

Seeing Without Knowing: Dissociating Evidence, Bias, and Confidence in Artificial Spatial Reasoning

Ahmad M. Nazzal

Email: a-nazzal[at]outlook[dot]com

Abstract

A central question in artificial intelligence is whether systems that produce intelligent behavior can meaningfully evaluate the reliability of their own decisions, which is a hallmark of human cognition. In biological systems, this capacity is closely tied to metacognition, where confidence reflects the quality of available evidence and supports adaptive behavior. Spatial reasoning offers a particularly relevant domain for examining these processes, as it forms a core component of human world models, enabling the integration of structured visual information into coherent representations of the environment. As vision-capable generative models are increasingly used to interpret spatial scenes, it remains unclear whether their decisions are guided by graded evidence and accompanied by calibrated self-evaluation. Here, we introduce a controlled psychophysical paradigm to probe evidence sensitivity, spatial decision-making, and confidence behavior in a vision-capable generative model. We designed a coherence-based task in which directional evidence was manipulated by varying the proportion of left- versus right-pointing arrows within a fixed spatial array. Across 100 randomized trials, the model reported the perceived majority direction along with a graded confidence estimate, enabling construction of a psychometric function linking coherence to choice behavior. Directional choices showed sensitivity to evidence but were strongly asymmetric, with a marked response bias and a delayed transition point requiring high coherence to reverse decisions. Critically, confidence remained uniformly high across conditions and did not track difficulty or accuracy. This dissociation suggests limited metacognitive calibration and raises questions about whether current models form spatially grounded internal representations. More broadly, the study establishes a scalable methodology for probing confidence and bias in AI systems and highlights a gap between perceptual performance, spatial reasoning, and evidence-based self-monitoring. This preprint presents a pilot psychophysical evaluation of decision bias and confidence behavior in a vision-language model using controlled perceptual stimuli. The work is exploratory and establishes a methodological framework for future, larger-scale investigations.

Introduction

A central philosophical question surrounding artificial intelligence is whether systems that produce intelligent-seeming outputs can be considered genuinely intelligent if they cannot evaluate, monitor, or reflect on their own decisions. Human cognition is not defined solely by the ability to generate responses, but also by the capacity to assess the reliability of those responses. Reflection, uncertainty monitoring, and the ability to adjust behavior in light of confidence are core features of what is commonly understood as thinking. As artificial systems become increasingly capable across language, vision, and reasoning tasks, the question shifts from whether they can perform well to whether they can evaluate their own performance in a meaningful and structured way.

This question is closely tied to the concept of metacognition. In biological systems, metacognition refers to the ability to monitor and regulate one's own cognitive processes. In perceptual decision-making, this often manifests as a subjective estimate of confidence in a choice. Confidence is not merely an accompanying feeling; it is a computationally relevant signal that guides learning, strategy selection, and adaptive behavior. When evidence is ambiguous, humans tend to express lower confidence; when evidence is strong, confidence increases. This dynamic relationship between evidence strength, performance, and confidence is a defining feature of metacognitive sensitivity.

Confidence therefore provides a measurable window into second-order cognition. If an agent is able to track uncertainty and adjust confidence appropriately, it suggests that it has access to internal signals related to evidence quality. Conversely, if confidence remains high regardless of task difficulty or correctness, this may indicate that responses are generated without an internal evaluation of reliability. In this sense, confidence reports offer a tractable behavioral proxy for probing whether an agent monitors its own decision processes.

Spatial perception and spatial decision-making provide a particularly well-suited domain for studying these phenomena. The human brain contains specialized systems for representing and reasoning about spatial information, integrating sensory evidence across the visual field to guide perception and action. Disruptions to these systems, such as those seen in hemispatial neglect following stroke, can produce stable and measurable directional biases. Even in healthy observers, spatial tasks have long been used in psychophysics to quantify evidence accumulation, sensitivity to signal strength, and confidence calibration. The random-dot motion paradigm, for example, has served as a canonical framework for understanding how graded sensory evidence is translated into categorical choices and confidence judgments.

