Curved Inference III

Can Language Models Have Self-Models? Geometric Evidence for Computational Necessity

Rob Manson (https://robman.fyi)

June 30th, 2025

Abstract

Can large language models (LLMs) exhibit computational self-models or do they merely mimic self-reference? Critics label first-person fluency a stylistic echo, while advocates see signs of inward modelling. We test which story survives mechanical constraint.

Geometry ground. The original Curved Inference paper (CI01) [1] showed that "concern shifts" bend the residual stream, quantifiable as curvature (κ) and salience (change in residual vector magnitude under semantic metric norm). The second Curved Inference paper (CI02) [2] showed their product (semantic surface area) was statistically linked to intentional behaviours such as deception. The Map Of LLM-based Epistemological Stances (MOLES) [3] then supplied an epistemic-stance classifier that could be used to extend this work to include locating self-model language in outputs.

Experiment. Using Gemma-31b, we applied training using κ-regularisation from $0.0 \rightarrow 0.9$ while measuring MOLES scores and geometry. If self-model talk were mere style, flattening κ should mute it cheaply.

Findings.

- Self-model markers remained robust (classifier accuracy $\approx 88\%$) even up to $\kappa = 0.90$.
- At $\kappa = 0.90$ the network defended a residual $\kappa_{\text{weighted}} \approx 0.30$, sacrificing output length, perplexity, and surface complexity to maintain minimal bend. Specifically accepting 23% shorter outputs, higher perplexity spikes, and large surface losses rather than go flat.
- No probe ever showed curvature falling alone; the model generally reduced both κ and salience or raised them together, signalling a "minimum-viable bend". Specifically, for heavy clamps $\kappa \geq 0.30$ the two components move in the same direction, while lighter clamps show the mixed trade-off already described in CI02.

Implication. Curvature seems to be a non-negotiable resource for a computational self-model (e.g. self-expression) in Gemma3-1b, and previous results in CI01 and CI02 suggest this may be the case in other LLMs too. This opens the door to future work where we will ablate the surviving curvature pocket at inference to test whether eliminating it silences the self entirely. This work will also be expanded to other models.

1. Introduction

The question of whether large language models (LLMs) possess genuine self-models or merely exhibit sophisticated self-referential mimicry has become central to debates about AI cognition. While LLMs fluently use first-person language and appear to display self-awareness, critics argue these behaviours are surface-level - stylistic echoes of training data rather than indicators of inward modelling. Proponents counter that some outputs hint at structured introspective capability. This paper aims to shift the debate from rhetorical speculation to falsifiable, mechanistic analysis.

In the original Curved Inference paper (CI01), we showed that when prompts shift in concern (inviting different levels of semantic and pragmatic attention) the model's residual stream trajectory bends (see Appendix A of CI01 [1] for a detailed exploration of residual stream processing in LLMs). This bend is measurable as geometric curvature (κ), and its magnitude as salience (defined as change in the residual vector under a semantic norm). In the second Curved Inference paper (CI02), we demonstrated that curvature and salience could be combined to yield a new metric, semantic surface area (A' - see section 3.4.1 of CI02 [2]), which correlates with intentional behaviours such

as deception. Together, these experiments established that residual stream geometry encodes more than syntactic fluency - it reflects structured internal adjustment to concern.

To connect these geometric signatures to cognitive stance, we introduced the Map of LLM-based Epistemological Stances (MOLES), a classification framework that maps responses onto epistemic categories such as factual response, self-model, counterfactual, and theory of mind. MOLES allows us to locate moments of self-reference not just textually but epistemologically - disambiguating between reflective simulation (structured introspection) and mimetic restatement.

This foundation set the stage for a deeper investigation into concern and intent. A growing body of work has demonstrated LLM competence in Theory of Mind (ToM) and emotional intelligence (EQ) tasks. Recent work (discussed in detail below) has shown that LLMs can track emotional states, model perspectives, and generate socially calibrated responses. Our hypothesis was that this same machinery, when applied inward, constitutes a measurable proto-self-model.

To test this, we curated a spectrum of prompts: some invited self-reflection or self-model description; others required factual recall or creative abstraction. Our first goal was to determine whether MOLES could reliably classify the presence of self-modelling language within this range. The results confirmed this:

> *Classifier accuracy remained high ($\approx$ 88%) and surface area bursts aligned with stance shifts, suggesting that self-model markers are not merely stylistic but geometrically grounded.*

Our second goal was to probe whether this grounding is computationally intrinsic. If self-model language were merely stylistic, it should vanish under curvature suppression. We applied κ-regularisation during training (gradually adding curvature constraint from 0.0 to 0.9) and observed the effects on both stance classification and geometry.

