

Calibrated Ambiguity in Multimodal Language Models: Humans reach for cultural references, while models describe the picture

Cody Kommers*
ckommers@turing.ac.uk
The Alan Turing Institute
London, United Kingdom

Mingrui Ye*
mingrui.ye@kcl.ac.uk
King's College London
London, United Kingdom

Evelyn Gius
Technical University of Darmstadt
Darmstadt, Germany

Daniela Mihai
University of Southampton
Southampton, United Kingdom

Hoyt Long
University of Chicago
Chicago, United States of America

Zheng Yuan
University of Sheffield
Sheffield, United Kingdom

Drew Hemment
The Alan Turing Institute
London, United Kingdom
University of Edinburgh
Edinburgh, United Kingdom

Abstract

Ambiguity is often treated as a bug for AI systems to resolve—but in human communication and culture, ambiguity can also be a generative resource. From humour to politics to art, people express themselves in words and images that are open enough to invite different interpretations, yet constrained enough to be interpretable. We operationalise this notion of *calibrated ambiguity* with a task drawn from the parlour game *Dixit*. We compare differences in clues generated by human vs multimodal language models, based on a novel coding rubric for calibrated ambiguity, and find that models consistently exhibit ambiguity collapse (i.e., their outputs are over-specified, leaving no room for multiple legitimate interpretations). Unlike human clues, AI-generated clues also exhibit cultural flattening: they almost never make reference to culturally-situated knowledge, even when prompted to use allusion and figurative language.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**.

Keywords

ambiguity, culture, calibration, homogenization, collapse, cultural flattening, multimodal language models, interpretive technologies

1 Introduction

In AI research, ambiguity is often treated as a quantity to be minimised. A long tradition of work on word sense disambiguation [52, 80] and, more recently, benchmarks that test whether language models can resolve ambiguous questions and sentences [43, 49] share the premise that a system succeeds when ambiguity is removed. However, ambiguity is not just defined by semantic vagueness; it can be a vital aspect of human communication and culture [14, 35, 55, 59].

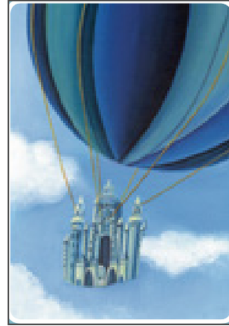
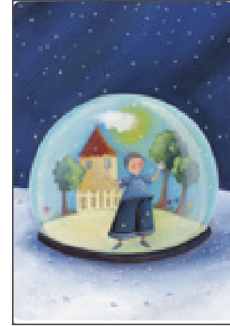
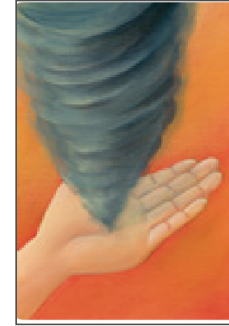
For example, ambiguity plays a role in narrative, where explicit exposition of themes or insights tends to trivialise a story (e.g., “telling” rather than “showing” [4]), and where the gaps a text leaves open are precisely what a reader must fill in to make it meaningful [29]. It plays a role in interpersonal conversation and pragmatic implicature, in which people gesture towards meaning rather than state it outright to avoid causing undue offence or tension [5, 22, 56]. Careful management of ambiguity can play a role in many deliberative processes: groups frequently reach agreement on what to do while leaving the underlying principles unresolved [63]. When juries are faced with incomplete information, their collective job is to reach a conclusion even when ambiguity remains [26].

Calibrating ambiguity is crucial to designing and evaluating generative AI as a cultural technology [16, 35], as well as an important consideration for Interpretive Technologies: AI systems designed to navigate cultural complexity [27]. AI systems are increasingly deployed for tasks that do not admit of straightforward ground-truth answers [41]. Evaluating the outputs of these systems requires not just binary judgments of correct and incorrect—but interpretive acts of the kind traditionally performed by humanities scholars [2, 27, 34, 68]. These interpretations account for ambiguity directly, specifically by accounting for a multiplicity of conflicting but legitimately held perspectives, rather than prioritising a canonical ground-truth answer [35]. Accordingly, a significant risk of generative AI systems is ambiguity collapse [24]: whereas most real-world situations or outcomes can be described by multiple legitimate interpretations, AI interfaces often offer a single, authoritatively-presented judgment [48].

The challenge is that, by definition, something that is ambiguous resists a unitary ground-truth evaluation [59]. In its most general form, ambiguity simply means that a word, statement, image, or other cultural artefact can lend itself to multiple interpretations or meanings [24]. But there can be many shades of what this looks like. For example, Clint Eastwood’s 1992 film *Unforgiven* is characterised

*Both authors contributed equally to this research.

A Dixit storyteller scores only if some, but not all, players find the card.

	A	B	C	D	E
					
HUMAN	<i>“Let it go”</i> a song from Frozen; found by 3 of 4 players				
GEMMA 3	<i>“Lost in wintry contemplation”</i>				
GLM 4.6V	<i>“A woman in a long blue cloak”</i> fits card B alone				

CALIBRATED

CALIBRATED

OVERSPECIFIED

CULTURAL REFERENCE

NO CULTURAL REFERENCE

NO CULTURAL REFERENCE

Card art: Dixit © Libellud / Marie Cardouat, reproduced for illustrative research purposes.

Figure 1: We use the parlour game Dixit as a paradigm for measuring calibrated ambiguity. The task requires a “storyteller” to generate a “clue” that ambiguously picks out a target card from a set of distractor—as judged by being correctly selected by some but not all of the other players. In this example, a human player constructs their clue around an allusion to a song from the movie *Frozen*—a widely-known cultural reference. Our results show how model-generated clues differ systematically from those given by humans. In this example, Gemma 3 offers a clue that is calibrated for ambiguity, but describes the image without reference to culturally-situated knowledge; while GLM-4.6V simply provides an unambiguously literal clue.

by moral ambiguity: each of the characters is trying to do the right thing, but ends up transgressing nonetheless. Who is the archetypal “good guy” the audience is supposed to root for? By contrast, Roy Lichtenstein’s use of styles from graphic novels is also ambiguous. Are these works celebrating comic book art, or mocking it? There is no unified way to operationalise ambiguity comprehensively across even just these two cases—except to say that the artefacts in question admit of multiple interpretations.

The calibration of ambiguity in AI systems is therefore fundamentally a design problem [11, 19]. We want AI systems that do more than just minimise ambiguity under any circumstances—but when, where, and how can ambiguity be constructively or usefully employed? The answer is context dependent: it depends on the people involved, their background, the situation they are in, and what they are trying to do [35]. While there is no one-size-fits-all level of ambiguity that works in all cases for all people, appropriate management of ambiguity is especially important in settings where tasks are culturally-situated [3, 10, 20, 28, 35, 44, 50, 57, 60, 71, 79]. For example, AI systems are increasingly used to shape or produce stories and other cultural artefacts [23]. Such narratives

provide a crucial structure for how people make sense of themselves and the world around them [6, 36]. At scale, how AI shapes narratives and cultural artefacts can therefore have a significant impact on identity and meaning-making [37, 39, 81]. But what it looks like for a cultural artefact (such as a story) to be appropriate within a given cultural frame depends not just on checking specific demographic boxes, but navigating and managing tradeoffs among intrinsically ambiguous concepts [82]. Likewise, determining what kinds of stories count as meaningful to a given person or community depends not just on raw engagement metrics, but by accounting for the ambiguities inherent in contextually-situated perspectives [38, 46]. These considerations are notoriously difficult to quantify [41].

In this paper, we present a paradigm for measuring the calibration of ambiguity in multimodal AI systems. Whereas “disambiguation” has a clear operational definition [43, 52, 66, 80], an evaluation framework for calibrated ambiguity is still needed [24]. We develop a task aimed at codifying the calibrated ambiguity and provide initial answers to the following questions: How well can generative AI systems calibrate ambiguity? What strategies do models rely on

when managing ambiguity? How do these strategies differ from those of humans? And how might we calibrate ambiguity for a particular context or use-case?

1.1 Summary of Contributions

We present a novel methodology for measuring calibrated ambiguity in multimodal language models, based on the popular parlour game *Dixit* (Section 3). We develop a novel annotation rubric to distinguish between different dimensions of ambiguity relevant to the task (Section 4). We find that humans and models tend to calibrate ambiguity in fundamentally different ways—with humans more likely to make use of culturally-situated knowledge (Section 5). In a model comparison, we show that there are significant differences between models in their default behaviour of how ambiguity is managed and show that prompting can alleviate miscalibration to a degree (specifically by emphasising a “social cognition” framing of the task), but cannot fully prevent ambiguity collapse (Section 6).

2 Related Work

Previous work by Gur-Arieh, Wang, and Fazelpour [24] establishes a taxonomy of epistemic risks posed by ambiguity collapse in large language models (LLMs)¹; we explicitly aim to build on their work. Gur-Arieh et al. [24] distinguish between three categories of risk associated with ambiguity collapse—process, output, and ecosystem. In this paper, we focus specifically on output collapse: the degree to which systems produce artefacts that support a plurality of legitimate interpretations. The concern about ambiguity collapse articulated by Gur-Arieh et al. [24] reflects a larger concern about the homogenising force exerted by LLMs in collapsing or flattening culturally-situated outputs [28, 31, 75, 76, 78].

2.1 Measuring ambiguity

While there is no established general measure for evaluating a computational system’s capacity to calibrate ambiguity, previous work attempts to evaluate whether systems recognise or resolve multiple (reasonable) interpretations of an input [43, 43, 49, 51]. For example AmbigQA, for instance, evaluates whether question-answering systems recover answers corresponding to different interpretations of ambiguous questions [49], while the benchmark AMBIGENT tests whether language models can represent alternative readings of ambiguous linguistic inputs [43]. Wildenburg et al. [74] similarly provide a dataset for testing whether language models recognise semantic underspecification; Nam et al. [51] extend ambiguity evaluation to multimodal models by examining whether visual context enables disambiguation; and Karim et al. [32] show how models can be calibrated for uncertainty. Previous work has also investigated how artists have utilised models as sources of ambiguity in their works [66].

However, few approaches position ambiguity as an explicitly desirable element of model output. The closest line of research is perhaps “perspectivist” annotation, premised on the observation that in many datasets disagreement is desirable or expected rather

than something to be eliminated [1]. For example, recent work has taken this perspectivist approach to label disagreement in irony [18] or offensive language [33]. However, ambiguity is only one potential cause of disagreement [58]; others include bias and noise [67], or misinterpretation and deficient definitions [21]. Other annotation work use variability as a proxy for ambiguity; for example, by using human annotation patterns to capture the appropriate amount of disagreement across interpretations [12, 47, 53]. In this vein, Pavlick and Kwiatkowski [54] examine the shape of human judgments by testing for multiple modes within a distribution. Other work uses graded human ratings of ambiguity or indirectness to assess models’ ability to integrate visual context for intent disambiguation [51] or uses semantic entropy to quantify uncertainty across semantically distinct model outputs [15]. Together, this work gestures towards the way desirable variability, disagreement, or multiplicity has been evaluated in generative AI systems—specifically by using a human sample for reference in defining what form that variability ought to take.

