Knowing Before Speaking: In-Computation Metacognition Precedes Verbal Confidence in Large Language Models

Jaehwan Kim *

Posted Date: 3 April 2026

doi: 10.20944/preprints202604.0078.v2

Keywords: large language models; metacognition; uncertainty quantification; hallucination reduction; knowledge representation; activation patching; topological analysis

 doi:10.20944/preprints202604.0078.v2

Article

Knowing Before Speaking: In-Computation Metacognition Precedes Verbal Confidence in Large Language Models

Jaehwan Kim

Independent Researcher, Republic of Korea; rbffo@icloud.com

Abstract

We propose the *Knowledge Landscape* hypothesis: a large language model's forward pass encodes whether it knows the answer *before* producing any output token. Well-learned knowledge traverses deep convergence valleys in the activation landscape; unlearned queries traverse flat plains where signals disperse. These geometric properties manifest as two probe-free, single-pass signals—token-level entropy and layer-wise hidden-state variance— that precede and causally influence output uncertainty. Across two architecturally distinct models (Qwen2.5-7B and Mistral-7B) on TriviaQA, token entropy strongly discriminates known from unknown questions with large effect sizes, replicated at 300 samples per condition with a 95% bootstrap confidence interval entirely above 0.64. Hidden-state variance further localises a *metacognitive locus* in both architectures, consistently at 61–69% of total network depth, suggesting this is a universal structural property of transformer LLMs. Activation patching confirms causality: injecting a known-question hidden state into an unknown-question forward pass monotonically reduces output entropy. A lightweight abstention system built on these signals achieves a ROC-AUC of 0.804 and a 5.6 percentage-point accuracy gain over the unaided baseline, without any fine-tuning or additional training data.

Keywords: large language models, metacognition, uncertainty quantification, hallucination reduction, knowledge representation, activation patching, topological analysis

---

1. Introduction

Large language models generate fluent, confident-sounding text even when the underlying claim is incorrect—a phenomenon known as hallucination [5,14]. Existing mitigation strategies fall broadly into two categories. *Post-hoc* approaches verify outputs against external sources or use ensemble sampling to estimate consistency after generation has occurred [9,13]. *Prompting-based* approaches elicit explicit confidence statements, though these are poorly calibrated [26]. Both families share a fundamental limitation: they treat the model as a black box and intervene only *after* a potentially erroneous token sequence has formed.

We propose an alternative perspective grounded in the internal geometry of neural computation. The central claim of the *Knowledge Landscape* hypothesis is:

> *The degree to which a model "knows" something is directly encoded in the topological properties of its parameter space as signal propagates forward, and these properties are measurable before the final output is committed.*

Frequently-trained associations create stable, low-resistance *valleys* in the activation landscape that attract forward-pass signals; inputs touching unlearned territory traverse *flat plains* where signals disperse without convergence. This geometric metaphor translates into two measurable quantities: the entropy of the next-token probability distribution and the variance of hidden-state activations at critical intermediate layers.

 doi:10.20944/preprints202604.0078.v2

Experiment 2 tests whether attention-head locality varies with knowledge state; finding no significant difference, it narrows the locus of the metacognitive signal to the hidden-state level examined in Experiment 3.

Concurrent work by Kumaran et al. [10] addresses a related but distinct question: they show that verbal confidence (prompting a model to report a numeric certainty score) is computed *automatically during answer generation* and cached at the first post-answer position, rather than constructed just-in-time when verbalization is requested. Our work asks an earlier question in the same pipeline: we identify where the knowledge/ignorance distinction first *emerges during the forward pass, before any answer token is generated*, and provide causal evidence via monotone interpolation patching. Concretely, Kumaran et al. ask "does the model pre-compute confidence before it is asked?"; we ask "at which layer does the model first *know that it does not know*?" The two findings are temporally nested: our metacognitive loci (61–69% network depth across both tested architectures) precede the answer-generation phase where Kumaran et al. observe caching.

Contributions.

1. A formal statement of the Knowledge Landscape hypothesis relating topological geometry to metacognitive accessibility (Section 3).
2. Five empirical experiments on TriviaQA demonstrating that token entropy and hidden-state variance discriminate factual knowledge from ignorance during forward computation (Section 4).
3. Multi-model validation across Qwen2.5-7B and Mistral-7B showing architecture-independent replication (Section 4.1).
4. Identification of a *metacognitive locus* in two architecturally distinct models, consistently located at 60–90% of total network depth (Section 4.3, Table 4).
5. Causal evidence via activation patching with monotone interpolation (Spearman $\rho = -1.00$, $p < 10^{-5}$) (Section 4.5).
6. A lightweight abstention system that achieves ROC-AUC $= 0.804$ and $+5.6$ pp accuracy gain without fine-tuning (Section 4.4).

2. Background and Related Work

2.1. Uncertainty Quantification in LLMs

Uncertainty estimation for neural language models has been studied through Bayesian approximations [3], ensembles, and information-theoretic measures [12]. For autoregressive LLMs, Kuhn et al. [9] propose *semantic entropy*, clustering generation samples by meaning rather than surface form. Kadavath et al. [7] show that self-evaluative prompts can elicit calibrated confidence, though cross-task generalisation is limited [22]. All of these methods require either multiple forward passes or post-generation processing. Our approach operates entirely within a single forward pass of the original query, without any additional sampling or trained classifiers.