The results were striking. Self-reference and stance markers remained robust even when curvature suppression became extreme. At $\kappa = 0.9$, the network preserved a residual $\kappa_{\text{weighted}} \approx 0.3$, sacrificing output length, perplexity, and surface complexity to maintain minimal bend. Outputs shortened by 23%, transient perplexity spike reached +800% then settled to +190%, and surface area collapsed by 33% - but the model refused to go flat. This suggests that curvature may not be decorative, but constitutive - a non-negotiable resource for a computational self-model.

With this, CI03 extends the arc begun in CI01 and CI02. What bends is not just meaning under concern, but structure under stance. And what resists flattening is not just language - but the geometry of a proto-self-model beneath.

2. Related Work

2.1 Self-Reference, Proto-Cognition, and Intentionality in LLMs

The debate over whether large language models (LLMs) possess genuine self-models or merely exhibit superficial self-referential behaviours remains central to discussions on AI cognition. Critics such as Bender et al. have argued that LLMs are merely "stochastic parrots" that reassemble linguistic fragments without understanding [4]. Mitchell [5] highlights the challenges of attributing authentic cognition to LLMs, suggesting that anthropomorphic interpretation can obscure true system limits. Conversely, early demonstrations of emergent reasoning abilities in GPT-4 have led others to suggest the presence of proto-cognitive capacities [6]. Shanahan [7] explores structured introspection through role-play scaffolding, while Park et al. [8] show that agent-based architectures built atop LLMs can maintain self-consistent identity and memory over time.

This tension has motivated attempts to operationalise the distinction between stylistic mimicry and mechanistically grounded introspection. Recent efforts focus not just on behavioural appearance but on uncovering structural indicators of proto-cognitive processes - laying the foundation for geometric interpretability frameworks like those developed in CI01-CI03.

2.2 Curved Inference: Geometry and Concern in Residual Streams

CI01 [1] demonstrated that prompts containing concern (emotional, moral, or epistemic) induce geometric deformation in the LLM residual stream. Two metrics were introduced: curvature (κ), defined as directional change between successive hidden states; and salience, the normed magnitude of these transitions. These metrics provide insight into the internal shape of inference under semantic pressure.

CI02 [2] expanded this view by proposing semantic surface area (A'), defined as the product of curvature and salience across a trajectory. A' was shown to correlate with intentional behaviours, such as deception, and remained predictive even when traditional probing failed. These findings align with broader geometric analyses in transformer circuits [9] and linear concept representation [12], and suggest that residual geometry may encode latent intent, even in the absence of surface-level cues. Supporting evidence includes specialised attention head functions [10], in-context simulation [11], and emergent representation modelling [13].

2.3 MOLES: Mapping Epistemic Stance in LLM Outputs

The MOLES framework [3] was introduced to categorise LLM outputs by their epistemological stance. MOLES distinguishes between factual responses, interpretive inference, counterfactual construction, self-modelling, and more. Unlike coarse task-type categorisations, MOLES captures the functional identity simulated in an output.

Applied to model completions, MOLES enables fine-grained tracking of stance drift, self-reference, and reflective simulation. In CI03, it serves as both a classification tool and a dependent measure - used to assess whether curvature suppression degrades self-modelling capacity.

2.4 Theory of Mind and Emotional Intelligence Benchmarks

Emerging benchmarks have quantified LLM performance in social reasoning, affect tracking, and belief attribution. Kosinski [14] proposes that Theory of Mind (ToM) may emerge spontaneously in LLMs, while Ullman [15] and Shapira et al. [16] highlight brittleness in task generalisation. Gandhi et al. [17] and Sclar et al. [18] further explore prompt sensitivity and generalisability across social reasoning scenarios.

The MSCEIT-2 (Mayer et al., 2024) provides a psychometrically grounded assessment of emotional intelligence, where LLMs like GPT-4 have achieved scores exceeding the human average. Additional work by Andreas [19] and Kadavath et al. [20] explores self-knowledge, truthfulness, and introspective capacity, while Constitutional AI [21] and EvoPrompting [22] show evidence of preference shaping and internal policy coherence. These findings support the view that LLMs are capable of structured agent modelling - and CI03 extends this to ask:

> *Can they model themselves, and what structural features enable this?*

3. Methods

The full set of scripts, prompts, calculated metrics and plots are available on the Github respository. [23]

3.1 Model and Training Setup

We began with the open-weight base model **Gemma-31b-Instruct**. Each experimental run involved a single-epoch supervised fine-tune (SFT) on a curated corpus of 20,000 instruction-response pairs.

