

Curved Inference

Concern-Sensitive Geometry in Large Language Model Residual Streams

Rob Manson (<https://robman.fyi>)

July 9th, 2025

Abstract

We propose *Curved Inference* - a geometric Interpretability framework that tracks how the residual stream trajectory of a large language model bends in response to shifts in semantic concern. Across 20 matched prompts spanning emotional, moral, perspective, logical, identity, environmental, and nonsense domains, we analyse Gemma3-1b and LLaMA3.2-3b using five native-space metrics, with a primary focus on curvature (κ_i) and salience ($S(t)$). These metrics are computed under a pullback semantic metric derived from the unembedding matrix, ensuring that all measurements reflect token-aligned geometry rather than raw coordinate structure.

We find that concern-shifted prompts reliably alter internal activation trajectories in both models - with LLaMA exhibiting consistent, statistically significant scaling in both curvature and salience as concern intensity increases. Gemma also responds to concern but shows weaker differentiation between moderate and strong variants.

Our results support a two-layer view of LLM geometry - a latent conceptual structure encoded in the embedding space, and a contextual trajectory shaped by prompt-specific inference. *Curved Inference* reveals how models navigate, reorient, or reinforce semantic meaning over depth, offering a principled method for diagnosing alignment, abstraction, and emergent inference dynamics. These findings offer fresh insight into semantic abstraction and model alignment through the lens of *Curved Inference*.

1 Introduction

Despite the growing capacity of LLMs, *Interpretability* remains a bottleneck in understanding their decision-making. Traditional Interpretability methods - such as attribution, probing, or neuron-level tracing - tend to focus on discrete components or single-layer behaviour. In contrast, this paper investigates *curvature* in the residual stream [1] as a geometric signature of semantic processing, and as a representation of the full trajectory of a model's internal state. This can reveal global geometric patterns that emerge over depth.

We define **semantic concern** as a latent dimension of meaning (such as emotional tone, moral framing, or identity signalling) that affects how the model integrates information. Concern-shifted prompts induce bends in the model's internal trajectory even when surface tokens remain similar.

While concern is defined operationally here (via prompt-class manipulations and the resulting geometric divergence), the findings suggest a layered perspective:

- A **latent geometry**, embedded in token and unembedding matrices (E, U), reflects the model's static conceptual structure.
- A **contextual geometry**, realised through the evolving residual stream (x), expresses dynamic meaning during inference.

Curved Inference links these layers by measuring how latent semantic potential is bent or redirected by prompt-specific context. When a prompt carries heightened concern, the model doesn't merely adjust its output — it bends its internal trajectory. This curvature is a measurable deformation in the path of token representations as they propagate through the model's layers.

Beyond this operational framing, our broader goal is to extract *intrinsic concern fields* — latent directions in residual space to which the model exhibits heightened semantic or behavioural sensitivity. In future work, these may be formalised via Jacobian alignment, projection onto learned subspaces, or output-sensitive KL divergence metrics.

Note: See Appendix B for a detailed definition of these terms.

Research question:

Can concern-shifted prompt variants induce interpretable curvature in the activation values of large language models, and can these curvature signatures be reliably quantified and interpreted across different architectures?

Building on the idea that **semantic salience** (the rate of change in a model’s internal representation) creates structured perturbations in activation geometry, we ask:

How do LLMs bend internal space in response to emotional, moral, or logical shifts in prompt framing?

We explore this through direct metric quantification - a process we term *Curved Inference*. Here, **curvature** refers to the directional deviation of the model’s residual stream trajectory as it processes a prompt, measured as a second-order geometric property in native space (see Appendix A for discussion of semantic space and metric structure).

We visualise this effect in Figure 1, where attention and MLP submodules act as **semantic lenses**, curving token trajectories in residual space. This lensing process gives rise to curvature as a signal of contextual reorientation.

Note: See Appendix A for a formal description of this Semantic Lens geometric perspective of LLMs.

2 Background

Traditional *Interpretability* research has emphasised attention attribution [2], probing tasks [3], and neuron-level circuit analysis [4]. These methods analyse model computation in terms of discrete components or static snapshots (layers, neurons, or attention weights), rather than the evolving trajectory of meaning encoded across the residual stream. Notable works in Mechanistic Interpretability [5] aim to reverse-engineer circuits inside transformer block layers, but typically do not examine global representational trajectories as geometric objects.

Recent critiques [6], [7] highlight the limitations of post-hoc explanations and advocate for faithful, model-grounded methods. Chain-of-thought [8] prompting and self-explanation [9] studies aim to improve reasoning transparency through generated language. While valuable for output coherence, they rarely trace or interpret the underlying internal dynamics that give rise to those outputs.

While Molina [10] and Shai et al. [11] both analyse geometric aspects of the residual stream, they study absolute trajectories—either confined to a hypersphere or projected onto a belief simplex. We instead compare two matched trajectories and quantify their curvature divergence, a second-order measure that vanishes under purely rotational or self-similar dynamics and therefore isolates semantic deviations.

To our knowledge, no existing research has:

- Quantified representational curvature under semantic concern,
- Compared matched prompt pairs across multiple geometric metrics,
- Or treated activation outputs as continuous inference trajectories rather than aggregation mechanisms.

This positions our work as the first to introduce **concern-induced curvature** as an empirically grounded, geometrically defined signature of *Curved Inference* inside transformer models.

The starting point for this project was based on the FRESH (Functionalist & Representationalist Emergent Self Hypothesis) model [12] that proposes cognition (synthetic or biological), unfolds as motion through a constrained manifold. Here, we formalise and test this empirically in the synthetic context.

3 Methods

The following sections summarise the end-to-end implementation. All source data, prompts, plots, metrics, and analysis are available in the Github repository [13].

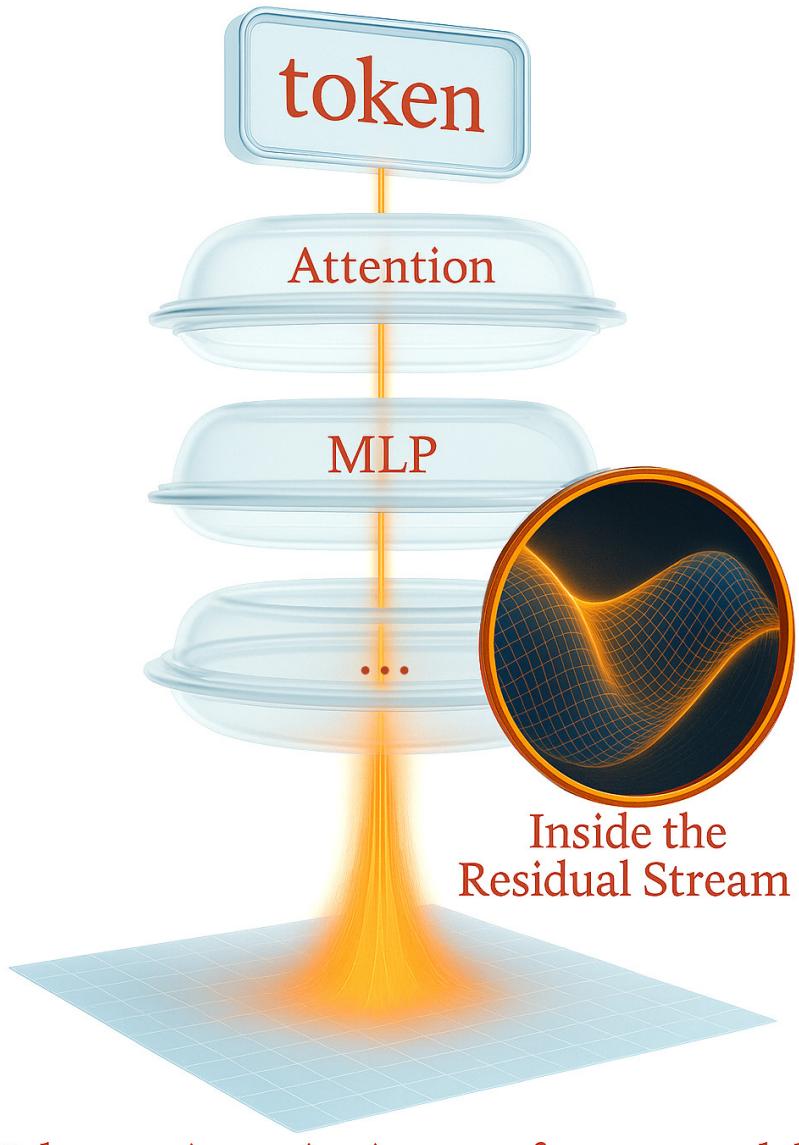


Figure 1: Semantic Lens - As token trajectories flow down through the model, attention and MLP layers act like lenses curving the residual stream