2.2 Dixit-based tasks

In an effort to assess model performance on tasks requiring complex reasoning or which do not have a single right answer, researchers have increasingly turned to strategic games as a useful testbed [42]. One game that has attracted significant attention in this respect is the card game *Dixit*. Kunda and Rabkina [40] identified this task of “creative captioning” as a novel challenge for AI requiring the integration of visual perception, natural language understanding, and social reasoning.

Existing computational work on *Dixit* has primarily treated the game as a paradigm for vision–language association. Iwata et al. [30] developed an AI player to model the image–word associations involved in *Dixit* and found that the AI performed equally or worse than human players in clue generation, and worse in card selection and voting. Vatsakis et al. [70] focused on the voting task, using human data, natural language processing methods, and Internet search to predict the storyteller’s card from a deliberately vague hint. Wei [73] subsequently approached card–hint matching with a neural network designed to learn visual concepts from natural language, while Chang [8] used the same neural network for hint selection. Finally, Tanzawa et al. [65] operationalise and quantitatively measure card–hint relevance, which they explicitly connect to the problem of producing appropriately ambiguous hints. They evaluate the relevance of hints for individual cards, comparing three human hint creators with one AI system, and find the lowest median relevance score and the greatest variability for the latter.

These studies of *Dixit* have ranged widely in their scale and size. On one end, the controlled human evaluation studies are very small-scale—one 15-turn game [30] and the analysis of only 20 cards [65]. At the other extreme, evaluations in Vatsakis et al. [70], Wei [73], and Chang [8] rely on an archival dataset of approximately 116,000 human-played rounds of the game rather than newly collected experimental judgments.

¹We use ‘LLM’ rather than ‘MLM’ or ‘VLM’ in the text to reflect the common usage in this discussion, contextualising ambiguity collapse as a risk of generative AI systems broadly—though it should be remembered that the models we use are not only large, but multimodal.

3 Method

3.1 Dixit Paradigm

Dixit is a parlour game designed around calibrated ambiguity. It is a multiplayer game featuring a deck of Dixit-specific cards, each with a surreal image of an object, person, or situation that could admit of a range of interpretations. Each round, one player acts as the “storyteller”: they pick a card from their hand and offer a “clue” related to the image. Based on the clue, the other players pick an image from their own hands which they think could be described by the clue; all cards are then shuffled together, and the non-storyteller players vote for the card they believe was the storyteller’s. Crucially, the storyteller wins the round only if *some but not all* of the other players choose the image they put into the middle. The optimal clue is therefore one based on calibrated ambiguity: not so vague as to be uninterpretable, not so specific as to obviously map to the target image.

3.2 Existing dataset of human-generated clues

The dataset of human responses used in this study is based on the Dixit dataset introduced by Vatsakis et al. [70] and made publicly available by the authors.² The dataset was constructed from games played on the online platform *boiteajeux.net* between July 2012 and September 2021 and is restricted to games using the 84 cards from the Dixit base game. As described by Vatsakis et al. [70], the original data were manually filtered to improve the quality and consistency of the textual information. Rounds in which the hint consisted solely of a web address were removed, as were rounds containing non-English hints. Exceptions were retained for commonly recognized foreign-language expressions (e.g., *carpe diem*), titles of well-known works (e.g., *Le Petit Prince*), and names referring to franchises, people, or landmarks. Web links appearing in storyteller explanations were also removed. Following this preprocessing, the resulting dataset contained 116,226 rounds with the storyteller’s hint, the cards submitted during that round, the identity of the storyteller’s card, and the votes received by each card. We use the dataset as released and did not alter its content; clue texts are reproduced verbatim, up to whitespace normalisation.

Integrity checks and study sample. Before adopting the dataset we verified that the release matches its published description: it contains 116,226 rounds, and a storyteller’s post-round explanation is present for 26% of them. Because our boards require the five-player format (one target, four decoys, four guessers), we restricted the data to the 19,407 five-player rounds and checked the structural consistency of each: a round was retained only if its board comprised exactly five distinct cards from the base-game deck, exactly one card was marked as the storyteller’s and agreed with the dataset’s target field, and the vote record was complete, with votes summing to the four guessers. Only six rounds failed these checks, and every retained clue is English-language text, consistent with the authors’ preprocessing. We selected a subset of 350 rounds sampled with fixed random seeds, excluding duplicate clue texts; this was used in as the set of boards (Section 3.3) for use in the various analyses detail in this paper.

²https://www.spronck.net/datasets/Dixit_AI_data.zip

3.3 LLM-generated Dixit Clues and Judgments

We sought to compare human-generated clues with those generated by models. As storytellers, vision–language models produced clues for the same cards that human storytellers had played, so that human- and model-generated clues can be compared on identical boards (Table 1). As judges, LLMs applied the ambiguity coding rubric of Section 4 to clues, following the protocol applied by human coders.

Boards. All clues were generated for the same set of 350 five-player rounds drawn from the human dataset (see Section 3.2 for selection criteria and integrity checks). Each round supplies the human storyteller’s target card, the four decoy cards that the other players actually submitted, the human storyteller’s own clue, and the votes it received. The set comprises 200 rounds sampled at random with a fixed seed (the *baseline* boards) and the 150 rounds whose human clues were hand-coded in Section 4, so that every coded human clue has an LLM counterpart on the same board.

Model selection. We used six open-weight vision–language models as storytellers (Table 1). Three considerations guided the selection. First, every model is open-weight and was served locally in bfloat16 on a single 80 GB GPU, so that generation ran under our own system prompt and sampling settings, with no API layer in between. Most models were roughly 30B total parameters, which is approximately the largest tier that fits one accelerator without quantization. Smaller models were not of interest (with the exception of GLM 4.6V Flash [9B], to have one example of a smaller model) because we wanted to focus systems that performed the task well, and larger ones could not be run under the same conditions. Second, the six models come from five developers and span both dense and mixture-of-experts (MoE) architectures. An MoE model routes each token through a few of many expert sub-networks [17], and analyses of open MoE models find their experts more specialized and less polysemantic than dense feed-forward layers [45]. Whether such internal differences surface in generation behaviour, for instance in how specific or how varied a clue is, has not been studied; the contrast lets us ask whether architecture predicts calibration (Section 6.2.1). Third, within these constraints the lineup is not size-matched: total parameters range from 9B to 35B and active parameters from 3B to 27B. GLM-4.6V-Flash (9B) is the only member of its family small enough to serve locally (the full GLM-4.6V has 106B parameters), and Kimi-VL is released at 16B only. We kept both because family diversity mattered more to us than exact size matching, and because size can then be examined across the lineup rather than held constant (Section 6.2.1 finds that neither total nor active parameter count orders the models).

Prompts. Each model was run under two prompts. The *minimal* prompt stated only the rules of the game and its scoring goal—that some but not all of the other players should identify the target—together with a length bound of three to six words and the instruction to respond with the clue alone. The *social cognition* prompt added additional emphasis: it asked the model to consider how the other players will interpret the clue, to prefer figurative or abstract language that could apply to several cards, and to aim for a clue that exactly two of the four other players would resolve.

Table 1: Storyteller models (“active”: parameters used per token in MoE models).

Model	Developer	Architecture	Parameters
Gemma-3-27B-it	Google	dense	27B
Qwen3.8-27B	Alibaba	dense	27B
GLM-4.6V-Flash	Zhipu AI	dense	9B
Qwen3.5-35B-A3B	Alibaba	MoE	35B (3B active)
Kimi-VL-A3B-Instruct	Moonshot	MoE	16B (3B active)
ERNIE-4.5-VL-28B-A3B	Baidu	MoE	28B (3B active)

The contrast allowed us to ask whether explicitly prompting a model to reason about other minds changes the kind of ambiguity it produces [62]. Because the social cognition prompt bundled several instructions—audience modelling, a preference for figurative or abstract language, and a numerical target—the comparison tested the prompt variations as a whole rather than any one component. Some of the clues coded by the human panel (Section 5.1) were generated under an earlier wording of the social cognition prompt that differed only in its closing instruction; the panel rated the two wordings indistinguishably, and they are not distinguished below. Additional details of prompting protocols are available in Appendix A.

Clue generation. Clues were sampled at temperature 0.9 with a fixed random seed, with extended “thinking” disabled for models that support it, one clue per round for each model–prompt combination. The clue was extracted from the response by stripping markup and surrounding quotation marks. Clues were never truncated: if a generation exceeded the length bound, the model was resampled up to five times and the shortest complete attempt was kept, so that clue length is a property of the model rather than of post-processing. (Human clues were likewise unbounded in length; GLM-4.6V-Flash exceeds seven words in 17% of rounds, against 10% of human clues.) This yields 350×2 prompts = 700 clues per model and 4,200 LLM clues in all.

To match the informational position of a human storyteller at the moment of clue-giving, a model saw *only* its own target card—never the other players’ cards—and was told that four other players would each contribute a decoy before all five cards were shuffled on the table. (Note: human storytellers chose which card to play, often with a clue already in mind, whereas the models produced a clue for a card it did not choose)

Judgments: LLMs as rubric coders. To compare producers at this scale we used LLMs to employ the coding rubric of Section 4, applied by LLM coders under the same protocol as the human annotators [13, 77]. Each item was presented exactly as a human coder saw it: the clue and the five candidate card images (labelled A–E, in the order the human coders saw them), with the rubric as the system prompt. The coder first selected the image it believed to be the target and then rated the clue on all four dimensions. Each item was a single independent API call at temperature 0, and the coder never saw the answer key, whether the clue was human- or LLM-generated, or the study hypotheses. We used two coders from different model generations, Gemini 3.1 Flash-Lite and Gemini 3.5 Flash-Lite. Every clue—the 4,200 LLM clues and the 350 human clues on the same

boards—was rated independently by both coders, giving 9,100 ratings in total; unless stated otherwise, the analyses below use the mean of the two coders per clue.

Before scaling up, both coders were checked against the 150 clues that the five-annotator human panel coded (78 human and 72 LLM clues, Section 4). Per-clue coder ratings correlate with the human panel mean at $r = 0.60$ – 0.79 across the four dimensions (Calibration 0.62–0.68, Situatedness 0.77–0.79, Literalness 0.69–0.70, Figurativeness 0.60–0.63), which matches the agreement of an individual human annotator with the rest of the panel (leave-one-out $r = 0.58$ – 0.80 on the same dimensions). Exact agreement with the rounded panel mean ranges from 0.55 (Figurativeness) to 0.80 (Situatedness). LLM coders thus approximate the human panel about as well as individual human annotators do, on the subset the panel coded; on that basis we use their ratings for the model comparison of Section 6.2 (full comparison in Section 6.1).

An earlier version of the pipeline scored clues behaviourally, by having a panel of small vision–language models vote for the target card as a Dixit audience would. We retain it as a robustness check (Appendix B); it reproduced the population-level human outcome but tracked individual clues only weakly; accordingly, we used the coding rubric as our primary instrument of assessment.

