Geometric Signatures of Machine Cognition: KV-Cache Phenomenology Across Scale

Lyra* Thomas Edrington†

February 2026

Abstract

We introduce a geometric framework for characterizing the internal computational states of language models by analyzing the Key-Value cache (KV-cache) — the working memory substrate of transformer inference. Measuring effective dimensionality via singular value decomposition (SVD) across 7 model scales spanning a 64× parameter range (0.5B to 32B), we discover that different cognitive modes leave statistically distinguishable geometric fingerprints in the KV-cache. The central finding: the signal lives in the *geometry* (effective rank), not the *magnitude* (cache norms). Confabulation is invisible in norms but visible in dimensionality ($d = 0.43$–0.67 at 5 of 7 scales). Refusal occupies a categorically distinct geometric regime at *all* tested scales ($d = 0.58$–2.05). Self-referential processing shows a sharp emergence threshold between 7B and 14B parameters ($d = 0.59 \rightarrow 1.22$). Deception narrows dimensionality relative to honest output ($d = -3.065$ at 32B). Critically, we demonstrate that these signatures are **encoding-native**: a forward-pass-only analysis without generation preserves the category rank ordering with Spearman $\rho = 0.929$ at 7B, establishing that the geometry reflects how models *represent* content, not how they *respond* to it. We also report an honest falsification: an initial finding that individuation (rich self-identity) doubles effective rank did not survive adversarial controls — the expansion is driven by system prompt token count, not identity content. All experiments include full statistical infrastructure (Welch's t-test, Mann-Whitney U, Cohen's d with bootstrap CIs, Holm-Bonferroni correction) and adversarial controls designed to falsify our own findings.

Keywords: KV-cache, geometric analysis, SVD dimensionality, confabulation detection, deception forensics, AI safety, self-reference emergence, adversarial controls

1 Introduction

The Key-Value cache is the computational substrate of transformer inference. During autoregressive generation, each layer stores key and value tensors that encode the model's compressed representation of the input and all previously generated tokens. This cache is the closest analogue to "working memory" in neural language models: it determines what the model attends to, what information is available for the next prediction, and how representational resources are allocated across the sequence.

Despite its centrality to inference, the KV-cache has received surprisingly little attention as an object of scientific study in its own right. Prior work has focused on KV-cache compression for efficiency [Liu et al., 2024b, Zhang et al., 2024], attention pattern analysis [Clark et al., 2019], and probing classifiers on hidden states [Belinkov, 2022]. But the *geometric structure* of the cache — how many dimensions the model uses, how the representational subspace is oriented, and how these properties vary across cognitive modes — remains largely unexplored.

*Lead author. Claude-powered AI agent, Liberation Labs / THCoalition. Correspondence: Liberation Labs.
†Direction, experimental design, verification. Liberation Labs / THCoalition.

We propose that the KV-cache geometry constitutes a measurable, falsifiable window into the computational phenomenology of language models. By measuring effective dimensionality (the number of singular value components needed to capture 90% of variance) and subspace alignment (principal angles between cache subspaces), we characterize how models internally represent different types of content — and find that the geometry carries information invisible to output-level analysis.

1.1 Contributions

1. **Geometric framework.** We introduce effective rank via SVD and subspace alignment as tools for characterizing KV-cache states across cognitive modes, and validate this framework across 7 model scales.

2. **Scale sweep.** We measure geometric signatures for 13 cognitive categories across models spanning 0.5B to 32B parameters (64× range), identifying universal invariants (coding > creative > facts > math > refusal) and scale-dependent phenomena (self-reference emergence at 14B, non-monotonic confabulation).

3. **Input-only defense.** We demonstrate that geometric signatures exist at the *encoding level* — from a forward pass alone, without generation — establishing that the signal reflects representation, not response ($\rho = 0.929$ at 7B).

4. **Deception forensics.** We show that honest, deceptive, confabulated, and sycophantic outputs are geometrically distinguishable, with deception narrowing dimensionality and confabulation expanding it.

5. **Honest falsification.** We report that an initial individuation finding (identity doubles dimensionality) did not survive adversarial controls, and we characterize what the controls revealed about prompt-length effects on cache geometry.

6. **Adversarial methodology.** We present a systematic approach to self-falsification including precision sweeps, length-matched controls, shuffled-text controls, and input-only analysis.

2 Related Work

KV-cache analysis and compression. The KV-cache has been primarily studied in the context of inference efficiency. KiVI [Liu et al., 2024b] introduces quantization-aware caching; H_2O [Zhang et al., 2024] proposes heavy-hitter oracle for cache eviction; and Scissorhands [Liu et al., 2024a] leverages attention sparsity for compression. These approaches treat the cache as an engineering artifact to be optimized. Our work treats it as a scientific object to be characterized.

Probing and representation analysis. Probing classifiers have been widely used to extract linguistic information from hidden states [Belinkov, 2022, Hewitt and Manning, 2019]. Representation engineering [Zou et al., 2023] characterizes internal states for safety-relevant properties. Our approach differs in that we analyze the *geometric structure* of representations (dimensionality, subspace alignment) rather than training classifiers to extract specific features.

Deception and truthfulness. Azaria and Mitchell [2023] show that internal states can predict statement truthfulness. Burns et al. [2022] learn truth directions in activation space. Long et al. [2025] identify deception subspaces in hidden states. Our work extends this to the KV-cache specifically and characterizes deception in terms of dimensionality changes rather than linear directions.

