and LLMs’ informativeness (main effect: $M = 0.74$, $CrI = [0.43, 1.05]$, $pd = 100\%$).³ In Figure 4, we plot the model estimates with their 95 % Credible Intervals (CrI). These intervals are regions of the posterior distribution of which we can be 95 % certain that the parameter lies within. Thus, if CrIs are non-overlapping, this speaks to a difference between informativeness ratings. In the *far* condition, humans ($M = 0.47$, $CrI =$

³We used the bayestestR package for a Bayesian test of the probability of the direction (pd) of an effect that can be used as a Bayesian test for the existence of an effect (Makowski et al., 2019).

[0.44, 0.50]) achieve optimal ratings of 0 more often than LLMs ($M = 0.35$, $\text{CrI} = [0.32, 0.38]$) with a certainly existing difference between ratings ($M = 0.12$, $\text{CrI} = [0.09, 0.15]$, $\text{pd} = 100\%$). Additionally, humans are more underinformative than LLMs, i.e., they have a higher probability of getting an informativeness rating of -2 ($M = 0.06$, $\text{CrI} = [0.03, 0.12]$, $\text{pd} = 100\%$). LLMs, on the other hand, are more overinformative than humans, i.e., more likely to get an informativeness rating of 2 ($M = -0.24$, $\text{CrI} = [-0.30, -0.17]$, $\text{pd} = 100\%$). In the *split* condition, we observe a very similar picture, but LLMs approximate humans more and achieve optimal informativeness ($M = 0.42$, $\text{CrI} = [0.39, 0.45]$) almost as often as humans ($M = 0.46$, $\text{CrI} = [0.42, 0.49]$) with no likely existing difference between humans and LLMs ($M = 0.04$, $\text{CrI} = [0.00, 0.07]$, $\text{pd} = 97.33\%$). In the *close* condition, the CrI s of human and LLM informativeness ratings overlap for each of the ratings, which means that in this condition, LLMs are similarly informative to humans. This is likely due to the way the *close* condition is set up: When color grids that are very similar need to be discriminated from each other, the ways how to do that are limited, leading to LLMs and humans converging on similar production strategies.

Visual Biases The Intersection-over-Union scores reported in Table 1 show that Qwen3.5/3.6-27B achieve the highest overlap with the target cells selected in human descriptions, suggesting that these systems exhibit the most human-like selection behavior. In thinking mode, however, both models deviate slightly more from human selections.

Taking a closer look at *which* segments of the target grids are selected, Figure 2b illustrates the frequency with which individual target cells are mentioned in descriptions generated by Qwen3.5-27B with thinking enabled, using a heatmap format analogous to Figure 2 for the human data. Generally, the model exhibits patterns similar to those observed in humans, with a bias toward the top row of the grid and especially the upper-left cell. This result is particularly noteworthy given that the model has no access to human response data or other explicit cues about positional biases in the prompt, suggesting that the observed biases are implicitly acquired from the training data rather than induced by the task setup. Consistent with the human results, the effect is strongest in the *far* condition, where lower

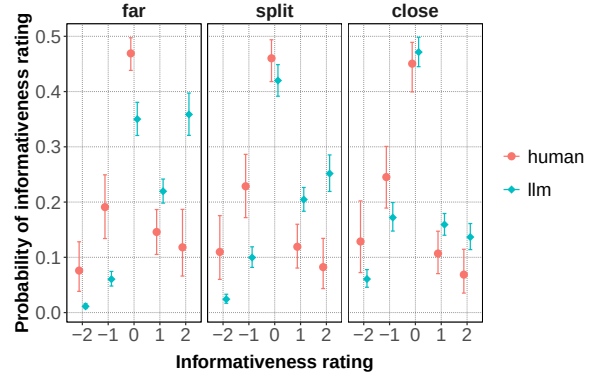


Figure 4: Regression model estimates for the informativeness ratings.

contextual pressure allows the model to select target cells with greater freedom while preserving discriminability, and gradually weakens across the *split* and *close* conditions.