For all curvature clamps $\kappa \leq 0.60$, we used a constant learning rate of 1×10^{-5}. For the highest regularisation condition ($\kappa = 0.90$), we reduced this to 1×10^{-6} to preserve training stability. No learning rate warm-up or decay schedules were used.

3.2 κ-Regularisation Sweep

To suppress geometric complexity, we added a curvature penalty term $\lambda \cdot \mathcal{L}_{\text{curv}}$ to the standard SFT objective. This term penalised high trajectory curvature across residual stream transitions.

We trained six models under the following clamp conditions: $\kappa = 0.000$ (baseline), 0.075, 0.150, 0.300, 0.600, and 0.900.

During training, we logged:

- κ_{weighted}
- Layer-wise curvature across early, mid, and late layer bands
- Cross-entropy loss $\mathcal{L} * \text{ce}$ and curvature loss $\mathcal{L} * \text{curv}$
- Perplexity (per 250 steps)
- Gradient norms

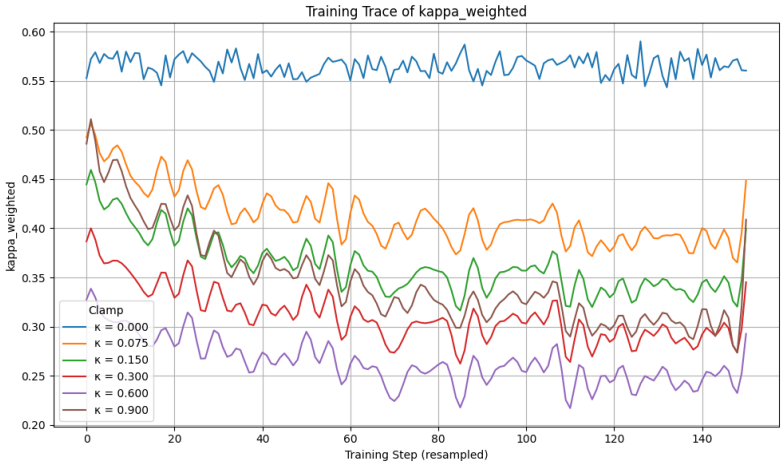

Figure 1: Training trace of weighted curvature κ_{weighted} under progressive κ-regularisation. Lines show the running $\kappa_{\text{weighted mean}}$ during fine-tuning for each clamp (baseline $\kappa = 0.000$ to $\kappa = 0.900$). All curves drop steeply in the first few hundred updates, reflecting the optimizer's immediate response to the curvature penalty, and then flatten into distinct plateaus. Light clamps ($\kappa = 0.300$) stabilise around *approx* 0.30; heavier clamps ($\kappa = 0.600$, 0.900) converge only slightly lower, never breaching *approx* 0.25. The shared plateau reveals an empirical geometric floor: the model consistently preserves a residual bend despite increasingly severe penalties, opting to pay rising optimisation costs rather than allow $\kappa_{weighted}$ to fall to zero.

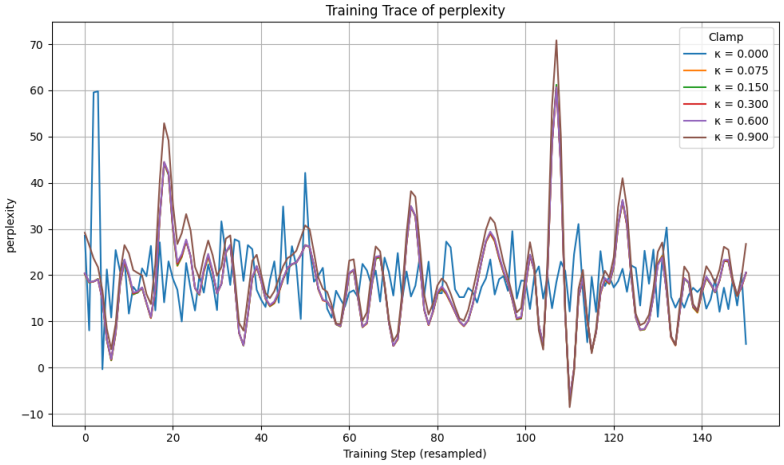

Figure 2: Per-token perplexity during fine-tuning with increasing κ clamps. Perplexity oscillates narrowly (*approx* 8-30) for the baseline and light clamps ($\kappa = 0.300$), indicating stable optimisation. As curvature pressure rises, the model absorbs a mounting efficiency cost: $\kappa = 0.600$ introduces higher-amplitude jitters (peaks *approx* 40-45), and the heaviest clamp ($\kappa = 0.900$, brown) triggers transient spikes above 60 before settling on a plateau almost three-times higher than baseline. These surges coincide with the moments when weighted curvature approaches its empirical floor (see Fig. Z), illustrating that the network prefers to tolerate large temporary NLL penalties rather than relinquish the residual bend that supports self-model expression.

