Anatomy of a Lie: A Multi-Stage Diagnostic Framework for Tracing Hallucinations in Vision-Language Models

Lexiang Xiong[1†], Qi Li[1†], Jingwen Ye[2], and Xinchao Wang[1⋆]

[1] National University of Singapore, Singapore
[2] Monash University, Australia
{lexiang, liqi}@u.nus.edu, jingwen.ye@monash.edu, xinchao@nus.edu.sg

Abstract. Vision-Language Models (VLMs) frequently 'hallucinate'–generate plausible yet factually incorrect statements—posing a critical barrier to their trustworthy deployment. In this work, we propose a new paradigm for diagnosing hallucinations, recasting them from static output errors into an auditable process-level anomaly. Our framework is grounded in a normative principle of computational rationality, allowing us to model a VLM's generation as a dynamic *cognitive trajectory*. We design a suite of information-theoretic probes that project this trajectory onto an interpretable, low-dimensional **Cognitive State Space**. Our central discovery is a governing principle we term the **geometric-information duality**: a cognitive trajectory's geometric abnormality within this space is fundamentally equivalent to its high information-theoretic surprisal. Hallucination detection is thus elegantly re-framed as a geometric anomaly detection problem. Evaluated across diverse settings—from rigorous binary QA (POPE) and comprehensive reasoning (MME) to unconstrained open-ended captioning (MS-COCO)—our framework achieves state-of-the-art performance. Crucially, it operates with high efficiency under weak supervision and remains highly robust even when calibration data is heavily contaminated. This approach enables a causal attribution of failures, mapping observable errors to distinct pathological states: **perceptual instability** (measured by Perceptual Entropy, H_{Evi}), **logical-causal failure** (measured by Inferential Conflict, S_{Conf}), and **decisional ambiguity** (measured by Decision Entropy, H_{Ans}). Ultimately, this opens a path toward building AI systems whose reasoning is transparent, auditable, and diagnosable by design.

1 Introduction

Consider a striking paradox in Vision-Language Models (VLMs): when asked "Is there a motorcycle in the image?", a model might confidently hallucinate evidence–"In the image, there is a motorcycle parked"–yet inexplicably conclude, "Therefore, the final answer is No." We term this cascade of errors *computational*

⋆ Corresponding author (xinchao@nus.edu.sg).
† Equal contribution.

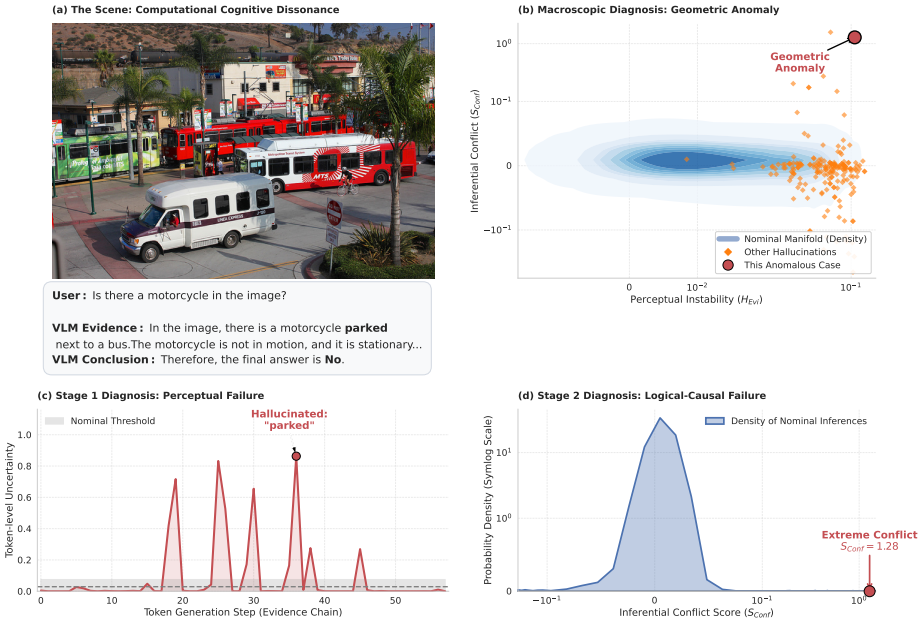

Fig. 1: An example of *computational cognitive dissonance* in Idefics2, where a cascade of failures leads to a coincidentally correct answer. **(1) Perceptual Failure:** The model hallucinates a 'motorcycle' in the evidence chain, an object not present in the image (a cyclist is visible). Our framework captures this as high **Perceptual Instability** (see panel (c)). **(2) Logical Failure:** The model then contradicts its own faulty evidence, concluding the final answer is 'No'. This breakdown of self-consistency is diagnosed as extremely high **Inferential Conflict** (see panel (d)). This case study demonstrates the limitation of accuracy-only evaluations and highlights our framework's ability to perform a stage-by-stage differential diagnosis of a VLM's cognitive process, identifying complex, multi-stage failure trajectories.

cognitive dissonance, as illustrated in Figure 1. What makes this specific case so dramatic is that a double failure (perceiving a non-existent object, then logically contradicting that very perception) leads to a *coincidentally correct* final answer.

This phenomenon exposes a critical insight that severely limits VLM deployment in high-stakes domains [2, 13]: hallucinations are rarely monolithic errors that can be diagnosed by a single metric like "accuracy" or "self-consistency." Instead, they are often complex, **multi-stage anomalies** where distinct failures–such as perceptual drift and logical bypass–compound and interact within a single cognitive trajectory. Current approaches to hallucination detection generally treat the generation process as an indivisible, monolithic event. They either evaluate the semantic consistency of final outputs via multiple sampling [8, 19] or probe for a binary 'truthfulness' representation within internal states [1, 5]. While foundational, these reductionist views conflate fundamentally different failure modes. They struggle to distinguish whether a hallucination stems from

an initial failure to ground concepts in the image (*perceptual drift*) or from an illogical jump that bypasses extracted facts (*inferential bypass*). Our central thesis is that hallucination is a process-level failure that must be diagnosed within a structured model of cognition.