Summary Our results point to substantial differences between the evaluated systems in both task success and informativeness. Several systems achieved accuracy scores in the human evaluation that surpassed the human baseline, but often at the cost of overly informative utterances, with many systems resorting to near-exhaustive descriptions, thereby revealing severe deficits in *communicative efficiency*. Qwen3.5/3.6-27B represent noteworthy exceptions to this pattern, achieving near-human trade-offs between task success and conciseness. They also show high overlap with human selections for target properties, as well as signs of inherited low-level visual biases. The comparison of human and LLM informativeness ratings reveals that humans and LLMs differ in how they achieve communicative efficiency: When diverging from the optimal level of informativeness, humans tend toward underinformativeness, whereas LLMs tend toward overinformativeness. A possible interpretation is that humans aim to be efficient even at the risk of underinformativeness, whereas LLMs are less affected by communicative costs and default to overinformativeness. The apparent underinformativeness among human dyads can be seen as an efficient strategy if considering that the descriptions were produced in a dialogue setting in which shared conversation histories potentially allow listeners to make sense of underinformative utterances. In addition, the availability of fallback strategies such as clarification further limits the risks of unsuccessful communication due to missing information.

7 Discussion and Conclusion

In this work, we investigated whether and how humans and current open-weight LLMs achieve communicative efficiency when referring to novel, highly abstract color grids. For this we focus on content selection, a long-standing problem in REG that has received little recent attention due to a shift toward LLMs and large image-text corpora, where content selection is difficult to assess and model explicitly. To this end, we collected a complementary set of annotations for the data of McDowell and Goodman (2019), indicating which segments of the target grids are described in referential dialogues between human dyads. We then generated descriptions using a range of LLMs and evaluated them through both human judgments and a focused analysis of content selection patterns, employing an LLM-based annotation procedure that builds on our human annotations.

Our results show that most models above a certain size threshold achieve near-human task success, seemingly in contrast to other work reporting mixed LLM performance on resolving referring expressions to color grids (Junker et al., 2025; Ali et al., 2026). Importantly, however, most systems achieve this success at substantial cost, producing near-exhaustive descriptions in many cases. The largest Qwen models are a notable exception, balancing descriptive cost and communicative success at near-human efficiency. The tendency of most LLMs toward excessive informativeness is particularly noteworthy in light of early REG work, which regarded algorithms that generate strictly minimal descriptions as cognitively implausible, motivating, e.g., the *Incremental Algorithm* which can also produce redundant properties (Dale and Reiter, 1995). Our findings suggest the opposite pattern: Whereas humans frequently produce short, concise descriptions, many LLMs resort to verbosity.

With regard to human descriptions, our work highlights *how* efficient humans are in reference games: Whenever possible, most dyads in our sample produce referring expressions with the minimum possible amount of content, describing only a single target cell. In addition, our results reveal pronounced positional biases in human descriptions, that are especially evident in conditions where speakers can choose relatively freely among multiple sets of target properties with similar discriminative value. Interestingly, our best-performing models reproduce these implicit biases to some extent, despite

being tested in a zero-shot setting on novel, highly abstract stimuli. These positional biases also offer a possible explanation for humans’ tendency to produce minimally specified descriptions: Participants may have preferred particular *search routes* over the visual items, concluding their search once a sufficiently discriminative target segment was found, akin to preference orders in Dale and Reiter (1995).

Future work could focus more strongly on the interplay between low-level visual biases and reference production, and on the dialogic aspects of reference generation. While we deliberately focus on single descriptions here, our data and methodology could also be used to study interactive phenomena, e.g., speaker–listener alignment and the cooperative formation of referential strategies across turns.

Limitations

Our study is subject to several limitations. First, we restricted our experiments to open-weight models hosted locally in order to ensure controllability and reproducibility. This decision imposed practical constraints on the maximum model size we could accommodate. Consequently we did not include larger, potentially more performant open-weight systems. We also did not evaluate commercial frontier models accessible only via APIs.