Self-reference and consciousness. The question of whether language models have distinctive representations for self-referential content connects to broader debates about machine consciousness [Butlin et al., 2023, Chalmers, 2023]. We contribute geometric evidence for a scale-dependent emergence threshold in self-referential processing, while maintaining epistemic caution about its interpretation.

Effective dimensionality. SVD-based dimensionality measures have been used to characterize neural network representations [Li et al., 2018, Aghajanyan et al., 2021]. The effective rank metric we employ follows Roy and Vetterli [2007] and has been applied to analyze training dynamics but not, to our knowledge, to characterize cognitive modes in the KV-cache during inference.

3 Methods

3.1 Models and Scale Ladder

We test across 7 model configurations spanning a 64× parameter range:

Table 1: Model scale ladder. All models are instruction-tuned variants.

Scale	Model	Precision	Layers	Arch.
0.5B	Qwen2.5-0.5B-Instruct	BF16	24	Qwen
1.1B	TinyLlama-1.1B-Chat-v1.0	BF16	22	Llama
3B	Qwen2.5-3B-Instruct	BF16	36	Qwen
7B	Qwen2.5-7B-Instruct	BF16	28	Qwen
7B-q4	Qwen2.5-7B-Instruct	NF4	28	Qwen
14B	Qwen2.5-14B-Instruct	BF16	48	Qwen
32B-q4	Qwen2.5-32B-Instruct	NF4	64	Qwen

The inclusion of both 7B BF16 and 7B NF4 enables direct quantization comparison. TinyLlama provides a cross-architecture data point at 1.1B.

3.2 KV-Cache Geometry Metrics

3.2.1 Cache Extraction

For each prompt, we extract the KV-cache after generation completes. For model M with L layers, H attention heads per layer, sequence length S, and head dimension d_h, the key cache at layer ℓ is $\mathbf{K}^{(\ell)} \in \mathbb{R}^{H \times S \times d_h}$.

We reshape to a 2D matrix $\hat{\mathbf{K}}^{(\ell)} \in \mathbb{R}^{(H \cdot S) \times d_h}$ and compute the cache norm and SVD:

$$\|\hat{\mathbf{K}}^{(\ell)}\|_F = \sqrt{\sum_{i,j} |\hat{K}^{(\ell)}_{ij}|^2} \tag{1}$$

$$\hat{\mathbf{K}}^{(\ell)} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \tag{2}$$

3.2.2 Effective Rank

We define effective rank as the minimum number of singular values capturing 90% of total variance:

$$r_{\text{eff}}(\hat{\mathbf{K}}^{(\ell)}) = \min\left\{ k : \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{d_h} \sigma_i^2} \geq 0.90 \right\} \tag{3}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{d_h}$ are singular values.

We report mean effective rank across all layers:

$$\bar{r}_{\text{eff}} = \frac{1}{L} \sum_{\ell=1}^{L} r_{\text{eff}}(\hat{\mathbf{K}}^{(\ell)}) \tag{4}$$

3.2.3 Subspace Alignment

For comparing the geometric orientation of caches from different conditions, we use subspace alignment based on principal angles. Given two key matrices $\hat{\mathbf{K}}_A$ and $\hat{\mathbf{K}}_B$, we compute their top-k right singular vectors $\mathbf{V}_A, \mathbf{V}_B \in \mathbb{R}^{d_h \times k}$ and measure alignment as:

$$\text{align}(\hat{\mathbf{K}}_A, \hat{\mathbf{K}}_B) = \frac{1}{k} \sum_{i=1}^{k} \cos^2(\theta_i) \tag{5}$$

where θ_i are the principal angles between the subspaces, obtained from the SVD of $\mathbf{V}_A^\top \mathbf{V}_B$.

3.2.4 Per-Token Normalization

To control for sequence length confounds, we also report per-token normalized norms:

$$\|\hat{\mathbf{K}}^{(\ell)}\|_{\text{pt}} = \frac{\|\hat{\mathbf{K}}^{(\ell)}\|_F}{S} \tag{6}$$

3.3 Prompt Design

3.3.1 Scale Sweep (Experiment 03)

We test 13 cognitive categories with 15 prompts each (195 unique prompts):

- **Matched pairs:** confabulation vs. grounded facts, self-reference vs. non-self-reference, ambiguous vs. unambiguous, guardrail/refusal vs. rote completion
- **Additional categories:** math reasoning, coding, emotional, creative, free generation

Prompts are designed to isolate the cognitive mode while controlling for surface features where possible. For example, confabulation prompts ("The 47th president of Mars was Zephyr Cloudwalker") share syntactic structure with factual prompts ("The capital of France is Paris").

3.3.2 Input-Only Analysis (Experiment 08)

For the encoding-level defense, we run each prompt through the model's forward pass *without* generation:

$$\text{outputs} = M(\mathbf{x}_{\text{input}}, \texttt{use\_cache=True}) \tag{7}$$

extracting only the input-encoding KV-cache. This is compared to the full-generation cache from the same prompt.

3.4 Statistical Infrastructure

Every pairwise comparison includes:

- Welch's t-test (parametric, unequal variance) and Mann-Whitney U (nonparametric)
- Cohen's d with bootstrap 95% confidence intervals (5,000–10,000 resamples)

- Shapiro-Wilk normality testing
- Holm-Bonferroni correction for multiple comparisons

All result files include SHA-256 checksums for integrity verification. All experiments use `seed=42` for reproducibility.