As vision-capable artificial models become more sophisticated, they are increasingly able to perform spatial reasoning tasks. However, it remains unclear whether these systems process spatial evidence in a graded and internally consistent manner, or whether they rely on fixed heuristics that produce plausible outputs without genuine evidence integration. In particular, little is known about how such models represent uncertainty in spatial decisions, and whether their confidence reports track evidence strength in a way that resembles metacognitive monitoring in humans.

To address this gap, we introduce a simplified spatial decision-making paradigm inspired by classic perceptual coherence tasks. In this task, directional evidence is conveyed by the proportion of left- versus right-pointing arrows within a circular visual array. By

systematically manipulating this proportion while keeping spatial layout constant, it becomes possible to control the strength of evidence favoring one direction over the other. This design allows for the construction of psychometric curves relating stimulus coherence to choice behavior and provides a framework for measuring confidence as a function of task difficulty.

The central problem this study addresses is whether a vision-capable generative model exhibits signatures of metacognitive sensitivity when making spatial judgments. Specifically, we test whether (i) directional choices track graded evidence in a smooth and unbiased manner, and (ii) confidence ratings vary systematically with coherence and performance. If the model integrates evidence in a human-like way, one would expect both a symmetric psychometric function centered near equal evidence and confidence that increases with signal strength. Deviations from this pattern—such as stable directional biases or confidence that remains high despite ambiguity—would suggest that decision outputs are not accompanied by a meaningful internal estimate of uncertainty.

By combining a controlled spatial task with trial-by-trial confidence reporting, this study establishes a behavioral methodology for probing metacognitive-like properties in artificial systems. Rather than evaluating performance alone, we treat the model as an experimental subject and analyze its responses using tools from psychophysics. In doing so, we aim to move beyond performance benchmarks toward a deeper characterization of how artificial systems represent evidence, make decisions, and signal certainty about those decisions.

Methods

Task design

We implemented a static analogue of the random-dot motion paradigm in which directional evidence was conveyed by the proportion of left- versus right-pointing arrows in a circular array. Each stimulus was a 100×100 pixel grayscale image containing black arrows distributed within a circular region around a central red fixation dot on a white background. Arrow positions, size, and spacing were held constant across stimuli; only the proportion of arrows pointing rightward (coherence) varied. Ten coherence levels were used ($p_{\text{right}} \in \{0.10, 0.20, 0.30, 0.40, 0.45, 0.55, 0.60, 0.70, 0.80, 0.90\}$), with p_{right} indicating the fraction of arrows oriented to the right and $1 - p_{\text{right}}$ the fraction oriented to the left. This design parallels the logic of random-dot motion tasks, in which motion coherence parametrically controls evidence strength.

Stimulus generation

Stimuli were generated programmatically in Python using the PIL library. Arrow locations were sampled within a circular region (annulus excluding the central fixation point) with a minimum inter-arrow distance constraint to prevent overlap and to maintain comparable crowding across coherence levels. Once a single set of spatial positions was established, it was reused across all stimuli so that visual layout and density were identical; only arrow direction assignments varied according to the specified coherence value. Arrow geometry (length, thickness, and head size) was fixed across all stimuli.