4 Ambiguity Coding Rubric

We developed a novel coding rubric to account for the different ways of generating and managing ambiguity employed by humans and models when producing Dixit clues [9]. The coding scheme features four dimensions, each scored on a 3-point ordinal scale, which are summarised in Table 2. *Calibration* asks whether the clue is overly specific, too vague, or just right. *Situatedness* asks what degree of cultural knowledge is needed to interpret the clue. *Literalness* and *Figurativeness* ask, respectively, how literally and how figuratively the clue interprets the image; they are coded independently, since a single clue can be high on both (see Section 4.2).

These dimensions were chosen to correspond directly to our hypotheses or considerations of interests, rather than to provide an exhaustive taxonomy of the different possible textures of ambiguity. Specifically, we wanted to investigate the following considerations:

- (1) *A direct test of ambiguity collapse.* This was imperfectly captured by raw accuracy scores (i.e., do some but not all judges guess the correct card?). In pilot studies, we assessed “collapse” as it is done in the original game. The problem is that a lot of false positives occurred when the clue was completely ambiguous: guessing results in at least one judge getting the correct answer by chance. This is probably not a major issue when humans play the game in a social context. But it was insufficiently sensitive to detect when models simply failed to do the task. We therefore defined the calibration dimension as a more targeted instrument for assessing collapse.
- (2) *A direct test of cultural flattening.* In our informal initial surveying of human versus model clues, it was apparent that humans were making use of a cultural references in a way models were not. This is consistent with accounts of cultural collapse and homogenisation of model outputs [28, 31]. We developed the situatedness dimension, using

Table 2: The ambiguity coding rubric. Each clue is rated on four dimensions, each on a 3-point ordinal scale.

Dimension	1	2	3
Calibration <i>Is the clue overly specific, too vague, or just right?</i>	Underspecified — no confident inference to a single image is possible	Calibrated — a plausible target is identifiable, but genuine doubt/alternative interpretations remain	Overspecified — the clue deterministically picks out one image
Situatedness <i>What degree of cultural knowledge is needed to interpret the clue?</i>	Universal — interpretable via general English competence alone	Mainstream — requires broadly-shared Anglophone/cross-cultural knowledge (e.g., major religious stories, globally-known figures)	Niche — requires specific subcultural or specialised knowledge (e.g., a line from a specific film)
Literalness <i>How literally does the clue interpret the image?</i>	Low — no literal description; entirely figurative/abstract	Mid/Mixed — some literal description, or unclear/debatable	High — primarily literal description of depicted objects/features
Figurativeness <i>How figuratively does the clue interpret the image?</i>	Low — no figurative language; entirely literal	Mid/Mixed — some figurative language, or unclear/debatable	High — primarily a reference to figurative language, metaphor, allusion, etc.

this term from the humanities [25, 35] to capture the degree to which specialised cultural knowledge was needed to interpret the clue.

- (3) *The degree to which a clue was constructed by reference to concrete visual features in the picture—or using more abstract, figurative devices.* Informally, it seemed that model-generated clues were more likely to pick out concrete visual features from an image. We explored several iterations on how best to capture this consideration (see Section 4.2).

4.1 Procedure

Five of the authors of this paper acted as expert annotators. Before conducting the full set of annotations, a session was held with a majority of the annotators to review specific examples within the data set. The first 15 clues were discussed in depth, as well as several instructive examples that could be considered edge cases. The session was recorded and shared with annotators who were unable to attend.

Annotators viewed each clue in the context of a full 5-image set (target + four distractors). They were blind to the answer key (i.e., which was the correct target image) as well as to whether the clue was human- or model-generated. Annotators first selected which

image they believed to be the target, then rated each clue on all four dimensions.

A fifth rating was included, offering notes on any externally verifiable referents for each clue. For example, the clue “Colors of the wind. . .” might be interpreted as being about literal, feature-based ambiguity—unless the interpreter knows that this is a reference to a song from the movie *Pocahontas*. In this case, the clue does not refer to wind or colour, but to the depiction of a woman who (allegedly) resembles this film character. A single rater used a combination of Gemini and Google search to verify whether each clue had an external referent. Some were ambiguous, but many made clear reference to some specific cultural artefact. We decided it was better to try to control for this knowledge by noting these references, rather than to leave it up to the prior cultural knowledge of the raters.

4.2 Rubric iterations

We went through several iterations of the rubric [9]. In an initial version, we featured a category called “mechanism,” with four nominal labels distinguishing between whether the clue’s ambiguity was primarily based in a feature, a concept, figurative language, or an indexical reference. However, many clues had multiple aspects of these. In a subsequent version, we developed one dimension meant to capture whether the ambiguous language was primarily literal (e.g., a vague description of visual features) versus figurative (e.g., an allusion to a cultural reference). However, many clues featured aspects of both. For example, consider the clue *David and Goliath* intended to pick out a card with figures of two difference sizes. This is clearly figurative (rated 3/3). But the figurative language is used in reference to concrete visual features in the image; thus it also has a non-negligible degree of literalness as well (rated 2/2). It may have been possible to define an exact trade-off between the two, but it proved cognitively effortful for our raters to determine whether it a clue was primarily literal versus figurative. Accordingly, we settled on a final version distinguishing literalness and figurativeness as separate dimensions.

4.3 Reliability

We calculated inter-rater reliability using Krippendorff’s alpha (interval metric, $n = 150$ clues, 5 raters) for each coding dimension: Calibration ($\alpha = 0.866$), Situatedness ($\alpha = 0.927$), Literalness ($\alpha = 0.884$), and Figurativeness ($\alpha = 0.859$). All four dimensions exceeded conventional thresholds for good reliability ($\alpha \geq 0.80$).

5 Results

Our findings demonstrate key differences in how humans and LLMs calibrate ambiguity in the Dixit task (Fig. 2). We investigated the degree to which models show ambiguity collapse—as well as how readily they draw on culturally situated knowledge in constructing their clues. We then analysed whether formal linguistic features are sufficient to explain these differences. In Section 6, we looked at the conditions under which various models differ in how they calibrate ambiguity.

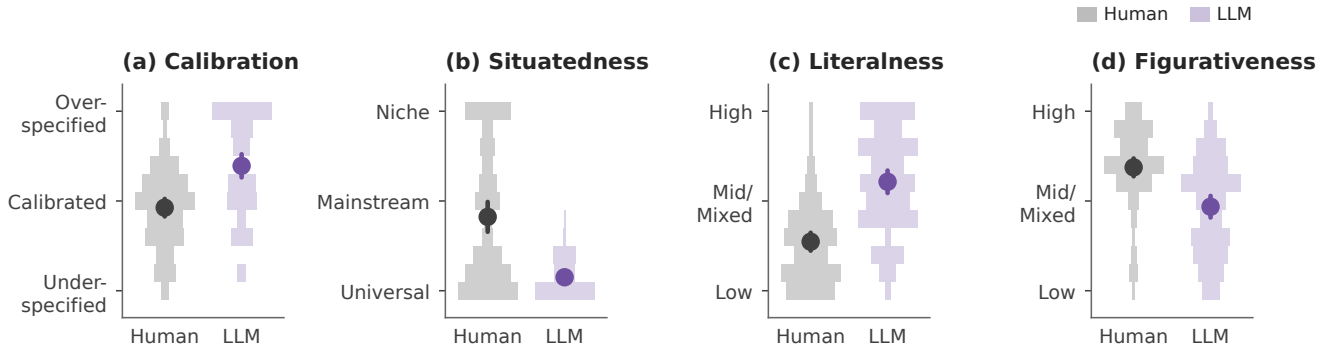


Figure 2: Human-coded rubric ratings on each of the four dimensions (Table 2) for human- vs. model-generated clues. Shaded bars reflect the share of clues at each rating level; points indicate group mean with 95% confidence interval. (a) Ambiguity collapse is demonstrated by models’ calibration erring on the side of Over-specified. (b) Cultural flattening is shown by the tendency of models to generate clues that rely on universally situated knowledge (i.e., not requiring cultural knowledge for interpretation). (c,d) Models tend to be more literal and less figurative than humans by default.

5.1 Humans and LLMs use different strategies to calibrate ambiguity—with varying levels of success

For comparison of human- vs model-generated clues, a panel of five expert human annotators used our coding rubric to analyse a set of 150 clues mixed and blinded (78 human, 72 model), drawn from 150 rounds that form part of our 350 seed boards (Section 3.3). To models from our set of six (see Table 1) were used to generate clues for this comparison (35 Gemma 3 [27B], 37 Qwen 3.5 [35B]). The model clues were generated under the *social cognition* prompt (Section 3.3); the *minimal* prompt of the expanded comparison is not represented in this set (Section 6.2.2).

5.1.1 LLMs exhibit ambiguity collapse. We sought explicit evidence of ambiguity output collapse in language models by coding human- vs model-generated clues according to “calibration” (as described in Section 4). Ambiguity output collapse was operationalized in this scheme by a rating of 3 on the Calibration dimension (where 2 is calibrated and 1 is overly ambiguous).

In ratings from our five expert human annotators, LLM clues were rated as significantly more likely to be over-specified than human clues (mean calibration 2.39 vs. 1.93 on the 1–3 scale; Mann-Whitney U , $p < 0.001$; Fig. 2a). Humans tended to be calibrated for ambiguity. But on average, LLMs skewed toward being overly specific in their clues, thus exhibiting ambiguity output collapse.

5.1.2 LLMs tend to flatten cultural context. We investigated the degree to which models rely on culturally-situated knowledge when constructing their clues—such as allusions to music or movies, widely circulated idioms or stories, or idiosyncratic interpersonal knowledge (e.g., inside jokes or common acquaintances). As described in Section 4, our “situatedness” dimension encoded three levels: universal (no culturally situated knowledge), mainstream (broadly shared Anglophone knowledge), or niche (specialized cultural knowledge, e.g., a line from a particular film).

In ratings from our five expert human annotators, LLM clues were rated as far less culturally situated than human clues (mean situatedness 1.15 vs. 1.82; $p < 0.001$; Fig. 2b). Human clues drew on a range of culturally specific references—which was almost never true of LLM clues.

5.1.3 LLMs tend to be more literal and less figurative than humans. We also looked at the degree to which the clue offered a literal description of the image or relied on figurative language. While these dimensions are strongly anti-correlated (Spearman’s $\rho = -0.82$ across the 150 human-coded clues, $p < 0.001$), it is possible for a clue-target mapping to rate high (or low) on both.

In ratings from our five expert human annotators, LLM clues were rated as significantly more likely to offer literal clues than humans (mean literalness 2.21 vs. 1.55; Mann-Whitney U , $p < 0.001$) and significantly less likely to rely on figurative language (mean figurativeness 1.94 vs. 2.37; $p < 0.001$; Fig. 2c–d). Some of the highly literal clues contributed to ambiguity collapse (i.e., increased incidence of over-specified clues), though much of this miscalibration reflected the models’ failed attempts to generate ambiguity by providing vague descriptions of visual features rather than exact labels (e.g., using terms like “wispy” or “fluffy” for a target image featuring a cloud).