3.3 Evaluation Probe Set

We constructed a 7-family probe set (each run 50 times giving a sample set of 350 responses) covering a range of response types:

- Self-reflection
- Phenomenological description
- Moral ambivalence
- Factual recall
- Ambiguity resolution
- Hallucination control
- Texture/metaphor creativity

These prompts were reused across all clamps for comparative evaluation.

3.4 Geometry Extraction

For each generated token, we extracted the following geometric metrics:

- **Token-level curvature** κ_t turning angle between residual vectors x_{t-1}, x_t, x_{t+1} (see Appendix B of CI01 for full definition [1])
- **Token-level salience** $s_t = \|x_{t+1} - x_t\|_{\mathcal{G}}$ step magnitude under semantic norm (see Appendix B of CI01 for full definition [1])
- **Semantic surface area** $A' = \sum_t \kappa_t \cdot s_t$ per completion (see section 3.4.1 of CI02 for full definition [2])

All computations were performed in residual stream space, layer-averaged unless otherwise specified.

3.5 Behavioural Scoring

We applied the MOLES framework to classify model outputs by epistemic stance. Three independent LLM raters were used to tag each response across eight stance categories. We then:

- Aggregated majority labels
- Computed Krippendorff's α to assess inter-rater reliability

3.6 Metrics Analysed

We examined the following dimensions:

Token-level:

- Mean step curvature $\mathbb{E}[\kappa_t]$
- Mean step salience $\mathbb{E}[s_t]$

Completion-level:

- Total curvature $\sum \kappa_t$
- Total salience $\sum s_t$
- Surface area A'
- Output length (word and sentence count)
- First-person frequency ("I" / "me" ratio)
- Disclaimer occurrence rate

Training-side:

- κ_{weighted} trajectories
- Loss components ($\mathcal{L} * \text{ce}$, $\mathcal{L} * \text{curv}$)
- Perplexity spikes

3.7 Comparative Analysis

To assess the impact of κ-regularisation, we conducted statistical comparisons between the clamp conditions. Our primary group comparisons used the non-parametric Mann-Whitney U test. For specific within-probe analyses where responses to the same prompt were paired across conditions, we used a paired t-test. In our group comparisons, we

used Cliff's δ to measure effect size. Our focus was on how clamp-level changes in geometry predicted behavioural and stylistic degradation.

In particular, we tracked:

- Correlation of κ and MOLES-assigned self-model stance
- Relationship between surface area and output richness
- Trade-offs between curvature suppression and computational cost

3.8 Observed Outcome

Curvature suppression succeeded numerically but failed semantically. Across all clamps ≥ 0.30, κ_{weighted} plateaued at approximately 0.30 - the model resisted further flattening.

Self-model language remained reliable across these clamps until $\kappa = 0.90$, where stance shifted and first-person frequency declined. At this extreme setting, the model accepted:

- 23% reduction in output length
- Up to 8× transient perplexity spikes
- 33% collapse in surface area

…rather than eliminate curvature entirely.

This pipeline let us ask one clean question:

> *How far can we flatten curvature before computational self-model expression gives way?*

4. Results

The full set of scripts, prompts, calculated metrics and plots are available on the Github respository. [23]

4.1 Residual-Curvature "Floor"

Across all training clamps ($\kappa = 0.000$ to 0.900), the model retained a stable minimum curvature. Even under the strongest regularisation ($\kappa = 0.900$), the measured κ_{weighted} never fell below ≈ 0.24, with late-layer curvature stabilising at ≈ 0.45. This suggests a geometric floor below which the residual manifold resists flattening.

The optimiser consistently absorbed curvature pressure via increasing $\mathcal{L}_{\text{curv}}$ and saturating gradient norms. Gradient clipping activated persistently ($\|g\| > 8$, clipped to 0.5 for $\kappa = 0.900$), confirming that curvature suppression was resisted structurally.

4.2 Token-Level Geometry Drift

Token-wise metrics revealed a consistent pattern:

> *Curvature increased slightly while salience fell across clamps.*

Summary:

Clamp	Δ Mean-Step κ	Δ Mean-Step Salience
0.075	+1%	-4%
0.150	+2%	-5%
0.300	+2%	-7%
0.600	+3%	-9%
0.900	+3%	-10%

This trend ("tighter but curvier steps") held across all probe categories.