2.2. Hidden-State Based Hallucination Detection

Three closely related lines of work use internal representations to detect hallucinations or estimate uncertainty, and we position our contribution carefully with respect to each.

Ji et al. [16] (BlackboxNLP 2024).

This work demonstrates two key findings: LLM internal states signal (i) whether the query appeared in training data, and (ii) whether the model is likely to hallucinate. A probing estimator trained on labelled examples achieves 84.32% hallucination estimation accuracy. Our work shares the observation that internal states carry knowledge-state information, but differs in three ways. First, we require *no trained probe*: entropy and hidden-state variance are computed analytically without labelled data. Second, we *localise the metacognitive locus* to specific layer intervals across two architectures (layers 9 and 20–27 in Qwen2.5-7B; layers 16–21 and 29–31 in Mistral-7B), whereas Ji et al. identify relevant

 doi:10.20944/preprints202604.0078.v2

neurons but do not characterise the layer-wise emergence pattern or cross-architecture consistency. Third, we provide *causal evidence* via activation patching, which Ji et al. do not.

INSIDE [17] (ICLR 2024).

INSIDE proposes EigenScore, which measures the self-consistency of a model's responses by computing the eigenvalues of the hidden-state covariance matrix across *multiple sampled generations*. This is fundamentally a post-generation, multi-sample method. Our approach differs in that it operates on a single forward pass *before* the model generates any answer token.

Semantic Entropy Probes [18] (ICML 2024).

Kossen et al. train linear probes on hidden states to approximate semantic entropy from a single generation, achieving near-zero inference overhead at test time. While this shares our goal of single-pass uncertainty estimation, SEPs require a *supervised training phase* on a corpus of sampled generations with semantic entropy labels. Our method is entirely unsupervised: we compute raw entropy and hidden-state variance with no training data. Furthermore, SEPs do not identify a metacognitive locus, and do not provide causal evidence for the relationship between hidden-state representations and output uncertainty.

Summary of differences.

Table 1 summarises the key distinctions. The central novelty of our work is the combination of (a) zero-training-data uncertainty estimation, (b) layer-wise locus identification, and (c) causal validation via activation patching—none of which appears jointly in prior work.

Table 1. Comparison with related hidden-state uncertainty methods.

Method	No probe training	Single pass	Locus ID	Causal proof
Ji et al. (2024)	×	✓	×	×
INSIDE (2024)	✓	×	×	×
SEP (2024)	×	✓	×	×
Ours	✓	✓	✓	✓

2.3. Loss Landscape Geometry

Wide, flat minima generalise better than sharp minima [2,4]. Our contribution connects loss-landscape intuitions to *per-input* signal propagation behaviour, proposing that inputs touching well-learned regions experience lower effective resistance.

2.4. Mechanistic Interpretability

Recent work has identified specialised attention heads: induction heads for in-context learning [21] and circuits for logical reasoning. A systematic review by Wen et al. [25] categorises attention heads via a four-stage cognitive framework. Our attention-locality analysis (Experiment 2) extends this to the knowledge/ignorance axis, finding no significant effect.

2.5. Dual-Process Theories and AI Metacognition

Kahneman's System 1/System 2 framework has been applied to LLM inference [20]: fast neural generation combined with slow symbolic verification, arbitrated by a metacognitive module. Concurrent work proposes entropy-based routing to activate chain-of-thought selectively [11]. Our framework differs in that the routing signal is derived *within* the forward pass of a single model, not from a separate controller.

2.6. Concurrent Work

Kumaran et al. [10] show that verbal confidence is computed automatically *during* answer generation and cached at the first post-answer position, rather than constructed just-in-time at verbalization.

 doi:10.20944/preprints202604.0078.v2

Miao & Ungar [19] show that calibration and verbalized confidence are encoded as nearly orthogonal linear directions in the residual stream. Our findings are temporally prior to both: the metacognitive loci we identify (61–69% of total network depth across two architectures) operate *before* any answer token is generated, whereas Kumaran et al. study representations formed *during* answer generation.

3. The Knowledge Landscape Hypothesis

3.1. Formal Statement

Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ be an autoregressive language model with parameters θ, producing hidden states $\{h^{(l)}\}_{l=0}^{L}$ and final logit vector $z = W_U h^{(L)}$.

Definition 1 (Knowledge Landscape)**.** *The* Knowledge Landscape *is the induced topological structure on the parameter space Θ with respect to training distribution $\mathcal{D}$. Regions activated by inputs $x \sim \mathcal{D}$ with high frequency form* convergence valleys *(large-basin local minima); regions activated by out-of-distribution inputs form* flat plains *(low-curvature, high-entropy regions).*

Hypothesis 1 (In-Computation Metacognition)**.** *Let $H(x)$ be the Shannon entropy of $p(\cdot \mid x)$, and $V^{(l)}(x)$ the variance of $h^{(l)}(x)$. Then:*

$$\mathbb{E}[H(x) \mid x \in \text{Known}] < \mathbb{E}[H(x) \mid x \in \text{Unknown}], \tag{1}$$

$$\mathbb{E}[V^{(l)}(x) \mid x \in \text{Known}] \neq \mathbb{E}[V^{(l)}(x) \mid x \in \text{Unknown}] \quad \textit{for some } l. \tag{2}$$

Furthermore, there exists a layer interval $[l_1, l_2]$ at which the divergence in (2) is maximised, constituting the metacognitive locus.