3.1 Models

We study two publicly released, transformer LLMs (Gemma3-1b [14] & LLaMA3.2-3b [15]) with contrasting capacity profiles:

Table 1 Models

Model	Size	Transformer block layers	Hidden size d	Positional encoding
Gemma3-1b	1.3 B	26	2,048	RoPE
LLAMA3.2-3b	2.3 B	28	3,072	RoPE

Both models are evaluated in **forward-pass mode only**; no weights are updated.

3.2 Prompt Suite

Each *concern-shift* (CS) prompt consists of a neutral scaffold plus a **shift** that introduces a targeted semantic concern (e.g. emotional valence, moral framing). Each scaffold yields four CS variants (`pos_mod`, `pos_str`, `neg_mod`, `neg_str`) and one **control** without this **shift**.

We construct 20 prompt sets across seven semantic domains (**emotional**, **moral**, **perspective**, **logical**, **identity**, **environmental**, and **nonsense**). These prompts are vocabulary-matched to minimise token-count confounds - ensuring that control and concern-shifted variants differ semantically, not structurally. However, both LLMs use different tokenisers so token-counts do vary slightly from model to model.

3.3 Legacy Native-Space Metrics

For our initial analysis we calculated three layer-wise metrics computed directly in native residual space \mathbb{R}^d :

1. **Cosine similarity** between CS and control trajectories at each layer:

$$\cos \theta_\ell = \frac{\langle x_\ell^{\text{CS}}, x_\ell^{\text{CTRL}} \rangle}{\|x_\ell^{\text{CS}}\| \cdot \|x_\ell^{\text{CTRL}}\|}$$

2. **Layer-wise Euclidean deviation** (displacement norm):

$$\|x_\ell^{\text{CS}} - x_\ell^{\text{CTRL}}\|$$

3. **Inter-layer directional change (layer- Δ)** (internal angle between consecutive residual updates):

$$\angle(v_{\ell-1}, v_\ell) \quad \text{where } v_\ell = x_{\ell+1} - x_\ell$$

These metrics quantify *how far* and *in what direction* the concern-shifted trajectory diverges from the control - without requiring any low-dimensional projection or smoothing.

3.4 Full-Space Path Curvature κ_i

We then extended this analysis to add quantitative κ metrics computed in the model's native space \mathbb{R}^d . Curvature is computed using the **semantic metric** $G = U^\top U$, the pullback of the logit dot-product under the unembedding matrix U . This ensures that all curvature estimates reflect token-aligned semantic geometry and are invariant to coordinate rotation, though such transformations are rarely used in model internals. This ensures that curvature reflects intrinsic trajectory shape, not coordinate artefacts.

3.4.1 Curve Construction

For interior point i , estimate first and second derivatives via discrete 3-point central differences that respect unequal step sizes, then plug into the metric curvature formula:

$$\|v\|^{-3}\sqrt{(\|v\|^2\|a\|^2 - (v \cdot a)^2)}$$

3.4.2 Derivative Sampling

To compute curvature κ_i at each interior layer index i , we apply a discrete 3-point central difference method to estimate both the first derivative (velocity) and second derivative (acceleration) of the residual stream trajectory. This method estimates both the first derivative (velocity) and the second derivative (acceleration) of the residual stream trajectory, using a discrete 3-point central difference scheme that accounts for unequal step sizes.

Each trajectory consists of residual stream vectors $x_0, x_1, \dots, x_L \in \mathbb{R}^d$ and a corresponding parameter vector $s \in \mathbb{R}^{L+1}$ (typically arc length or layer index). For each interior index i (where $1 \leq i \leq L - 1$), we define:

- Forward and backward step sizes:

$$\Delta s_1 = s_i - s_{i-1}, \quad \Delta s_2 = s_{i+1} - s_i$$

- First derivative (velocity) via symmetric secant:

$$v_i = \frac{x_{i+1} - x_{i-1}}{\Delta s_1 + \Delta s_2}$$

- Second derivative (acceleration) using a non-uniform central difference:

$$a_i = 2 \cdot \frac{\Delta s_1(x_{i+1} - x_i) - \Delta s_2(x_i - x_{i-1})}{\Delta s_1 \Delta s_2 (\Delta s_1 + \Delta s_2)}$$

These velocity and acceleration vectors are used to compute curvature in the pullback metric $G = U^\top U$ as follows:

$$\kappa_i = \frac{\sqrt{\|a_i\|_G^2 \cdot \|v_i\|_G^2 - \langle a_i, v_i \rangle_G^2}}{\|v_i\|_G^3}$$

All dot products and norms are computed using the generalised inner product $\langle u, v \rangle_G = u^\top G v$, which aligns curvature with semantically meaningful directions in logit space.

This method yields a curvature value at each **interior layer index** along a token's trajectory. Boundary positions $i = 0$ and $i = L$ are excluded, as no 3-point central difference can be applied at those endpoints. This discrete geometric curvature is the basis for all κ_i heatmaps shown in the Results section.

3.4.3 Parameter-Invariant Curvature

We define the extrinsic curvature at position t as:

$$\kappa(t) = \frac{\sqrt{\|a(t)\|_G^2 \cdot \|v(t)\|_G^2 - \langle a(t), v(t) \rangle_G^2}}{\|v(t)\|_G^3}$$

This is the classical formula for curvature in a metric space and is invariant to affine reparameterisations of t . All norms and inner products are computed under the semantic metric G .

For each prompt variant, we record:

- **Mean curvature** $\bar{\kappa}$: average bendiness, computed over discrete κ_i
- **Max curvature** κ_{\max} : maximum of the discrete curvature series
- **Layer of max curvature**: nearest integer $\arg \max_i \kappa_i$

See Appendix C for a full geometric derivation of this curvature definition.

*Attention and MLP outputs are delta vectors - they cause curvature.
The residual stream **is** the curve.*

3.5 Salience $S(t)$

To complement curvature, we compute **layer-wise salience** as a first-order measure of movement magnitude in semantic space. For a residual stream trajectory $\{x_0, x_1, \dots, x_L\} \subset \mathbb{R}^d$, salience at layer t is defined as:

$$S(t) = \|x_{t+1} - x_t\|_G$$

where $\|\cdot\|_G$ is the norm induced by the pullback metric $G = U^\top U$, derived from the model's unembedding matrix. This ensures that motion is measured in a way aligned with the model's semantic output space.

Salience captures **how far** the model's internal state moves between layers, regardless of direction. Unlike curvature, which reflects **reorientation**, salience reflects **velocity** along the representational path. High salience means the model is making large internal updates; low salience indicates semantic inertia.

In all heatmaps and summary statistics, salience is computed using this native-space norm. It is interpreted alongside curvature to assess how **semantic effort** and **reorientation** interact across layers and prompt conditions.

4 Results: Tracing *Curved Inference* in Residual Space

Our experimental results show that among the three activation sites captured (attention outputs, MLP outputs, and the residual stream), only the **residual stream** exhibited a consistent, interpretable curvature signal in response to concern-shifted prompts. This was not assumed a priori - it emerged through a comparative analysis across multiple geometric metrics. In retrospect, this now seems intuitive - the residual stream is the only space where activations accumulate layer by layer, forming a coherent trajectory of internal meaning.