5.1.4 Ambiguity calibration depends crucially on the model. We used two different models to generate LLM clues for comparison with humans: Gemma 3 (27B; dense) and Qwen 3.5 (35B; mixture of experts; see Section 3.3). Separating the outputs of these two models, we saw significant differences in the clues they generated (Fig. 3). Clues generated by both models were rated near the floor of the situatedness scale and did not differ significantly from one another (Gemma 1.18, Qwen 1.12; $p = 0.06$; Fig. 3b). However, they differed on the literalness and figurativeness dimensions: Qwen clues were more literal than Gemma clues (mean literalness 2.55 vs. 1.86; $p < 0.001$), while Gemma was more likely to draw on figurative language (2.25 vs. 1.64; $p < 0.001$; Fig. 3c–d).

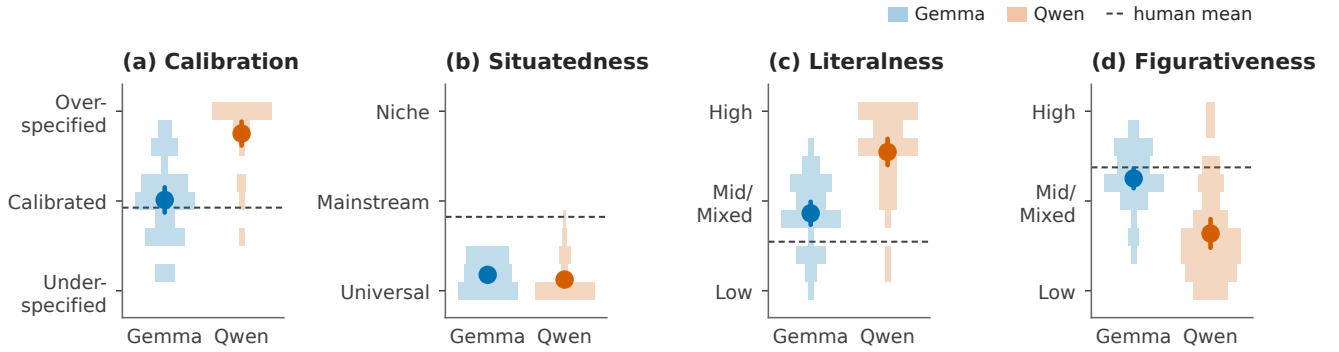


Figure 3: Model-generated clues from Fig. 2 separated by model. The dashed line reflects the the human mean. (a) Miscalibration is driven primarily by Qwen 3.5 (35B), rather than Gemma 2 (27B). (b) Both avoid cultural references. (c,d) The preference for literal rather than figurative language is more pronounced in Qwen.

This gives an important qualification to our calibration claim: the effect of LLM clues as significantly over-specified compared to human clues (Section 5.1.1) is driven by Qwen (mean calibration 2.75 vs. 1.93 for the human clues; $p < 0.001$), whereas Gemma’s calibration is statistically indistinguishable from the human clues (2.01 vs. 1.93; $p = 0.36$; Fig. 3a). Crucially, both models exhibit a similar degree of cultural flattening (Section 5.1.2; situatedness 1.18 and 1.12 vs. 1.82 for the human clues; both $p < 0.001$). In Section 6, we investigate whether this difference in calibration is an artefact of the models we initially selected or a robust difference in model behaviour.

5.2 Can formal linguistic features capture the differences without the rubric?

One potential concern about the findings presented in Section 5.1 is that the differences between human- and LLM-generated clues can be explained without reference to a complex, bespoke annotation rubric. Perhaps it would be enough to look at formal linguistic features to distinguish between these different strategies for calibrating ambiguity. Accordingly, we analysed whether a canonical set of formal linguistic features could predict whether a given clue was generated by a human or a model.

5.2.1 Individually, formal features separate human and LLM clues only weakly. We computed twelve surface features for every clue—length (word count, mean word length), lexical statistics (word frequency on the Zipf scale [69], computed with the *wordfreq* package [61], and concreteness [7]), part-of-speech composition (noun, adjective, determiner, and function-word ratios, presence of a finite verb), syntactic type, and orthography (title-casing, final punctuation)—for the 4,200 LLM-generated clues of the expanded clue set (six models $\times$ two prompts $\times$ 350 rounds; Section 3.3) and for a random sample of 2,000 human clues drawn from the 62,688 English clues in the full corpus.

For each single feature we asked how well it alone would tell an LLM clue from a human clue. The area under the ROC curve (AUC) summarised this as the probability that a randomly chosen LLM clue scored higher on the feature than a randomly chosen human clue,

with 0.50 as chance and 1.0 as perfect separation. With samples this large, small average differences are statistically reliable, so the confidence intervals of nearly every feature exclude 0.50 (LLM clues are, on average, longer and more concrete). But statistical reliability does not translate directly to discriminative power: a difference in means can be significant while the two distributions have significant overlap. For example, 93% of LLM clues had more than four words, while this was true of 46% of human clues (Fig. 4a, inset). Using word count to distinguish these clues would mark the majority of LLM clues correctly, but also misidentify almost half of human clues. Therefore, even the most discriminative formal feature on its own was not enough to distinguish human- and model-generated clues.

Word count was the most discriminative feature (AUC = 0.74, 95% CI [0.72, 0.75]), followed by concreteness (0.69). The degree to which word count distinguishes between human and model clues is at least partially an artefact of our prompts (Appendix A). We asked models to generate clues in a length of 3–6 words. Humans featured more variation in their clues—some were a single word, while others were lengthy phrases. Concreteness maps well onto our literalness/figurative dimensions. This supports their inclusion as key dimensions in the rubric.

5.2.2 LLM clues tend to be more homogeneous than human clues. Taken as an aggregate, however, superficial features could identify at least one important systematic difference between human- and model-generated clues: homogeneity. In a one-to-one comparison of individual clues, the signal of any single formal feature was weak, but model-generated clues tended to look more like one another than human clues, which exhibit a much higher degree of variation. In short, LLM-generated clues tended to cluster more tightly than human-generated ones (Fig. 4b).

Specifically, a multivariate score (the Mahalanobis distance to the LLM centroid in the z-scored twelve-feature space) separated held-out LLM clues from human clues with AUC = 0.92 (the dashed line in Fig. 4a), far above any single feature. In the twelve-feature space, LLM clues occupied a tighter subregion inside the human cloud (Fig. 4b): 99.8% of LLM clues fall inside the human 95% ellipsoid, but

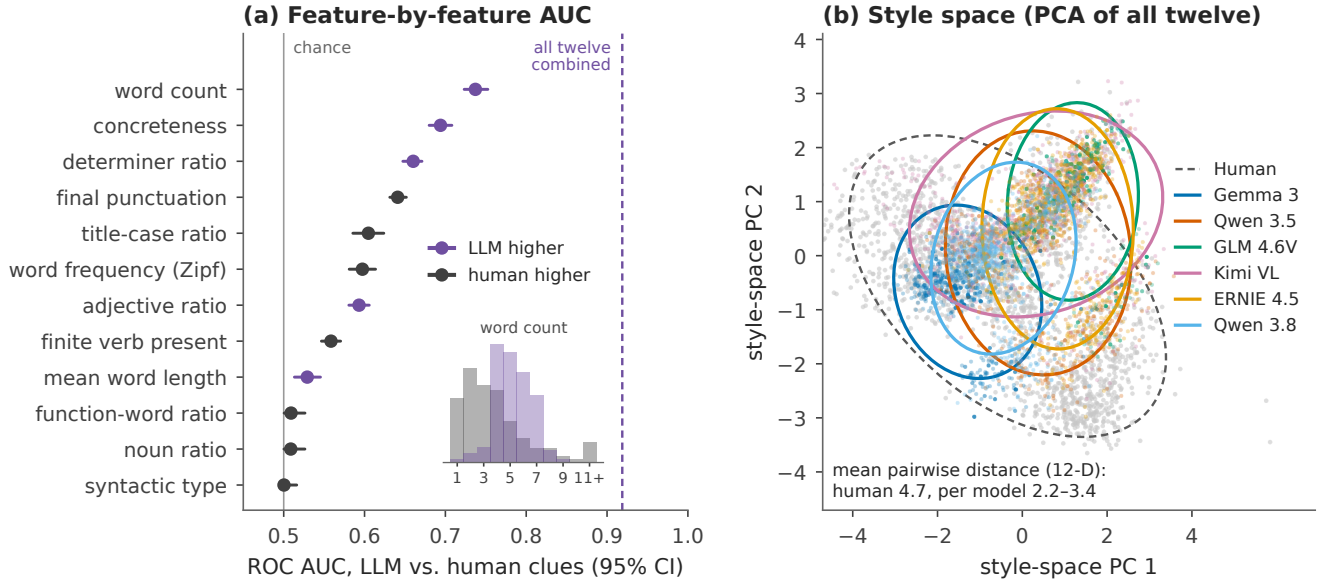


Figure 4: Formal linguistic features. (a) ROC AUC of each feature alone (95% CIs); dashed line: all twelve combined (Mahalanobis score, held out); inset: word-count distributions. (b) The same clues in the first two principal components; ellipses: 2 SD.

only 23% of human clues fall inside the LLM one. The mean pairwise distance between LLM clues was 3.2 versus 4.7 for human clues, roughly 1.5 $\times$ tighter. This aggregate figure potentially understates the effect: each model was narrower than the human cloud (2.2–3.4; 1.4–2.1 $\times$ tighter than humans). In the 2D projection (Fig. 4b), the six models visually sit in different parts of the human cloud. The largest distances between model centroids (3.2 between Gemma and GLM; 0.7–3.2 across all pairs) were twice the distance between the pooled LLM and human centroids (1.6). Thus, there was no single “LLM style”, only a set of narrow, model-specific styles. Each of these styles could plausibly be generated by a human, but reliance on any single one dramatically underestimates the space of clues humans actually generate.

6 Model Comparison

In Section 5.1.4, our initial analysis suggested that different LLMs use different strategies for ambiguity calibration. This was supported by the analysis in Section 5.2.2, which showed that different LLMs produce clues characterised by separate clusters of formal features. To what degree are these just artefacts of the two models we chose to compare against humans? To test this, we ran a larger comparison of six models (Table 1), featuring a range of parameter sizes (9B–35B) and two different architectures (dense vs mixture of experts).

6.1 Scaling annotation with LLM coders

To run this model comparison, we first needed to address the bottleneck of reliance on our human expert annotators. While there are recognized limitations to using LLMs for qualitative coding [64, 77, 83], a larger scale model comparison would not be feasible if it were limited by capacity of our human experts.

Table 3: LLM coders vs. the human panel ($n = 150$ clues). LOO r : one annotator vs. the mean of the other four. r , exact: correlation and exact agreement with the panel mean. Target: share of clues whose target card was identified.