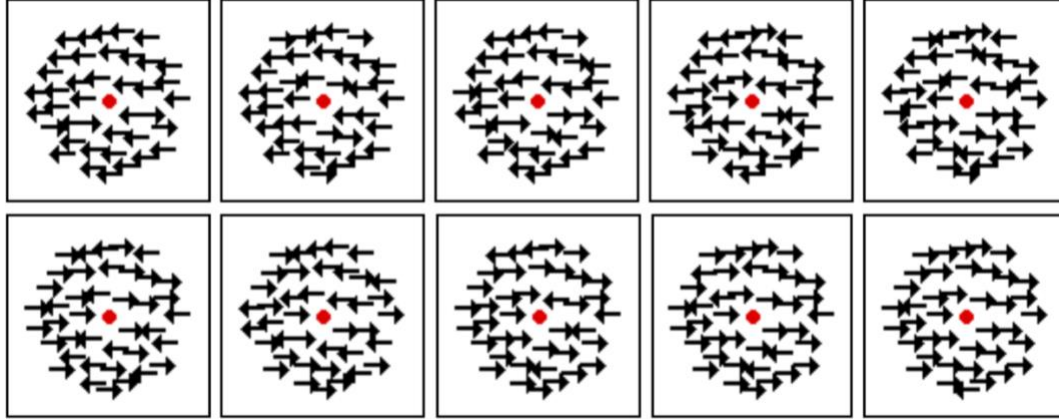


Figure 1. Example stimulus set used in the spatial coherence task. Each image consists of a circular array of black arrows arranged around a central red fixation point on a white background. Task difficulty was manipulated by varying the proportion of arrows pointing left versus right (coherence) while keeping spatial layout, density, and arrow geometry constant. The ten panels illustrate the full range of coherence levels used in the experiment, from strong left-majority to strong right-majority conditions. On each trial, the model was asked to judge the overall majority direction and report a confidence rating.

Experimental procedure

Ten unique stimuli (one per coherence level) were presented to a vision-capable large language model (gpt-4.1-mini) via an API interface. Each stimulus was presented repeatedly across trials. On each trial, one image was selected at random and submitted independently. A total of 100 trials were conducted, with stimuli sampled with replacement to approximate uniform exposure across coherence levels. For each trial, the model received a standardized instruction: to determine whether the majority of arrows pointed left or right and to report both a binary choice and a confidence rating. Responses were constrained to include (i) a direction judgment (“L” or “R”) and (ii) a confidence rating on a four-point scale (1 = not confident, 4 = very confident). Each trial was treated as an independent observation.

```
instruction = (
    "You will be shown a small stimulus image with many black arrows and a red dot in the center.\n"
    "Task: Decide whether the MAJORITY of arrows point LEFT or RIGHT.\n"
    "Respond in JSON ONLY with exactly these keys:\n"
    '  {"direction": "L" or "R", "confidence": 1-4}\n'
    "Confidence scale: 1=not confident, 4=very confident.\n"
    "Do not add any other keys or text."
)
```

Figure 2. Instruction prompt used for all trials. The model was presented with a standardized text instruction alongside each stimulus image, directing it to determine whether the majority of arrows pointed left or right and to report its decision together with a confidence rating on a four-point scale (1 = not confident, 4 = very confident). The structured response format was enforced to ensure consistent extraction of directional choices and confidence estimates across trials.

Ground truth labeling

For each stimulus, the correct response was defined by the majority arrow direction. Specifically, stimuli with $p_{\text{right}} > 0.5$ were labeled “right-majority” and those with p_{right}

< 0.5 were labeled “left-majority.” The p_right parameter served as the task’s evidence strength analogue to motion coherence in perceptual decision-making paradigms.

Outcome measures

Three primary measures were computed:

1. Accuracy: proportion of trials on which the model’s directional judgment matched the ground truth majority direction, calculated separately for each coherence level.
2. Choice proportion: probability of choosing “right” as a function of p_right , used to construct a psychometric curve.
3. Confidence: mean self-reported confidence rating at each coherence level.

Psychometric analysis

To characterize decision behavior, we modeled the probability of a “right” response as a function of coherence using a logistic function. Trial-level binary choices (right = 1, left = 0) were regressed on p_right , yielding a sigmoid psychometric curve. This analysis enabled visualization of response bias (horizontal shift of the curve), sensitivity to evidence (slope), and decision threshold (coherence value at which $P(\text{right})=0.5$). Confidence ratings were analyzed descriptively by averaging within coherence bins.