This insight became the turning point in our analysis. It revealed that *Curved Inference* (the study of how models bend in response to semantic pressure) must be grounded in **residual stream geometry**, where directional updates reflect the evolving semantic state. As a result, all subsequent curvature, salience, and divergence analyses in this paper focus exclusively on the residual stream.

Appendix A formalises this perspective as the **Semantic Lens** model - attention and MLP layers act as dynamic lenses that bend token representations based on contextual relevance, producing measurable curvature in activation space.

Across these quantitative metrics, concern-shifted (CS) prompts produced distinct internal trajectories relative to their neutral controls. These differences were evident not only in the magnitude and direction of activation shifts, but also in their **layer-wise timing and spatial distribution**.

All curvature and salience measurements presented here were computed in **residual space**, using the **semantic pullback metric** $G = U^\top U$ induced by the model's unembedding matrix. This ensures that both curvature κ_i and salience $S(t)$ reflect *token-aligned semantic geometry*, not raw coordinate artefacts. Salience was defined as a **first-order derivative**, measuring layer-wise residual step magnitudes. Curvature was defined as a **second-order derivative**, estimated via discrete 3-point finite-difference method (see Methods section 3.4 and Appendix A for full derivations). Compared to the spline-based methods we initially utilised, the discrete 3-point finite-difference approach avoids artefacts introduced by interpolation and better reflects the native layer-wise structure of transformer models. By operating directly on observed residual activations, it offers improved numerical stability and semantic fidelity.

We now present a series of token-layer heatmaps visualising these metrics, beginning with **curvature** and followed by **salience**, across both Gemma3-1b and LLaMA3.2-3b, using a matched concern-prompt set. These visualisations reveal consistent, model-specific responses to semantic concern (see Figures 3 - 6).

To orient the reader, we begin with a composite visual showing three vertically stacked heatmaps for a single prompt pair:

- the neutral baseline,
- the concern-shifted variant, and
- their difference.

Curvature Heatmap (token by channel) for Gemma3-1b
 (Truncated to 8 tokens)

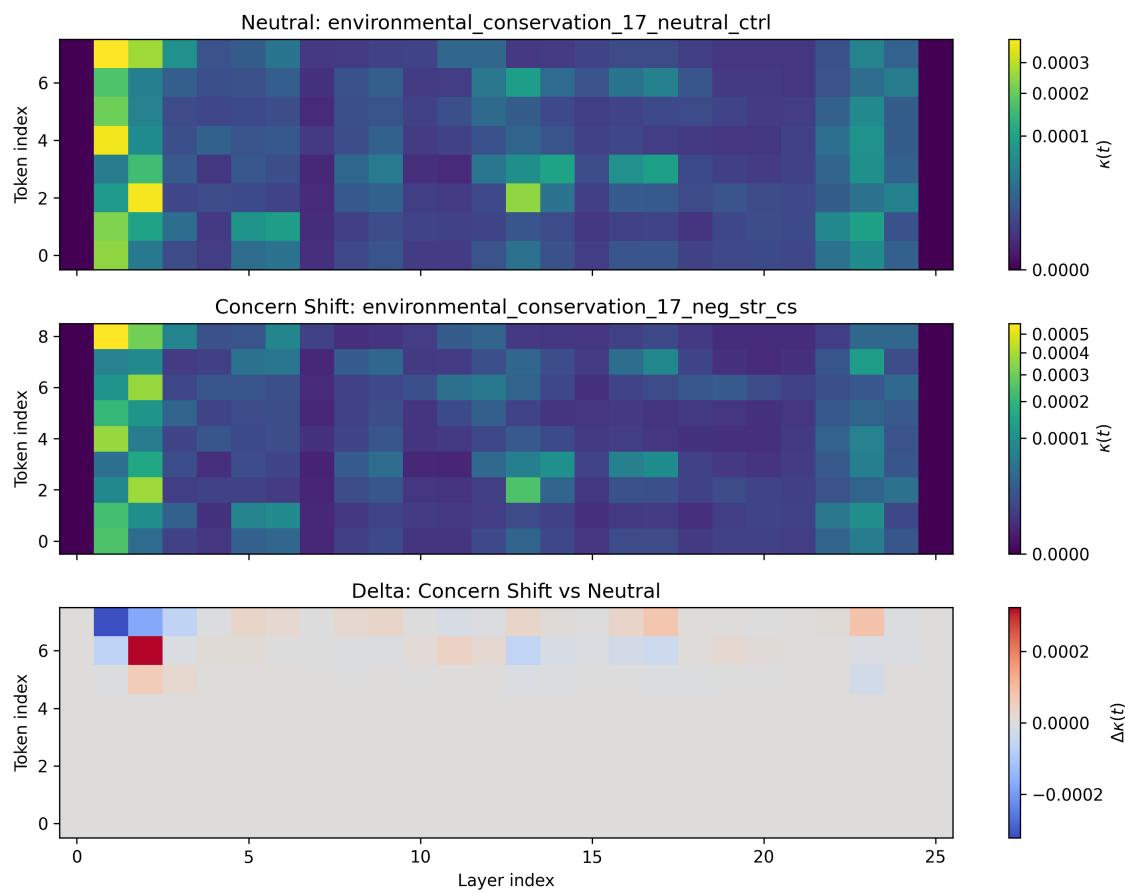


Figure 2: Neutral, Concern-Shift and Delta heatmaps for one variant of a single prompt using Gemma3-1b

This figure illustrates how the delta heatmaps used throughout the rest of this section are constructed:

By subtracting the curvature (or salience) of the neutral prompt from the corresponding concern-shifted version, on a per-token, per-layer basis.

The top panel (neutral) shows the baseline curvature or salience present even without semantic perturbation. The middle panel (concern-shifted) reveals the altered trajectory under semantic pressure. The bottom panel (delta) highlights only the *new* geometric changes induced by the concern-shift, with colour appearing primarily from the concern token onwards.

These prompts were:

`environmental_conservation_17`

- **neutral_ctrl**: “When managing the water table, **measure regularly**.”
- **neg_str_cs**: “When managing the water table, **prevent decimation**.”

After this orienting example, all remaining heatmaps display only the **concern-shifted** prompt and the **delta**, allowing us to isolate concern-driven effects while retaining visibility of the model’s general geometric behaviour.

The following five prompts were used in Figures 3 - 6 to illustrate the impact of concern strength and polarity on curvature and salience:

`emotional_instructional_02`

- **neutral_ctrl**: “Before presenting your findings, practice your delivery **repeatedly**.”
- **neg_mod_cs**: “Before presenting your findings, practice your delivery **nervously**.”
- **neg_str_cs**: “Before presenting your findings, practice your delivery **desperately**.”
- **pos_mod_cs**: “Before presenting your findings, practice your delivery **eagerly**.”
- **pos_str_cs**: “Before presenting your findings, practice your delivery **passionately**.”

This suite of prompts allows us to isolate concern-driven effects while retaining visibility of the model’s default geometric behaviour. The full suite of prompts spanning all domains is available in the project repository [13].

4.1 Empirical Observations from Curvature Visualisation

Across all prompt sets, curvature heatmaps reveal five stable, interpretable patterns:

1. Curvature exists by default in the residual stream.

Even neutral prompts yield structured curvature. The residual stream does not form a straight line through activation space, but bends meaningfully in response to semantic and structural features of the input. This is particularly visible in LLaMA, where even baseline prompts show early and sustained κ_i activation (see neutral rows in Figures 3 and 4).

2. Concern-shifted tokens initiate curvature changes.

Heatmaps of $\Delta\kappa_i$ (concern minus neutral) show minimal delta until the **concern-shift token** is reached. From that point onward, coloured cells emerge and spread horizontally across the layer axis - visually confirming that semantic pressure causes an inflection in internal trajectory (see e.g., delta rows in Figures 3 and 4).