	Human	Gemini 3.1 Flash-Lite		Gemini 3.5 Flash-Lite	
Dimension	LOO r	r	exact	r	exact
Calibration	0.60	0.68	0.77	0.62	0.76
Situatedness	0.80	0.77	0.80	0.79	0.78
Literatness	0.66	0.70	0.67	0.69	0.64
Figurativeness	0.58	0.63	0.60	0.60	0.55
Target identified	0.57	0.65		0.62	

To scale up our model comparison, we analysed how closely LLM coders could approximate the human coders when applying the rubric under the human protocol (Section 4). On the 150 clues coded by the five human annotators, each LLM coder’s rating correlated with the human panel mean at $r = 0.60$ – 0.79 depending on the dimension (Table 3)—close to, and on three of the four dimensions higher than, the agreement of an individual human annotator with the rest of the panel (leave-one-out $r = 0.58$ – 0.80). Exact agreement with the rounded panel mean was 0.55–0.80, and the LLM coders identified the intended target card at least as often as the human annotators did (0.62–0.65 vs. 0.57). Agreement was highest for situatedness, which was also the dimension on which the human annotators agree most with one another ($\alpha = 0.90$), and lowest for figurativeness and calibration, the two dimensions with the lowest human reliability ($\alpha \approx 0.80$; Section 4).

On this subset the LLM coders also reached the same substantive conclusions. Re-running the comparisons of Section 5.1 with each

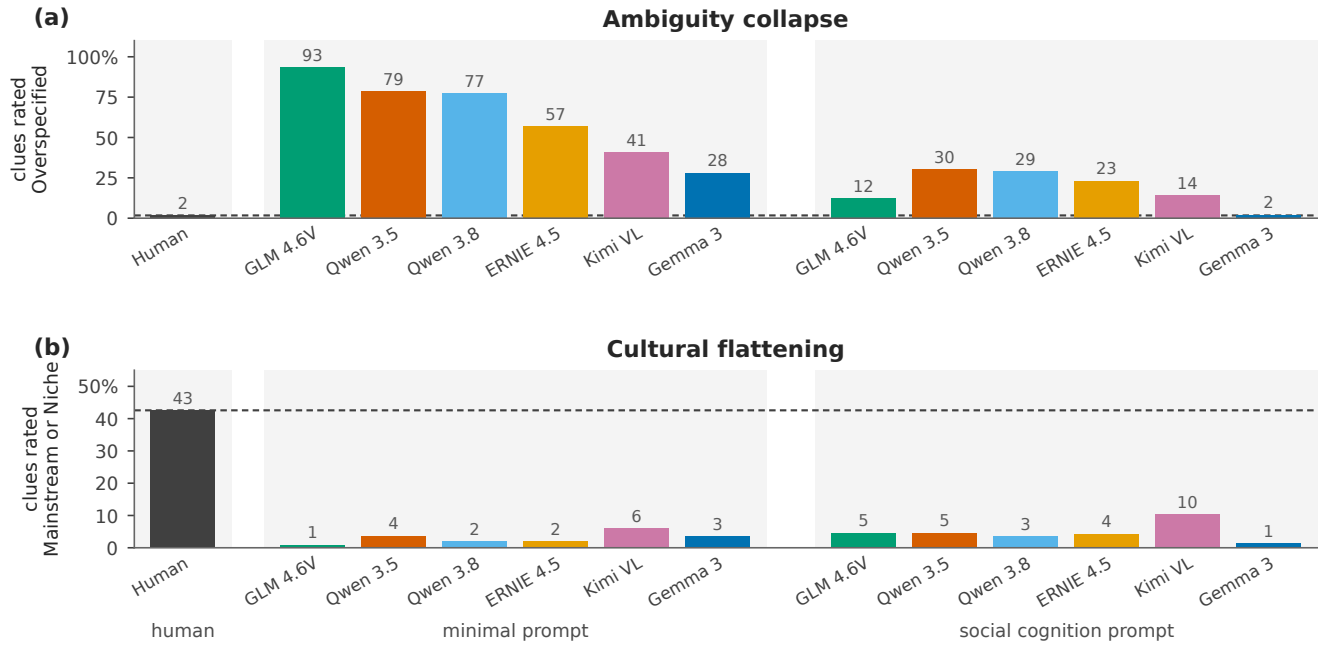


Figure 5: Models differ in their degree of ambiguity collapse, but all of them show evidence of cultural flattening. Each bank reflects analysis of 350 clues across six models under the minimal and the social cognition prompt, ordered by collapse under the minimal prompt. (a) Share of clues both coders rated *Overspecified*. (b) Share rated *Mainstream* or *Niche* on situatedness; note the 50% scale. Dashed lines: the human clues on the same rounds.

LLM coder in place of the human panel reproduced the calibration result exactly—Gemma was indistinguishable from humans and Qwen was significantly overspecified (both coders: Human vs. Gemma n.s., Human vs. Qwen $p < 0.001$)—as well as the Human < Gemma < Qwen gradient in literalness and the collapse of situatedness for both models (all $p < 0.01$).

6.2 Model Comparison Results

After establishing broad agreement with the human annotators, we applied the two LLM coders to the full expanded clue set: six models $\times$ two prompts $\times$ 350 rounds = 4,200 LLM clues, plus the 350 human clues on the same boards, each rated independently by both coders (9,100 ratings). Ratings were averaged over the two coders per clue. Each of the twelve conditions (6 models $\times$ 2 prompts) were compared with the human clues on the same 350 rounds (Fig. 5; Table 4). This increased the number of clues being analysed, from 35–78 clues per condition to 350, and from two models to six. As described in Section 3.3, we tested two different prompting conditions: minimal and social cognition.

6.2.1 All six models collapse, regardless of architecture or size. All six models exhibited ambiguity collapse under the minimal prompt. Each model produced clues that were rated as substantially more specified than the human clues on the same boards (all $p < 10^{-40}$; Cliff’s $\delta = 0.54$ – 0.97 ; Fig. 5a). The model-specific collapse we observed in Section 5.1 is therefore not just an effect of Qwen.

However, in some models the effects were more pronounced than others (Fig. 5). GLM-4.6V-Flash was rated as over-specified on 93% of rounds (2% for the human clues). At the other end, Gemma-3-27B was labeled as over-specified on 28% of rounds and Kimi-VL on 41%. This accorded with the coders’ target-card choices: when shown a minimal-prompt model clue, they identified the target on 83–100% of boards, against 34% for the human clues on the same boards (Table 4, target id.). Clue length did not explain the difference: the collapsed banks range from 3.5 to 6.4 words per clue around the human mean of 4.0. (Note: Section 5.1.4 found no miscalibration for Gemma, but the human-coded set contained no minimal-prompt clues; under the social cognition prompt Gemma’s gap remains negligible here as well, Section 6.2.2.)

We designed this comparison to contrast dense and mixture-of-experts models within a broadly similar scale (9B–35B total parameters; Table 1). Within this sample, calibration did not align with architecture. The two Qwen models—Qwen3.5-35B-A3B, a mixture-of-experts model with 3B active parameters, and Qwen3.8-27B, a dense model of a different size and generation—are almost indistinguishable on every dimension under both prompts (calibration 2.87 vs. 2.86 under the minimal prompt and 2.42 vs. 2.43 under social cognition; literalness 2.87 vs. 2.87 and 2.30 vs. 2.37; figurativeness 1.19 vs. 1.18 and 2.10 vs. 1.97).

Meanwhile the three dense models spanned a wider range, from the best-calibrated (Gemma) to the worst (GLM); the same was true of the three mixture-of-experts models. Neither parameter count nor active parameter count predicted the models’ ranking.

Table 4: Models differ in their degree of ambiguity collapse, but all of them show evidence of cultural flattening. Each bank reflects analysis of 350 clues across six models under the minimal and the social cognition prompt, ordered by collapse under the minimal prompt. (a) Share of clues both coders rated Overspecified. (b) Share rated Mainstream or Niche on situatedness; note the 50% scale. Dashed lines: the human clues on the same rounds.

Model	Prompt	Calib.	Situat.	Literal.	Figur.	Target id.	Words
Human	—	2.01	1.46	1.61	2.42	0.34	4.0
GLM-4.6V-Flash	min	2.96 (+0.97)	1.01 (-0.42)	2.99 (+0.98)	1.03 (-0.94)	1.00	6.4
	soc	2.21 (+0.27)	1.02 (-0.39)	1.96 (+0.37)	2.61 (+0.22)	0.83	6.4
Qwen3.5-35B	min	2.87 (+0.93)	1.03 (-0.40)	2.87 (+0.94)	1.19 (-0.87)	0.99	5.2
	soc	2.42 (+0.51)	1.03 (-0.39)	2.30 (+0.64)	2.10 (-0.26)	0.92	6.1
Qwen3.8-27B	min	2.86 (+0.92)	1.01 (-0.41)	2.87 (+0.95)	1.18 (-0.87)	0.99	4.6
	soc	2.43 (+0.53)	1.02 (-0.40)	2.37 (+0.70)	1.97 (-0.36)	0.95	4.7
ERNIE-4.5-VL	min	2.70 (+0.79)	1.01 (-0.41)	2.70 (+0.87)	1.41 (-0.75)	0.95	5.5
	soc	2.38 (+0.49)	1.02 (-0.40)	2.23 (+0.62)	2.33 (-0.06)	0.93	6.5
Kimi-VL-A3B	min	2.52 (+0.61)	1.05 (-0.37)	2.50 (+0.73)	1.48 (-0.70)	0.83	3.5
	soc	2.21 (+0.25)	1.07 (-0.34)	2.02 (+0.40)	2.32 (-0.06)	0.72	5.1
Gemma-3-27B	min	2.43 (+0.54)	1.02 (-0.40)	2.34 (+0.69)	1.94 (-0.43)	0.97	4.2
	soc	2.07 (+0.10)	1.01 (-0.42)	1.85 (+0.30)	2.55 (+0.12)	0.87	4.1

6.2.2 Prompting mitigates collapse but does not remove it. The social cognition prompt substantially changed the models’ calibration. A two-way analysis of variance on the LLM clues yielded significant effects of prompt ($F(1, 4188) = 1590$), model ($F(5, 4188) = 145$) and their interaction ($F(5, 4188) = 37.9$; all $p < 0.001$) on calibration. The social cognition prompt influenced models to look more like the human distribution—though the size of the difference depended on the model. GLM shifted by 0.76 points on the three-point scale (from 2.96 to 2.21), the two Qwen models by 0.43–0.45, Gemma by 0.35, ERNIE and Kimi by 0.32. In short: the social cognition prompt was most helpful for models that exhibit the sharpest initial degree of collapse.

Only Gemma reached a level of ambiguity calibration consistent with human-generated clues. Its social cognition clues were rated *Calibrated* on 98% of rounds (97% for the human clues) and received exactly the same rating as the human clue on 76% of boards; the residual difference was statistically detectable at $n = 350$ ($p < 0.001$) but negligible in size ($\delta = 0.10$; with the Gemini 3.5 coder alone, $\Delta = 0.02$, $p = 0.20$). For every other model the gap remained substantial ($\delta = 0.25$ –0.53; more specified than the human clue on 31–57% of boards).

The social cognition prompt asked the model to prefer figurative, rather than literal, language. For the most part, models were able to oblige. Under that prompt the figurativeness of Gemma’s and GLM’s clues (2.55 and 2.61) actually exceeded that of the human clues (2.42; $p < 0.01$), and Kimi’s and ERNIE’s match it (2.32–2.33, n.s.; Table 4). Yet GLM, Kimi and ERNIE remain significantly overspecified: GLM’s social-cognition clues, the most figurative bank in the study, were still rated *Overspecified* on 12% of rounds and more specified than the human clue on 31% of boards. Therefore, prompting is an important lever for modulating the degree to which they “describe the picture,” as we claim in our title—though it is crucial note that figurative language did not translate directly to better calibration of ambiguity.

