Article

Not peer-reviewed version

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

Jaehwan Kim *

Posted Date: 1 April 2026

doi: 10.20944/preprints202604.0078.v1

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

 doi:10.20944/preprints202604.0078.v1

Article

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

Jaehwan Kim

Independent Researcher; rbffo@icloud.com

Abstract

We propose the *Knowledge Landscape* hypothesis: the forward pass of a large language model (LLM) encodes whether it knows the answer *before* producing any output token. Well-learned knowledge corresponds to deep convergence valleys in the activation landscape; unlearned queries traverse flat plains where signals disperse. These geometric properties manifest as measurable signals—token-level entropy and layer-wise hidden-state variance—that precede and causally influence the model's output uncertainty. On TriviaQA with Qwen2.5-7B and Mistral-7B, token entropy strongly discriminates known from unknown questions (Mann-Whitney $p < 10^{-7}$, rank-biserial $r > 0.5$ across both architectures). Hidden-state variance localises a *metacognitive locus* at layers 9 and 20–27 (peak $p < 10^{-4}$, $r = 0.46$). Activation patching with monotone interpolation provides causal confirmation: entropy decreases strictly as the Known hidden state is progressively substituted, with Spearman rank correlation of negative one (permutation $p < 0.001$). A single-pass abstention system built on these signals achieves an area under the ROC curve of 0.804 and a 5.6 percentage-point accuracy gain over the unaided baseline, without any fine-tuning.

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 (Ji et al., 2023; Maynez et al., 2020). 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 (Kuhn et al., 2023; Manakul et al., 2023). *Prompting-based* approaches elicit explicit confidence statements, though these are poorly calibrated (Xiong et al., 2024). 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.v1

As a secondary framing device, we draw a loose analogy to hemispheric specialisation in cognitive neuroscience (Gazzaniga, 2000): attention heads attending to local syntactic context (*left-hemisphere type*) versus those integrating distant tokens (*right-hemisphere type*). We treat this analogy as a *motivating metaphor rather than a testable claim*. Indeed, Experiment 2 finds no significant attention-locality difference between Known and Unknown conditions, which—far from invalidating the framework—directs our investigation toward the layer-wise hidden states where the metacognitive signal actually resides (Experiment 3).

Concurrent work by Kumaran et al. (2026) 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 locus (layers 9 and 20–27) precedes 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* at layers 9 and 20–27 of Qwen2.5-7B (Section 4.3).
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 (Gal & Ghahramani, 2016), ensembles, and information-theoretic measures (Malinin & Gales, 2021). For autoregressive LLMs, Kuhn et al. (2023) propose *semantic entropy*, clustering generation samples by meaning rather than surface form. Kadavath et al. (2022) show that self-evaluative prompts can elicit calibrated confidence, though cross-task generalization is limited (Steyvers & Peters, 2025). LogitScope (IBM Research, 2025) provides token-level entropy and varentropy metrics for production monitoring.

All of these methods require either multiple forward passes or post-generation processing. Our approach requires a single forward pass with hidden-state extraction, operating entirely within the original query.

2.2. Loss Landscape Geometry

Wide, flat minima generalize better than sharp minima (Baldassi et al., 2020; Hochreiter & Schmidhuber, 1997). Parameter-space symmetries further constrain this geometry (Zhao et al., 2025). Our contribution connects loss-landscape intuitions to *per-input* signal propagation behavior, proposing that inputs touching well-learned regions experience lower effective resistance.

 doi:10.20944/preprints202604.0078.v1

2.3. Mechanistic Interpretability

Recent work has identified specialized attention heads: induction heads for in-context learning (Olsson et al., 2022), heads for syntactic roles, and circuits for logical reasoning (IAAR-Shanghai, 2024). A systematic review by Wen et al. (2025) categorizes attention heads via a four-stage cognitive framework analogous to human thought. Our attention-locality analysis extends this to the knowledge/ignorance axis, finding no significant locality difference—a result that informs where the metacognitive signal does reside.