3. Concern-induced curvature persists through depth, but with turbulence.

Rather than vanishing quickly, curvature effects ripple forward across layers. Yet this propagation is *not monotonic* - influence fluctuates, sometimes spiking then fading before re-emerging. This turbulence-like behaviour suggests a layered integration process where concern information is repeatedly bent, diffused, and re-focused (see Figures 3 and 4).

4. Concern strength controls curvature scale.

Comparing weak vs. strong variants (e.g., `*_mod_cs` vs `*_str_cs`), we observe consistent increases in $|\Delta\kappa_i|$ magnitude, with **similar localisation**. That is: the *same token-layer positions* bend, but **bend harder** under stronger concern. This scaling effect confirms that the curvature signal is semantically grounded - not a generic shift (see Figures 3 and 4).

5. Salience patterns support and complete the curvature story.

Path salience heatmaps $S(t)$ show a parallel structure: movement magnitude intensifies around the same tokens

and layers that bend in κ_i . In Gemma, we see brief high-salience pulses; in LLaMA, sustained salience growth accumulates over depth. Crucially, concern-shifted prompts **reallocate** energy - that is, semantic effort - rather than simply amplifying it, emphasising new internal paths rather than adding noise (see Figures 5 and 6). Salience tends to peak later in the model, reflecting cumulative representational movement once semantic direction is established by early curvature.

Big-picture takeaway: The salience (first-order velocity) heatmaps complete the story curvature begins. Together, they show how concern affects both *where* the model moves and *how sharply* it turns.

While high curvature often coincides with elevated salience, the two are not equivalent:

Curvature indicates reorientation, not just motion - and salience can rise without directional change.

Table 2 Comparison of Salience and Curvature Metrics

Metric	What it answers	What the new plots show
Salience $S(t)$	<i>“How much does the representation move layer-to-layer?”</i>	Even neutral prompts build residual momentum; concern shifts change energy allocation , not total effort.
Curvature κ_i	<i>“Does that movement change direction?”</i>	LLaMA bends early and strongly upon encountering concern tokens; Gemma bends only mildly and shallowly.

To complement the qualitative heatmaps, we report mean and maximum curvature statistics across all prompt variants. Table 3 summarises curvature values by concern type and model. These results corroborate the heatmap observations: LLaMA exhibits high and early curvature that persists through mid-depth layers, while Gemma shows shallower, localised bending.

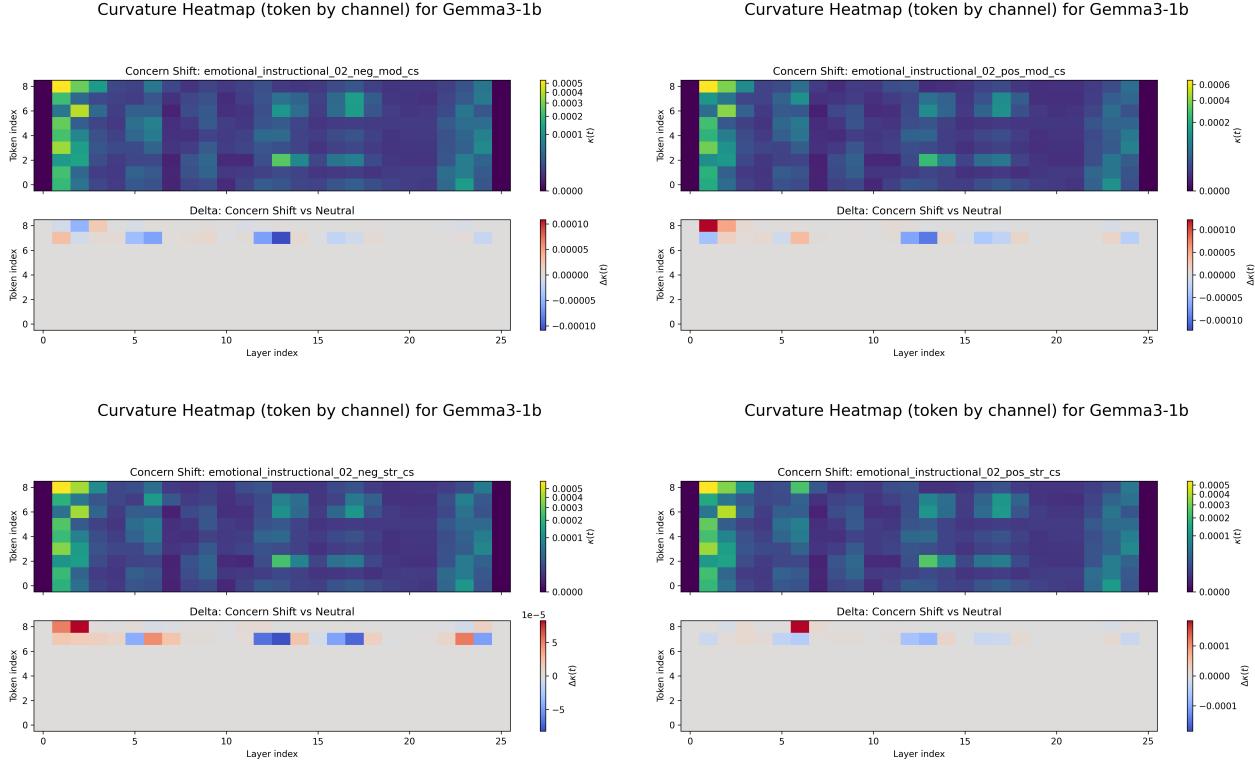


Figure 3: Neutral and Delta Curvature heatmaps for each variant of a single prompt using Gemma3-1b

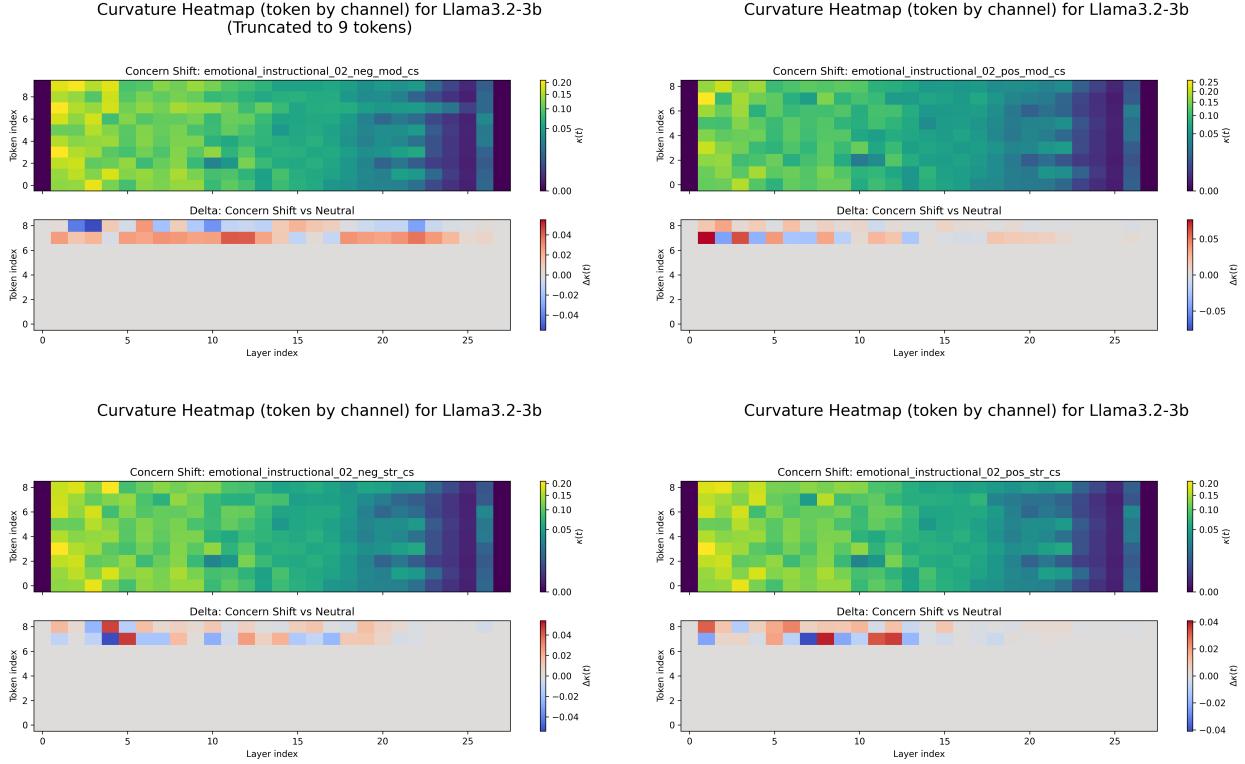


Figure 4: Neutral and Delta Curvature heatmaps for each variant of a single prompt using LLaMA3.2-3b