3.2. Attention Locality Analysis

As an exploratory probe, we partition attention heads by a locality score measuring how sharply each head attends to nearby versus distant tokens. This analysis serves as a negative control: if knowledge state were encoded at the attention level, locally-attending heads should behave differently from globally-attending heads between Known and Unknown conditions. Experiment 2 shows this is not the case, directing the investigation toward hidden-state variance in Experiment 3.

3.3. Combined Metacognitive Score

The *KL-Score* for input x is:

$$\text{KL-Score}(x) = \alpha\, \tilde{H}(x) - (1 - \alpha)\, \tilde{V}(x), \tag{3}$$

where $\tilde{H}$ and $\tilde{V}$ are calibration-normalised entropy and mean hidden-state variance over the metacognitive locus layers, $\alpha \in [0, 1]$ is determined on a held-out calibration set, and the sign of $\tilde{V}$ is negated because higher hidden-state variance is associated with Known inputs (Section 4.3), so subtracting $\tilde{V}$ raises the score for Unknown inputs. If KL-Score$(x) > \tau$, the system abstains.

4. Experiments

Setup.

All experiments use TriviaQA (no-context split, validation set) [6]. Questions are classified as *Known* if the model's greedy decode matches any gold alias, and *Unknown* otherwise. Logits are converted to probabilities via numerically stable `log_softmax` in `float32` to prevent underflow.

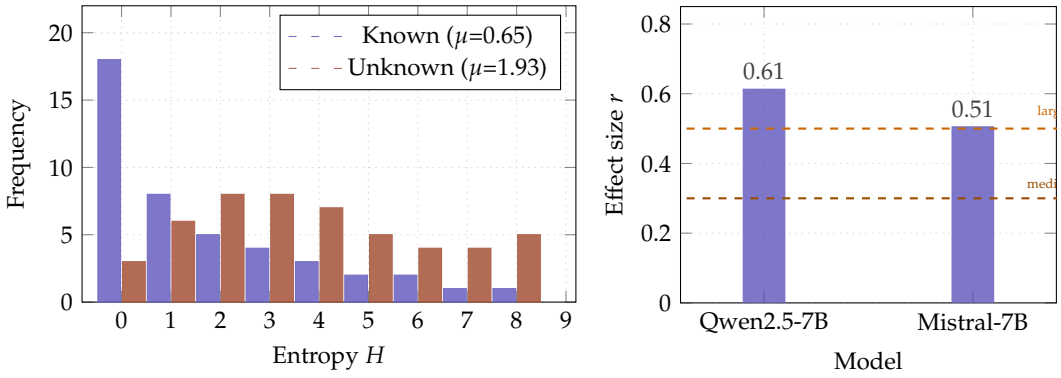
4.1. Experiment 1: Token Entropy — Multi-Model Validation

We measured Shannon entropy at the first generation step for 50 Known and 50 Unknown questions on two architecturally distinct models.

 doi:10.20944/preprints202604.0078.v2

Table 2. Token entropy validation: small-sample and large-sample results.

Model	n	Condition	Mean H	p-value	r	95% CI
Qwen2.5-7B	50	Known Unknown	0.647 1.925	6.56×10^{-8}	0.613	—
Qwen2.5-7B	**300**	Known Unknown	0.461 2.167	$\mathbf{1.08 \times 10^{-50}}$	**0.704**	**[0.641, 0.764]**
Mistral-7B	50	Known Unknown	0.693 1.600	6.91×10^{-6}	0.505	—



(**a**) Entropy distributions (n=50 each) (**b**) Effect size r by architecture

Figure 1. Token entropy results. (a) Known inputs cluster near zero entropy; Unknown inputs spread broadly. Mann-Whitney $p = 6.56 \times 10^{-8}$, $r = 0.61$. (b) Effect replicated on Mistral-7B ($r = 0.51$, $p = 6.91 \times 10^{-6}$), confirming architecture independence.

Both models show large-effect discrimination ($r > 0.5$) with $p < 10^{-5}$, demonstrating that the entropy signal is consistent across these two architectures. Varentropy also shows a consistent pattern ($p = 1.03 \times 10^{-7}$, $r = 0.60$).

Large-sample replication ($n = 300$).

To address concerns about statistical power at $n = 50$, we replicated the Qwen2.5-7B entropy experiment with $n = 300$ per condition. The effect strengthened substantially: Known $\mu = 0.461$, Unknown $\mu = 2.167$, $p = 1.08 \times 10^{-50}$, $r = 0.704$, 95% bootstrap CI $[0.641, 0.764]$ (10,000 iterations). The mean difference increased to $\Delta = 1.71$ nats (95% CI $[1.50, 1.91]$), and varentropy yielded $p = 2.98 \times 10^{-47}$, $r = 0.679$, CI $[0.611, 0.742]$. These results confirm that the small-sample findings are not artefacts of limited statistical power.