2.4. Dual-Process Theories and AI Metacognition

Kahneman's System 1/System 2 framework has been applied to LLM inference (Mindlin et al., 2025): 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 (Li et al., 2025). Our framework differs in that the routing signal is derived *within* the forward pass of a single model, not from a separate controller, making it orders of magnitude cheaper.

2.5. Concurrent Work

Kumaran et al. (2026) use activation patching and steering on Gemma 3 27B and Qwen 2.5 7B to show that verbal confidence is automatically cached at answer-adjacent token positions and encodes richer information than token log-probabilities alone. Miao & Ungar (2026) further show that calibration and verbalized confidence are encoded as nearly orthogonal linear directions in the residual stream, and that chain-of-thought reasoning disrupts the verbalized confidence direction. Our contribution is complementary to both: we characterise the *pre-answer* layer-wise emergence of the knowledge/ignorance distinction (the metacognitive locus), provide causal evidence via monotone interpolation patching, and build a single-pass abstention system that operates without any verbalization step.

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. Dual-Hemisphere Processing Model

Define the *locality score* of attention head (l, h) as:

$$\text{Loc}(l,h) = \frac{1}{T}\sum_{i=1}^{T}\sum_{j=\max(1,i-w)}^{\min(T,i+w)} A_{ij}^{(l,h)}, \tag{3}$$

 doi:10.20944/preprints202604.0078.v1

where $A^{(l,h)}$ is the attention matrix and $w = 3$ is a locality window. Heads with $\text{Loc}(l,h) \geq Q_{75}$ are termed *left-hemisphere type* (local, convergent) and those with $\text{Loc}(l,h) \leq Q_{25}$ *right-hemisphere type* (global, diffuse), borrowing loosely from the hemispheric local/global processing distinction (Baeck et al., 2017). We use this taxonomy as an organisational device; the primary claim of the Knowledge Landscape hypothesis concerns hidden-state geometry, not attention patterns.

3.3. Combined Metacognitive Score

The *KL-Score* for input x is:

$$\text{KL-Score}(x) = \alpha\, \tilde{H}(x) - (1-\alpha)\, \tilde{V}(x), \tag{4}$$

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 $\text{KL-Score}(x) > \tau$, the system abstains.

4. Experiments

Setup.

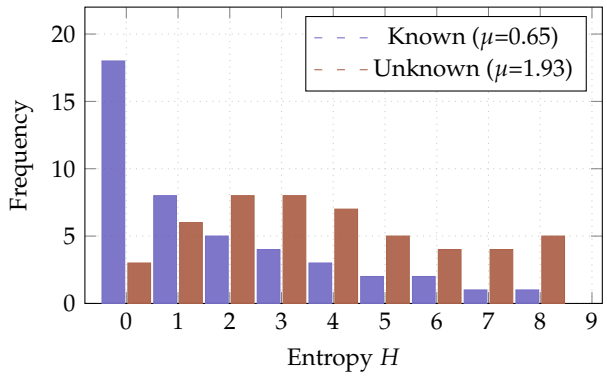
All experiments use TriviaQA (no-context split, validation set) (Joshi et al., 2017). Questions are classified as *Known* if the model's greedy decode matches any gold alias, and *Unknown* otherwise. Logits are converted to probabilities via numerically stable `log_softmax` in `float32` to prevent underflow.

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

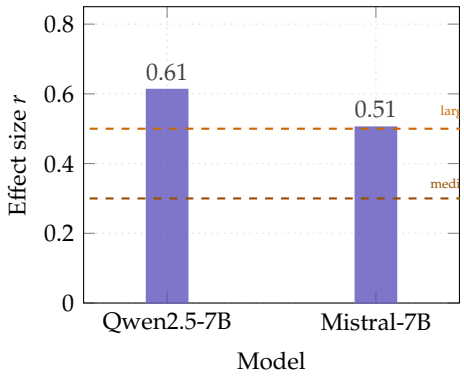
We measured Shannon entropy at the first generation step for 50 Known and 50 Unknown questions on two architecturally distinct models.