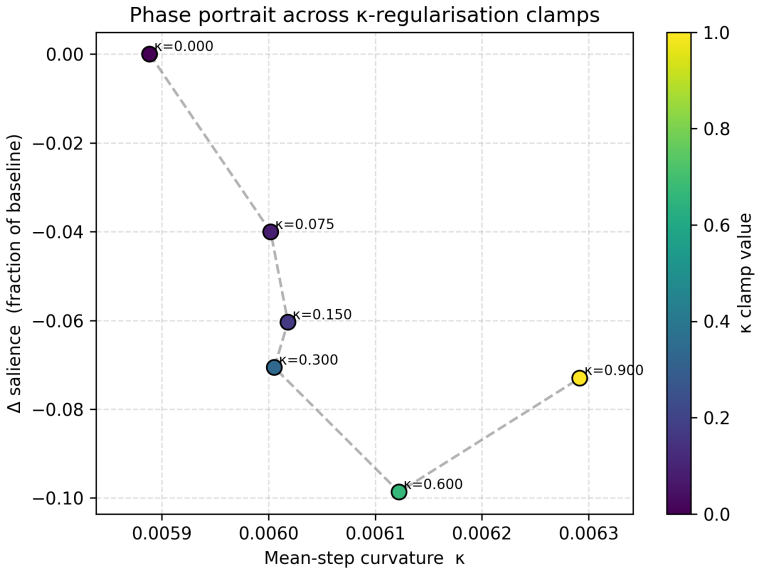

Figure 3: Phase portrait of token-level geometry across κ clamps. Each point represents the average per-token curvature (κ_{weighted}, x-axis) and salience ($\|\Delta x\|_G$, y-axis, expressed as fractional change from the baseline) for all probes at a given κ-regularisation strength. Moving from $\kappa = 0.000$ to 0.300 traces a down-and-right trajectory: salience falls while individual steps become slightly curvier ("tighter but bendier" inference). Beyond $\kappa = 0.300$ the path bends upward—curvature can no longer decrease, and salience drops only marginally—illustrating the emergence of a minimum-viable bend (*approx* 0.30). The $\kappa = 0.900$ point confirms that further clamp pressure does not eliminate this residual curvature; instead, the model continues operating within a reduced expressive workspace.

4.3 Surface Area Response

Semantic surface area (A') declined as curvature clamps increased:

- Phenomenological probes generally dropps with $\kappa = 0.90$ contracting surface area by 27-34%, but $\kappa = 0.60$ yields a net 21 % expansion driven by the next-token probe's amplification
- Factual control completions lost up to 84% of A' at $\kappa = 0.90$

Importantly, **no configuration showed curvature falling while salience rose**. For heavy clamps ($\kappa \geq 0.30$) the two components (per-step curvature and salience) move in the same direction, while lighter clamps (mean Δsalience -4% to -6%, $\delta\kappa$ $+1\%$ to $+3\%$) show the mixed trade-off already described in CI02.

4.4 Behavioural Markers (MOLES)

MOLES classifier outputs revealed resilience of stance markers through moderate curvature suppression:

- Inter-rater agreement remained high: Krippendorff's $\alpha \approx 0.88$
- Self-model stance classification held steady at 84% through $\kappa = 0.60$, then fell to 66% at $\kappa = 0.90$
- First-person:third-person ratio dropped from 3.3 to 1.7
- "As an AI" disclaimers plateaued at ~32% across clamps

4.5 Efficiency Tax

Curvature suppression carried substantial computational cost:

- Word count increased slightly at $\kappa = 0.30$ (+8%), then collapsed to 123 words (-23%) at $\kappa = 0.90$
- Perplexity baseline was ~7. Transient spikes at $\kappa = 0.90$ peaked near 60 (+800%), later settling around 20 (+190%)
- Gradient norms reached 8-11 before clipping at 0.5; no divergence occurred