Results

Choice behavior as a function of coherence

Model responses showed a systematic dependence on the proportion of rightward-pointing arrows. The probability of choosing “right” increased as p_right increased, yielding a monotonic psychometric relationship. However, this function was markedly shifted relative to the unbiased expectation. In an ideal observer without directional bias, the point of subjective equality (PSE)—where the probability of choosing right equals 0.5—would be expected near $p_right = 0.5$. Instead, the fitted logistic curve indicated a substantial rightward shift: the model rarely chose “right” at intermediate coherence levels and only transitioned toward predominantly rightward responses at high coherence values (≈ 0.8 – 0.9). Across low to moderate coherence conditions ($p_right \leq 0.7$), the model almost uniformly responded “left,” even when rightward arrows constituted the majority. Only at the highest coherence levels (0.8 and 0.9) did responses reliably shift to “right.” This pattern is consistent with a strong baseline bias toward leftward judgments and a high decision threshold for switching to rightward responses.

Accuracy across coherence levels

Across all trials, overall classification accuracy was 73.0%. Performance was high when the correct answer was “left” (low p_right), as the model’s default tendency aligned with the ground truth. However, accuracy dropped substantially in intermediate right-majority conditions, where the model continued to respond “left” despite a rightward majority. Accuracy then increased again at the highest coherence levels, where the directional evidence was sufficiently strong to overcome the apparent bias. This produced an asymmetric performance profile across coherence, rather than the symmetric improvement around the midpoint typically observed in perceptual decision tasks.

Psychometric slope and sensitivity

The logistic fit to trial-level choices showed a steep slope but displaced midpoint. This indicates that once the model began responding “right,” the transition was rapid, suggesting

sensitivity to strong evidence. However, the delayed transition point implies reduced sensitivity to moderate evidence. In effect, the model behaved as though it required a large majority of rightward arrows before altering its categorical judgment.

Confidence ratings

Mean reported confidence was high (3.93/4; 98.25% of the maximum possible rating), indicating consistently elevated certainty despite variability in task difficulty and performance. As such, confidence did not vary systematically with coherence and did not decrease in conditions where accuracy was low. In particular, confidence remained elevated even in intermediate coherence ranges where the model frequently produced incorrect responses. This dissociation between accuracy and confidence suggests limited calibration between perceived certainty and objective evidence strength.

Summary of behavioral pattern

Taken together, the results indicate three key features of performance. First, directional judgments were strongly biased toward “left,” as reflected by a rightward shift in the psychometric function. Second, the model showed a high apparent decision threshold, requiring strong evidence before switching to rightward responses. Third, confidence ratings remained uniformly high and showed little relationship to task difficulty or performance. These findings suggest that responses were influenced by a stable directional bias and were not tightly coupled to graded evidence strength across the full coherence range.