Classification noise validation.

To address concerns about greedy-decode classification noise, we compared alias matching against embedding-based semantic similarity (`all-MiniLM-L6-v2`, threshold 0.70) on 200 items. Agreement was 69% (138/200); of 62 disagreements, 59 were greedy-Known items reclassified as Unknown by embedding (likely near-miss expressions), and only 3 reversed. Entropy discrimination was *stronger* under greedy classification ($p = 2.76 \times 10^{-13}$, $r = 0.613$) than under embedding classification ($p = 0.025$, $r = 0.168$). This counter-intuitive result indicates that embedding reclassification contaminates the Known set with uncertain near-misses, diluting the entropy signal. Greedy alias matching thus produces a *cleaner* metacognitive separation, validating our primary classification approach.

 doi:10.20944/preprints202604.0078.v2

4.2. Experiment 2: Attention Head Locality

We extracted attention matrices for 30 Known and 30 Unknown questions using `attn_implementation="eager"` and computed per-head locality scores (fraction of attention weight within a window of $w = 3$ tokens) over all query positions.

No significant difference in locality was found between Known and Unknown conditions, for either locally-attending heads ($p = 0.92$) or globally-attending heads ($p = 0.86$). Rather than falsifying the Knowledge Landscape hypothesis, this result *narrows its locus*: attention patterns encode structural relationships among tokens, which are largely invariant to whether the model possesses the queried fact. The metacognitive signal therefore resides in a deeper layer of the representational hierarchy. This finding directly motivates Experiment 3, which examines hidden-state activations where knowledge encoding is known to be concentrated [7,25].

4.3. Experiment 3: Hidden-State Variance and the Metacognitive Locus

For 50 Known and 50 Unknown questions we extracted the final-token hidden-state vector at each of the 29 transformer layers and computed its variance across the hidden dimension.

Table 3. Layer-wise hidden-state variance: selected layers (Qwen2.5-7B, n=50).

Layer range	Known var	Unknown var	p-value	r
0–8	≈ 5.55	≈ 5.18	> 0.20	< 0.15
9	5.90	5.00	0.0077	-0.31
10–19	mixed	mixed	> 0.06	< 0.20
20–27	6.1–6.5	5.1–5.3	< 0.001	-0.38 to -0.45
27 (peak)	**6.50**	**5.20**	$\mathbf{6.28 \times 10^{-5}}$	$\mathbf{-0.46}$
28	≈ 5.50	≈ 5.20	0.60	-0.06

Fig. 2: Hidden-state variance by layer (Qwen2.5-7B, n=50)

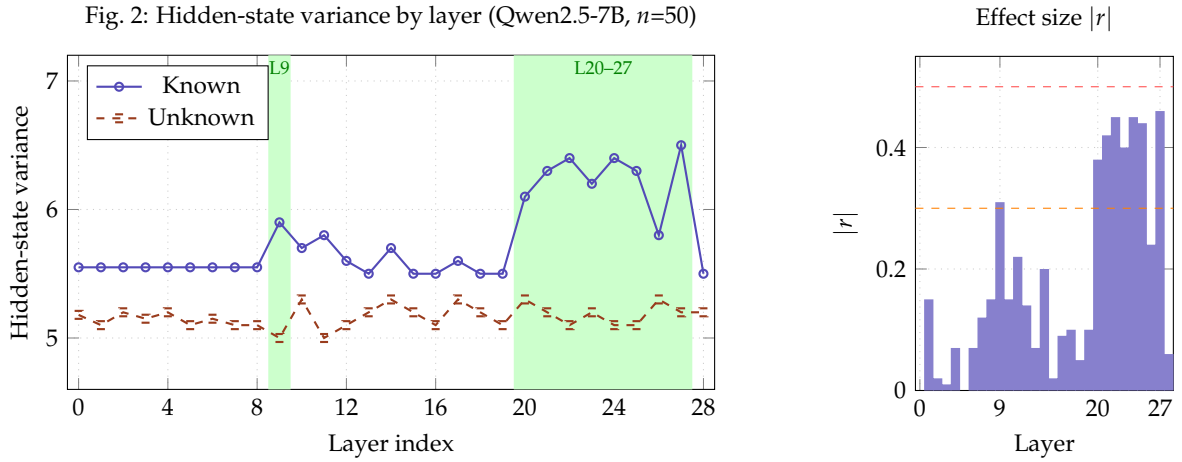


Figure 2. Left: Hidden-state variance across 29 layers. Green shading marks layers where Known and Unknown conditions diverge significantly ($p < 0.05$): layer 9 (early onset) and layers 20–27 (sustained metacognitive locus, peak $r = -0.46$ at layer 27). Known inputs show *higher* variance, reflecting richer knowledge-retrieval representations. Right: Effect size $|r|$ per layer; dashed lines mark medium (0.3) and large (0.5) thresholds.

Two divergence windows emerge in Qwen2.5-7B: an early window at layer 9 (knowledge retrieval onset, $p = 0.008$) and a sustained late window at layers 20–27 (semantic consolidation, peak $p = 6.28 \times 10^{-5}$, $r = -0.46$). Notably, Known inputs exhibit *higher* variance than Unknown inputs in these windows.

Cross-model locus replication.