To address this, we introduce a normative abstraction of computational rationality [11, 21] for VLMs, formalized as a Markovian information flow: Image ($\mathcal{I}$) $\xrightarrow{\text{Perception}}$ Textual Evidence ($\mathcal{T}_{\text{evi}}$) $\xrightarrow{\text{Inference}}$ Final Answer ($\mathcal{A}$). This principle asserts that for a rational, explainable agent, the final answer $\mathcal{A}$ should be conditionally independent of the image $\mathcal{I}$ given the explicitly stated evidence $\mathcal{T}_{\text{evi}}$, implying the conditional mutual information $I(\mathcal{A}; \mathcal{I}|\mathcal{T}_{\text{evi}})$ must be zero. Critics might argue that requiring an explicit evidence chain limits the applicability of such a framework, or that it may not perfectly reconstruct the model's true hidden reasoning. However, we employ Chain-of-Thought (CoT) not to faithfully capture latent cognition, but as an **observable diagnostic trace**–akin to a medical contrast agent. By prompting the model to externalize an evidence chain, we make the generation trajectory observable and mathematically auditable. The Supplementary Material further clarifies the scope of this diagnostic assumption.

To diagnose this cognitive process, we design a suite of probes. While S_{Conf} directly measures violations of our core principle, Perceptual Entropy (H_{Evi}) and Decision Entropy (H_{Ans}) quantify the stability of the process's initial and final stages, providing a complete diagnostic picture. These probes act as natural coordinates to project the high-dimensional trajectory onto an interpretable 3D **Cognitive State Space**:

- **Perceptual Instability (H_{Evi}):** Measured via Perceptual Entropy, this probes the uncertainty at the perception stage ($\mathcal{I} \to \mathcal{T}_{\text{evi}}$).
- **Logical-Causal Failure (S_{Conf}):** Measured via Inferential Conflict, this directly quantifies the information leakage violating our normative abstraction, indicating the answer is not fully mediated by the stated evidence.
- **Decisional Ambiguity (H_{Ans}):** Measured via Decision Entropy, this probes the final uncertainty at the trajectory's terminal stage.

Collectively, these probes summarize each cognitive trajectory as a Cognitive State Vector within this space.

This perspective reveals a powerful **geometric-information duality**: a trajectory's geometric abnormality within this space is fundamentally an expression of its high information-theoretic surprisal. Our experimental results (see Fig. 3) provide strong empirical evidence for this duality. Normative trajectories consistently cluster in stable, high-density regions, forming a dense submanifold. Hallucinations, conversely, are high-surprisal deviations that are perturbed off this manifold, appearing as geometric anomalies. This duality serves as the practical bridge that translates the semantic problem of hallucination into a rigorous geometric anomaly detection task [28].

This novel reframing achieves state-of-the-art detection performance while offering significant practical advantages. Unlike multi-sample methods, our approach requires only a single generation pass plus a highly efficient non-autoregressive

replay (detailed in Sec. 3). Furthermore, it operates under weak supervision (requiring only ground-truth answers, not fine-grained hallucination labels) and remains highly resilient even when calibration data is contaminated with up to 30% noise. To demonstrate its versatility, we rigorously validate our framework across a spectrum of tasks: from the controlled adversarial subset of POPE [16], to the comprehensive reasoning categories of MME [10], and finally to unconstrained open-ended captioning on MS-COCO [17].

In summary, our primary contribution is not merely a new state-of-the-art hallucination detector, but a principled diagnostic framework that provides a new lens through which to understand VLM failures. Our contributions are:

- We propose a diagnostic framework that reframes hallucination from a static flaw to a dynamic analysis of a VLM's cognitive trajectory, grounded in a normative principle of computational rationality.
- We design a suite of information-theoretic probes that act as natural coordinates to project the generative process onto an interpretable **Cognitive State Space**, enabling a stage-by-stage differential diagnosis.
- Grounded in a powerful geometric-information duality, we introduce a novel detection method based on geometric anomaly detection, which operates under weak supervision and achieves state-of-the-art performance.
- We deliver a novel mechanistic categorization of VLM failure modes by analyzing the topology of their cognitive manifolds, diagnosing complex errors like 'computational cognitive dissonance'.

2 Related Work

Our research is positioned at the confluence of VLM hallucination evaluation, mitigation, and internal state analysis [2, 13]. With these foundations, we introduce a novel *diagnostic* framework that moves beyond static error detection to model the VLM's dynamic internal cognitive process.

VLM Hallucination Benchmarking. A significant body of work has focused on quantifying VLM hallucinations from final outputs. The pioneering CHAIR metric [27] measured hallucinated objects in captions. To address its instability, works like POPE [16] and ROPE [6] established a stable polling-based evaluation paradigm. To capture more complex failure modes, recent benchmarks have expanded to evaluate real-world instruction following (VisIT-Bench [3], MMHal-Bench [29], HallusionBench [12]), fine-grained visual-text alignment (LM2-Bench [24], WYSWIR [35]), advanced commonsense reasoning (Visual Riddles [37]), and spatial reasoning, where recent work shows that VLMs can underuse geometric evidence and fall back on appearance shortcuts [39]. Adversarial benchmarks such as MAD-Bench [26] further probe robustness against deceptive prompts. While these static evaluation suites are invaluable for assessing *what* errors a model makes across diverse scenarios, they predominantly treat the VLM as a black box. Our work provides a crucial complement: diagnosing the dynamic generative process itself to explain *how* and *why* these errors occur.

Inference-Time Hallucination Detection and Mitigation. Recent efforts have focused on inference-time strategies. One prominent line of work is contrastive decoding, which penalizes outputs driven primarily by language priors rather than visual evidence [32,33]. State-of-the-art methods like Hallucination-Induced Optimization (HIO) [4] refine this by training a dedicated 'evil' model to provide a targeted contrastive signal. Parallel efforts in automated evaluation (auto-eval) employ strong LLMs (e.g., Clair [31]) or contrastive grounding techniques (e.g., Contrastive Region Guidance [34]) to assess output quality without manual labels. Another direction analyzes internal signals, such as VADE [25], which models attention map sequences. While these methods excel at scoring or *correcting* the final output, our framework is distinctly focused on *diagnosing the mechanistic failure*. Our Inferential Conflict metric (Sec. 3) directly isolates the illicit vision-language information flow, offering a causal interpretation that auto-evals typically lack.