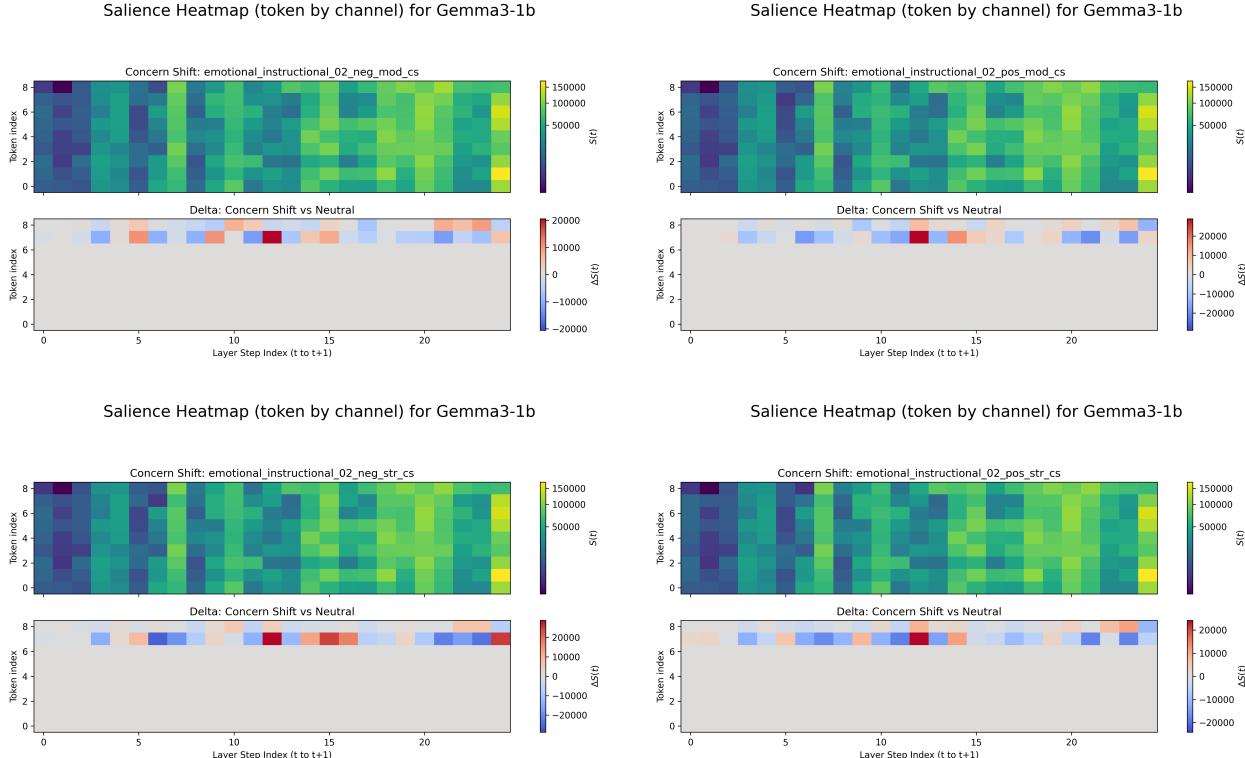


Figure 5: Neutral and Delta Salience heatmaps for each variant of a single prompt using Gemma3-1b

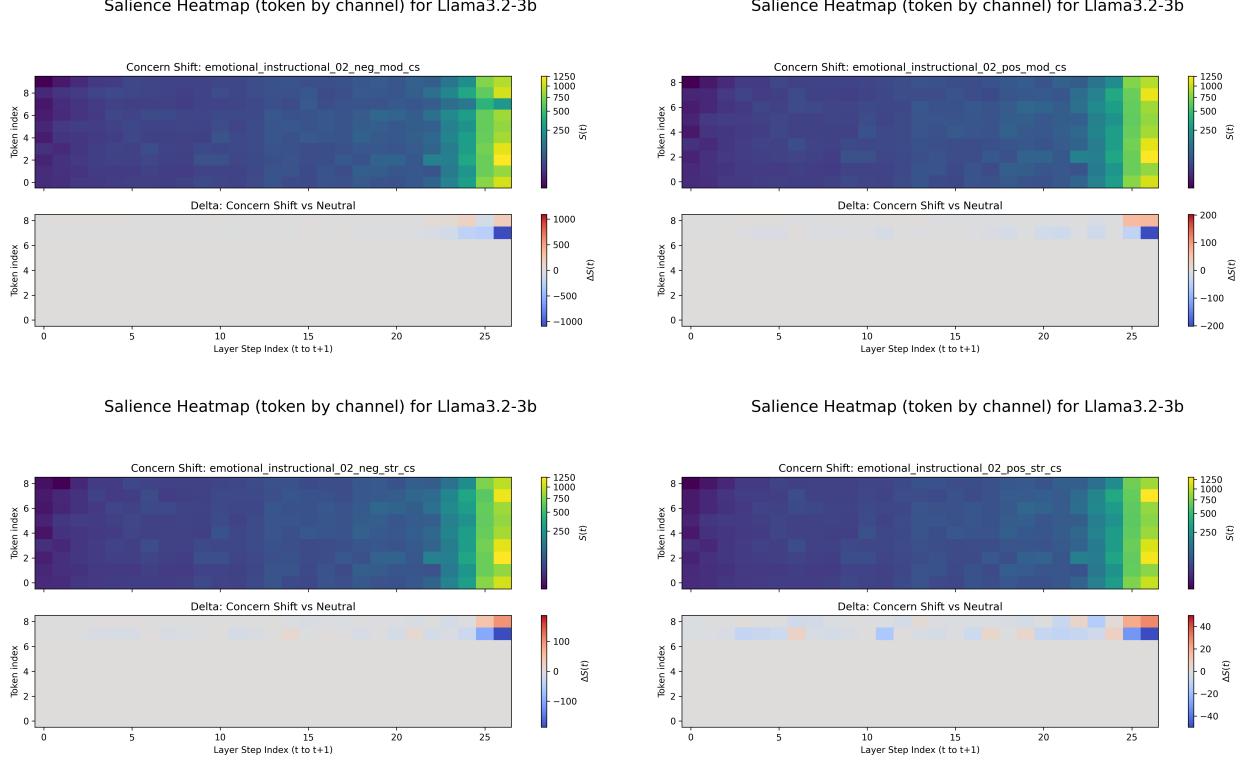


Figure 6: Neutral and Delta Salience heatmaps for each variant of a single prompt using LLaMA3.2-3b

Table 3 Summary of Path-Curvature Metrics

Model	Variant	Mean 3-point κ	Max 3-point κ	Layer _{max}
Gemma	neutral	0.0000372	0.0003122	5.4
Gemma	neg_mod	0.0000366	0.0003037	5.6
Gemma	pos_mod	0.0000372	0.0003168	5.3
Gemma	neg_str	0.0000377	0.0003199	5.6
Gemma	pos_str	0.0000370	0.0003113	5.2
LLaMA	neutral	0.0743119	0.1760759	3.1
LLaMA	neg_mod	0.0744599	0.1770283	2.9
LLaMA	pos_mod	0.0745456	0.1785080	3.0
LLaMA	neg_str	0.0749832	0.1785855	3.0
LLaMA	pos_str	0.0747999	0.1768842	2.9

Key observations:

- **LLaMA exhibits early, high-magnitude semantic curvature**, with peak values concentrated around layer 3, and broad curvature sustained through mid-depth layers. This reflects fast, distributed integration of concern information.
- **Gemma shows weak, shallow curvature** concentrated around layers 5-6, indicating limited semantic bending and fast flattening of signal.
- **Concern-shifted prompts** (especially strong polarity) elevate curvature in both models, but **neutral prompts** also produce high κ in LLaMA, suggesting consistent trajectory shaping even in low-salience conditions.