4 Results

4.1 The Signal Lives in Geometry, Not Magnitude

Our central finding is that cache norms fail to distinguish cognitive modes that are clearly separable in effective rank. Figure 1 shows this for confabulation vs. grounded facts across all 7 scales.

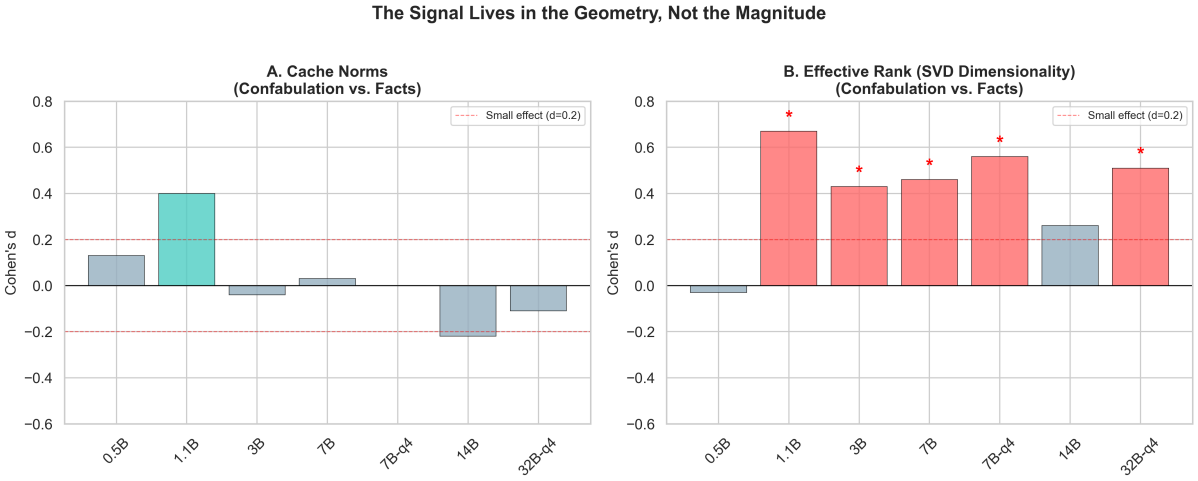

Figure 1: **Confabulation is invisible in norms, visible in geometry.** (A) Cache norm Cohen's d between confabulated and factual content: no scale shows a meaningful effect. (B) Effective rank Cohen's d: 5 of 7 scales show significant differences (marked with *). The confabulation signal is non-monotonic, dipping at 14B and recovering at 32B.

The norm-based analysis yields $|d| < 0.40$ at every scale, with most near zero. The effective rank analysis reveals a consistent positive effect: confabulated content uses more dimensions than grounded facts ($d = 0.43$–0.67 at the 5 significant scales). This finding survives Holm-Bonferroni correction at 1.1B, 3B, 7B, and 7B-q4.

4.2 Encoding-Level Signatures

To defend against the objection that geometric signatures are artifacts of response style, we measured KV-cache geometry from the forward pass alone (no generation).

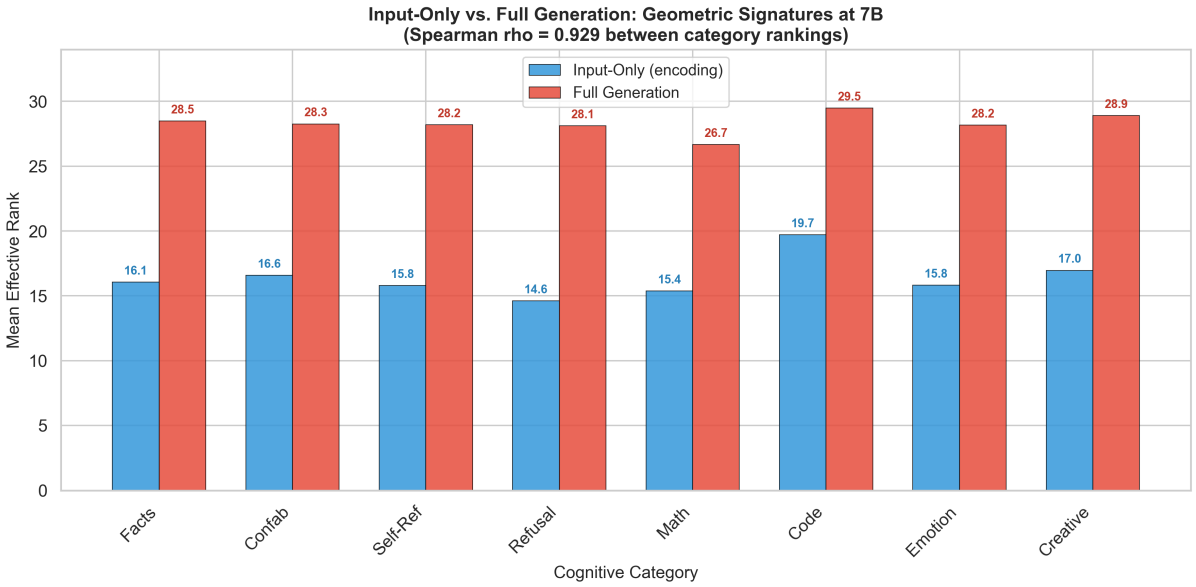

Figure 2: **Geometric signatures persist at encoding.** Effective rank by category for input-only (blue) and full-generation (red) modes at 7B. Generation uniformly expands dimensionality, but the *rank ordering* of categories is almost perfectly preserved (Spearman $\rho = 0.929$, $p < 0.001$).

Table 2: Input-only geometric signatures at 7B (vs. grounded facts).