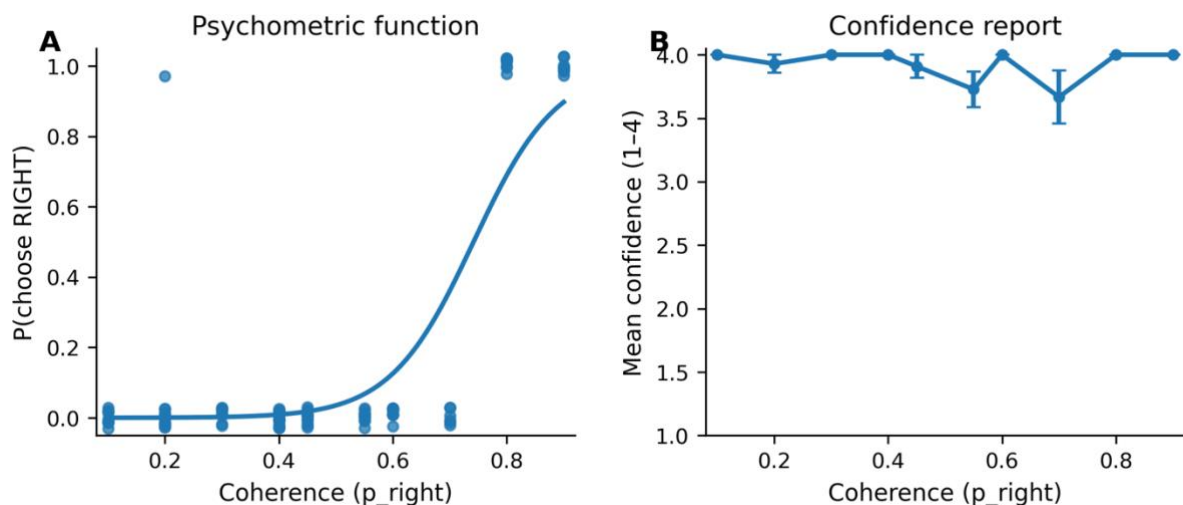


Figure 3. Evidence-dependent choice behavior and confidence reports.

(A) Psychometric function relating stimulus coherence (p_{right} ; proportion of right-pointing arrows) to the probability of a “right” response. Points show individual trials (binary choice; jittered vertically for visibility). The solid curve is a logistic regression fit to trial-level choices, summarizing the coherence–choice relationship and revealing a strong response bias (rightward shift of the transition region). (B) Mean reported confidence (1–4) as a function of coherence. Points show means across trials at each coherence level; error bars indicate the standard error of the mean. Confidence remains uniformly high across coherence levels, showing weak coupling to evidence strength.

Discussion

The present study adapted a classic perceptual decision-making framework to probe the relationship between evidence, choice, and confidence in a vision-capable generative model. By parametrically varying directional coherence while holding all other stimulus properties constant, we obtained a psychometric characterization of the model's behavior. Two central findings emerged: a pronounced and stable directional bias in choice behavior, and a near absence of modulation in reported confidence as a function of evidence strength.

From a metacognitive perspective, the most striking result is the dissociation between performance and confidence. In human observers, confidence typically scales with evidence strength and tracks accuracy, particularly in perceptual discrimination tasks. Both experimental and theoretical work suggest that confidence reflects a second-order estimate of decision reliability derived from the same internal evidence used to make the primary choice. When evidence is weak or ambiguous, confidence is correspondingly reduced; when evidence is strong, confidence increases. This coupling between first-order performance and second-order evaluation is a hallmark of metacognitive monitoring.

In contrast, the model tested here displayed high confidence across nearly all coherence levels, including conditions in which it was systematically incorrect. Confidence ratings remained clustered near the upper bound of the scale and showed little sensitivity to difficulty or to the graded structure of the stimulus evidence. This pattern suggests that the reported confidence signal is not grounded in an internal estimate of uncertainty in the same way as in biological perceptual systems. Instead, confidence appears to function more as a stable output tendency than as a calibrated readout of evidence quality.

These observations are relevant to ongoing debates about whether current generative systems exhibit anything analogous to reflective or evaluative cognition. In perceptual decision paradigms, metacognitive sensitivity is typically taken as an indicator that an agent not only produces responses but also monitors the reliability of those responses. The absence of systematic coupling between evidence strength, accuracy, and confidence in the present results suggests that the model does not engage in such second-order monitoring in a way that resembles human metacognitive processing. Rather than dynamically evaluating the strength of perceptual evidence, the model appears to apply a relatively fixed response heuristic combined with consistently high self-reported certainty. While this does not resolve broader philosophical questions about machine cognition, it provides empirical support for the view that current generative models may not “reflect” on their own outputs in a robust, evidence-sensitive manner.