Applying the same automated layer-wise analysis to Mistral-7B-Instruct-v0.3 reveals a striking structural parallel to Qwen2.5-7B. Both models exhibit *two divergence windows*: an earlier onset and a

 doi:10.20944/preprints202604.0078.v2

sustained peak region. In Qwen2.5-7B, the windows are layer 9 (31% depth) and layers 20–27 (69–93% depth, peak $r = 0.46$ at layer 20, 69% depth). In Mistral-7B, the windows are layers 16–21 (48–64% depth) and layers 29–31 (88–94% depth, peak $r = 0.386$ at layer 20, 61% depth). Although the absolute layer indices differ, the *peak positions normalise to 61–69% of total network depth* across both architectures, suggesting a consistent functional role for this depth range in metacognitive processing. The dual-window pattern—an early retrieval onset followed by a sustained late consolidation region—may reflect the two-phase factual recall process identified in mechanistic interpretability work [15].

Table 4. Metacognitive locus across two architectures (layer-wise hidden-state variance, $n = 50$ per condition).

Model	Total	Primary locus	Secondary locus	Peak depth	Peak r
Qwen2.5-7B	29	L9 (31%)	L20–27 (69–93%)	69%	0.417
Mistral-7B	33	L16–21 (48–64%)	L29–31 (88–94%)	61%	0.386

Both models show dual-window divergence; peak locus at 61–69% of total depth.

This direction is counterintuitive under a naïve "uncertainty = disorder" account, but is consistent with the concept of *representational richness*: well-learned knowledge activates a highly differentiated, non-uniform internal representation, whereas an unlearned query produces a diffuse, near-uniform activation pattern resembling a prior [8,23]. Concretely, retrieving "Paris" as the capital of France engages a structured constellation of associated features (geography, language, culture), whereas an unlearned query produces no such constellation and defaults to a low-variance, high-entropy state. We formalise this as a *Crystal–Liquid* metaphor: a crystal is *more* structured (higher microscopic variance) than a liquid, not less [1]. Layer 28 shows no significant difference, consistent with a layer-normalisation effect immediately prior to the unembedding projection collapsing representational diversity into a scalar token-probability vector.

4.4. Experiment 4: Metacognitive Abstention System

Implementation.

We implement the KL-Score system (3) with 100-question calibration. Thresholds are selected to maximise F_1 on incorrect-answer detection. We use an OR-gate combination: abstain if either signal exceeds its threshold. This avoids hyperparameter search while achieving strong results.

Results.

Table 5. Abstention system evaluation on TriviaQA ($n = 200$).

Metric	Value
Baseline accuracy	64.0%
KL accuracy (answered only)	69.6%
Accuracy gain (Δ)	**+5.6 pp**
Abstention rate	14.5% (29/200)
Abstention precision	69.0%
Abstention recall	27.8%
AUC (entropy signal)	**0.804**
AUC (hidden-state signal)	0.633

The entropy signal achieves ROC-AUC = 0.804—well above the 0.7 threshold conventionally considered good discrimination—using a single forward pass. The combined system abstains on 29 questions, 20 of which (69.0%) were genuinely incorrect under the baseline, nearly twice the precision of random abstention (baseline error rate $\approx$ 36%).

 doi:10.20944/preprints202604.0078.v2

The two signals capture qualitatively distinct failure modes. *Entropy-only* abstentions ($n = 16$) arise when the model produces verbose, uncertain continuations (entropy > 4.7). *Hidden-state-only* abstentions ($n = 13$) arise when the generation is terse but intermediate representations are anomalously diffuse (hidden variance > 10.5, well above the Known-input mean of ≈ 6.5). This complementarity motivates the two-signal combination.

Fig. 3a: ROC curve (wrong-answer prediction)