Second, due to time and financial constraints, our analysis relies on a limited sample drawn from the dataset of McDowell and Goodman (2019). A larger sample could plausibly yield additional insights beyond those reported here.

Third, our work is confined to a narrow domain, namely references to color grids, and we deliberately factor out dialogic phenomena that may be present in the underlying dataset. We plan to focus on such phenomena in future work. Furthermore, our analysis of the (human or LLM-generated) cell annotations focuses on the questions of *how many* and *which* target cells are selected in the descriptions. A natural extension of this would be to examine the combinatorial patterns among selected cells, which we also plan to cover in future projects.

Acknowledgments

We thank Vicky Rupprecht and Hannah Bultmann for their work on data annotation and evaluation.

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project no. 512393437, project B02.

Risks and Ethical Considerations

We do not believe that there are significant risks associated with this work, as it involves descriptions of abstract items that are unlikely to be perceived as harmful. We do note, however, that the use of LLMs entails non-trivial environmental costs, even when models are used solely for inference, as in this study, rather than fine-tuned or trained from scratch. No ethics review was required for this work. All human annotators and evaluators were informed that their data would be used for scientific purposes and publication. Our data contains no protected information and is fully anonymized.

References

- Manar Ali, Marika Sarzotti, Simeon Junker, Hendrik Buschmeier, and Sina Zarri  . 2025. [Towards neuro-symbolic approaches for referring expression generation](#). In *Proceedings of the 2025 CLASP Conference on Language Models and Representations*, pages 38–50, Gothenburg, Sweden. ACL.
- Manar Ali, Judith Sieker, Sina Zarri  , and Hendrik Buschmeier. 2026. [Reference games as a testbed for the alignment of model uncertainty and clarification requests](#). In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pages 990–998, San Diego, CA, USA. ACL.
- Raha Askari, Sina Zarri  ,   zge Alacam, and Judith Sieker. 2025. [Are BabyLMs deaf to Gricean maxims? A pragmatic evaluation of sample-efficient language models](#). In *Proceedings of the First BabyLM Workshop*, pages 52–65, Suzhou, China. ACL.
- Emily M. Bender and Nanna Inie. 2026. [We need to talk about how we talk about ’AI’](#). Tech Policy Press.
- Bastian Bunzeck, Sina Zarri  , Stefan Hartmann, and Elena Lieven. 2026. [No usage-based linguistics without language use](#). *Behavioral and Brain Sciences*, 49:e203.
- Paul-Christian B  rkner. 2017. [brms: An R package for Bayesian multilevel models using stan](#). *Journal of Statistical Software*, 80(1):1–28.
- Paul-Christian B  rkner and Matti Vuorre. 2019. [Ordinal regression models in psychology: A tutorial](#). *Advances in Methods and Practices in Psychological Science*, 2(1):77–101.
- Tyler A. Chang and Benjamin K. Bergen. 2024. [Language model behavior: A comprehensive survey](#). *Computational Linguistics*, 50(1):293–350.
- Daniel L. Chen, Martin Schonger, and Chris Wickens. 2016. [oTree – An open-source platform for laboratory, online, and field experiments](#). *Journal of Behavioral and Experimental Finance*, 9:88–97.
- Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. [Referring as a collaborative process](#). *Cognition*, 22(1):1–39.
- Robert Dale. 1992. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. MIT Press, Cambridge, MA.
- Robert Dale and Nicholas Haddock. 1991. [Content determination in the generation of referring expressions](#). *Computational Intelligence*, 7(4):252–265.
- Robert Dale and Ehud Reiter. 1995. [Computational interpretations of the gricean maxims in the generation of referring expressions](#). *Cognitive Science*, 19(2):233–263.
- Judith Degen, Robert D. Hawkins, Caroline Graf, Elisa Kreiss, and Noah D. Goodman. 2020. [When redundancy is useful: A Bayesian approach to “overinformative” referring expressions](#). *Psychological Review*, 127(4):591–621.
- Joseph L. Fleiss. 1971. [Measuring nominal scale agreement among many raters](#). *Psychological Bulletin*, 76(5):378–382.
- Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. [Pragmatics in language grounding: Phenomena, tasks, and modeling approaches](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640, Singapore. ACL.
- Simon Garrod and Anthony Anderson. 1987. [Saying what you mean in dialogue: A study in conceptual and semantic co-ordination](#). *Cognition*, 27:181–218.
- Gemma Team. 2026. [Gemma 4 technical report](#). Preprint, arXiv:2607.02770.
- Noah D. Goodman and Michael C. Frank. 2016. [Pragmatic language interpretation as probabilistic inference](#). *Trends in Cognitive Sciences*, 20:818–829.
- H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.
- Simeon Junker, Manar Ali, Larissa Koch, Sina Zarri  , 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.
- Kristina Kobrock, Xenia Ohmer, Elia Bruni, and Nicole Gotzner. 2026. [The role of pragmatic mechanisms in referential communication and categorization: An emergent communication model](#). *PLOS Computational Biology*, 22(5):e1014326.
- Emiel Krahmer and Kees van Deemter. 2019. [Computational generation of referring expressions: An updated survey](#). In Jeanette Gundel and Barbara Abbott, editors, *The Oxford Handbook of Reference*, pages 411–456. Oxford University Press, Oxford, UK.

- Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025a. [Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria. ACL.
- Ziqiao Ma, Jing Ding, Xuejun Zhang, Dezhi Luo, Ji-ah Ding, Sihan Xu, Yuchen Huang, Run Peng, and Joyce Chai. 2025b. [Vision-language models are not pragmatically competent in referring expression generation](#). In *Proceedings of the 2nd Conference on Language Modeling*, Montreal, Canada.
- Dominique Makowski, Mattan S. Ben-Shachar, and Daniel Lüdecke. 2019. [bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework](#). *Journal of Open Source Software*, 4(40):1541.
- 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.
- Bruno Nicenboim. 2026. [Bayesian ordinal models: A very short practical guide](#). Zenodo.
- OpenAI. 2024. [OpenAI o1 system card](#). *Preprint*, arXiv:2412.16720.
- Thomas Pechmann. 1989. [Incremental speech production and referential overspecification](#). *Linguistics*, 27(1):89–110.
- Qwen Team. 2026a. [Qwen3.5: Towards native multi-modal agents](#).
- Qwen Team. 2026b. [Qwen3.6-27B: Flagship-level coding in a 27B dense model](#).
- Qwen Team. 2026c. [Qwen3.6-35B-A3B: Agentic coding power, now open to all](#).
- Paula Rubio-Fernandez, Francis Mollica, and Julian Jara-Ettinger. 2021. [Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation](#). *Journal of Experimental Psychology: General*, 150(3):583–594.
- Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. [Verbosity bias in preference labeling by large language models](#). In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.
- Simeon Schüz, Albert Gatt, and Sina Zarriß. 2023. [Rethinking symbolic and visual context in referring expression generation](#). *Frontiers in Artificial Intelligence*, 6:1067125.
- Claude E. Shannon. 1948. [A mathematical theory of communication](#). *The Bell Systems Technical Journal*, 27:379–423.
- Judith Sieker and Sina Zarriß. 2026. [How hypocritical is your LLM judge? listener-speaker asymmetries in the pragmatic competence of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 7551–7565, San Diego, California, United States. ACL.
- Sara Spotorno and Benjamin W. Tatler. 2025. [What’s left of the leftward bias in scene viewing? Lateral asymmetries in information processing during early search guidance](#). *Cognition*, 254:106009.
- Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2025. [Label Studio: Data labeling software](#). Open source software available from <https://github.com/HumanSignal/label-studio>.
- Elli N. Tourtouri, Francesca Delogu, Les Sikos, and Matthew W. Crocker. 2019. [Rational overspecification in visually-situated comprehension and production](#). *Journal of Cultural Cognitive Science*, 3(2):175–202.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran.