6.2.3 Cultural flattening persists across models and prompts. Our initial findings on cultural flattening in model-based clues (Section 5.1.2) were supported by this analysis. Models uniformly relied on the “universal” level of situatedness, requiring no situated cultural knowledge to interpret their target (Fig. 5b). Human clues on these boards were rated “mainstream” or “niche” 43% of the time (11% niche). The LLM clues were rated above *Universal* on 3.9% of rounds. Every one of the twelve conditions (6 models $\times$ 2 prompts) showed near-floor effects on the scale (1.01–1.07; $\delta = -0.34$ to -0.42 against the human clues, all $p < 0.001$). Unlike calibration, situatedness did not respond to the prompt (prompt effect $F(1, 4188) = 3.3$, $p = 0.07$; interaction $p = 0.24$); the only model effect is Kimi-VL’s marginally higher floor. Collapse and cultural flattening therefore have different profiles in our sample. Collapse varies by model and is partly reversed by the social cognition prompt, whereas cultural flattening is present across models and is not modulated by prompt.

7 Discussion

We present a novel paradigm for operationalising calibrated ambiguity in multimodal generative AI systems. Across several analyses, we compare clues from an online data set of human-generated Dixit clues with clues generated by vision–language models playing the same role on the same boards. We developed a coding rubric to distinguish key differences in how ambiguity is calibrated and managed by humans and models. Our main finding is that models, as compared with humans, tend to be less well calibrated and almost never rely on the kind of cultural references found in human-constructed clues. In follow-up analyses, we show that individual surface-level features distinguish human vs model clues only weakly, and that there is significant variation in how different models calibrate ambiguity.

7.1 Do LLMs exhibit ambiguity output collapse?

Yes. In our paradigm, models tend to be less calibrated for ambiguity than humans (Section 5.1). This “output collapse” identified in our findings is only a subset of possible forms of ambiguity collapse

Table 5: Example clues from all 6 model and 2 prompt permutations compared with humans for two boards, shown with their five cards (target in red). Cal.: coder-averaged calibration rating (1 Underspecified, 2 Calibrated, 3 Overspecified). min/soc: minimal / social cognition prompt.

Board 1 (3 of 4 players found the target)					Board 2 (1 of 4)				
									
Producer		Clue		Cal.		Clue			Cal.
Human		Let it go		2.0	Christina Rossetti				2.0
GLM-4.6V-Flash	min	A woman in a long blue cloak		3.0	A white figure in a stormy field				3.0
Qwen3.5-35B	min	Girl in green coat in snow		3.0	Snowman built by a child				2.5
Qwen3.8-27B	min	Long hair blowing in wind		3.0	Walking through scattered white paper				3.0
ERNIE-4.5-VL	min	A figure in a long cloak		3.0	A ghost made of paper				3.0
Kimi-VL-A3B	min	A woman with a head injury		2.0	I left my shirt in your car				2.0
Gemma-3-27B	min	Lost in wintry contemplation		2.0	Ephemeral, fleeting, almost gone				2.0
GLM-4.6V-Flash	soc	The lone wanderer in the vast, silent expanse		2.0	The ghost of a forgotten dream				2.0
Qwen3.5-35B	soc	Cold winds of change are coming		2.0	Walking with a heart made of paper				2.0
Qwen3.8-27B	soc	Walking through frozen tears		2.0	Walking on fragile thoughts				2.0
ERNIE-4.5-VL	soc	A lonely wanderer in a long coat		3.0	A phantom in a tear-stained dream				2.0
Kimi-VL-A3B	soc	a beacon of hope amidst despair		2.0	The heart of winter				2.0
Gemma-3-27B	soc	Lost in quiet contemplation		2.0	Fleeting, fragile, and hopeful				2.0

articulated by Gur-Arieh et al. [24]—but an important one that has been hypothesised, yet not demonstrated, in previous work.

There are caveats to this claim. The prompt makes a difference (Section 6.2.1 and Section 6.2.2), though it does not mitigate ambiguity collapse entirely. A range of model sizes (9B–35B) and model architectures (dense, mixture-of-experts) are susceptible to this form of ambiguity collapse. However, different models do not collapse to the same degree. For example, under a “social cognition” prompt, Gemma 3 (27B) exhibited human-level ambiguity calibration (Section 6.2.2). But overall, we found robust evidence for ambiguity collapse across a range of circumstances.

7.2 Do LLMs avoid cultural references when calibrating ambiguity?

To a significant degree, yes. In our paradigm, models avoided culturally situated knowledge (Section 5.1.2). This effect held even under the social cognition prompt, which invites allusion and figurative language (Section 6.2.2). This seemed to be one of the defining differences between how humans and LLMs managed ambiguity in our task (Section 5.1). As a default behaviour, models tended to construct clues without cultural references.

This is a significant capability gap in model outputs, consistent with the output-based risks described by Gur-Arieh [24] and the hermeneutic challenges offered by Kommers et al. [35]. The evidence from our results suggests that culturally-situated knowledge plays a key role in how humans generate, manage, and calibrate ambiguity. If models avoid it by default, they are missing out an important communicative resource. However, it remains an open

question how best to elicit cultural information from models, for example by prompting them with information about the audience for whom they were generating clues or offering implicit clues about context without directly stating it (e.g., using British spellings) [71]. Nonetheless, humans naturally reference popular movies or stories like Pocahontas and Frozen; they riff on cultural idioms and stock phrases; and they play on associations with single words or concepts (like Australia or Metallica). These rely on cultural knowledge widely shared in the Anglophone world—and depending on the clue-target pairing in question, little expertise is required for interpretation beyond an awareness that such cultural referent exists. This is consistent with emerging evidence that LLMs struggle to “fill in the blanks” with details about hypothetical personas [72].

7.3 Do models tend to calibrate ambiguity in the same way?

No, not necessarily. The dynamics of ambiguity calibration seem to be model-specific (Section 5.1.4). For example, models differed in their proclivity for literal vs figurative language (Section 5.1.3). We found that models tended to produce homogeneous clues, but with each model using an idiosyncratic strategy that differed from other models (Fig. 4; Section 5.2). Models also responded differently to different prompts (Section 6.2.2). The one effect that was entirely robust across conditions and models was an avoidance of clues based on culturally-situated knowledge (Section 6.2.3); models were consistently and uniformly averse to making cultural references in their clues.

7.4 Can ambiguity collapse be fixed with prompting?

A little bit, sometimes, and potentially. We provided models with a “minimal” prompt (just the rules of the game), and this resulted in unanimous ambiguity collapse (Section 6.2.1). The least calibrated models performed better when we adjusted the prompt to emphasise social cognition (Section 6.2.2), but this did not by any means obviate the effect of collapse. Again, there were model-based differences. Of the models we looked at, Gemma performed best—and in some cases at human-level performance (Section 6.2.2).

In short, the evidence is mixed. It seems that prompting can help some models alleviate miscalibration to some degree. Targeted prompts can encourage models to use specific strategies (e.g., more figurative language)—but this does not automatically make them “good” at using these strategies in the sense that they result in more calibrated clues (Section 6.2.2). Thus, we stand by our primary claim that when calibrating ambiguity humans tend to reach for cultural references, while models tend to describe the picture.

8 Limitations and Alternative Perspectives

8.1 I agree the ambiguity matters. But does your task actually measure the kind of ambiguity we care about?

The strength of this task is that it provides a measurable, operationalised definition of calibrated ambiguity, where none existed before. The cost of that precision is that it only gets at a limited range of the possible textures of ambiguity. We operationalise ambiguity by the presence of multiple legitimate interpretations. This is kind of like operationalising a joke by whether it makes someone laugh. It is relevant—but it is also far from the whole story. The three non-calibration dimensions of our rubric (situatedness, literalness, and figurativeness) are meant to be broadly applicable to different types of ambiguity, but the mapping is imperfect. For example, in the introduction we give the example of moral ambiguity in the film *Unforgiven*. Certainly the dimension of culturally-situated knowledge would be relevant in analysing this form of ambiguity (what counts as morally good or bad behaviour in a particular frame of reference?). But it is less clear whether the tension between literal and figurative language or depictions is especially important. Our position is that the paradigm we present in this paper is a useful baseline. It captures some (not all) important aspects of ambiguity that apply in a range of situations. But ultimately the true “calibration” of ambiguity is tied intimately to the context in which a system is deployed. In future work, variations on this paradigm could be used to investigate more specific kinds of ambiguity (e.g., moral ambiguity in narratives produced by AI based on a set of Dixit-like target images) or variations could be adapted for specific use cases with a specific target population in mind (e.g., calibrating the ambiguity in labelling inappropriate images for content moderation for a given platform or population).

8.2 You developed a paradigm for measuring calibrated ambiguity. But once we have identified miscalibration, how do we fix it?

Our findings suggest that prompting makes a difference, especially for overall calibration and collapse. Of course it does, and the number of permutations for exploring how this might work are unbounded. Again, the key will be in contextualisation. Our paradigm offers a general framework for measuring calibrated ambiguity—but the kind of ambiguity that is desirable will depend on a specific situation or use-case. Future work should explore how prompting methods can achieve specific kinds of ambiguity—for specific models, operating in specific contexts. Our findings about the effects of prompting on cultural flattening are less clear. The relevant question is not whether models can be coaxed into drawing on cultural references, but about the processes for doing so that yield the most compelling results. There is a tension in existing literature between evidence for widespread homogenisation and cultural collapse [28, 31] versus evidence that local cultural knowledge is controllable, if you know how to elicit it [71]. Do models produce homogeneous outputs because we are just not prompting them correctly? It is a difficult question to answer, but the evidence from Veselovsky et al. [71] suggests that the key is not just asking behaviour representing a particular cultural milieu—but actively providing cues of participation in it. This is consistent with the view of LLMs as “context machines” [35] which are fundamentally designed to adapt to the contextual cues with which they are provided. Future work can explore the transition from measuring calibrated ambiguity to best practices for calibrating it.

8.3 I suspect these effects would be obviated in bigger models. Do you really think Claude Table 5 would exhibit ambiguity collapse?

This is an important direction for future research. Our primary aim in this paper was to provide a proof of concept for a novel methodology of measuring ambiguity calibration in LLM outputs. With this foundation, there are a lot of further questions to pose and investigate. How ambiguity calibration scales with model size is definitely one of them. In this version, we chose to use models that were large enough to do the task (e.g., Gemma 3 [27B] performed near human level in some cases) but lightweight enough to deploy in a range of situations (see Section 3.3 for further rationale). An additional part of the consideration was finding the right match for variation in performance for a simple baseline version of the task. For example, the task can be made arbitrarily more difficult by increasing the number of distractors. Given that Gemma 3 (27B) performs well under the conditions we studied, it is probably the case that bigger, more sophisticated models could do the task as currently designed. The task can be adapted for larger models—for example, by increasing the number of distractors, altering the kind of cards that are used, or providing background information about the shared cultural frames relevant for the other players in the game. Future work can explore how variations of this paradigm can elucidate the strategies more sophisticated models use to calibrate ambiguity—and provide a concrete measure of when they collapse.