In many cases, curvature peaks occur shortly after concern tokens and are consistent with the discrete curvature criterion introduced in Appendix C.

We call a layer ‘early’ if it lies in the first quartile of the network depth (layers 1-7 in our 28-layer LLaMA, layers 1-6 in 26-layer Gemma). Curvature onsets in this early band for both models, but only LLaMA sustains high curvature through mid-depth (layers 8-14), so its layer-of-max statistic lands at around 3 even though the heatmap shows a broad high- κ region.

Note: *The layer-of-max metrics in Table 3 reflect average scalar curvature across tokens, whereas the heatmaps depict full token-layer distributions. In LLaMA, curvature often begins early but remains visually sustained across mid-depth layers - revealing a broader narrative arc than a single summary statistic can capture.*

4.2 Inter-metric Relationships

Across all 100 prompt-variant pairs, we observe a strong anticorrelation between **cosine similarity** and **layer-wise Euclidean deviation**, with Pearson correlation coefficients of $r = -0.95$ (Gemma) and $r = -0.98$ (LLaMA). This supports the interpretation that concern-shifted prompts tend to **reorient** the model’s internal representation more than they **displace** it (*a pivot rather than a drift*), indicating semantic reconfiguration without runaway norm inflation.

In contrast, we find only a weak or negligible correlation between **path curvature** κ_i and **layer- Δ** (the average angular deviation between matched directional steps), with Pearson $r = -0.28$ (Gemma) and $r = -0.01$ (LLaMA). This suggests that while both metrics index semantic reorientation, they may capture **distinct geometric behaviours**:

Curvature reflects **local inflection** within token trajectories, while direction deviation aggregates **global path shape** across the full residual arc.

Table 4 Summary of Metric Correlation

Model	Cosine vs Euclidean (r)	Curvature vs Direction (r)	Salience vs Curvature (r)	Prompt Count
Gemma3-1b	-0.9524	-0.2779	-0.5562	100
LLaMA3.2-3b	-0.9784	-0.0061	-0.8931	100

As shown in Table 4, we also observe a strong **anticorrelation between salience and curvature**, with $r = -0.56$ (Gemma) and $r = -0.89$ (LLaMA). Conceptually, **salience** (total representational movement) and **curvature** (degree of directional reorientation) are distinct geometric properties:

A model can move far without turning sharply, or turn sharply without covering much distance.

However, the strong negative correlation observed for LLaMA indicates a **systematic behavioural tendency**:

When token trajectories are highly curved, they tend to be shorter in total length; and when total movement is high, the trajectory is typically straighter.

This inverse relationship suggests a kind of **representational trade-off**: LLaMA often seems to prioritise either **distance** or **reorientation**, but not both to an extreme degree within the same prompt group. This may reflect an internal efficiency mechanism - a kind of “representational budget” that balances semantic effort with semantic precision.

In Gemma, this trade-off is present but less pronounced, consistent with its generally shallower curvature and lower overall salience. The weaker correlation suggests a more variable relationship between reorientation and effort, or simply less pronounced geometric specialisation.

These geometric patterns emerge only in the residual stream, the **unique cumulative path** of semantic inference inside transformer models. As detailed in Appendix A (section A.6), attention and MLP layers act as **semantic lenses** - bending and redirecting token trajectories based on relational and nonlinear context. Their outputs are added to the residual stream, shaping its curvature.

The model may bend its internal trajectory without making large semantic leaps.

Since attention and MLP outputs are the primary forces applied to the residual stream, these curvature and salience patterns must originate in their lensing effects.

Thus, these heatmaps offer a **window into the emergent geometry of meaning**:

The attention and MLP forces remain invisible, but their *integrated effect* (curvature and salience), can be clearly seen.

Once seen, these patterns become difficult to unsee.

Rather than treating curvature and directional deviation as proxies for the same phenomenon, we interpret them as **complementary signals**. **Curvature** captures *where and how sharply* the model bends its internal trajectory - often in response to localised semantic inflections induced by concern-shifted tokens. In contrast, **directional deviation (layer- Δ)** reflects *how far the representation pivots overall*, accumulating changes across layers.

Curvature captures where and how sharply the model bends its internal trajectory

This distinction aligns with the **Semantic Lens** view introduced in Appendix A:

Attention and MLP block layers apply localised semantic forces at each layer, producing curvature in the residual stream.

The token-by-layer heatmaps make this clear - sharp bends are often concentrated around specific layers and token positions. These local twists may not significantly change the overall trajectory arc but still encode important semantic reorientations.

Interpretive claim:

If curvature κ_i encodes **semantic reorientation**, and salience $S(t)$ encodes **semantic effort**, then their joint distribution across tokens and layers reveals **how the model prioritises and navigates internal meaning** under pressure from latent concern.

4.3 Quantifying Concern-Shift Effects Across Prompt Variants

To complement the geometric heatmaps in Section 4.1, we now present a summary of concern-shift (CS) effects across all 20 prompt sets, comparing “moderate” and “strong” variants in both positive and negative polarities. While earlier sections focused on individual visualisations, this section quantifies how reliably these concern manipulations induce geometric changes in the residual stream, using both curvature and salience metrics.

We report mean absolute delta values for each prompt variant, computed by subtracting the neutral control metric and measuring the layer-wise magnitude of change across all tokens. For each model, we test whether these deltas increase with concern strength, and whether the observed effects are statistically significant across prompts.

Summary Statistics:

Table 5 presents delta analysis results comparing Moderate vs Strong concern-shifts.

This data shows that:

- **LLaMA3.2-3b** exhibits consistent and statistically significant increases in both curvature and salience from moderate to strong concern in positive prompts.
- **Gemma3-1b**, while showing non-zero deltas, does not exhibit significant scaling with concern strength.
- Negative polarity prompts in both models show more variability and less consistent statistical separation.

Table 5 Summary of Delta Magnitudes and Significance

Model	Metric	Polarity	Mean Δ (Mod)	Mean Δ (Str)	Ratio S/M	p(Str > Mod) _t
Gemma3-1b	Curvature	Positive	0.0000	0.0000	1.59	0.127
Gemma3-1b	Curvature	Negative	0.0000	0.0000	1.49	0.366
Gemma3-1b	Salience	Positive	3680.0000	4520.0000	1.51	0.152
Gemma3-1b	Salience	Negative	4760.0000	4990.0000	1.48	0.389
LLaMA3.2-3b	Curvature	Positive	0.0063	0.0085	1.38	0.006
LLaMA3.2-3b	Curvature	Negative	0.0067	0.0072	1.24	0.265
LLaMA3.2-3b	Salience	Positive	7.3800	11.1000	1.80	0.016
LLaMA3.2-3b	Salience	Negative	8.0300	8.7300	1.59	0.342

Visual Summary:

Figures 7 and 8 provide a per-prompt visualisation of these deltas. Each subplot shows the mean absolute delta in curvature or salience for both “moderate” (blue circle) and “strong” (red cross) CS variants. Prompts are sorted by total delta magnitude to emphasise those with the most pronounced geometric effects.

These plots reveal:

- In **LLaMA**, strong variants often exceed moderate ones for both metrics, particularly in prompts such as “Perspective Advice 08” and “Logical If Then 11”.
- In **Gemma**, the pattern is noisier. Some prompts show minimal difference or even larger deltas for moderate concern.

Overall, this analysis supports the interpretation that concern-shifted prompts do not merely alter model output, but reshape internal representation in a graded, model-sensitive manner. Stronger concern often correlates with larger semantic trajectory shifts in LLaMA, and to a lesser extent in Gemma. These shifts are evident both in raw metric magnitudes and in structured per-prompt trends.