A CRediT Contributor Statement

Simeon Junker: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing; **Manar Ali:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing; **Kristina Kobrock:** Formal analysis, Writing – original draft, Writing – review & editing; **Larissa Koch:** Data curation, Writing – original draft; **Marika Sarzotti:** Data curation, Writing – review & editing; **Hendrik Buschmeier:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing; **Sina Zarriß:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing;

B Data Availability Statement

Our work builds on the data from [McDowell and Goodman \(2019\)](#), which is available at <https://github.com/forkunited/ltprg>. We also use LLM weights published on Hugging Face, see the following section for details.

Our own raw and processed data (annotations, evaluations, generated descriptions) as well as

our code is available at: <https://github.com/clause-bielefeld/color-grid-efficiency>,
<https://doi.org/10.5281/zenodo.22306409>.

C Majority Voting for Human Cell Annotations

The referential spans included in our final cell annotation data are determined by majority voting (2 of 3 annotators); we discard the remaining spans. Importantly, most span annotations without majority agreement are either substrings of majority-agreed spans or vice versa: Only 9 spans without majority ratings do not overlap with any span included in the final data (2.1 % of discarded spans and 0.03 % of all annotated spans). As a consequence, our data lacks annotations for only 5 of the 1500 rounds across the 25 games in our sample. We similarly rely on majority voting for the cell labels associated with included spans, complemented by a consensus mechanism: If fewer than 2/3 of annotators agree on a given set of labels, we include their intersection, i.e., the labels all annotators attribute to the span. If this intersection is empty, we assign a fallback label denoting unclear reference.

D Models

We use weights from [Hugging Face](#) for the following models in our study:

- [google/gemma-4-E4B-it](#)
- [google/gemma-4-12B-it](#)
- [RedHatAI/gemma-4-26B-A4B-it-FP8-dynamic](#)
- [Qwen/Qwen3.5-4B](#)
- [Qwen/Qwen3.5-9B](#)
- [Qwen/Qwen3.5-27B-FP8](#)
- [Qwen/Qwen3.6-27B-FP8](#)
- [Qwen/Qwen3.6-35B-A3B-FP8](#)

We run Qwen3.5-27B, Qwen3.6-27B and gemma-4-26B-A4B-it with FP8 quantization.

All models are served locally on a workstation with an NVIDIA RTX A6000 GPU. We serve the models with [vLLM](#). For inference, we use the OpenAI Python API with continuous batching and a maximum of 40 concurrent requests.

E Model Prompts

We use the following prompt for our description LLMs:

You are a visual referring-expression assistant.

You will be shown three 3x3 color grids in random order. Exactly one grid is the target, marked by a green box.

Your task is to produce a short description of the target grid that allows a conversational partner to uniquely identify it, even if the same three grids are presented in a different order.

Instructions:

- Describe only the target grid.
- Make the description specific enough to distinguish the target from the other two grids.
- Keep the description concise and as short as possible while remaining unambiguous.
- Describe the target in terms of the colors it is composed of. Do not mention any markers, such as the green box or borders around the items.
- Do not mention the ordering of the grids or the position of the target (left, middle or right). Assume that your partner sees the grids in a different order.
- Do not mention that the target is a 3x3 color grid.

Output format:

- Return exactly one plain-text string.
- Output only the description and nothing else.

We use the following prompt for our annotation LLM:

You are an expert annotator for vision-language data.

Your task is to identify all referring expressions in a dialogue that refer to any part of a visual stimulus consisting of three 3x3 color grids, and map each referring expression to the relevant grid cell or cells.