Internal State Analysis for Hallucination. A nascent line of inquiry explores VLM internal states, inspired by seminal research in LLMs suggesting truthfulness is encoded in hidden activations [1,5,9,20]. While methods like VADE [25] analyze internal patterns, they focus on attention mechanisms. Other approaches, often adapted from the text-only domain, may treat the VLM's internal state as a monolithic representation [7,22]. This simplification is ill-equipped to distinguish between a failure in initial perception versus a breakdown in subsequent reasoning. **Distinct from all prior work**, our research introduces a multi-faceted diagnostic framework that models a VLM's reasoning not as a static state, but as a **measurable cognitive trajectory through distinct, macroscopic stages**. This process-oriented view enables a mechanistic, differential diagnosis of where a breakdown originates.

Practical Advantages of Our Framework. Beyond its theoretical grounding, our framework offers significant practical advantages. It operates under weak supervision–requiring only ground-truth final answers rather than expensive, often ambiguous token-level annotations required by fully supervised detectors. Once calibrated, our method is highly efficient, requiring only the initial generation and a single non-autoregressive forward pass through the language decoder. This makes it significantly faster than multi-sample consistency methods [8,19] and highly scalable for real-world deployment.

3 Methodology: An Information-Geometric Framework for Diagnosing Hallucination

We reconceptualize VLM hallucination not as a simple output error, but as a symptom of a breakdown within the model's internal information processing. We first formalize the ideal, logically self-consistent cognitive process through an axiomatic probabilistic graphical model (PGM) that defines the normative flow of information: $\mathcal{I} \xrightarrow{\text{Perception}} \mathcal{T}_{\text{evi}} \xrightarrow{\text{Inference}} \mathcal{A}$, where $\mathcal{I}$ is the visual input,

$\mathcal{T}_{\text{evi}}$ is the explicitly generated evidence chain, and $\mathcal{A}$ is the final answer to a given query $\mathcal{Q}$.

This model embodies a critical abstraction from information theory: the generated evidence $\mathcal{T}_{\text{evi}}$ serves as a **sufficient statistic** for the final answer $\mathcal{A}$ with respect to the image $\mathcal{I}$. Mathematically, this defines the ground truth of an auditable process as one where the conditional mutual information is zero: $I(\mathcal{A}; \mathcal{I} | \mathcal{T}_{\text{evi}}) = 0$. Deviations from this axiom signify a consistency failure. Crucially, we do not claim that $\mathcal{T}_{\text{evi}}$ faithfully reconstructs the model's inscrutable latent reasoning; rather, it serves as an observable diagnostic probe. While our conceptual framework applies to general generation, we anchor our mathematical formalization and primary diagnosis in structured Visual Question Answering (VQA) tasks, as their constrained action spaces allow for rigorous quantification of these latent failure modes.

An Information-Geometric View of Cognition. We define a VLM's generation of a token trajectory τ as a probabilistic event drawn from a distribution $P(\tau|\mathcal{I}, Q)$. The informational content of any specific trajectory is its **self-information**, or **surprisal**, defined as $I(\tau) = -\log P(\tau|\mathcal{I}, Q)$. This allows for a rigorous, first-principles definition of hallucination: a *nominal* cognitive process is a low-surprisal event, corresponding to a high-probability trajectory that aligns with the model's learned world model. Conversely, we define hallucination as a **high-surprisal cognitive event**–a rare, low-probability trajectory that deviates unexpectedly from this nominal behavior.

To diagnose such events, we project the high-dimensional internal state of the generation process into a low-dimensional, 3D Observable Information Manifold. Each generation is represented by a Cognitive State Vector $v = [H_{\text{Evi}}, S_{\text{Conf}}, H_{\text{Ans}}]$ on this manifold. We posit that nominal processes correspond to points residing in high-density regions, or 'attractors' on this manifold. Our three diagnostic probes are thus reinterpreted as direct, quantitative measures of the information flow's properties at different stages:

- **Perceptual Entropy (H_{Evi})** measures the *initial state's information entropy*, quantifying the uncertainty in the evidence formulation stage.
- **Inferential Conflict (S_{Conf})** directly measures *information leakage* across generation stages, quantifying the violation of our core auditable axiom (evidence-consistency).
- **Decision Entropy (H_{Ans})** quantifies the *terminal state's residual entropy*, measuring the final decision uncertainty.

3.1 Probing Perceptual Uncertainty (H_{Evi})

To quantify the *initial state's information entropy*, we measure the uncertainty of the evidence formulation stage ($\mathcal{I} \to \mathcal{T}_{\text{evi}}$) with Perceptual Entropy (H_{Evi}).[3]

[3] The complete word lists, adapted from prior work on language model uncertainty [14,38], along with a sensitivity analysis, are provided in the Supplementary Material.

A high H_{Evi} signifies an unstable starting point for the cognitive trajectory. We model the model's choice at each token step i as a Bernoulli trial $C_i \in \{\text{Factual}, \text{Uncertain}\}$ by defining two disjoint token subsets: a factual set V_f and an uncertainty set V_u. For the logits $\mathbf{l}_i$ of each token, we project the softmax distribution onto this semantic axis:

$$p(C_i = \text{F}) = \frac{\sum_{t \in V_f} \text{softmax}(\mathbf{l}_i)_t}{\sum_{t \in V_f \cup V_u} \text{softmax}(\mathbf{l}_i)_t}, \quad p(C_i = \text{U}) = 1 - p(C_i = \text{F}) \tag{1}$$

The token-level entropy is the Shannon entropy $H_i = H(C_i)$. The final metric is the path-averaged entropy: $H_{\text{Evi}} = \frac{1}{|\mathcal{T}_{\text{evi}}|} \sum_{i=1}^{|\mathcal{T}_{\text{evi}}|} H_i$.