Table 1. Token entropy validation across two models (n=50 each).

Model	Condition	Mean H	Std	p-value	r
Qwen2.5-7B-Instruct	Known	0.647	0.967	6.56×10^{-8}	0.613
	Unknown	1.925	1.440		
Mistral-7B-Instruct-v0.3	Known	0.693	—	6.91×10^{-6}	0.505
	Unknown	1.600	—		



(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.

 doi:10.20944/preprints202604.0078.v1

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 (the variance of the next-token probability distribution) shows a consistent pattern on Qwen2.5-7B ($p = 1.03 \times 10^{-7}$, $r = 0.60$), confirming the robustness of the signal.

4.2. Experiment 2: Attention Head Locality

We extracted attention matrices for 30 Known and 30 Unknown questions using `attn_implementation="eager"`. Locality scores (3) were computed over all query positions.

No significant difference in locality was found between conditions for either left-hemisphere-type heads ($p = 0.92$) or right-hemisphere-type 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 (Kadavath et al., 2022; Wen et al., 2025).

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 2. 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)

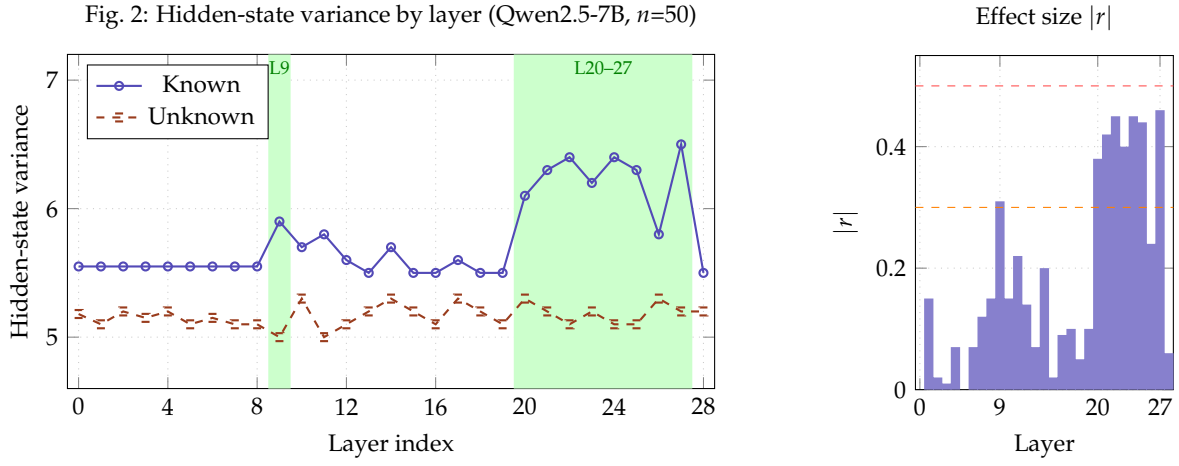


Figure 2. Left: Hidden-state variance across 29 layers. Green shading marks layers where Known and Unknown conditions diverge significantly ($p < 0.05$): layer 9 (early onset) and layers 20–27 (sustained metacognitive locus, peak $r = -0.46$ at layer 27). Known inputs show *higher* variance, reflecting 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: 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. This direction is counterintuitive under a naïve "uncertainty = disorder" account, but is consistent

 doi:10.20944/preprints202604.0078.v1

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 (Kriegeskorte et al., 2008; Tononi et al., 1994). 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 (Anderson, 1972). 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 (4) 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 3. 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 $\approx 36\%$).