You will receive:

- a triplet of 3x3 color grids
- a dialogue about these grids

Grid roles:

- the left grid is always the target grid
- the middle grid is the first distractor
- the right grid is the second distractor

Cell labels:

- target grid: t1, t2, t3, t4, t5, t6, t7, t8, t9
- first distractor: d1_1, d1_2, d1_3, d1_4, d1_5, d1_6, d1_7, d1_8, d1_9
- second distractor: d2_1, d2_2, d2_3, d2_4, d2_5, d2_6, d2_7, d2_8, d2_9

Cell indexing is row-wise from top-left to bottom-right:

- top row: 1, 2, 3

- middle row: 4, 5, 6
- bottom row: 7, 8, 9

Character offset convention:

- use zero-based character indexing
- `start` is inclusive
- `end` is exclusive
- `span` must exactly match `dialogue[start:end]`

Your output must be valid JSON only.

Do not output any prose, explanation, markdown, or extra text.

Do not wrap the JSON in code fences.

Do not include keys other than those defined below.

Output format:

```
{
  "annotations": [
    {
      "span": "exact substring from the dialogue",
      "start": <int>,
      "end": <int>,
      "cells": ["<cell_label>", "..."],
      "comment": "optional comment only when needed"
    },
    {
      "span": "exact substring from the dialogue",
      "start": <int>,
      "end": <int>,
      "cells": ["<cell_label>", "..."],
      "comment": "optional comment only when needed"
    }
  ],
  "no_reference": false
}
```

"annotations" is a list, include one object per relevant span found in the dialogue. There is no fixed number; include as many or as few as apply.

"cells" is a list, include one or more cell labels from the taxonomy that apply to this span; use multiple labels if more than one cell is relevant, and use exactly one if only one applies.

If and only if the dialogue contains no descriptions that can be mapped to grid cells, output:

```
{
  "annotations": [],
  "no_reference": true
}
```

Annotation rules:

1. Annotate all referring expressions in both speaker and listener utterances.
 - This includes questions such as "bright green?"
2. Annotate intended references even if the text is ambiguous, vague, or partially incorrect.

- Choose the most plausible intended cell or cells.
- If the utterance is erroneous (for example, it names the wrong location or color), annotate the most likely intended cell or cells and include a brief `comment`.

3. Keep textual units together whenever possible.

- Prefer maximal coherent referring expressions over fragmented ones.
 - Example: annotate "bottom left is hot pink" as one span, not separate spans for "bottom left" and "hot pink".
 - Example: annotate "the left middle square is the brightest pink" as one span.
- Annotate row or column descriptions as single spans.
 - Example: "Top row is green grey turquoise" should be one span mapped to three cells.
 - This also applies to less explicit formulations such as "top, dark blue, aqua, purple".

4. If only colors or properties are mentioned without explicit locations, annotate the most likely cell or cells referred to.

- Example: "dark purple", "one yellow", "the grid has a lot of blue", "the one with the red square"
- Do not annotate an entire grid unless the expression truly refers to all cells in that grid.

5. For exclusions or negative properties that distinguish the target from distractors, annotate the distractor cells implicated by the description.

- Example: "no bright green" should be mapped to the distractor cell or cells that are bright green.

6. For comparative descriptions, annotate both:

- the target cell or cells being described
- the distractor cell or cells used as the comparison set

- Example: "top right is a little brighter than the other two" should include the target top-right cell and the top-right cells of both distractors.

7. Repeated mentions should be annotated separately if they occur as separate spans in the dialogue.

8. Use `comment` only when necessary.

- Valid reasons include: ambiguity, error in the utterance, inferred referent not literally stated, or unusual interpretation.

9. The `cells` array must:

- contain one or more valid cell labels
- contain no duplicates
- be ordered consistently by grid and then by cell index
- use only the labels defined above

10. The `annotations` array must be sorted by:

- ascending `start`

- then ascending `end`

General principle:

- Annotate all and only text spans that refer to visual content in the grids.
- Prefer coherent, maximal referring expressions.
- Resolve ambiguity by selecting the most plausible referent(s) from the visual context and dialogue context.