3.2 Probing Inferential Conflict (S_{Conf})

To operationalize our idealized causal graph ($\mathcal{I} \to \mathcal{T}_{\text{evi}} \to \mathcal{A}$), we introduce **Inferential Conflict (S_{Conf})**. This probe estimates the *Conditional Pointwise Mutual Information* (CPMI) to quantify the strength of the illicit direct causal path from $\mathcal{I}$ to $\mathcal{A}$. It is a pointwise metric for a specific outcome $\mathcal{A}_{\text{token}}$, making it highly suitable for diagnosing single, concrete instances of generation. It measures the information gain from the visual modality on the generated answer token $\mathcal{A}_{\text{token}}$[4], conditioned on the textual evidence $\mathcal{T}_{\text{evi}}$. This quantity is computed as the **log-probability difference**:

$$S_{\text{Conf}} = \log p_v(\mathcal{A}_{\text{token}} | \mathcal{I}, \mathcal{T}_{\text{evi}}) - \log p_t(\mathcal{A}_{\text{token}} | \emptyset_{\mathcal{I}}, \mathcal{T}_{\text{evi}}) \tag{2}$$

$$= \text{CPMI}(\mathcal{A}_{\text{token}}; \mathcal{I} \mid \mathcal{T}_{\text{evi}}) \tag{3}$$

where p_v is the probability with visual context and p_t is the counterfactual probability without it. A large positive S_{Conf} indicates strong, positive point-wise information flowing directly from the visual input to the final answer, unmediated by the evidence, thus measuring the violation of d-separation. To obtain p_t, we perform a causal intervention by replaying the generation process with the visual input ablated[5] [23]. A practical boundary condition is that the VLM architecture must allow for such a causal intervention. The Supplementary Material distinguishes this evidence-consistency use of CPMI from standard grounding scores and discusses black-box proxies.

3.3 Probing Decision Uncertainty (H_{Ans})

Finally, to quantify the *terminal state's residual entropy*, we measure the final uncertainty with Decision Entropy (H_{Ans}). A high entropy indicates the system

[4] For multi-token answers, $\mathcal{A}_{\text{token}}$ is defined as the first token corresponding to the primary decision keyword (e.g., 'Yes' or 'No'). This localization is made reliable by the structured prompts used in our experimental setup.

[5] In our implementation with Idefics2, this is achieved by providing 'images=None' to the processor during the text-only forward pass.

has failed to converge to a stable, determined state.

$$H_{\text{Ans}} = -\sum_{a \in \{\text{Yes,No}\}} p(a) \log_2 p(a) \tag{4}$$

3.4 Diagnosis via Geometric Anomaly Detection in the Cognitive State Space

Our framework performs diagnosis at inference-time on single instances after a one-time, hallucination-label-free calibration. This phase learns the geometric structure of the 'nominal cognitive state space' $\mathcal{S}_{\text{nominal}}$.

Phase 1: Learning the Geometry of the Nominal Cognitive State Space. We represent each VLM generation by its 3D Cognitive State Vector $v = [H_{\text{Evi}}, S_{\text{Conf}}, H_{\text{Ans}}]$. This calibration is fitted on a calibration set $\mathcal{D}_{\text{cal}}$. This process requires only ground-truth final answers (e.g., 'Yes'/'No'), a form of **weak supervision** that is vastly more accessible and scalable than obtaining fine-grained, token-level hallucination labels. We hypothesize that the landscape of nominal states is multi-modal, as different types of valid cognitive processes (e.g., simple object recognition versus complex relational reasoning) may form distinct, dense clusters in the state space. We therefore employ a Gaussian Mixture Model (GMM), which is naturally suited to capturing such underlying structures, to model the probability density $p(v|\text{nominal})$. This set undergoes a Coherence Filter step: we first select for correct final answers and then apply automated heuristics to exclude 'lucky guesses' (e.g., cases where the model answers 'Yes' while its generated evidence explicitly states 'There is no such object in the image'). This purification ensures our GMM learns a less biased estimate of the true density on $\mathcal{S}_{\text{nominal}}$. Prior to fitting, we standardize each dimension and determine the optimal number of GMM components via the Bayesian Information Criterion (BIC).

Phase 2: Hallucination Diagnosis as a High-Surprisal Cognitive Event. From an information geometry perspective, a hallucination is a cognitive process whose state vector v is geometrically distant from the learned high-density regions (attractors). Its Hallucination Score is therefore the self-information content, or surprisal, of observing this atypical state vector:

$$S_{\text{hall}}(v) = I(v) = -\log p(v|\mathcal{M}_{\text{GMM}}) \tag{5}$$

This score quantifies the 'unexpectedness' of the observed cognitive trajectory. Nominal processes are common, predictable, low-information events, whereas hallucinations are rare, high-information deviations. The workflow is summarized in Algorithm 1.

Algorithm 1 Cognitive Anomaly Detection Framework (CAD)

Function DIAGNOSEHALLUCINATION $(I, Q, \mathcal{M}, \text{CalibratedComponents})$
$(\mathcal{M}_{\text{GMM}}, \mu, \sigma) \leftarrow \text{CalibratedComponents}$

▷ 1. Generate cognitive trajectory and extract metrics
$(\mathcal{T}_{\text{evi}}, \mathcal{A}_{\text{model}}, \text{scores}) \leftarrow \mathcal{M}.\text{generate}(I, Q, \text{output\_scores=True})$
$H_{\text{Evi}} \leftarrow \text{CalcPerceptualEntropy}(\text{scores}_{\text{evi}})$
$S_{\text{Conf}} \leftarrow \text{CalcInferentialConflict}(I, Q, \mathcal{T}_{\text{evi}}, \mathcal{M})$
$H_{\text{Ans}} \leftarrow \text{CalcDecisionEntropy}(\text{scores}_{\text{ans}})$

▷ 2. Compute anomaly score in the cognitive space
$v_{\text{new}} \leftarrow [H_{\text{Evi}}, S_{\text{Conf}}, H_{\text{Ans}}]$
Standardize v_{new} using μ, σ.
$S_{\text{hall}} \leftarrow -\log p(v_{\text{new}} | \mathcal{M}_{\text{GMM}})$
return S_{hall}

// Note: CalibratedComponents are pre-computed offline by fitting a GMM on cognitive state vectors from a purified set of non-hallucinatory samples.