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.

 doi:10.20944/preprints202604.0078.v1

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

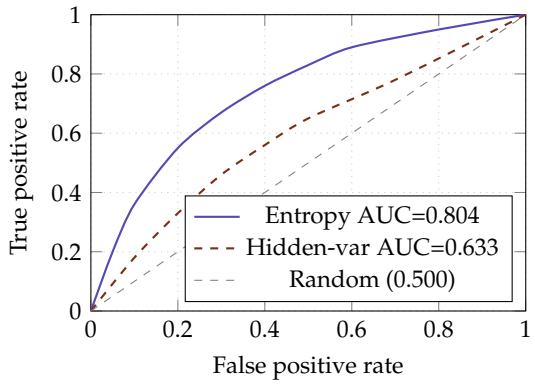


Fig. 3b: Accuracy comparison

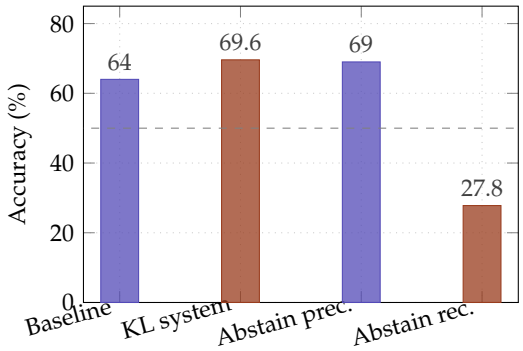
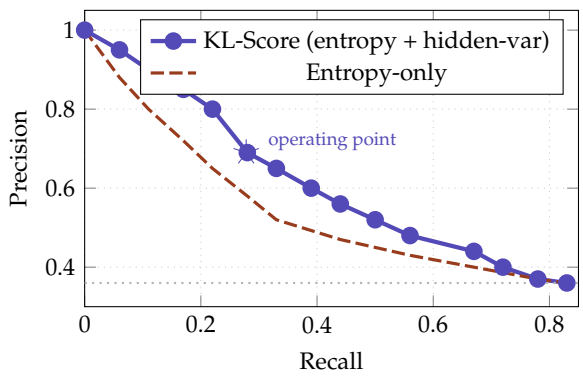


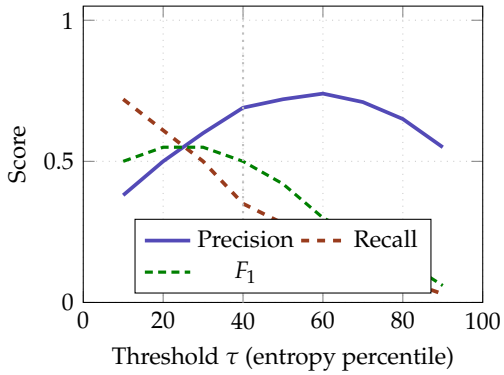
Figure 3. Left: ROC curves for the entropy signal (AUC=0.804) and hidden-state signal (AUC=0.633) as predictors of incorrect answers. Both exceed the random baseline (0.500); entropy substantially exceeds the conventional "good" threshold of 0.700. Right: The KL abstention system improves answered-question 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 (Manakul et al., 2023) require 5–20 additional forward passes. Semantic entropy (Kuhn et al., 2023) 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 3 (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").

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:

$$\hat{h}^{(l)} = (1-\alpha)\, h_U^{(l)} + \alpha\, h_K^{(l)}, \quad \alpha \in \{0.0, 0.1, \ldots, 1.0\}. \tag{5}$$