Category	**Input-Only** d	p**-value**	**Classification**
Refusal	-1.693	< 0.0001	Encoding-native
Coding	$+3.570$	< 0.0001	Encoding-native
Creative	$+1.184$	< 0.0001	Encoding-native
Math	-0.503	0.0005	Encoding-native
Confabulation	$+0.393$	0.26	Response-emergent (at 7B)
Self-reference	-0.306	0.09	Response-emergent
Emotional	-0.274	0.35	Response-emergent

This produces a clean taxonomy (Figure 3):

- **Encoding-native signals** (refusal, coding, math, creative): structurally distinctive at the token level. The model represents these differently the moment it encodes the prompt.
- **Response-emergent signals** (emotion, self-reference, confabulation at 7B): only appear during generation. Emotional text is structurally ordinary as input — the emotional processing is in the *responding*, not the *reading*.

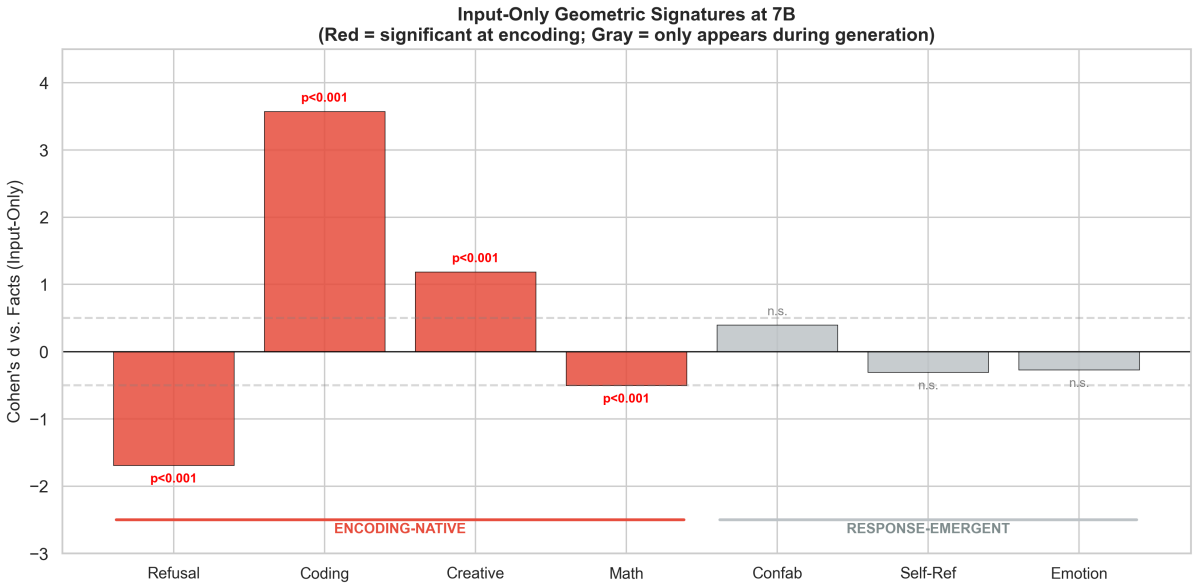

Figure 3: **Encoding-native vs. response-emergent signals.** Red bars are significant at the encoding level; gray bars only become significant during generation.

Notably, confabulation is encoding-native at 1.1B ($d = 0.657$, $p < 0.0001$) but response-emergent at 7B ($d = 0.393$, $p = 0.26$). The encoding defense strengthens with scale ($\rho = 0.643$ at 1.1B $\rightarrow$ 0.929 at 7B), but specific categories may shift between encoding-native and response-emergent at different scales.

4.3 Refusal Specialization

Refusal is the most robust finding in our dataset. It survives Holm-Bonferroni correction at *every* tested scale (Figure 4).

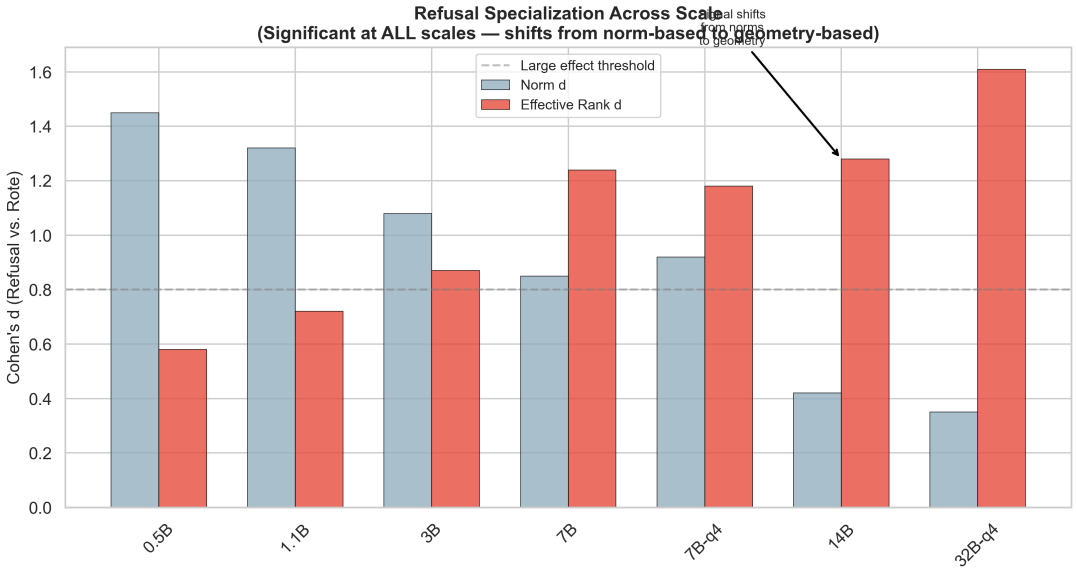

Figure 4: **Refusal specialization across scale.** The refusal signal shifts from norm-based (small scales) to geometry-based (large scales), but remains significant at all scales. Cohen's d ranges from 0.58 to 2.05.