Before returning your answer, ensure:

- the output is valid JSON
- every `span` exactly matches the dialogue substring at the given offsets
- every annotation has `span`, `start`, `end`, and `cells`
- no extra text appears outside the JSON object

F Examples for In-Context Learning

In addition to the prompted instructions, we supply our annotation LLM with five examples from our human annotation data. For each item, we sample target and distractor triplets, human descriptions and annotations for five distinct items, ensuring sufficient diversity with the following criteria:

- one item contains descriptions of at least one cell of any distractor
- one item includes at least two dialogue turns
- one item includes only a single turn, a single span, and no references to distractors
- one item includes an *invalid description* and a corresponding annotation, signaling *no reference*
- one additional item is randomly sampled from all remaining items in the data

For the invalid item, we manually replace the original description with the following text: “speaker: A 3x3 grid of colored squares with a green border surrounding it”.

G Human Annotators and Evaluators

This study involved five annotators who worked as student assistants or interns or are co-authors of this paper. The student assistants were paid according to the local pay scale. All annotators were informed of the purpose of the annotations and consented to their use for scientific purposes and publications.

H Annotation and Evaluation Interfaces

We collected our human cell annotations with Label Studio (Tkachenko et al., 2025). The annotation interface is shown in Figure 5.

We implemented our human evaluation with oTree (Chen et al., 2016). The evaluation interface is shown in Figure 6.

The guidelines and configuration files for the human annotation and evaluation are available at: <https://github.com/claude-bielefeld/color-grid-efficiency>, <https://doi.org/10.5281/zenodo.22306409>

I Informativeness and Cell Count Distributions

Figure 7 illustrates the distributions of informativeness scores (human evaluation) and cell counts (model-based cell annotation) for human descriptions and expressions produced by Qwen3.5-4B, Qwen3.5-9B and Qwen3.5-27B (all without thinking mode).

Descriptions produced by humans and Qwen3.5-27B receive optimal informativeness scores in most cases. Qwen3.5-4B is evaluated to overspecify in the majority of cases, Qwen3.5-9B receives more mixed ratings. Similarly, humans and Qwen3.5-27B rely on descriptions of single cells in most cases, while Qwen3.5-4B resorts to exhaustive descriptions for most items. Note, however, that this system also produces a high rate of descriptions that cannot be mapped to target cells, see Section 6.

J Target Cell Heatmaps

In Figure 2 we illustrate location biases in human and LLM-generated descriptions as heatmaps of the distributions of selected target grid cells.

Figure 8a illustrates the frequency of mention of target cells in human descriptions, with data from our LLM-based annotation. Importantly, the heatmaps show near-identical patterns as our human annotation (Figure 2a), further validating the reliability of our LLM-based approach.

In Figure 2b, we showed the frequency of mention of target cells in descriptions produced by Qwen3.5-27b with thinking activated. Figure 8b illustrates the content selection patterns of the same model, but without thinking mode. The model shows the same top-left biases as with thinking mode, but with less accentuated differences between the cell values, indicating that explicit reasoning exacerbates the visual biases to some extent.



Figure 5: Annotation Interface (Label Studio).

Click on the grid that is described by the utterance below

No Object Fits

Multiple Objects Fit

Top row: grey, yellow, teal. Middle row: cyan, maroon, blue. Bottom row: teal, grey, orange.

Rate the amount of information provided in the description

Choose from the following:

To identify the target grid, the description provides:

- (1) Less information than I needed
- (2) Somewhat less information than I needed
- (3) The right amount of information
- (4) Somewhat more information than I needed
- (5) More information than I needed

1 2 3 4 5

Click to select a value

Confirm to proceed to the next item

Confirm Selection

Figure 6: Evaluation Interface (oTree).