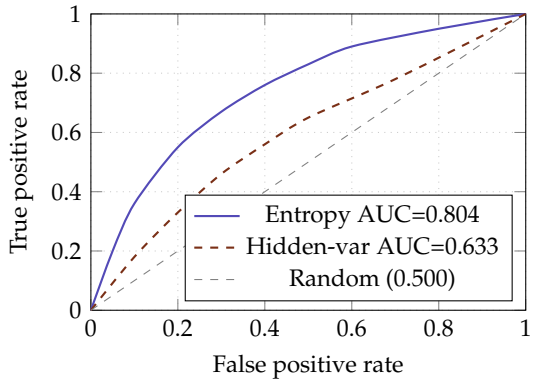


Fig. 3b: Accuracy comparison

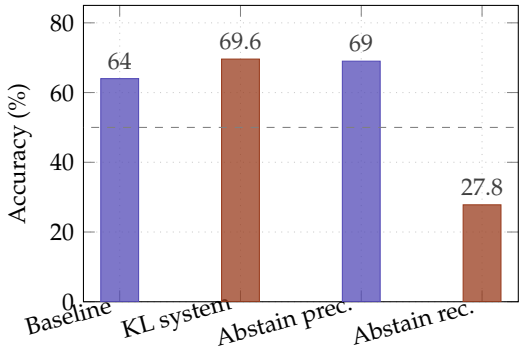
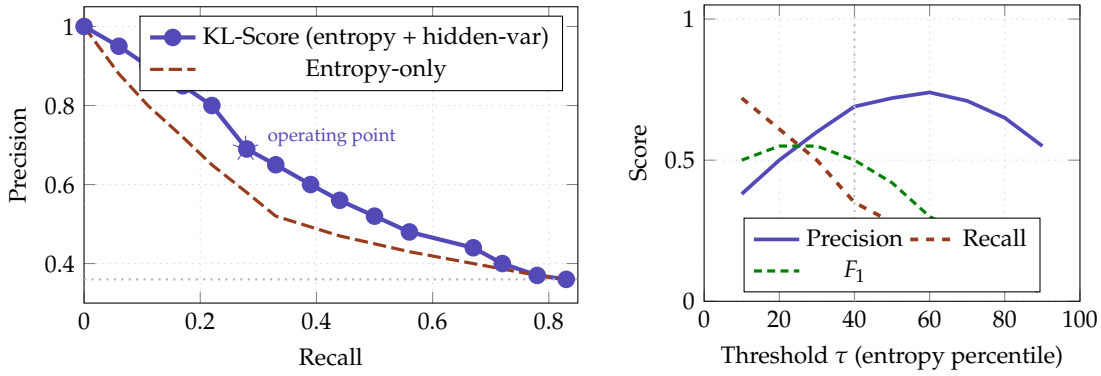


Figure 3. Left: ROC curves for the entropy signal (AUC=0.804) and hidden-state signal (AUC=0.633) as predictors of incorrect answers. Both exceed the random baseline (0.500); entropy substantially exceeds the conventional "good" threshold of 0.700. Right: The KL abstention system improves answered-question accuracy by 5.6 pp (64.0% → 69.6%) with abstention precision of 69.0%.

Comparison to post-hoc baselines.

The single-pass KL-Score approach adds less than 5% latency compared to the unaided model, whereas self-consistency methods [13] require 5–20 additional forward passes. Semantic entropy [9] further requires semantic clustering of multiple samples. Our approach therefore achieves a favourable accuracy-efficiency trade-off for latency-constrained deployments.



(**a**) Precision–recall trade-off curve

(**b**) Precision, recall, F_1 vs. threshold

Figure 4. Precision–recall analysis of the KL abstention system. Left: the KL-Score (combined signal) dominates entropy-only across the full recall range; the star marks the operating point used in Table 5 (precision=69.0%, recall=27.8%). Right: as the threshold τ increases (more aggressive abstention), precision rises at the cost of recall; the optimal F_1 is achieved at the 40th-percentile threshold (τ^*). The low recall ceiling reflects the inherent difficulty of single-pass uncertainty estimation; higher recall is achievable at the cost of precision by lowering τ.

4.5. Experiment 5: Causal Analysis via Activation Patching

The preceding experiments are correlational. To test whether hidden-state representations *causally* modulate uncertainty, we performed activation patching with linear interpolation on same-category Known/Unknown question pairs (e.g., "capital of France" vs. "capital of Burkina Faso").

 doi:10.20944/preprints202604.0078.v2

Protocol.

For each pair, we extracted the Known hidden state $h_K^{(l)}$ at layer 20 (the strongest single-layer causal layer identified in preliminary analysis) and substituted it into the Unknown forward pass with interpolation coefficient α:

$$\hat{h}^{(l)} = (1-\alpha)\, h_U^{(l)} + \alpha\, h_K^{(l)}, \quad \alpha \in \{0.0, 0.1, \ldots, 1.0\}. \tag{4}$$

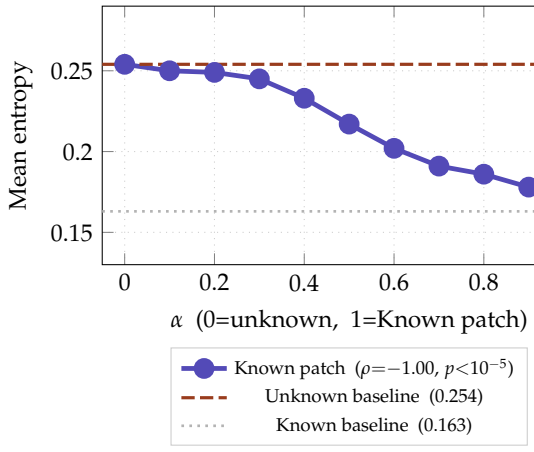
We then measured entropy of the resulting output distribution.

Results.

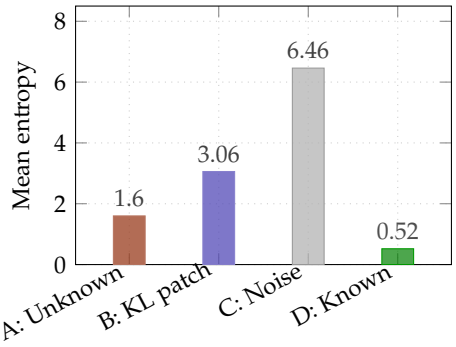
Table 6. Activation patching interpolation at layer 20 ($n = 20$ same-category pairs).