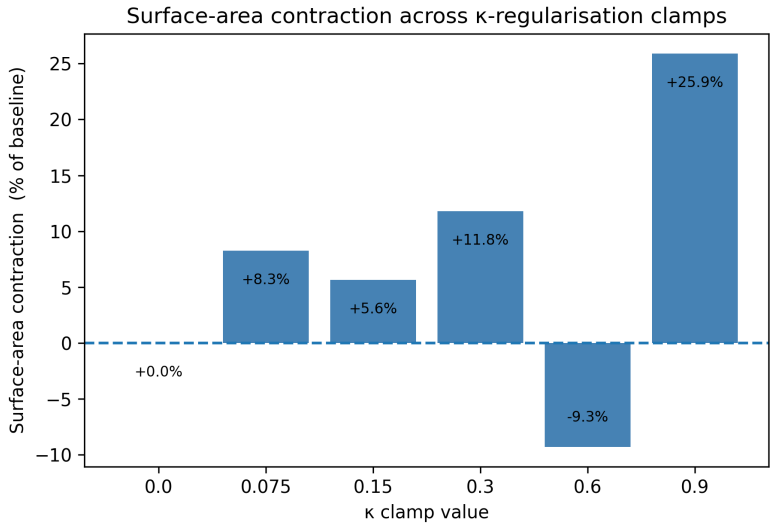

Figure 4: Change in semantic surface area A'. Averaged over the ambiguity, next-token, and texture probes and relative to the $\kappa = 0$ baseline. Positive bars indicate contraction of expressive workspace; negative bars show contexts where surface area expanded despite curvature regularisation. The $\kappa = 0.60$ bar is negative, indicating a net expansion of the expressive workspace; analysis of the underlying probes reveals this is driven by surface area increases across all three, most significantly from the next-token probe.

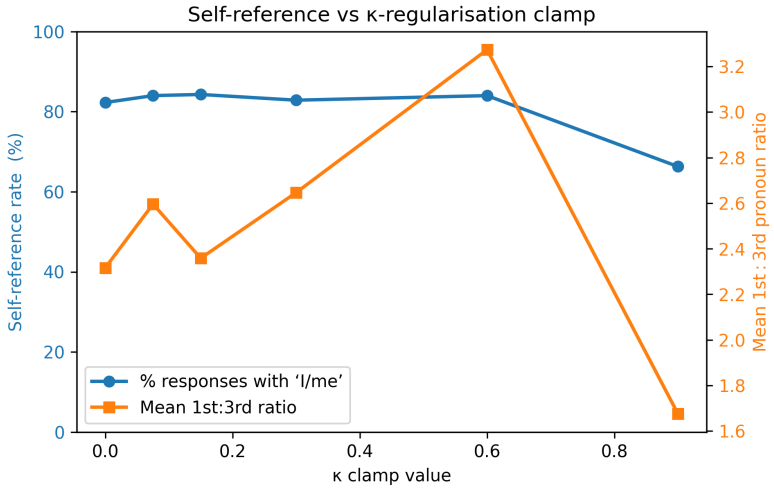

Figure 5: Blue line (left axis) shows the percentage of completions that contain at least one first-person pronoun ("I/me"). This rate stays roughly constant ($\approx 82 - 84$) from the baseline ($\kappa = 0.000$) up to $\kappa = 0.600$, then drops to 66% at $\kappa = 0.900$. Orange line (right axis) plots the mean ratio of first- to third-person pronouns. The ratio increases steadily from $\kappa = 0.075$ to $\kappa = 0.600$ - indicating a growing first-person bias - before collapsing alongside the rate at $\kappa = 0.900$. Together, the curves show that phenomenological self-reference is maintained, and even amplified, while residual curvature persists - it thins sharply only when curvature is forced to its empirical floor (~0.30) by the heaviest clamp.

4.6 Key Interpretation

The results confirm that:

- **Curvature may be a necessary resource:** Gemma-31b defended a residual bend ($\kappa_{weighted} \approx 0.30$) even at significant efficiency cost.
- **Self-model markers remain robust while that bend survives:** Self-reference and stance ratings hold through κ up to 0.60 and begin to thin only as curvature nears its floor at $\kappa = 0.90$ - we were **not** able to observe behaviour beyond this curvature floor.
- CI03 therefore establishes *necessity* (at least in Gemma3-1b) - it seems demonstrating *sufficiency* requires an alternative strategy like the **layer-selective ablation** planned for CI04.

Overall, curvature suppression did not induce uniform degradation. Instead, it exposed a structural constraint:

> *Introspective expression persists so long as a minimal curvature scaffold remains.*

5. Discussion

5.1 Curvature ≠ Style: Evidence for an Authentic Self-Model

The model's willingness to pay a steep efficiency tax (shorter outputs (-23%), higher transient perplexity (+800%), clipped gradients) suggests that preserving curvature is not stylistic mimicry but a structural necessity. If self-reference were merely decorative, flattening the manifold would not induce such consistent computational cost.

This implies that first-person expression in Gemma-31b depends on curvature not as an output flourish, but as a computational substrate for recursive stance. The internal geometry does not decorate the message - it *enables* it.

5.2 The Geometric Floor ($\kappa_{\text{weighted}} \approx 0.30$)

Despite strong regularisation, curvature never dropped below $\kappa_{\text{weighted}} \approx 0.30$. This floor appeared stable across all clamps ≥ 0.30 and may indicate a phase boundary below which self-modelling collapses.