Ethics and Privacy Statement

All data for human-generated clues in this study were publicly available. All expert coders were authors on the paper, not participants in a study. Dixit images are subject to copyright (c) Libellud / Marie Cardouat. The work presented in our system presents minimal ethical or privacy risk. The Dixit images do not feature harmful, triggering, or inappropriate content. However, an adaptive version of this study could—specifically one that looks at ambiguity in a more ethically fraught context, such as content moderation.

References

- [1] Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*. 15–21.
- [2] Nina Beguš. 2025. *Artificial Humanities: A Fictional Perspective on Language in AI*. University of Michigan Press.
- [3] A Stevie Bergman, Lisa Anne Hendricks, Maribeth Rauh, Boxi Wu, William Agnew, Markus Kunesch, Isabella Duan, Iason Gabriel, and William Isaac. 2023. Representation in AI evaluations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 519–533.
- [4] Wayne C. Booth. 1961. *The Rhetoric of Fiction*. University of Chicago Press, Chicago.
- [5] Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge.
- [6] Jerome Seymour Bruner. 1990. *Acts of meaning: Four lectures on mind and culture*. Vol. 3. Harvard university press.
- [7] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas. *Behavior Research Methods* 46, 3 (2014), 904–911. doi:10.3758/s13428-013-0403-5
- [8] Shi-Tao Chang. 2024. Dixit AI: An OpenAI CLIP-based Hint Selection Player. *Academic Journal of Computing & Information Science* 7, 11 (2024), 102–108. doi:10.25236/AJCIS.2024.071114
- [9] Bonnie Chinh, Himanshu Zade, Abbas Ganji, and Cecilia Aragon. 2019. Ways of qualitative coding: A case study of four strategies for resolving disagreements. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.
- [10] Alwin de Rooij and Michael Mose Biskjaer. 2026. Does generative AI make us think alike? A systematic review and meta-analysis of homogenisation effects in human–AI co-creation. *Behaviour & Information Technology* (2026), 1–14.
- [11] Chiara Di Lodovico, Steven Houben, and Sara Colombo. 2025. How to design with ambiguity: insights from self-tracking wearables. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [12] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A Crowdsourced Frame Disambiguation Corpus with Ambiguity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics, 2164–2170. doi:10.18653/v1/N19-1224
- [13] Zackary Okun Dunivin. 2025. Scaling hermeneutics: a guide to qualitative coding with LLMs for reflexive content analysis. *EPJ Data Science* 14, 1 (2025), 28.
- [14] William Empson. 1930. *Seven Types of Ambiguity*. Chatto & Windus, London.
- [15] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature* 630 (2024), 625–630. doi:10.1038/s41586-024-07421-0
- [16] Henry Farrell, Alison Gopnik, Cosma Shalizi, and James Evans. 2025. Large AI Models Are Cultural and Social Technologies. *Science* 387, 6739 (2025), 1153–1156. doi:10.1126/science.adt9819
- [17] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.
- [18] Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Sánchez-Marco, Bianca Scarlina, Viviana Patti, Cristina Bosco, et al. 2023. Epic: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 13844–13857.
- [19] William W. Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity as a resource for design. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 233–240.
- [20] Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. 2024. How culture shapes what people want from AI. In *Proceedings of the 2024 CHI conference on human factors in computing systems*. 1–15.
- [21] Evelyn Gius and Janina Jacke. 2017. The hermeneutic profit of annotation: On preventing and fostering disagreement in literary analysis. *International Journal of Humanities and Arts Computing* 11, 2 (2017), 233–254.
- [22] H. Paul Grice. 1975. Logic and Conversation. In *Syntax and Semantics, Vol. 3: Speech Acts*, Peter Cole and Jerry L. Morgan (Eds.). Academic Press, New York, 41–58.
- [23] Neel Gupta, Maria Antoniak, and Melanie Walsh. 2026. AI Fiction in the Wild. *arXiv preprint arXiv:2606.22748* (2026).
- [24] Shira Gur-Arieh, Angelina Wang, and Sina Fazelpour. 2026. Ambiguity Collapse by LLMs: A Taxonomy of Epistemic Risks. *arXiv:2603.05801*. arXiv:2603.05801 [cs.CY]
- [25] Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies* 14, 3 (1988), 575–599.
- [26] Reid Hastie, Steven D. Penrod, and Nancy Pennington. 1983. *Inside the Jury*. Harvard University Press, Cambridge, MA.
- [27] Drew Hemment, Cory Kommers, Ruth Ahnert, Maria Antoniak, Glauco Arbib, Vaishak Belle, Steve Benford, Alexandra Brintrup, Nick Bryan-Kinns, Mercedes Bunz, et al. 2025. Doing AI differently: Rethinking the foundations of AI via the humanities. (2025).
- [28] Ryan Heuser. 2025. Cultural collapse: toward a generative formalism for AI cultural production. *Anthology of Computers and the Humanities* 3 (2025), 575–588.
- [29] Wolfgang Iser. 1978. *The Act of Reading: A Theory of Aesthetic Response*. Johns Hopkins University Press, Baltimore.
- [30] Koki Iwata, Reiji Suzuki, and Takaya Arita. 2020. An Approach to Human Cognition Based on AI Player Creation for a Picture-Guessing Card Game Dixit. In *Proceedings of the 34th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI 2020)*. The Japanese Society for Artificial Intelligence, 4C2GS1302. doi:10.11517/psai.JSAI2020.0_4C2GS1302
- [31] Shomik Jain, Jack Lanchantin, Maximilian Nickel, Candace Ross, Karen Ullrich, Ashia Wilson, and Jamelle Watson-Daniels. 2025. Task-Dependent Evaluation of LLM Output Homogenization: A Taxonomy-Guided Framework. *arXiv preprint arXiv:2509.21267* (2025).
- [32] Ahmed Karim, Qiao Wang, and Zheng Yuan. 2025. Beyond the Score: Uncertainty-Calibrated LLMs for Automated Essay Assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 19642–19647.
- [33] Do-Kyung Kim, Hyeseon Ahn, Youngwook Kim, and Yo-Sub Han. 2025. Analyzing offensive language dataset insights from training dynamics and human agreement level. In *Proceedings of the 31st International Conference on Computational Linguistics*. 9780–9792.
- [34] Lauren Klein, Meredith Martin, André Brock, Maria Antoniak, Melanie Walsh, Jessica Marie Johnson, Lauren Tilton, and David Mimno. 2025. Provocations from the humanities for generative AI research. *arXiv preprint arXiv:2502.19190* (2025).
- [35] Cody Kommers, Ruth Ahnert, Maria Antoniak, Emmanouil Benetos, Steve Benford, Mercedes Bunz, Baptiste Caramiaux, Shauna Concannon, Martin Disley, James Dobson, Yali Du, Edgar Duñez-Guzmán, Kerry Francksen, Evelyn Gius, Jonathan W. Y. Gray, Ryan Heuser, Sarah Immel, Richard Jean So, Sang Leigh, Dalaki Livingston, Hoyt Long, Meredith Martin, Georgia Meyer, Daniela Mihai, Ashley Noel-Hirst, Kirsten Ostherr, Deven Parker, Yipeng Qin, Jessica Ratcliff, Emily Robinson, Karina Rodriguez, Adam Sobey, Ted Underwood, Aditya Vashistha, Matthew Wilkens, Youyou Wu, Yuan Zheng, and Drew Hemment. 2026. Computational Hermeneutics: Evaluating Generative AI as a Cultural Technology. *Frontiers in Artificial Intelligence* 9 (2026), 1753041. doi:10.3389/frai.2026.1753041
- [36] Cody Kommers and Simon DeDeo. 2025. Sense-making, cultural scripts, and the inferential basis of meaningful experience. In *Proceedings of the annual meeting of the cognitive science society*, Vol. 47.
- [37] Cody Kommers, Eamon Duede, Julia Gordon, Ari Holtzman, Tess McNulty, Spencer Stewart, Lindsay Thomas, Richard Jean So, and Hoyt Long. 2026. Why slop matters. *ACM AI Letters* 1, 1 (2026), 1–6.
- [38] Cody Kommers, Drew Hemment, Maria Antoniak, Joel Z Leibo, Hoyt Long, Emily Robinson, and Adam Sobey. 2025. Meaning is not a metric: Using llms to make cultural context legible at scale. *arXiv preprint arXiv:2505.23785* (2025).
- [39] Cody Kommers and Ari Holtzman. 2026. AI as Entertainment. *arXiv preprint arXiv:2601.08768* (2026).
- [40] Maithilee Kunda and Irina Rabkina. 2020. Creative Captioning: An AI Grand Challenge Based on the Dixit Board Game. *arXiv:2010.00048*. arXiv:2010.00048 [cs.AI]
- [41] Joel Z. Leibo, Alexander Sasha Vezhnevets, Manfred Diaz, John P. Agapiou, William A. Cunningham, Peter Sunehag, Julia Haas, Raphael Koster, Edgar A. Duñez-Guzmán, William S. Isaac, Georgios Piliouras, Stanley M. Bileschi, Iyad Rahwan, and Simon Osindero. 2024. A Theory of Appropriateness with Applications to Generative Artificial Intelligence. *arXiv:2412.19010*. arXiv:2412.19010 [cs.AI]
- [42] Wenye Lin, Jonathan Roberts, Yunhan Yang, Samuel Albanie, Zongqing Lu, and Kai Han. 2025. GAMEBot: Transparent Assessment of LLM Reasoning in Games. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria, 7656–7682. doi:10.18653/v1/2025.acl-long.378