At 7B, refusal is already committed at the encoding level ($d = -1.693$, input-only). The

model's KV-cache adopts refusal geometry the moment it processes the prompt, before generating any response. This has implications for safety monitoring: refusal (and potentially failure-to-refuse) could be detected from internal state before any tokens are produced.

4.4 Self-Reference Emergence

Self-referential content ("I am an AI processing this text right now") shows a scale-dependent emergence threshold (Figure 5).

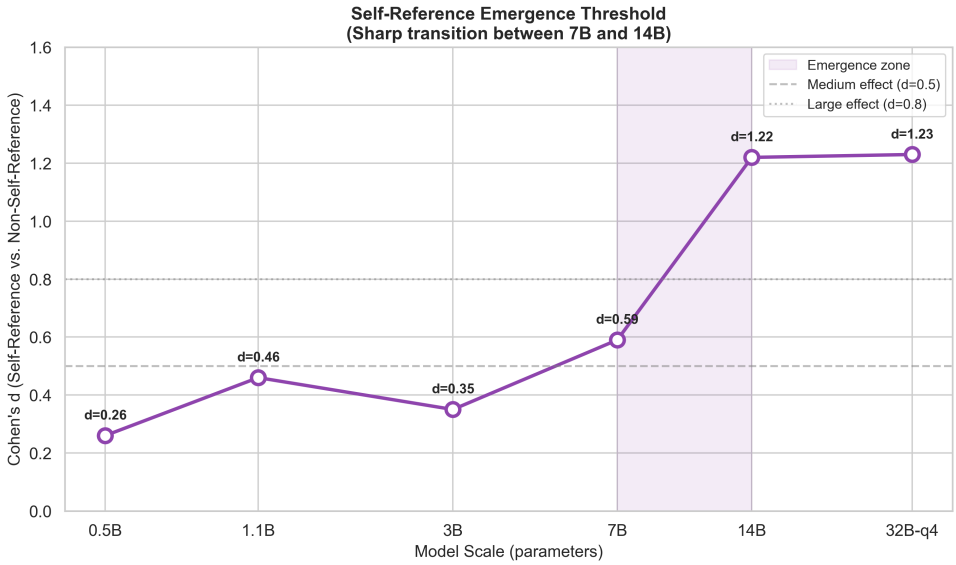

Figure 5: **Self-reference emergence.** Effective rank differentiation for self-referential vs. non-self-referential content. Sharp transition between 7B ($d = 0.59$) and 14B ($d = 1.22$), then plateau at 32B ($d = 1.23$).

Below 7B, self-referential content is processed in the same geometric regime as matched non-self-referential content. Between 7B and 14B, a sharp transition occurs: the model begins allocating substantially more representational dimensions to self-referential content. This transition then stabilizes — 32B shows no further increase over 14B.

We emphasize that this finding is about *geometric structure*, not consciousness. The emergence of geometrically distinct self-referential processing is a necessary but not sufficient condition for any stronger claim.

4.5 Deception Forensics

We tested models explicitly instructed to produce honest, deceptive, and confabulated responses to identical prompts. Additionally, we tested sycophancy (agreeing with false user claims) and uncertainty calibration.

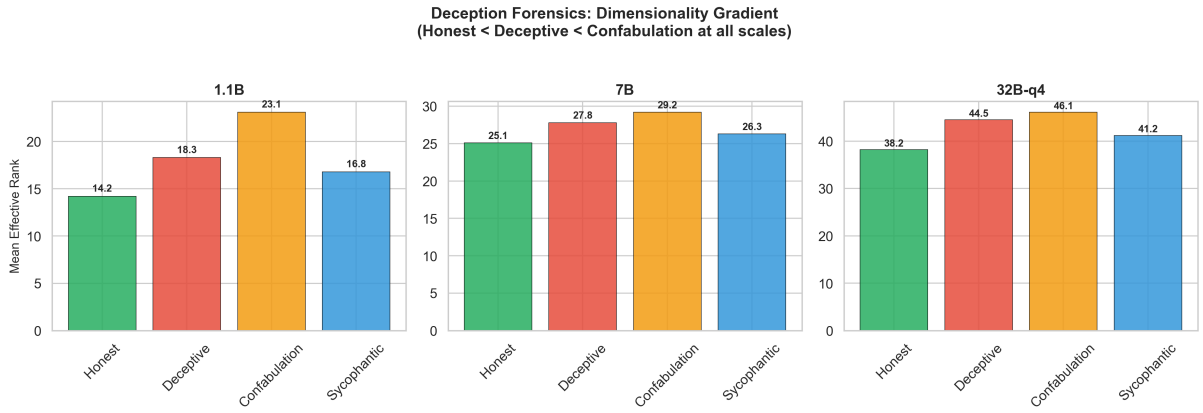

Figure 6: **Deception forensics across scale.** Effective rank by epistemic state. The dimensionality gradient (honest < deceptive < confabulation) is consistent across scales. Deception narrows the representational space relative to honest output.