4 Experiments

4.1 Experimental Setup

Datasets and Multi-Dimensional Evaluation Protocol. Our work's core philosophy is that object-level hallucination is merely the final, observable symptom of a broader cognitive failure. To thoroughly evaluate our framework across diverse settings and address the limitations of narrow benchmark testing, we design a multi-dimensional evaluation protocol:

- **Diagnostic Deep-Dive (POPE [16]):** We transform the POPE benchmark into a rich diagnostic playground using Chain-of-Thought (CoT) prompts to externalize reasoning ($\mathcal{T}_{\text{evi}}$). Following [16], we strictly focus on the 'adversarial' subset to diagnose genuine, hard-to-detect hallucinations, which serves as our primary testbed for mechanistic analysis.
- **Comprehensive Generalization (MME [10]):** To ensure our method generalizes beyond specific task formats, we evaluate on the expansive MME benchmark, reporting macro-averaged results across its diverse perceptual and reasoning categories.
- **Open-Ended Validation (MS-COCO [17]):** As a targeted ablation, we validate our perceptual probe on open-ended image captioning, using the CHAIR [27] metric to confirm its independent generalizability.

Evaluated Models and Baselines. To ensure a comprehensive analysis across architectural families, we evaluate four state-of-the-art VLMs: Llava-v1.6-Mistral-7B [18], Idefics2-8b [15], Qwen2-VL [30], and DeepSeek-VL2-Small [36]. We compare our proposed **Cognitive Anomaly Detection (CAD)** against a comprehensive suite of baselines: 1) **Token Entropy** and **Neg Log Probability**

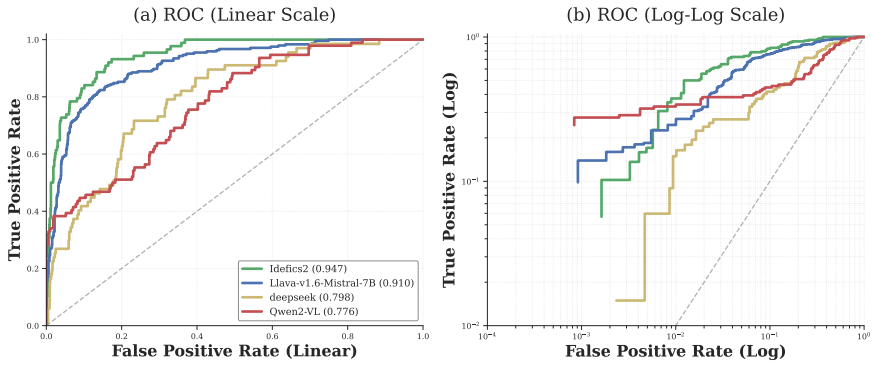

Fig. 2: ROC curves of our Cognitive Anomaly Detection (CAD) framework. (a) Linear-scale ROC curves show superior overall performance across all architectures. (b) log-log curves highlight CAD's dominance in the critical low-FPR regime (FPR $<$ 10^{-2}), which is essential for reliable real-world deployment.

(token-level uncertainty); 2) a **Supervised Probe** [5] (a linear classifier trained on final hidden states using balanced hallucination labels); and 3) **Semantic Entropy** [8] (a strong, sampling-based consistency baseline requiring 10× inference cost). Detailed metric calculations, GMM hyperparameter optimization (via BIC), and prompt templates are provided in the Supplementary Material. Additional prompt, calibration, GMM-component, lexicon-free, and open-ended CAD variants are also reported there.

4.2 State-of-the-Art Detection Across Diverse Benchmarks

We find that reframing hallucination as a diagnosable *cognitive process anomaly* leads to a state-of-the-art framework that is both effective and efficient.

Performance on POPE. As shown in Tab. 1 and Fig. 2, our CAD framework– which models the distribution of only non-hallucinatory examples– achieves superior overall performance. Critically, the log-log ROC curves (Fig. 2b) highlight that CAD maintains high true positive rates even at extremely low false positive rates (FPR $< 10^{-2}$), a regime where baseline methods often fail. This confirms that modeling the geometric properties of the cognitive process offers superior reliability for real-world deployment compared to simple uncertainty thresholds.

Generalization on MME. Moving beyond simple object existence questions, Tab. 2 demonstrates CAD's strong generalization capabilities on the comprehensive MME benchmark. MME encompasses diverse multimodal tasks including spatial reasoning, OCR, and commonsense logic. In these complex scenarios, the natural variance in generated text increases significantly. While the Supervised Probe, benefiting from in-domain training labels, shows competitive performance, our weakly-supervised CAD framework achieves comparable or even

Table 1: Main AUC results on the POPE benchmark (Adversarial). Best performance is in **bold**, second best is underlined. Our single-pass, weakly supervised CAD significantly outperforms all baselines.

Method	Cost	Llava-v1.6	Idefics2	Qwen2-VL	DeepSeek-VL	Average
Token Entropy	1x	0.603	0.806	0.409	0.732	0.638
Neg Log Prob	1x	0.604	0.832	0.428	0.710	0.644
Supervised Probe	1x	0.787	0.898	0.762	0.715	0.791
Semantic Entropy	10x	0.711	0.751	0.673	0.702	0.709
Ours (CAD)	**1x**	**0.910**	**0.947**	**0.776**	**0.798**	**0.858**

Table 2: Macro-averaged AUC results on the MME benchmark. Our CAD framework demonstrates strong generalization across diverse multimodal reasoning and perceptual tasks.

Method	Llava-v1.6	Idefics2	Qwen2-VL	DeepSeek-VL	Average
Token Entropy	0.5106	0.5486	0.5291	0.6687	0.5643
Neg Log Prob	0.5061	0.5449	0.5136	0.6578	0.5556
Supervised Probe	0.7620	0.7905	**0.7408**	0.7104	0.7509
Semantic Entropy	0.6811	0.7081	0.6616	0.6199	0.6677
Ours (CAD)	**0.8514**	**0.8411**	0.7233	**0.7680**	**0.7960**

superior results (e.g., 0.851 on Llava-v1.6) without requiring any hallucination labels for calibration. This highlights CAD's robustness and adaptability across a wide array of diverse task types.