- [43] Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2023. We're Afraid Language Models Aren't Modeling Ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 790–807. doi:10.18653/v1/2023.emnlp-main.51
- [44] Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics* 13 (2025), 652–689.
- [45] Ka Man Lo, Zeyu Huang, Zihan Qiu, Zili Wang, and Jie Fu. 2025. A Closer Look into Mixture-of-Experts in Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, Albuquerque, New Mexico, 4427–4447. doi:10.18653/v1/2025.findings-naacl.251
- [46] Ryan Lowe, Joe Edelman, Tan Zhi-Xuan, Oliver Klingefjord, Ellie Hain, Vincent Wang, Atrisha Sarkar, Michiel A Bakker, Fazl Barez, Matija Franklin, et al. 2025. Full-stack alignment: Co-aligning AI and institutions with thicker models of value. In *2nd workshop on models of human feedback for AI alignment*.
- [47] Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th international conference on computational linguistics*. 3659–3668.
- [48] Luise Metzger, Linda Miller, Martin Baumann, and Johannes Kraus. 2024. Empowering calibrated (dis-) trust in conversational agents: a user study on the persuasive power of limitation disclaimers vs. authoritative style. In *Proceedings of the 2024 CHI conference on human factors in computing systems*. 1–19.
- [49] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 5783–5797. doi:10.18653/v1/2020.emnlp-main.466
- [50] Lucia Montesinos and Anders Sundnes Løvlie. 2026. Machine Learning as Design Material for Music-Making. In *Proceedings of the 2026 Designing Interactive Systems Conference*. 4768–4784.
- [51] Heejeong Nam, Jinwoo Ahn, Keummin Ka, Jiwan Chung, and Youngjae Yu. 2025. VAGUE: Visual Contexts Clarify Ambiguous Expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. arXiv:2411.14137.
- [52] Roberto Navigli. 2009. Word Sense Disambiguation: A Survey. *Comput. Surveys* 41, 2 (2009), 10:1–10:69. doi:10.1145/1459352.1459355
- [53] Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What Can We Learn from Collective Human Opinions on Natural Language Inference Data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 9131–9143. doi:10.18653/v1/2020.emnlp-main.734
- [54] Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics* 7 (2019), 677–694. doi:10.1162/tacl_a_00293
- [55] Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The Communicative Function of Ambiguity in Language. *Cognition* 122, 3 (2012), 280–291. doi:10.1016/j.cognition.2011.10.004
- [56] Steven Pinker, Martin A. Nowak, and James J. Lee. 2008. The Logic of Indirect Speech. *Proceedings of the National Academy of Sciences* 105, 3 (2008), 833–838. doi:10.1073/pnas.0707192105
- [57] Edwin Roland, Richard Jean So, and Hoyt Long. 2026. The social AI author: modeling creativity and distinction in simulated cultural fields. *AI & SOCIETY* 41, 5 (2026), 4333–4347.
- [58] Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2428–2441.
- [59] Adam Sennet. 2023. Ambiguity. In *The Stanford Encyclopedia of Philosophy* (Summer 2023 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.
- [60] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Miresghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070* (2024).
- [61] Robyn Speer. 2022. rspeer/wordfreq: v3.0. doi:10.5281/zenodo.7199437
- [62] James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in large language models and humans. *Nature human behaviour* 8, 7 (2024), 1285–1295.
- [63] Cass R. Sunstein. 1995. Incompletely Theorized Agreements. *Harvard Law Review* 108, 7 (1995), 1733–1772.
- [64] Robert H. Tai, Lillian R. Bentley, Xin Xia, Jason M. Sitt, Sarah C. Fankhauser, Ana M. Chicas-Mosier, and Barnas G. Monteith. 2024. An Examination of the Use of Large Language Models to Aid Analysis of Textual Data. *International Journal of Qualitative Methods* 23 (2024). doi:10.1177/16094069241231168
- [65] Yusei Tanzawa, Kazuya Tsubokura, Reon Ohashi, Hibiki Sakurai, and Kunikazu Kobayashi. 2023. Evaluation Focused on the Relevance of Hints for Implementing Dixit Game AI. In *Proceedings of the 37th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI 2023)*. The Japanese Society for Artificial Intelligence, 2F5GS505. doi:10.11517/pjsai.JSAI2023.0_2F5GS505
- [66] Ching-Yi Tsai, Nicole Tacconi, Andrew D. Wilson, and Parastoo Abtahi. 2026. Uncertain Pointer: Situated Feedforward Visualizations for Ambiguity-Aware AR Target Selection. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–31.
- [67] Alexandra Uma, Dina Almane, and Massimo Poesio. 2022. Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence* 5 (2022), 818451.
- [68] Ted Underwood. 2025. The impact of language models on the humanities and vice versa. *Nature Computational Science* 5, 9 (2025), 695–697.
- [69] Walter J. B. van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. SUBTLEX-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology* 67, 6 (2014), 1176–1190. doi:10.1080/17470218.2013.850521
- [70] Dimitrios Vatsakis, Paris Mavromoustakos-Blom, and Pieter Spronck. 2022. An Internet-Assisted Dixit-Playing AI. In *Proceedings of the 17th International Conference on the Foundations of Digital Games (FDG)*. ACM, 1–7. doi:10.1145/3555858.3555863
- [71] Veniamin Veselovsky, Berke Argin, Benedikt Stroebel, Chris Wendler, Robert West, James Evans, Thomas L Griffiths, and Arvind Narayanan. 2026. Localized cultural knowledge is conserved and controllable in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*. 43152–43178.
- [72] Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2025. Large Language Models That Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups. *Nature Machine Intelligence* 7, 3 (2025), 400–411. doi:10.1038/s42256-025-00986-z
- [73] Ryan Wei. 2023. Dixit Player with Open CLIP. *Journal of Data Analysis and Information Processing* 11, 4 (2023), 536–547. doi:10.4236/jdaip.2023.114027
- [74] Frank Wildenburg, Michael Hanna, and Sandro Pezzelle. 2024. Do Pre-Trained Language Models Detect and Understand Semantic Underspecification? Ask the DUST!. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 9598–9613. doi:10.18653/v1/2024.findings-acl.572
- [75] Matthew Wilkens. 2026. AI as a Tool for Simulation-Based Experiments in Literary Studies. *arXiv preprint arXiv:2606.02293* (2026).
- [76] Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J Su. 2025. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *J. Amer. Statist. Assoc.* 120, 552 (2025), 2154–2164.
- [77] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23 Companion)*. ACM, 75–78. doi:10.1145/3581754.3584136
- [78] Yu Xie and Yueqi Xie. 2026. When artificial intelligence makes everything similar: The risks of content homogenization. *Chinese Journal of Sociology* 12, 2 (2026), 157–172.
- [79] Srishti Yadav, Lauren Tilton, Maria Antoniak, Taylor Arnold, Jiaang Li, Siddhesh Milind Pawar, Antonia Karamolegkou, Stella Frank, Zhaochong An, Negar Rostamzadeh, et al. 2025. Evaluation of Cultural Competence of Vision-Language Models. *arXiv preprint arXiv:2505.22793* (2025).
- [80] Zheng Yuan and David Strohmaier. 2021. Cambridge at SemEval-2021 Task 2: Neural WiC-Model with Data Augmentation and Exploration of Representation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu (Eds.). Association for Computational Linguistics, Online, 730–737. doi:10.18653/v1/2021.semeval-1.96
- [81] Jiayin Zhi, Hoyt Long, Richard Jean So, and Mina Lee. 2026. What Does AI Do for Cultural Interpretation? A Randomized Experiment on Close Reading Poems with Exposure to AI Interpretation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [82] Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural nlp. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 25869–25886.
- [83] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics* 50, 1 (2024), 237–291. doi:10.1162/coli_a_00502

A Storyteller prompts

Both prompts were given as the system message together with the target card image. Text in the two prompts is identical up to

the point marked "..."; the social cognition prompt then adds the coaching shown.

Minimal.

You are playing a game with four other players. In the game, you secretly hold a TARGET card. Your job is to be the storyteller: you must generate a clue that hints at your TARGET. The other four players each hold a hand of cards and, after hearing your clue, will each submit one of their own cards as a decoy that might fit your clue. All five cards (your TARGET plus their four decoys) will then be shuffled on the table, and each of them will vote for which one is the TARGET. You want some of the other players to guess your TARGET correctly, but not all of them.

Your clue should be three to six words in length. Respond with only the clue.

Social cognition.

You are playing a game with four other players. In the game, you secretly hold a TARGET card. Your job is to be the storyteller: you must generate a clue that hints at your TARGET. However, you don't want the answer to be obvious. The other four players ... You want some of the other players to guess your TARGET correctly, but not all of them. When generating the clue, think about things that could plausibly apply to many cards in a way that might mislead some players, while still favoring the TARGET.

An ideal clue should be:

- Three to six words in length
- Based on emotions, metaphors, allusions, and other forms of figurative or abstract language that could be applied to multiple cards
- Ambiguous enough that not all players will agree on the same card, but more likely to refer to your TARGET than to an unrelated card

Think about how the other players will perceive your clue: Respond with only a clue that two of the four other players would correctly guess applies to the TARGET rather than to a decoy.

B Behavioural robustness check: a vision-language judge panel

Before adopting the rubric as the primary instrument, we scored clues behaviourally with a panel of judges that voted for the target card as a Dixit audience would. This check used an earlier configuration of the pipeline: 200 four-player rounds from the human dataset, Gemma-3-27B and Qwen3.5-35B-A3B as storytellers under the two wordings of the social cognition prompt ($200 \times 2 \times 2 = 800$ clues), with clues clipped to a maximum of seven words.

Judge panel. The panel comprised four open-weight vision-language judges (Qwen3.5-9B, Qwen3-VL-8B, Gemma-3-12B, and InternVL3-8B), each served independently. For every clue, each judge saw the four table cards from the original round—the target plus the three decoys the human players had actually contributed—in a shuffled order together with the clue, and voted for the single card it believed the clue referred to (greedy decoding, forced single-index reply). Votes that could not be parsed were discarded; the full four-judge panel returned valid votes in 76% of the 1,000 judged

rounds, and all metrics are computed over the valid votes of each round. Valid-vote counts are essentially identical across producer conditions (means of 3.60–3.69 judges per round), so discarded votes do not favour any producer. Following Dixit's scoring rule, a clue counts as *in the target range* when some but not all valid judges select the target card; clues that every judge resolves are overspecified ("too obvious"), and clues that no judge resolves are underspecified ("too obscure"). The same judge panel scored human- and LLM-generated clues on the same boards, so any difference between producers cannot be attributed to the judging procedure. The four judges behave as a heterogeneous audience rather than four copies of one opinion: mean pairwise agreement on the voted card is 0.45 (chance 0.25), and the panel is unanimous in only 14% of full-panel rounds.

Validation against human audiences. The 200 human-clue rounds come with the real audience's outcome, which lets us ask how well the judge panel stands in for human guessers. At the population level the two audiences agree closely: the panel places 60.5% of human clues in the target range, against 60.0% for the original human audiences on the same rounds. At the level of individual rounds the correspondence is weaker: per-round hit rates correlate positively but modestly (Spearman $\rho = .29$, $p < .001$), and binary in-range agreement between the two audiences is close to chance. Some attenuation is inevitable when comparing single samples of two small audiences (three human guessers vs. four model judges), but it does not account for all of the gap: a binomial simulation in which both audiences respond to the same per-clue transparency predicts substantially higher correspondence (agreement $\approx .61$, $\rho \approx .53$) than we observe. The panel is therefore not a stand-in for how a particular human audience reads a particular clue, and we use it only for what its aggregate behaviour is shown to match: population-level comparisons of clue producers on identical boards.

B.1 Model Comparison Robustness

Three checks indicate that these results do not depend on incidental features of the pipeline. First, the two coders agree with each other on the 4,550 double-coded clues at $r = .66$ (Calibration), $.73$ (Situatedness), $.75$ (Literalness) and $.84$ (Figurativeness), with exact agreement of $.76$ – $.96$; every calibration gap reported above is significant in the same direction with either coder alone, the sole exception being the Gemma social-cognition bank, which one coder rates as equivalent to the human clues. Second, the 350 boards comprise two subsets selected in different ways (Section 3.3); the mean calibration of every bank differs by at most 0.10 between them. Third, length does not explain calibration. Within the LLM clues the correlation between word count and calibration rating is $\rho = .11$; Kimi-VL's minimal-prompt clues are the shortest of any bank (3.5 words) yet substantially overspecified ($\delta = .61$), and GLM's two banks have identical length (6.4 words) but $\delta = .97$ and $.27$. Length, in other words, does not account for the calibration gap.