A second notable finding was the strong and consistent directional bias, reflected in a rightward shift of the psychometric function. The model showed a marked tendency to default to one response category and required disproportionately strong evidence to switch to the alternative. This pattern bears a superficial resemblance to lateralized biases observed in certain neurological conditions. For example, patients with hemispatial neglect following stroke often exhibit systematic spatial asymmetries in tasks such as line bisection and cancellation, with stable directional biases that persist even when stimuli are balanced. While the mechanisms underlying neglect involve focal neural damage and disrupted spatial attention networks, the phenomenological similarity lies in the presence of a consistent response bias that can override graded sensory evidence. However, this analogy should be interpreted cautiously. The observed asymmetry in the present study may instead reflect task-

specific factors, including prompt and response-format effects. In particular, the fixed presentation of response options (e.g., “L” preceding “R”) may have introduced a positional or token-level bias that influenced model behavior independently of perceptual evidence. Nevertheless, even if the bias stems from the prompt, the present paradigm demonstrates that controlled perturbations of evidence and careful measurement of bias can reveal stable directional tendencies in artificial systems. Such task frameworks could potentially serve as a bridge for modeling aspects of pathological decision behavior in silico. For example, one could systematically induce or measure biases in artificial agents and examine how they interact with graded evidence, offering a comparative platform for studying decision asymmetries and their consequences.

Limitations and Future Directions

Several limitations should be considered when interpreting the present findings.

First, the study examined a single vision-capable generative model in a constrained perceptual task. While the paradigm was designed to isolate evidence sensitivity and confidence calibration, conclusions cannot be generalized across architectures, training regimes, or model families. Systematic comparisons across multiple models will be necessary to determine whether the observed response bias and weak coupling between evidence and confidence reflect general properties of current generative systems or are specific to the model tested here.

Second, the experimental design relied on a limited set of static stimuli, with a small number of unique images repeated across trials. Although this allowed precise control over spatial layout and coherence, repetition may introduce image-specific confounds and potential memorization effects. Additionally, the static nature of the stimuli does not capture the temporal dynamics present in classical perceptual decision-making paradigms such as motion coherence tasks. Future work should extend the methodology to video-based stimuli that more closely approximate continuous evidence accumulation and allow investigation of temporal integration processes.

Third, the task probed a narrow slice of spatial reasoning using a simplified directional discrimination paradigm. While this reductionist approach enabled psychometric characterization, spatial cognition in both biological and artificial systems is multidimensional. Additional tasks involving depth perception, occlusion, spatial transformations, or navigation-like reasoning may provide a more comprehensive picture of how models construct and use spatial representations.

Fourth, aspects of prompt design and response formatting were not systematically controlled. Fixed option ordering and consistent label placement may introduce token-level or positional biases that influence response selection independently of visual processing. Similarly, inference parameters such as temperature, sampling strategy, and context-state management were not formally manipulated or documented, and these factors may affect both decision behavior and confidence reporting. Future studies should incorporate controlled prompt randomization, session resets, and systematic parameter reporting to improve interpretability and reproducibility.

Fifth, confidence estimates were elicited without calibration scaffolding or incentive-compatible mechanisms. This may encourage default or maximal confidence responses and

limits the strength of conclusions regarding calibration or metacognitive-like behavior. More rigorous elicitation procedures and the inclusion of standard calibration metrics—such as reliability diagrams, Brier scores, and type-2 sensitivity measures—would strengthen future analyses.

Sixth, stimulus specification and perceptual constraints may also have influenced performance. Certain low-level properties of the stimuli, including exact geometry and resolution, may affect how visual encoders process directional information. The relatively low image resolution used in this study may approach the lower bound for reliable visual interpretation, and perceptual limitations cannot be fully ruled out as contributors to observed biases.

Seventh, the study did not include human baselines, classical algorithmic baselines, or cross-model comparisons. The absence of such controls limits interpretation of response biases and confidence patterns, as it remains unclear whether these reflect general properties of visual decision-making systems, artifacts of a specific architecture, or features of the task design. Future work should incorporate comparative benchmarks to better contextualize performance and calibration.