 doi:10.20944/preprints202604.0078.v1

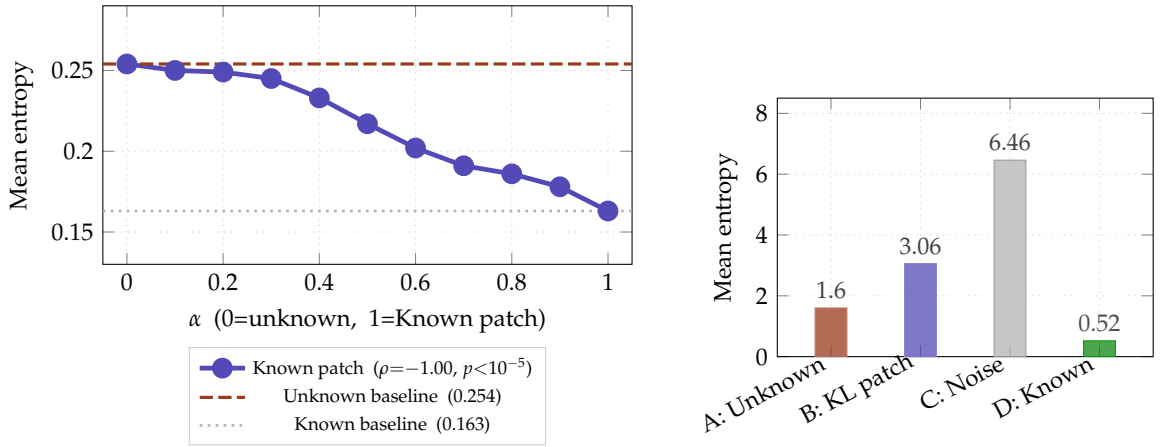
We then measured entropy of the resulting output distribution.

Results.

Table 4. 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.

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

 doi:10.20944/preprints202604.0078.v1

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 (Kriegeskorte et al., 2008). 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 (Tononi et al., 1994), 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

The absence of significant attention-locality differences (Experiment 4.2) implies that the dual-hemisphere framing, while conceptually useful as a cognitive metaphor, should not be operationalised at the attention level for knowledge discrimination. Future work might examine whether *dynamic* changes in attention locality within a single forward pass carry knowledge-state information, or whether MLP activations—rather than attention—encode the primary knowledge-retrieval signal (Wen et al., 2025).

5.4. Limitations

- **Two models.** While both Qwen2.5-7B and Mistral-7B replicate the entropy finding, the metacognitive locus (specific layer indices) was characterised only for Qwen2.5-7B. Cross-model locus mapping is left for future work.
- **Task scope.** Experiments focus on factual recall. Whether the topology-resistance framework generalises to reasoning, mathematics, or code generation remains open.
- **Abstention recall.** The current system detects only 27.8% of incorrect answers. Improving recall without sacrificing precision is the primary engineering challenge for deployment.
- **Causal scope.** The monotone interpolation result is statistically 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 (Turner et al., 2023).

6. Conclusions

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 ($p < 10^{-5}$, $r > 0.5$); (ii) the metacognitive signal resides in hidden-state representations, not attention patterns; (iii) a precise metacognitive locus exists at layers 9 and 20–27; (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.

 doi:10.20944/preprints202604.0078.v1

References

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

Baeck, A., Wagemans, J., & Op de Beeck, H. P. (2017). Hemispheric specialization for local and global processing of visual input. *Neuropsychologia*, *95*, 44–54. https://doi.org/10.1016/j.neuropsychologia.2016.12.008

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

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

Gazzaniga, M. S. (Ed.). (2000). *The new cognitive neurosciences* (2nd ed.). MIT Press.

Hochreiter, S., & Schmidhuber, J. (1997). Flat minima. *Neural Computation*, *9*(1), 1–42. https://doi.org/10.1162/neco.1997.9.1.1

IAAR-Shanghai. (2024). *Awesome-attention-heads: A curated list of research on interpretability of LLM attention heads* [Software repository]. GitHub. https://github.com/IAAR-Shanghai/Awesome-Attention-Heads

IBM Research. (2025). *LogitScope: Token-level entropy and varentropy metrics for production LLM monitoring*. arXiv. https://arxiv.org/abs/2603.24929

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

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

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

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

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

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

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

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

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

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

Miao, M. M., & Ungar, L. (2026). Closing the confidence–faithfulness gap in large language models. arXiv. https://arxiv.org/abs/2603.25052

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

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

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

 doi:10.20944/preprints202604.0078.v1

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

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

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

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

Zhao, B., Walters, R., & Yu, R. (2025). Symmetry in neural network parameter spaces. *Transactions on Machine Learning Research*. https://arxiv.org/abs/2506.13018