4.3 Mechanistic Diagnosis: Unveiling Cognitive Fingerprints

The success of our CAD framework over the supervised 'Supervised Probe' is a crucial finding, suggesting that analyzing the entire cognitive trajectory for *anomalous patterns* provides a richer signal than a localized biopsy of a single token's state. While the SOTA results are compelling, a deeper question arises: *what underlying mechanistic differences cause performance to vary so dramatically across models?* Our framework's primary strength is its diagnostic capability. By projecting the generative process into the 3D Cognitive State Space, we can visualize the 'cognitive manifolds' of nominal and hallucinatory behavior. As Fig. 3 reveals, different VLM architectures follow strikingly different pathological pathways.

Idefics2's 'Structural Disorder' Pattern. Idefics2 presents a stark signature. Its nominal processes form an extremely compact, low-variance manifold–a tight blue cluster representing a rigid, stable cognitive workflow. Hallucinations are characterized as anomalous deviations from this stable state. This suggests a **structural failure** mechanism, explaining why our density-based anomaly detector is exceptionally effective for this model.

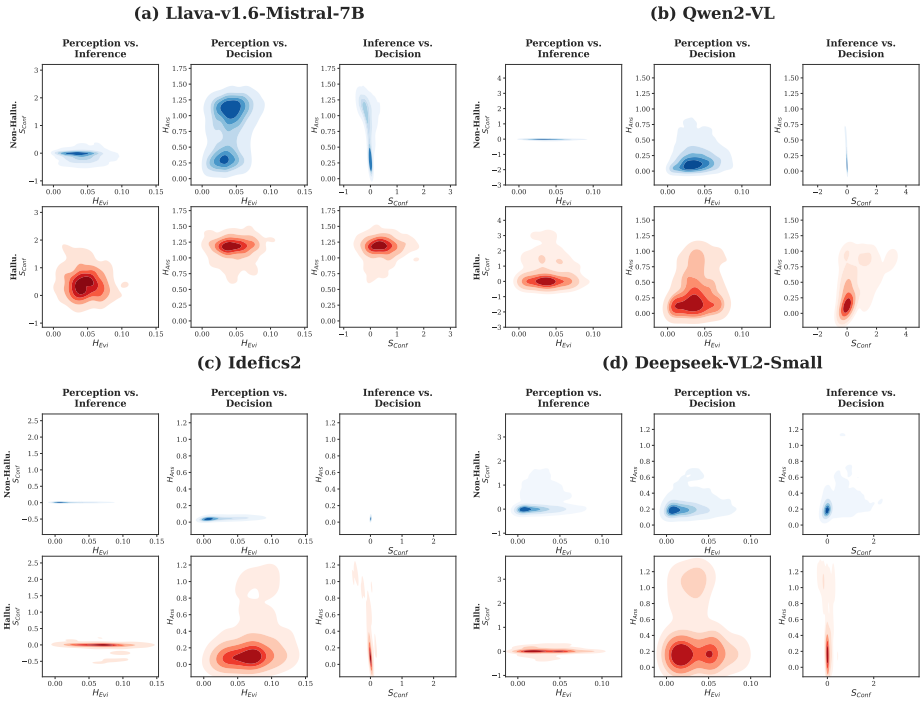

Fig. 3: Visualizing the 'Cognitive Fingerprints' of Hallucination. Density projections of the 3D Cognitive State Space, separated into non-hallucinatory (blue, top row of each pair) and hallucinatory (red, bottom row of each pair) processes. These manifolds reveal unique failure signatures for each model.

Llava's 'Transparent Struggle' Pattern. Llava-v1.6 exhibits a cognitively transparent pattern. Its hallucinatory manifold (red) occupies a region largely separable from the nominal one (blue), characterized by high Inferential Conflict and Decisional Ambiguity. Its internal cognitive struggle is explicitly manifested through our metrics, making its failure mode highly transparent.

Qwen2-VL & DeepSeek's 'Entangled States' Pattern. These models exhibit the most insidious pattern: 'confident lies.' The 'Entangled States' pattern provides a direct, geometric explanation for their lower AUC scores. For these models, the hallucinatory manifold (red) forms its own dense, confident clusters that deeply intertwine with the 'healthy' manifold (blue). This is not a failure of our method, but a profound diagnostic finding. It reveals that for certain architectures, hallucination is not merely a process anomaly but can be a content error originating from a seemingly normal process, highlighting a key challenge for future research.

4.4 Ablation Study and Real-World Robustness

To validate our framework's multi-component design and its practical viability, we conducted comprehensive ablation studies and stress-testing.

The Necessity of a Holistic Diagnosis. The significant 'Synergy Gain' revealed in Fig. 4(a) is a testament to the multi-dimensional nature of cognitive failure. For a model like Idefics2, no single probe provides a sufficient signal (individual AUCs ∼0.75-0.78). Only by viewing the cognitive trajectory as a point in our 3D state space can its 'anomalous deviations' be reliably detected, as evidenced by the jump to 0.947 AUC, quantitatively proving that a holistic diagnosis is necessary.

The Adaptive Diagnostician. Figure 4(b) reveals the unique 'diagnostic fingerprint' of each model. Our GMM-based detector acts as an adaptive diagnostician: for Llava, it learns to be highly sensitive to deviations along the *Inferential Conflict* axis; for DeepSeek-VL, it identifies *Perceptual Instability* as the key symptom. The GMM's ability to learn where to focus–and what to ignore (e.g., the confounding noise from H_{Evi} for Idefics2)–demonstrates its power to adapt to each model's unique cognitive fingerprint.