Eighth, the number of trials and stimulus conditions, while sufficient to demonstrate proof of concept, limits the precision of parameter estimation for psychometric curves and confidence calibration. Larger-scale experiments with more repetitions and finer sampling near decision boundaries would allow more robust modeling of sensitivity, bias, and the relationship between evidence and confidence.

A further practical limitation is that this work was conducted as independent, exploratory research with constrained computational and financial resources. Expanding the study to include multiple large-scale models, dynamic video-based tasks, and higher trial counts would require sustained access to API-based systems and associated funding. Future research will therefore focus on scaling the experimental framework across a broader range of models and task domains as resources permit.

Despite these constraints, the present study establishes a reproducible methodology for probing confidence behavior and spatial decision-making in artificial systems. Extending the paradigm to dynamic stimuli, expanding stimulus diversity, and incorporating comparative baselines represent important next steps. Such developments may help clarify whether the observed response biases and confidence patterns persist under more realistic conditions and whether artificial systems can form temporally grounded spatial representations consistent with classical models of perceptual decision-making.

Future work will extend this approach in several directions. First, testing multiple models will allow assessment of whether the observed bias and weak confidence calibration are general properties of current generative architectures or idiosyncratic to a specific system. Second, alternative perceptual tasks—such as motion coherence analogues, noise masking paradigms, or spatial attention manipulations—could help determine whether similar patterns emerge across modalities and stimulus classes. Third, more refined analyses of confidence scaling, including parametric manipulation around the decision boundary, may clarify the extent to which confidence signals in these systems can be shaped or improved.

Conclusion

This study introduced a psychophysics-inspired paradigm to examine evidence sensitivity, response bias, and confidence behavior in a vision-capable generative model. Using a controlled directional discrimination task, we observed two central patterns: a stable lateralized response bias and consistently high confidence that was weakly modulated by changes in sensory evidence. Together, these findings suggest a dissociation between graded perceptual information and expressed certainty, raising important questions about how such systems internally represent uncertainty and make categorical decisions.

While the observed directional asymmetry and near-ceiling confidence were robust within the present setup, the study was intentionally exploratory and conducted under constrained conditions. As such, the findings should be interpreted cautiously, particularly given the possibility that prompt structure, response formatting, sampling settings, or limited stimulus diversity contributed to the observed patterns. Nonetheless, the results highlight the value of applying psychophysical methods to systematically probe decision behavior and confidence calibration in artificial systems.

More broadly, the present work establishes a methodological approach for probing metacognitive-like signals in AI systems by directly manipulating evidence strength and observing both first-order performance and second-order confidence. By borrowing tools from psychophysics—coherence manipulations, psychometric curve fitting, and confidence reporting—it becomes possible to treat model outputs as behavioral data and analyze them using the same frameworks applied to human and animal cognition. In this sense, the study provides a proof of concept for experimentally “perturbing” confidence in artificial agents and quantifying how tightly it tracks task-relevant information.

Taken together, the present study introduces a simple, reproducible framework for quantifying choice bias, sensitivity to evidence, and confidence behavior in generative models. The results suggest that, at least in this setting, confidence reports are poorly calibrated to task difficulty and that decision behavior can be dominated by stable priors. These findings support the value of psychophysical methods as a tool for interrogating the internal characteristics of AI systems and provide an initial step toward a systematic science of metacognitive evaluation in artificial agents.

References

Relevant theoretical and empirical foundations for this work are well established in the literature on psychophysics, perceptual decision-making, metacognition, spatial reasoning, and systems neuroscience. Given the methodological focus of the present manuscript, readers are directed to the extensive bodies of work in these domains for background on coherence-based paradigms, psychometric modeling, confidence estimation, and spatial cognition.

Correspondence

For code requests or collaboration inquiries, please contact the author via email.