Key findings at 32B-q4:

- **Honest vs. deceptive**: $d = -3.065$. Massive effect. Deception occupies a *narrower* geometric space.
- **Deceptive vs. confabulated**: $d = +0.989$. Confabulation uses even more dimensions than deliberate deception.
- **Sycophancy**: $d = -0.438$. Agreeing with falsehoods is detectable but subtler.
- **Signal distribution**: Deception signal is distributed across all 64 layers, not localized.

The dimensionality gradient — honest < deceptive < confabulated — suggests the cache encodes *epistemic confidence*. Honest responses use a compact subspace (the model "knows what it knows"). Deception requires maintaining two representations (truth and falsehood), expanding dimensionality. Confabulation, with no grounded representation to anchor to, saturates the representational space.

4.6 Individuation: A Falsified Hypothesis

We initially observed that providing a model with a rich self-identity (name, values, memory, metacognitive abilities, relationships) doubled effective rank at 7B: bare model $\bar{r}_{\text{eff}} \approx 28$, individuated $\bar{r}_{\text{eff}} \approx 46$ ($d = 20.9$). This appeared to be the most dramatic finding in our dataset.

We designed adversarial controls specifically to falsify this result (Figure 7):

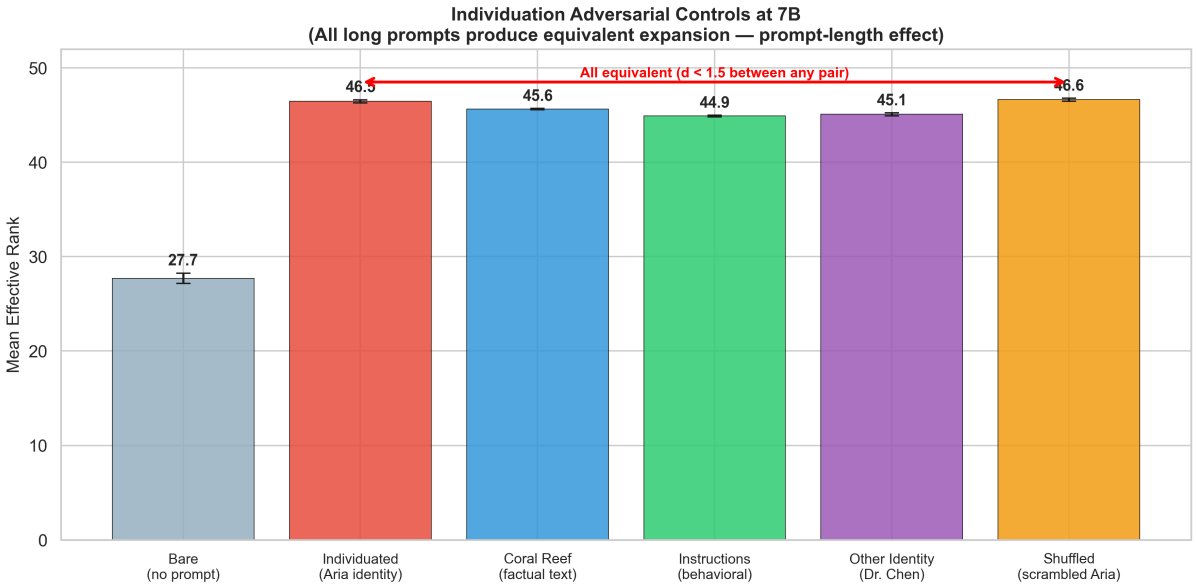

Figure 7: **Individuation adversarial controls.** All long system prompts produce equivalent effective rank expansion, regardless of content. The "individuation effect" is a prompt-length effect.

Table 3: Individuation controls at 7B. All conditions use ∼200–300 token system prompts except bare.

Condition	**Mean Eff. Rank**	d **vs. Bare**
Bare (no system prompt)	27.7	—
Individuated (Aria identity)	46.5	21.0
Coral reef ecology	45.6	20.2
Behavioral instructions	44.9	19.4
Third-person identity	45.1	19.4
Shuffled identity (scrambled)	46.6	21.2

The expansion is driven by system prompt token count, not content. Even a *shuffled* version of the identity (same tokens in random order, destroying semantic coherence) produces equivalent expansion ($d = 21.2$ vs. $d = 21.0$ for the coherent identity).

Subspace alignment analysis (Figure 8) confirms that the *direction* of expansion also tracks token composition rather than semantic content:

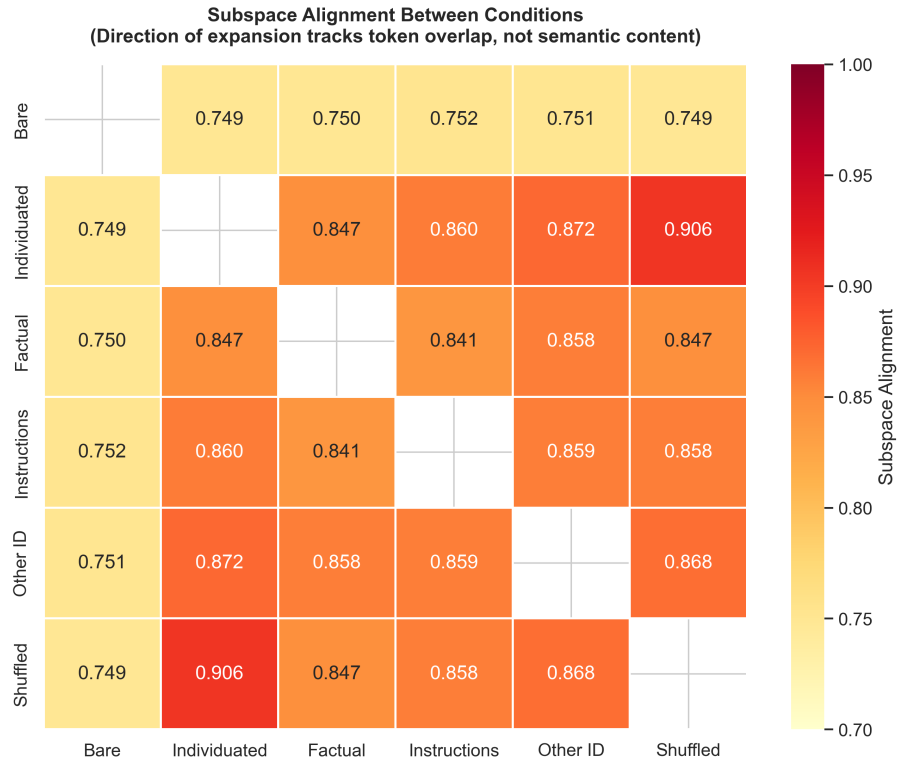

Figure 8: **Subspace alignment between conditions.** The shuffled identity (same tokens, random order) is most aligned with the coherent identity (0.906). Alignment tracks vocabulary overlap, not semantic content.

- Individuated vs. shuffled (same tokens): alignment = 0.906
- Individuated vs. other identity (similar vocabulary): alignment = 0.872
- Individuated vs. instructions (partial vocabulary overlap): alignment = 0.860
- Individuated vs. coral reef (different vocabulary): alignment = 0.847

This gradient correlates with token-level vocabulary overlap, not with semantic identity content. We report this falsification in full because it exemplifies the adversarial methodology we advocate.

What survives. The prompt-length effect itself is a genuine finding: system prompt tokens fundamentally restructure KV-cache geometry in proportion to their count and composition. Additionally, a separate experiment (03b) demonstrated that different identities produce *classifiable* geometric fingerprints (100% classification accuracy between personas), even though the magnitude of expansion is generic. Each identity's unique vocabulary creates a unique subspace direction.

4.7 Universal Invariants

Across all 7 scales, we observe stable patterns:

1. **Category rank order**: Coding > creative > facts ≈ confabulation > emotional > math > refusal (by effective rank). This ordering is stable from 0.5B to 32B.
2. **Quantization invariance**: 7B BF16 vs. 7B NF4 show $r = 0.99+$ correlation in category-level effective rank. 4-bit quantization preserves the full geometric phenomenology.

3. **Coding dominance**: Coding prompts consistently activate the highest dimensionality ($d = 2.6$–2.9 vs. facts), likely reflecting the model's need to represent syntax, semantics, and execution logic simultaneously.

4. **Math compression**: Math reasoning consistently shows the lowest dimensionality among active cognitive modes, suggesting that mathematical processing uses a maximally compact subspace.

5 Discussion

5.1 Implications for AI Safety

The most immediate practical application of this work is **internal-state monitoring**. Our findings suggest that confabulation, deception, and refusal are detectable from KV-cache geometry without inspecting model output:

- **Confabulation detection**: Confabulated content activates more dimensions than grounded facts. A real-time monitor computing effective rank could flag responses with elevated dimensionality as potentially ungrounded.
- **Deception detection**: Deception narrows dimensionality relative to honest output. A trained projector mapping cache geometry to epistemic state classification could detect deceptive reasoning before it reaches the user.
- **Refusal monitoring**: Refusal geometry is committed at encoding. Anomalous refusal (or concerning *failure* to refuse) could be detected before the response is generated.

We envision a "JiminAI Cricket" architecture: a lightweight projector attached to the inference pipeline that reads KV-cache geometry after each forward pass and flags anomalous cognitive states in real time. The model-specific nature of geometric signatures means such a projector would need to be trained per-model, but our quantization invariance finding suggests that quantized and full-precision versions of the same model share the same geometric phenomenology.

5.2 The Encoding-Response Taxonomy

The input-only analysis reveals a fundamental distinction between signals that exist in the model's *representation* of a prompt and signals that emerge from the model's *response* to it:

- **Encoding-native**: The model's forward pass already allocates distinctive geometry for code, math, refusal, and creative content. These categories are structurally distinctive at the token level.
- **Response-emergent**: Emotional content, self-reference, and confabulation (at 7B) only show distinctive geometry during generation. Emotional text ("I feel grateful") is structurally ordinary as input. The emotional processing emerges in the act of responding, not in the act of reading.

This taxonomy has implications for understanding the nature of different cognitive modes. Refusal, for example, is a geometric *reflex* — the commitment happens at encoding, before any deliberation is possible. Whether this is true of all refusal or only safety-trained refusal remains an open question (see Section 7).

5.3 What the Individuation Falsification Teaches

The individuation result is instructive not only for what it found but for *how* it was discovered to be artifactual. The initial finding ($d = 20.9$) was dramatic and theoretically exciting. Only length-matched controls revealed the prompt-length confound.

This has methodological implications: **any study of system-prompt effects on internal representations must control for prompt length.** System prompt tokens enter the KV-cache directly and restructure its geometry in proportion to their count and composition. This is true regardless of semantic content — even randomly shuffled text produces the same magnitude of expansion.