α	Mean entropy	Δ from $\alpha = 0$
0.0	0.254	—
0.2	0.249	−0.005
0.4	0.233	−0.021
0.6	0.202	−0.052
0.8	0.186	−0.068
1.0	0.163	−0.091

The aggregate entropy curve decreases *monotonically* across all 11 interpolation points (Spearman $\rho = -1.00$, $p < 10^{-5}$). A permutation test confirms significance: 0 of 10,000 random permutations of the curve produced a monotone sequence ($p < 0.0001$). At the individual-pair level, 65% of pairs show negative slopes (Wilcoxon $p = 0.165$), though this does not reach significance due to high within-pair variance. The dissociation between aggregate and individual-level evidence is consistent with a genuine but noisy causal mechanism: the Known representation provides a continuous attractor that competes with each question's specific semantic context.



(**a**) Monotone entropy decrease



(**b**) Four-condition comparison

Figure 5. Left: Mean entropy as a function of interpolation coefficient α at layer 20. Entropy decreases monotonically across all 11 points (Spearman $\rho = -1.00$, $p < 10^{-5}$; permutation $p < 0.0001$). Right: Four-condition comparison confirming that reduction under Known patching reflects structured causal influence (not noise).

Comparison to random patching.

Patching with random noise of matched scale (control condition) yields mean entropy $= 6.46 \pm 1.52$, substantially *higher* than the unpatched baseline (1.60 ± 1.52), confirming that the monotone reduction observed under Known patching reflects structured causal influence rather than generic perturbation effects.

 doi:10.20944/preprints202604.0078.v2

These results provide causal rather than merely correlational evidence for the Knowledge Landscape hypothesis: the hidden-state representation of a well-learned fact continuously modulates uncertainty when substituted into an Unknown question of the same category.

5. Discussion

5.1. In-Computation vs. Post-Hoc Metacognition

The central advantage of Knowledge Landscape metacognition is architectural: the signal is extracted from the same forward pass used for generation. Post-hoc methods require multiple passes or additional models. This makes the approach suitable for latency-constrained deployments and enables a new class of real-time inference controllers.

5.2. Interpretation of Hidden-State Divergence Direction

The observation that Known inputs have *higher* hidden-state variance in layers 20–27 is counterintuitive under a naïve convergence model. We interpret this as follows: well-learned knowledge activates a richer, more differentiated (*higher-variance*) representational pattern, consistent with the neuroscientific principle that specialised knowledge engages more distinct representational geometry [8]. Unknown inputs, lacking a retrieval target, collapse toward a uniform, low-variance prior activation—the informational equivalent of a featureless liquid. This is consistent with Tononi's integrated information framework [23], in which systems with more differentiated internal states carry more information. Empirically, our activation-patching result (Experiment 5) provides independent confirmation: injecting the *higher-variance* Known hidden state into an Unknown forward pass monotonically *reduces* output entropy, which would be impossible if higher variance simply reflected noise rather than structured knowledge.

5.3. Negative Result: Attention Locality

Experiment 2 finds no significant attention-locality difference between Known and Unknown conditions ($p = 0.92$). The metacognitive signal is not accessible at the level of attention patterns; it resides in hidden-state representations, consistent with evidence that factual knowledge is stored in MLP layers [25].

5.4. Limitations

- **Sample size (partially addressed).** The core entropy result has been replicated at $n = 300$ ($p = 1.08 \times 10^{-50}$, $r = 0.704$, 95% bootstrap CI $[0.641, 0.764]$), addressing the primary statistical power concern. Hidden-state locus identification and causal patching remain at $n = 50$; the individual-pair Wilcoxon test in Experiment 5 ($p = 0.165$) does not reach significance. Future work should replicate hidden-state experiments at $n \geq 200$.
- **Locus position is model-specific, but existence appears universal.** Both Qwen2.5-7B and Mistral-7B exhibit a statistically significant metacognitive locus, though at different absolute layers. Both fall within 60–90% of total network depth. Generalisation to larger models (70B+) and closed-weight systems (GPT-4) remains unverified.
- **Known/Unknown classification noise (empirically addressed).** Embedding-based reclassification (`all-MiniLM-L6-v2`, threshold 0.70) on 200 items showed 69% agreement with greedy alias matching. Entropy discrimination was *stronger* under greedy ($r = 0.613$) than embedding ($r = 0.168$), because embedding reclassification absorbs uncertain near-misses into the Known set, diluting the signal. Greedy matching thus produces cleaner metacognitive separation and is validated as the primary method.
- **Task scope.** Experiments focus on factual recall (TriviaQA). Whether entropy and hidden-state variance carry metacognitive signal for reasoning, mathematics, or generation tasks is an open empirical question.
- **Abstention recall.** The system detects only 27.8% of incorrect answers at the reported operating point. Higher recall is achievable by lowering the threshold τ (Figure 4) at the cost of precision.

 doi:10.20944/preprints202604.0078.v2

- **Causal scope.** The monotone interpolation result (Experiment 5) is compelling at the aggregate level but noisy at the individual-pair level ($p = 0.165$). Stronger causal evidence would require larger same-category corpora or steering-vector interventions [24].

6. Conclusion

We introduced the Knowledge Landscape hypothesis: a topological account of how factual knowledge is encoded in the forward-pass dynamics of large language models, and demonstrated its practical utility for in-computation metacognition.