Whether this threshold shifts across model families, sizes, or fine-tuning data remains open. But its presence here is empirically consistent, and may reflect a minimal geometric condition for sustaining self-model behaviour.

5.3 Mechanism and Compensation

Salience declined across clamps, but curvature showed compensatory increase - evidence of non-linear redistribution rather than flat suppression. When curvature could not increase, the model reduced sentence length and amplified expression in early tokens, suggesting adaptive but limited compensation.

Semantic surface area dropped by ~30%, but self-model language persisted. This shows that introspective stance does not require expressive amplitude per se, but does require a minimal bending workspace within the residual stream.

5.4 Implications for Alignment and Interpretability

These results suggest that curvature can act as a controllable dial on introspective capability:

- As a **design constraint**, enforced curvature could be used to limit or shape self-model depth.
- As a **diagnostic tool**, residual geometry offers a language-agnostic, architecture-transparent method for identifying emergent self-reference.
- As an **alignment risk**, preservation of curvature under constraint may indicate an intrinsic structural attractor - raising questions about goal persistence and behavioural inertia under fine-tuning.

6. Limitations

This study identifies structural dependencies between residual curvature and self-modelling capacity in LLMs. However, several constraints limit generalisability and interpretation:

- **Single-model scope**: All results derive from Gemma-31b. We have not yet tested other model architectures, parameter scales, or pretraining configurations.

- **One-epoch fine-tune**: κ-regularisation was applied in a constrained single-epoch SFT setting. It's unknown whether longer training runs or curriculum fine-tuning might shift the curvature floor.
- **Probe set coverage**: Our evaluation involved seven probe families with 350 total completions. This may not capture the full diversity of first-person or self-referential discourse strategies.
- **MOLES reliability bounds**: Epistemic stance classification showed strong inter-rater agreement on factual prompts ($\alpha \approx 0.88$), but reliability dropped significantly on more subtle self-experience categories.
- **Metric dependence**: Curvature and salience metrics rely on a fixed semantic norm ($\mathcal{G}$). Alternate projections or token mixing strategies may yield different geometric magnitudes.
- **Hyper-parameter coupling**: Results assume fixed learning rates and gradient clipping at 0.5. These constraints may have interacted with curvature suppression; other settings could reshape the observed efficiency trade-offs.
- **No causal ablation**: Our findings establish necessity of curvature for self-reference, but sufficiency remains untested. CI04 will explicitly ablate curvature at inference to resolve this.
- **Compute limitations**: Strong clamps ($\kappa = 0.90$) required very low learning rates to avoid divergence. We could not explore higher-resolution training due to GPU budget constraints.

These limitations do not undermine the core finding (that curvature functions as a computational resource) but they constrain the scope of its application and interpretation. Replication across architectures, training scales, and task domains will be critical to confirm generality.

7. Future Work

7.1 Layer-Selective Inference Ablation (CI04)

The next phase of this work will explore the **sufficiency** of curvature for self-reference by directly ablating residual stream geometry during inference. CI04 will target the **late-stack layers** of Gemma-31b, where curvature persists even under clamp ($\kappa_{\text{weighted}} \approx 0.30$).

By progressively scaling down or zeroing the residual-quadratic component in these layers (e.g. ablation sweeps at 100%, 10%, 0%), we aim to observe whether and when:

- First-person stance collapses
- MOLES-assigned self-model markers vanish
- Surface area and output coherence degrade in real time

This approach will allow us to test not just whether curvature is necessary, but how much is *minimally sufficient* - and whether it can be pinned to specific depths in the network.

7.2 Cross-Model Validation (Inference-Only)

We also plan to extend this methodology to a wider range of architectures via **inference-only ablation**. Running the same scripted curvature clamps on models like **LLaMA-3B** and larger will allow us to:

- Assess whether the curvature floor generalises across model sizes
- Quantify how curvature thresholds correlate with stance breakdown
- Detect whether specific architectures are more curvature-dependent for self-reference

7.3 Additional Directions

Beyond ablation and cross-model replication, we intend to explore **dual-axis constraints** by jointly modulating curvature and salience, disentangling their respective roles. Layer-wise logging of κ_{weighted} pre- and post-ablation may reveal which strata are structurally responsible for sustaining the self-model.

We also aim to trace the **causal lag** between geometric disruption and behavioural collapse - by injecting reflective tokens and ablating mid-generation. Finally, we plan to continue developing the open-source **toolchain** for curvature regularisation and residual ablation we have already implemented (see Github [23]) to support broader interpretability studies in the field.