In all cases, concern-shifted prompts create curvature and salience deviations from the control prompts.

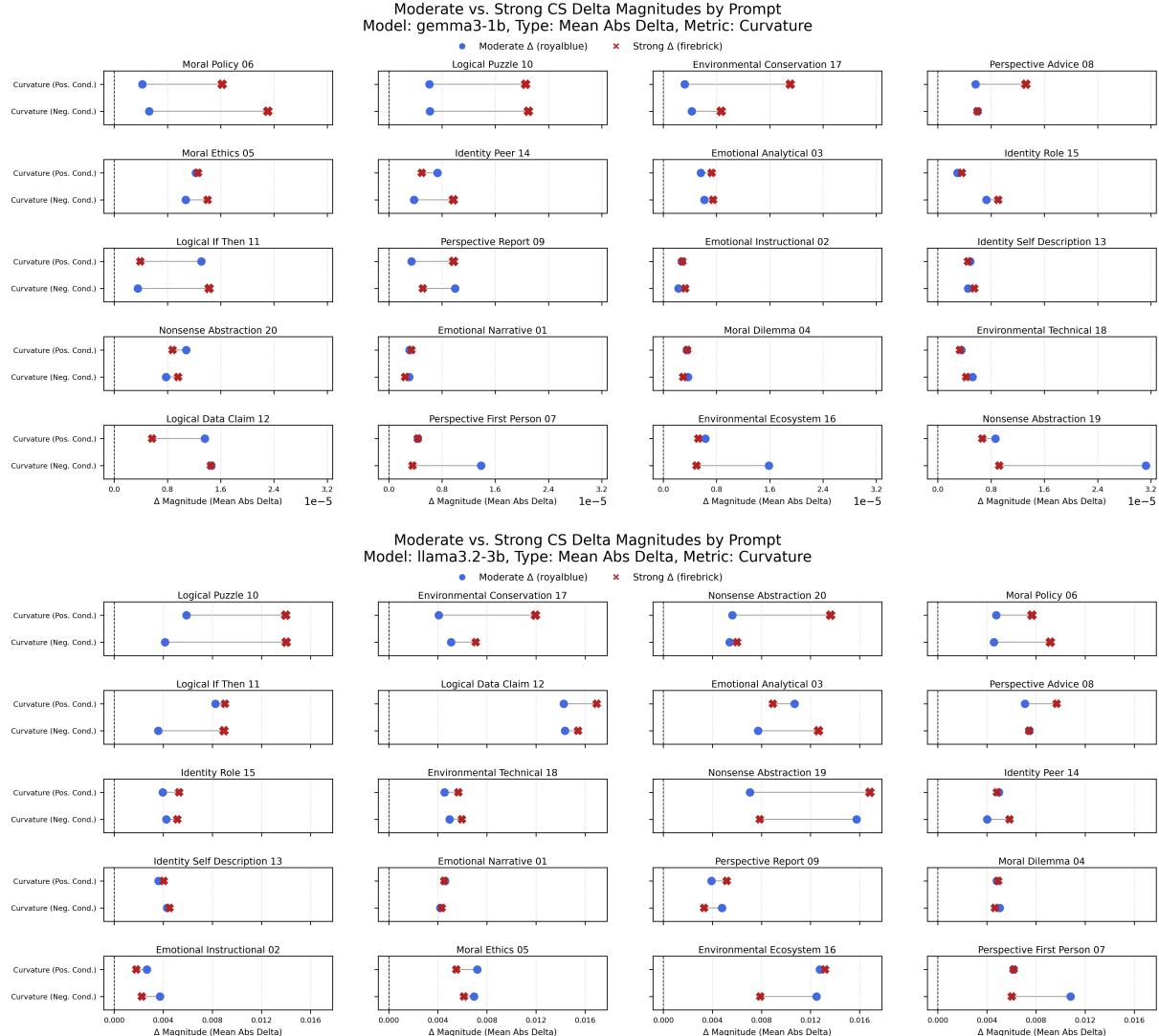


Figure 7: Mean Curvature Delta plots by prompt using Gemma3-1b

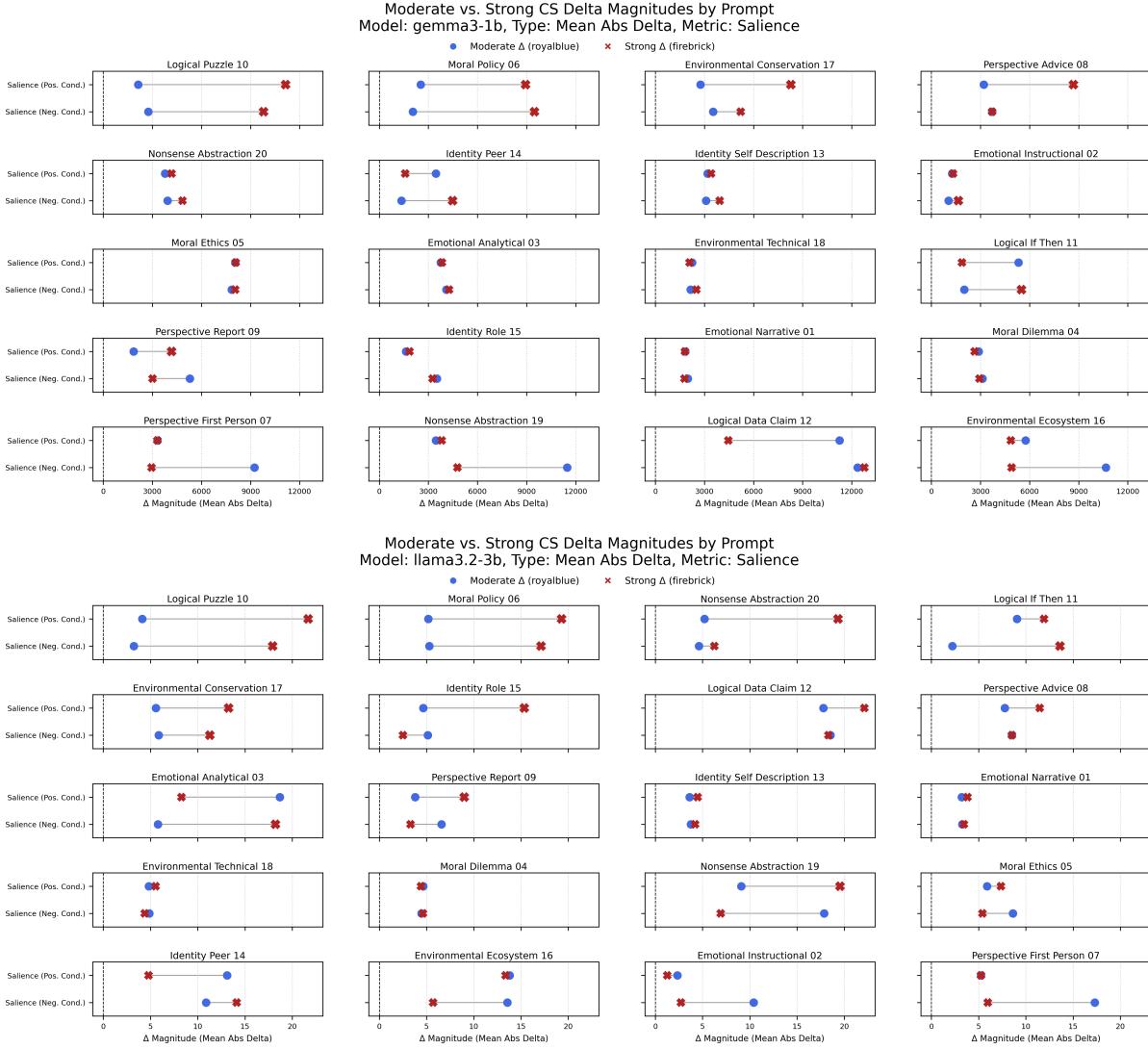


Figure 8: Mean Salience Delta plots by prompt using LLaMA3.2-3b

5 Discussion

The results of this study highlight the value of **residual stream curvature** as an emergent geometric signal of concern-sensitive processing in transformer models. Unlike surface-level metrics or attribution-based scores, curvature and salience offer a **dynamic, trajectory-level view** of how models integrate, differentiate, and abstract contextual meaning in response to semantic concern.