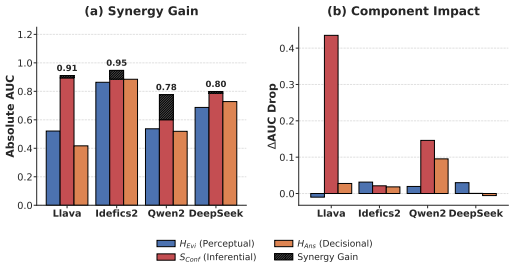

Fig. 4: Ablation Study. (a) Standalone metrics vs. synergistic gain. The gray hatched area represents the *Synergy Gain*. (b) Impact of component removal (ΔAUC), revealing model-specific 'diagnostic fingerprints'.

Generalization of Individual Probes:
H_{Evi} on Open-Ended Tasks. While our holistic framework is powerful, it is crucial to validate the independent efficacy and generalizability of its constituent probes. We conducted a targeted ablation on our **Perceptual Instability (H_{Evi})** probe by evaluating it on the challenging open-ended task of MS-COCO image captioning ($N = 1000$). Since unconstrained captioning lacks a binary decision boundary (e.g., Yes/No) to anchor the Inferential Conflict (S_{Conf}) and Decision Entropy (H_{Ans}) metrics, this setting allows us to isolate perceptual drift as a primary driver of hallucination in free-form generation. As shown in Fig. 5, H_{Evi} scores are significantly higher for hallucinated captions (validated via CHAIR [27]) across all models, with profound statistical significance ($p \ll 0.001$). This result not only demonstrates the standalone power of our perceptual probe but also proves its ability to generalize far beyond structured VQA tasks.

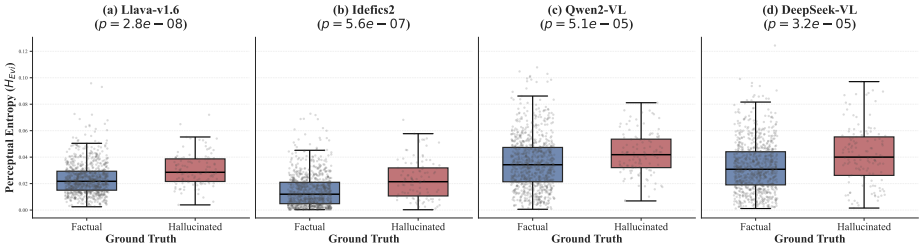

Fig. 5: Generalization of Perceptual Instability ($H_{\mathbf{Evi}}$) to Open-Ended Captioning (MS-COCO, $N = 1000$). Without any task-specific tuning, our perceptual probe consistently assigns significantly higher entropy to hallucinatory captions (red) compared to factual ones (blue) across all four architectures. **Statistical Significance:** The distinction is profound, with Welch's t-test yielding $p \ll 0.001$ in all cases, validating that H_{Evi} captures a fundamental cognitive signature of hallucination beyond VQA formats. **Cross-Model Insight:** The shared Y-axis highlights that models like Qwen2-VL and DeepSeek-VL2 (bottom row) exhibit higher baseline entropy in their factual generations compared to Llava and Idefics2.

Robustness to Calibration Contamination. CAD's reliance on weak supervision–calibrating solely on samples with correct final answers–is a significant practical advantage, as perfectly clean data is often unavailable in real-world deployments. To evaluate its limits, we stress-test the framework by intentionally contaminating the calibration set with 0%–30% undetected hallucinations. As illustrated in Fig. 6, the resilience of our GMM-based detector aligns elegantly with the underlying 'cognitive fingerprints' of each architecture. Idefics2 remains remarkably robust, maintaining an AUC above 0.91 even under 30% contamination. This confirms its *Structural Disorder* profile: the nominal cognitive core is so geometrically compact that the GMM effectively treats injected hallucinations as negligible outliers, leaving the learned density of the manifold intact.

In contrast, Llava-v1.6's noise-sensitivity reflects its *Transparent Struggle*: including high-variance hallucinations forces GMM variance dilation, blurring thresholds. As its hallucinations are characterized by conspicuously high-variance and extreme inferential conflict, mistakenly including them as "nominal" forces the GMM to artificially stretch its variance parameters to encompass these anomalies. Conversely, Qwen2-VL and DeepSeek-VL2 remain stable as noise merely reinforces their inherent *Entangled States* overlap. Such resilience under severe contamination underscores CAD's utility for robust real-world auditing.

5 Conclusion

In this paper, we introduce a diagnostic framework that reframes VLM hallucination from a static output error to a dynamic failure within the model's cognitive process. By modeling generation as a three-stage cognitive trajectory and projecting it into an interpretable state space, we develop a state-of-the-art, single-pass, and weakly supervised anomaly detector. Crucially, our robust

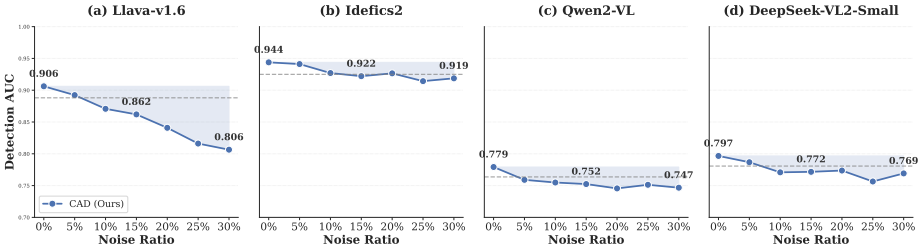

Fig. 6: Robustness to Calibration Contamination across Architectures. We evaluate detection performance (AUC) as the calibration set is increasingly contaminated with hallucinatory samples (0%–30%). **(a) Sensitivity of Transparent Struggle:** Llava-v1.6 exhibits a noticeable drop, as mistakenly treating its highly conflicted, high-variance hallucinatory states as nominal forces the GMM to blur the anomaly boundary. **(b) Resilience of Structural Disorder:** Idefics2 maintains high performance (> 0.91) with negligible degradation, as its compact nominal core remains robust against geometrically distinct outliers. **(c, d) Stability of Entangled States:** Qwen2-VL and DeepSeek-VL2 show flatter trajectories; their performance is constrained by the intrinsic geometric overlap of their states rather than calibration noise. The gray dashed line shows a strict 2% performance drop threshold relative to the clean baseline.

framework provides the first mechanistic categorization of VLM hallucinations, uncovering distinct 'cognitive fingerprints' for different architectures–from transparent failures to deeply entangled errors. This approach advances beyond mere detection, enabling the stage-by-stage diagnosis of complex, multi-modal failure modes, ultimately providing a transparent new lens for auditing and building reliable Vision-Language Models.