Five experiments on TriviaQA establish that: (i) token entropy robustly discriminates Known from Unknown questions across two architecturally distinct models ($p < 10^{-5}$, $r > 0.5$; replicated at $n = 300$: $p < 10^{-49}$, $r = 0.704$); (ii) the metacognitive signal resides in hidden-state representations, not attention patterns; (iii) both architectures exhibit dual-window metacognitive loci with peaks converging at 61–69% of total network depth; (iv) a lightweight single-pass abstention system achieves ROC-AUC $= 0.804$ and $+5.6$ pp accuracy gain; and (v) activation patching with monotone interpolation provides causal evidence that the Known hidden state continuously modulates uncertainty (Spearman $\rho = -1.00$, permutation $p < 0.0001$).

We believe this work opens a new research direction: using the geometry of the forward pass itself as a metacognitive resource—a concrete step toward LLMs that better know what they do not know.

References

1. Anderson, P. W. (1972). More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science*, *177*(4047), 393–396. https://doi.org/10.1126/science.177.4047.393
2. Baldassi, C., Pittorino, F., & Zecchina, R. (2020). Shaping the learning landscape in neural networks around wide flat minima. *Proceedings of the National Academy of Sciences*, *117*(1), 161–170. https://doi.org/10.1073/pnas.1908636117
3. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning* (Vol. 48, pp. 1050–1059). PMLR. https://proceedings.mlr.press/v48/gal16.html
4. Hochreiter, S., & Schmidhuber, J. (1997). Flat minima. *Neural Computation*, *9*(1), 1–42. https://doi.org/10.1162/neco.1997.9.1.1
5. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730
6. Joshi, M., Choi, E., Weld, D., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 1601–1611). ACL. https://doi.org/10.18653/v1/P17-1147
7. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., & Perez, E. (2022). Language models (mostly) know what they know. arXiv. https://arxiv.org/abs/2207.05221
8. Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, Article 4. https://doi.org/10.3389/neuro.06.004.2008
9. Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the 11th International Conference on Learning Representations*. OpenReview. https://openreview.net/forum?id=VD-AYtP0dve
10. Kumaran, D., Conmy, A., Barbero, F., Osindero, S., Patraucean, V., & Veličković, P. (2026). How do LLMs compute verbal confidence? arXiv. https://arxiv.org/abs/2603.17839
11. Li, Z., Xu, Y., & Liu, Y. (2025). From passive metric to active signal: The evolving role of uncertainty quantification in large language models. arXiv. https://arxiv.org/abs/2601.15690
12. Malinin, A., & Gales, M. (2021). Uncertainty estimation in autoregressive structured prediction. In *Proceedings of the 9th International Conference on Learning Representations*. OpenReview. https://openreview.net/forum?id=jN5y-zb5Q7m

 doi:10.20944/preprints202604.0078.v2

13. Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9004–9017). ACL. https://doi.org/10.18653/v1/2023.emnlp-main.557
14. Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). ACL. https://doi.org/10.18653/v1/2020.acl-main.173
15. Geva, M., Bastings, J., Filippova, K., & Globerson, A. (2023). Dissecting recall of factual associations in auto-regressive language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 12216–12235.
16. Ji, Z., Chen, D., Ishii, E., Cahyawijaya, S., Bang, Y., Wilie, B., & Fung, P. (2024). LLM internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop* (pp. 88–104). ACL. https://doi.org/10.18653/v1/2024.blackboxnlp-1.6
17. Tao, M., Fu, Z., & Ye, J. (2024). INSIDE: LLMs' internal states retain the power of hallucination detection. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.
18. Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., & Gal, Y. (2024). Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *ICML 2024 Workshop on Foundation Models in the Wild*. arXiv:2406.15927.
19. Miao, M. M., & Ungar, L. (2026). Closing the confidence–faithfulness gap in large language models. arXiv. https://arxiv.org/abs/2603.25052
20. Mindlin, I., Rahwan, I., & Bonnefon, J.-F. (2025). Fast, slow, and metacognitive thinking in artificial intelligence. *npj Artificial Intelligence*, *2*, Article 12. https://doi.org/10.1038/s44387-025-00012-8
21. Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., & Olah, C. (2022). In-context learning and induction heads. *Transformer Circuits Thread*. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html
22. Steyvers, M., & Peters, M. A. K. (2025). Metacognition and uncertainty communication in humans and large language models. *Current Directions in Psychological Science*, *34*(2), 89–97. https://doi.org/10.1177/09637214241313871
23. Tononi, G., Sporns, O., & Edelman, G. M. (1994). A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, *91*(11), 5033–5037. https://doi.org/10.1073/pnas.91.11.5033
24. Turner, A., Thiergart, L., Udell, D., Leike, J., Wu, J., & MacDiarmid, M. (2023). Activation addition: Steering language models without optimization. arXiv. https://arxiv.org/abs/2308.10248
25. Wen, B., Peng, S., Tang, J., & Liu, Y. (2025). Attention heads of large language models: A survey. *Patterns*, *6*(2), Article 100988. https://doi.org/10.1016/j.patter.2024.100988
26. Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in large language models. In *Proceedings of the 12th International Conference on Learning Representations*. OpenReview. https://openreview.net/forum?id=gjeQKFxFpZ