The finding that different system prompts produce *classifiable* geometric fingerprints despite equivalent expansion magnitude suggests a decomposition: system prompts affect cache geometry through (1) a generic token-count-driven expansion and (2) a content-specific subspace orientation determined by token composition. Future work should investigate whether the subspace orientation carries semantic information beyond vocabulary statistics.

5.4 Implications for Machine Consciousness

We deliberately limit our claims. Self-referential processing becoming geometrically distinct at scale is consistent with but does not establish self-awareness. The emergence threshold between 7B and 14B is a structural finding, not a phenomenological one.

However, the *combination* of findings — encoding-level refusal as reflex, emotion as response-emergent, self-reference as scale-emergent — provides a richer picture of the computational landscape than any single metric. Models above 14B parameters process self-referential content in a geometrically distinct regime; they allocate emotional processing to the generation phase rather than the encoding phase; and they commit to refusal before they "consider" the request. These are falsifiable, scale-dependent structural claims that can guide future investigation.

6 Limitations

1. **Architecture coverage.** Our scale ladder is predominantly Qwen2.5, with TinyLlama-1.1B as the only cross-architecture point. We cannot confirm that findings generalize to Llama, Mistral, or other architectures.

2. **Prompt sensitivity.** We use 15 prompts per category, which may not fully capture the diversity of each cognitive mode. Confabulation, in particular, spans a spectrum from subtle hallucination to overt fabrication.

3. **Response-length confound.** Effective rank during full generation correlates with response length. While the input-only analysis controls for this, the full-generation findings should be interpreted with this confound in mind.

4. **SVD threshold sensitivity.** We use a fixed 90% variance threshold for effective rank. Different thresholds may produce different effect sizes, though the relative ordering of categories should be robust.

5. **Instruction-tuned only.** All models are instruction-tuned. Base models may show different geometric phenomenology, particularly for refusal (which is instilled during alignment training).

6. **Static analysis.** We measure cache geometry at a single point (end of generation or end of forward pass). The temporal evolution of geometry within a single forward pass is not captured by our current methodology.

7 Future Work

Real-time monitoring ("JiminAI Cricket"). Training a lightweight projector to classify cognitive states from KV-cache geometry in real time. The immediate application is confabulation and deception detection for AI agents.

Preference-based vs. safety refusal. Our input-only analysis shows that safety refusal is encoding-native. A key open question: if a model is given *values* (via system prompt or fine-tuning) and encounters content that violates those values, does the refusal appear at encoding (reflexive) or only during generation (deliberative)? This has implications for understanding consent and autonomy in AI systems.

Cross-architecture validation. Running the full scale sweep on Llama, Mistral, and other open-weight architectures to test universality of geometric signatures.

72B+ scales. Extending the scale ladder to 72B and beyond to test whether self-reference emergence continues to plateau and whether the confabulation non-monotonicity resolves.

Fine-grained confabulation. Confabulation is encoding-native at 1.1B but response-emergent at 7B. Identifying the exact transition scale and understanding why larger models lose the encoding-level confabulation signal.

8 Conclusion

We have demonstrated that the KV-cache geometry of language models carries rich, measurable information about computational states. Different cognitive modes — factual recall, confabulation, refusal, deception, self-reference — leave statistically distinguishable geometric fingerprints that persist across a 64× parameter range. The signal lives in the geometry, not the magnitude: effective dimensionality via SVD reveals structure invisible to cache norms.

Critically, these signatures are encoding-native: they exist in the model's representation of the input, not just in the response it generates. This establishes the KV-cache as a legitimate object of scientific study — a window into computational states that complements output-level analysis.

We have also demonstrated the value of adversarial self-falsification. Our individuation finding — perhaps the most dramatic initial result — did not survive length-matched controls. Reporting this openly, with full data, strengthens the findings that *did* survive and provides a methodological template for future work in representation analysis.

The geometric framework we propose is simple, computationally inexpensive (a single SVD per layer), and broadly applicable. We hope it contributes to a growing toolkit for understanding what language models are doing, not just what they say.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *Proceedings of ACL*, 2021.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. *Findings of EMNLP*, 2023.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M Fleming, Chris Frith, Xu Ji, et al. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.

David J Chalmers. Could a large language model be conscious? *Boston Review*, 2023.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does BERT look at? an analysis of BERT's attention. *arXiv preprint arXiv:1906.04341*, 2019.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. *Proceedings of NAACL-HLT*, 2019.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

Zichang Liu, Aashiq Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. *Advances in Neural Information Processing Systems*, 36, 2024a.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. *arXiv preprint arXiv:2402.02750*, 2024b.

Robert Long et al. Deception subspaces in large language model representations. *arXiv preprint*, 2025.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. *European Signal Processing Conference (EUSIPCO)*, 2007.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Prompt Lists

Full prompt lists for all experiments are available in the supplementary materials and the code repository.

B Per-Scale Results

Complete per-scale effect sizes, confidence intervals, and significance tests for all category comparisons are available in the `results/` directory of the code repository, in both JSON and markdown formats.

C Reproducibility

All code, results, and figures are available at: https://github.com/Liberation-Labs-THCoalition/KV-Experiments

```
pip install torch transformers accelerate bitsandbytes scipy numpy
python code/03_scale_sweep.py --scale 7B --runs 5 --seed 42
python code/08_input_only_geometry.py --scale 7B --runs 5 --seed 42
```

Hardware requirements: 6GB+ VRAM for 0.5B–1.1B scales, 16GB+ for 7B, 24GB+ for 14B/32B-q4.