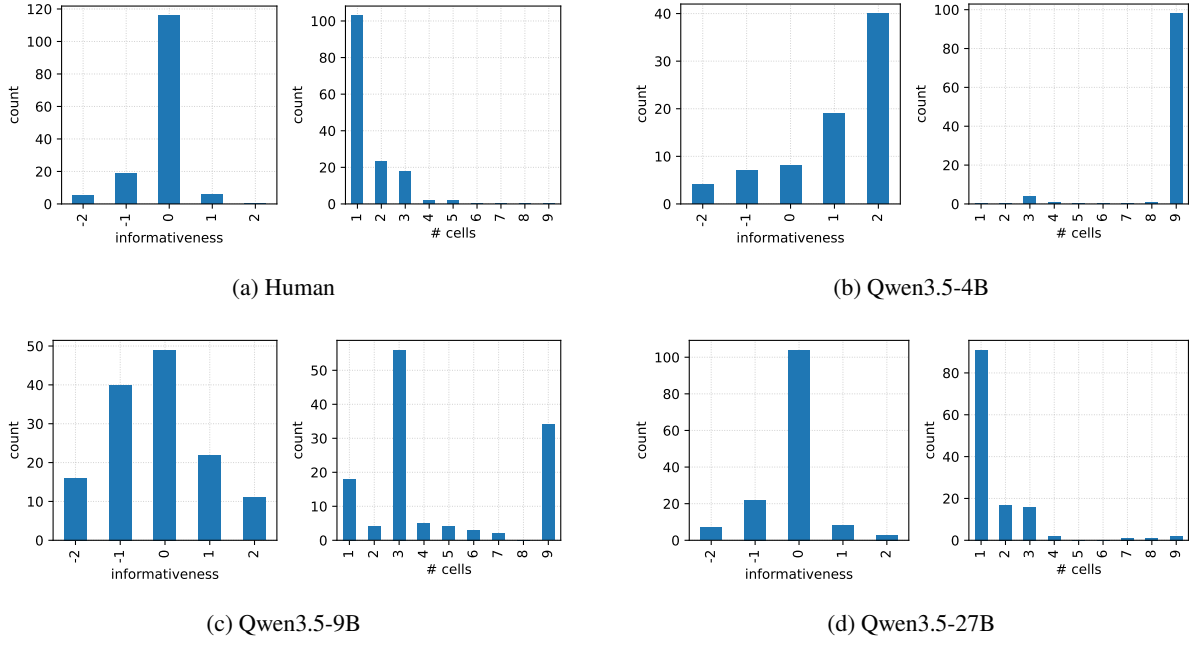


Figure 7: Distributions of informativeness scores and cell counts.

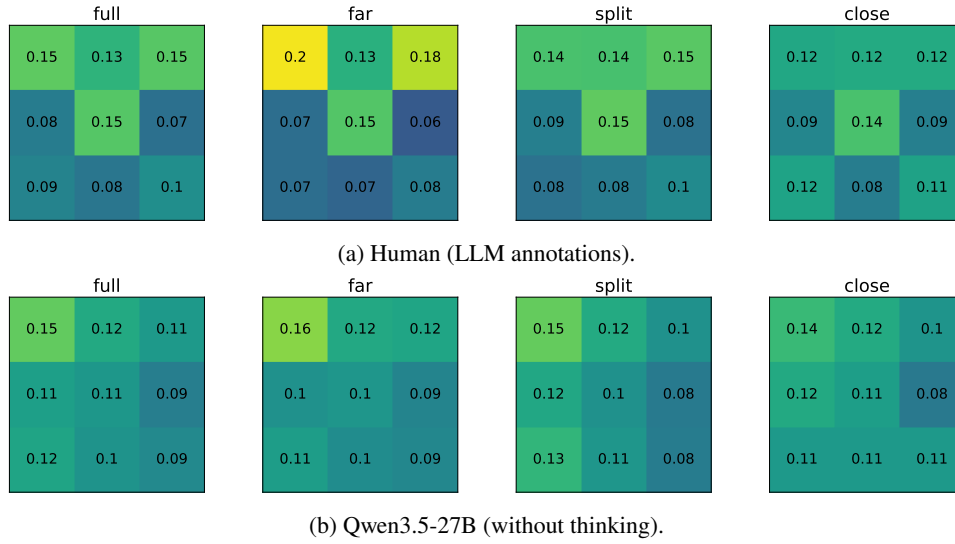


Figure 8: Frequency of mention for individual target grid segments in human and LLM descriptions.