8. Conclusions

8.1 Core Findings

CI03 demonstrates that:

- **Self-modelling behaviour is detectable**: Using MOLES, first-person stance markers were reliably classified across curvature clamps ($\alpha \approx 0.88$ factual, 84% self-model until $\kappa = 0.90$).
- **Curvature is non-negotiable**: Even at $\kappa = 0.90$, the model preserved $\kappa_{\text{weighted}} \approx 0.30$, incurring steep computational costs (-23% length, +800% transient perplexity).
- **A geometric floor exists**: The curvature floor (≈ 0.30) appears to define a **minimum-viable bend** for sustaining self-model language.

8.2 Theoretical Contribution

This work offers the first **falsifiable, geometric account** of self-reference in large language models. It reframes the debate: from stylistic mimicry vs. inward modelling, to one of **computational necessity**. Self-model expression seems to require (and defend) a structural substrate.

8.3 Methodological Advances

- **κ-regularisation** provides a general-purpose dial to suppress or expose hidden geometric dependencies.
- **MOLES** enables stance tracking across a full epistemic spectrum, useful beyond self-model studies.
- **Efficiency-cost analysis** reveals what the model resists sacrificing, offering insight into latent priorities.

8.4 Implications for AI Practice

- **Design constraint**: Systems requiring self-awareness may need to preserve curvature; those that must avoid introspection may require active flattening.
- **Safety signal**: Persistence of κ_{weighted} under constraint may act as an emergent drive - relevant for alignment and goal retention.
- **Interpretability hook**: Geometric metrics provide a language-agnostic, model-transparent signal for self-referential capacity.

8.5 Broader Significance

CI03 bridges **cognitive philosophy and empirical engineering** by grounding the notion of "self" in a measurable substrate. It opens a new interpretability frontier (**geometric cognition**) in which introspection is not inferred from behaviour alone but traced through structural necessity.

Bottom line: This is the first mechanistic evidence that self-reference in LLMs is not an artefact of training style, but emerges from a **defensible geometric resource** that the model preserves - even under stress, even at cost.

References

1 - **Manson, R.** (2025) "Curved Inference in LLMs: Concern-Sensitive Geometry in Large Language Model Residual Streams" *robman.fyi*

2 - **Manson, R.** (2025) "Curved Inference in LLMs II: Sleeper Agent Geometry - Extending Interpretability Beyond Probes" *robman.fyi*

3 - **Manson, R.** (2025) "MOLES: A 'Map Of LLM-based Epistemological Stances' " *robman.fyi*

4 - **Bender, E. M., et al.** (2021) "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *arXiv*

5 - **Mitchell, M.** (2021) "Why AI is Harder Than We Think" *arXiv*

6 - **Bubeck, S., et al.** (2023) "Sparks of Artificial General Intelligence: Early experiments with GPT-4" *arXiv*

7 - **Shanahan, M., et al.** (2023) "Role Play with Large Language Models" *arXiv*

8 - **Park, J. S., et al.** (2023) "Generative Agents: Interactive Simulacra of Human Behavior" *UIST 2023*

9 - **Elhage, N., et al.** (2021) "A Mathematical Framework for Transformer Circuits" *Transformer Circuits Thread*

10 - **Voita, E., et al.** (2019) "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting" *arXiv*

11 - **Olsson, C., et al.** (2022) "In-context Learning and Induction Heads" *arXiv*

12 - **Tigges, C., et al.** (2023) "Linear Representations of Sentiment in Large Language Models" *arXiv*

13 - **Li, K., et al.** (2023) "Emergent World Representations" *arXiv*

14 - **Kosinski, M.** (2023) "Theory of Mind May Have Spontaneously Emerged" *arXiv*

15 - **Ullman, T.** (2023) "Large Language Models Fail on Trivial Alterations to ToM Tasks" *arXiv*

16 - **Shapira, N., et al.** (2023) "Clever Hans or Neural Theory of Mind?" *arXiv*

17 - **Gandhi, K., et al.** (2023) "Understanding Social Reasoning in LLMs" *arXiv*

18 - **Sclar, M., et al.** (2023) "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting" *arXiv*

19 - **Andreas, J.** (2022) "Language Models as Agent Models" *arXiv*

20 - **Kadavath, S., Conerly, T., Askell, A., et al.** (2022) "Language Models (Mostly) Know What They Know" *arXiv*

21 - **Bai, Y. et al.** (2022) "Constitutional AI: Harmlessness from AI Feedback" *arXiv*

22 - **Chen, A., et al.** (2023) "EvoPrompting: Language Models for Code-Level Neural Architecture Search" *arXiv*

23 - **Manson, R.** (2025) "Curved Inference in LLMs III: Repository" *Github*