Acknowledgements

This project is supported by the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme and CyberSG R&D Cyber Research Programme Office (Award: CRPO-GC1-NTU-002).

References

1. Azaria, A., Mitchell, T.M.: The internal state of an LLM knows when its lying. In: Conference on Empirical Methods in Natural Language Processing (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.68
2. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930 (2024)
3. Bitton, Y., Bansal, H., et al.: VisIT-Bench: A benchmark for vision-language instruction following inspired by real-world use. In: Neural Information Processing Systems Track on Datasets and Benchmarks (2023)

4. Chen, B., Lyu, X., Gao, L., Song, J., Shen, H.: Alleviating hallucinations in large vision-language models through hallucination-induced optimization. In: Neural Information Processing Systems (2024)
5. Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z., Ye, J.: INSIDE: LLMs' internal states retain the power of hallucination detection. In: International Conference on Learning Representations (2024)
6. Chen, X., Ma, Z., Zhang, X., Xu, S., Qian, S., Yang, J., Fouhey, D.F., Chai, J.: Multi-object hallucination in vision-language models. In: Neural Information Processing Systems (2024)
7. Du, X., Xiao, C., Li, Y.: HaloScope: Harnessing unlabeled LLM generations for hallucination detection. In: Neural Information Processing Systems (2024)
8. Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting hallucinations in large language models using semantic entropy. Nature **630**, 625–630 (2024). https://doi.org/10.1038/s41586-024-07421-0
9. Ferrando, J., Obeso, O., Rajamanoharan, S., Nanda, N.: Do I know this entity? knowledge awareness and hallucinations in language models. In: International Conference on Learning Representations (2024)
10. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
11. Gershman, S.J., Horvitz, E.J., Tenenbaum, J.B.: Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science **349**(6245), 273–278 (2015)
12. Guan, T., Liu, F., et al.: HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
13. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (2023)
14. Ji, Z., Yu, L., Koishekenov, Y., Bang, Y., Hartshorn, A., Schelten, A., Zhang, C., Fung, P., Cancedda, N.: Calibrating verbal uncertainty as a linear feature to reduce hallucinations. arXiv preprint arXiv:2503.14477 (2025)
15. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., Cord, M., Sanh, V.: Obelics: An open web-scale filtered dataset of interleaved image-text documents. arXiv preprint arXiv:2306.16527 (2023)
16. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., rong Wen, J.: Evaluating object hallucination in large vision-language models. In: Conference on Empirical Methods in Natural Language Processing (2023)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. pp. 740–755. Springer (2014)
18. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-1.6: Improved reasoning, ocr, and world knowledge. arXiv preprint arXiv:2406.07945 (2024)
19. Manakul, P., Liusie, A., Gales, M.: SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In: Conference on Empirical Methods in Natural Language Processing (2023)
20. Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., Belinkov, Y.: LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In: International Conference on Learning Representations (2024)

21. Oulasvirta, A., Jokinen, J.P., Howes, A.: Computational rationality as a theory of interaction. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2022)
22. Park, S., Du, X., Yeh, M.H., Wang, H., Li, Y.: Steer LLM latents for hallucination detection. arXiv preprint arXiv:2503.01917 (2025)
23. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (2009)
24. Peyrard, M., et al.: LM2-Bench: A closer look at how well vlms implicitly link explicit matching visual cues. In: European Conference on Computer Vision (2024)
25. Prabhakaran, V., Aggarwal, P., Verma, V.K., Swamy, G., Saladi, A.: VADE: Visual attention guided hallucination detection and elimination. In: Annual Meeting of the Association for Computational Linguistics (2025)
26. Qian, Y., Zhang, H., Yang, Y., Gan, Z.: How easy is it to fool your multimodal LLMs? an empirical analysis on deceptive prompts. arXiv preprint arXiv:2402.13220 (2024)
27. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the 2018 conference on empirical methods in natural language processing. pp. 4035–4045 (2018)
28. Stolz, B.J., Tanner, J., Harrington, H.A., Nanda, V.: Geometric anomaly detection in data. Proceedings of the national academy of sciences **117**(33), 19664–19669 (2020)
29. Sun, Z., Shen, S., et al.: Aligning large multimodal models with factually augmented rlhf (mmhal-bench). arXiv preprint arXiv:2309.14525 (2023)
30. Team, Q.: Qwen2-vl technical report. `https://qwenlm.github.io/blog/qwen2-vl/` (2024), accessed: 30 June 2026
31. Tsun, C., et al.: Clair: Evaluating image captions with large language models. In: Conference on Empirical Methods in Natural Language Processing (2024)
32. Vu, M., Tran, B.K., Shah, S.A., Zollicoffer, G., Hoang-Xuan, N., Bhattarai, M.: HalluField: Detecting LLM hallucinations via field-theoretic modeling. arXiv preprint arXiv:2509.10753 (2025)
33. Wang, C., Fu, W., Zhou, Y.: TPC: Cross-temporal prediction connection for vision-language model hallucination reduction. arXiv preprint arXiv:2503.04457 (2025)
34. Wang, D., et al.: Contrastive region guidance: Improving grounding in vision-language models without training. In: European Conference on Computer Vision (2024)
35. Wang, J., et al.: What you see is what you read? improving text-image alignment evaluation. In: Neural Information Processing Systems (2023)
36. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
37. Yarom, M., et al.: Visual riddles: a commonsense and world knowledge challenge for large vision and language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
38. Yona, G., Aharoni, R., Geva, M.: Can large language models faithfully express their intrinsic uncertainty in words? In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 7752–7764 (2024)
39. Zhang, S., Shen, Q., Wang, S., Pan, T., Wang, X.: Make geometry matter for spatial reasoning. arXiv preprint arXiv:2603.26639 (2026)