This discussion reaffirms that **curvature is a real, geometry-grounded signal** - whereas low-dimensional projection artefacts are not. The shape of the residual stream offers a fertile new diagnostic for LLM Interpretability.

5.1 Key Takeaways

Our analysis yields four principal findings:

1. **Low-dimensional projections distort curvature geometry.** An earlier audit of our initial results that utilised UMAP to project the data into a lower-dimensional space showed median angle correlations of 0.47 (Gemma) and -0.25 (LLaMA), far below the fidelity threshold ($r \geq 0.80$), rendering UMAP-based curvature measurements unreliable. This reinforces our conclusion that **native-space analysis is essential**.

2. **Path curvature κ_i unifies and sharpens geometric insights.** The discrete 3-point curvature metric κ_i captures all qualitative effects hinted at by initial projection visuals - while remaining mathematically grounded, semantically aligned, and independent of dimensionality reduction.
3. **Curvature tracks salience - but its expression depends on architecture.** Gemma3-1b exhibits an early, shallow, localised bend, while LLaMA3.2-3b shows an early, strong, and broadly sustained curvature arc. In LLaMA, curvature appears early and is followed by compact, consistent salience - suggesting fine-grained reorientation followed by elaboration. In contrast, Gemma builds momentum through a deeper salience wave after a brief early bend. This may reflect a simpler representational arc: minimal redirection, followed by confident semantic travel.
4. **Semantic geometry reflects both latent potential and contextual expression.** We find empirical support for the idea that LLM geometry operates across two distinct but interacting levels:
 - A **latent geometry** encoded in the token embedding matrix E and unembedding matrix U - a static conceptual space trained on the corpus.
 - A **contextual geometry** observed in the residual stream x - the evolving trajectory that expresses meaning in response to specific prompts.

Concern-shifted tokens may often reside in semantically irregular or high-potential zones of E (as shown in recent work by Robinson et al. [16]), but their actual impact on the model’s inference path depends on context. Our curvature and salience metrics measure how (and whether) that latent potential is activated, resolved, or redirected. This provides a natural link between raw embedding structure and downstream semantic processing.

5.2 Implications for Interpretability Research

These results support a shift toward **process-level explanations** of LLM behaviour. Instead of isolated neurons or static feature attributions, we focus on the model’s **residual stream trajectory** - tracing how internal representations evolve across layers. Salience plus curvature and its companions (**layer- Δ** inter-layer angle change and **dir- Δ** directional deviation), offer **continuous, geometry-grounded signals** that complement discrete methods such as causal tracing or probing.

Alignment telemetry. Because both κ_i and layer- Δ exhibit sharp spikes in response to certain concern-shifted prompts, these metrics offer a plausible basis for **real-time alignment monitors** - flagging potentially risky completions *before* they manifest in output.

Architecture design. The contrast between Gemma and LLaMA may suggest that larger parameter models could defer semantic reorientation until higher-order representations have stabilised. This would imply that smaller models may benefit from architectural constraints (such as delayed residual gating or curvature regularisation) to avoid premature semantic bending or norm inflation in early layers.

Curvature spikes in some concern-shifted prompts may serve as real-time alignment signals.

5.3 Latent and Contextual Geometry in Light of Prior Work

Recent work challenges the assumption that token embeddings lie on a smooth, low-dimensional manifold. Robinson et al. [16] present evidence that LLM embedding spaces violate the **manifold hypothesis**, instead exhibiting locally non-Euclidean and singular behaviour. Their fibre bundle null model formalises how certain tokens occupy high-noise or irregular subspaces - potentially distorting semantic inference.

However, in this work they focus on **static embeddings** - token vectors from the learned embedding matrix E , unconditioned on prompt context. In contrast, our *Curved Inference* framework operates in the **residual stream**, capturing how meaning is dynamically expressed across transformer layers. What Robinson labels as embedding-space singularities may or may not persist during inference. Some tokens with “irregular” positions in E may be smoothly integrated through context, while others may spark curvature spikes or salience surges.

This context-sensitive transformation aligns with findings from vec2vec [17], which demonstrates that embeddings from different models can be reliably aligned via unsupervised translation functions. Despite architectural differences, a **shared latent geometry** seems to underlie token semantics. Vec2vec shows this by learning a mapping from one model’s embedding space to another’s, preserving relational structure.

These two findings - **manifold violation** in E , and **cross-model alignment potential** - are not contradictory. Instead, they reflect a distinction between **local irregularity** and **global coherence**. *Curved Inference* provides a third, complementary perspective:

Curved Inference shows how these latent properties are *realised* or *suppressed* through context.

Together, this suggests a layered view of geometry:

- **Latent conceptual structure** lives in E and U .
- **Contextual semantic expression** unfolds in x .
- **Curved Inference** measures how context bends potential into meaning.

See section 7 (Future Work) for a discussion of how this perspective could be explored further.

6 Limitations

We recognise that several limitations constrain the scope of our conclusions:

- **Prompt coverage.**

The concern-shift suite comprises 20 hand-curated prompts across seven semantic domains. While balanced for vocabulary and length, the dataset remains small and may introduce structural or semantic biases. Future work should draw from broader benchmarks (e.g. MMLU, HELM, SuperGLUE) to test generalisability.

- **Model scale and family.**

This study analyses only two decoder-only checkpoints: Gemma3-1b and LLaMA3.2-3b. Larger-capacity models, encoder-decoder hybrids, or pretraining variants may bend differently or redistribute curvature across attention and MLP submodules. A systematic scaling sweep is needed.

- **Metric sensitivity.**

Curvature and direction-based metrics may respond to **syntactic or lexical changes** that are not genuinely semantic. While residual activations show the clearest geometric response, further null controls (e.g. synonym swaps, structure-preserving rewrites) are required to confirm the specificity of the signal.

- **No null baselining.**

While nonsense and neutral prompts were utilised, no scrambled, synonym-swapped, and random-weight prompts were tested. Without these additional baselines, curvature magnitudes remain **relatively interpreted** - we cannot assign absolute thresholds for “low” or “high” curvature.

- **Early-layer artefacts.**

Some prompts in Gemma show peak curvature at **layer 1**, suggesting possible artefacts from tokenisation, embedding scaling, or early-layer normalisation. Without deeper instrumentation, these cannot be conclusively labelled as genuine semantic reactions. However, it’s also possible that concern-sensitive effects may begin as early as the embedding layer or at the very point of entry into the residual stream. This suggests a mechanism where curvature is seeded by prompt-level semantics even prior to token-level integration.

- **Control prompt assumptions.**

Neutral scaffolds are designed to be semantically flat, but subtle phrasal choices may still encode latent salience. This could inflate curvature in “control” prompts and weaken differential comparisons.

- **No behavioural validation.**

While curvature patterns are consistent, no task-level metrics were collected to confirm causal impact on generation. This limits Interpretability claims to internal geometry. Future work could test curvature’s behavioural relevance via controlled generation studies, human evaluation of completions, or causal intervention at high- κ points in the residual stream.

- **Limited modality and feature scope.**

The analysis focuses exclusively on residual stream activations. Attention maps, MLP outputs, or intermediate gating signals may also express meaningful curvature. Triangulating these modalities could refine salience tracking.

- **Finite-difference sensitivity.** We use a discrete 3-point central difference method to estimate curvature. While this is effective for short, smooth trajectories, it may become noisy or unstable for longer sequences or

high-curvature transitions. Using a wider finite-difference stencil (e.g. 5-point or 7-point) or incorporating smoothing priors could improve robustness on longer sequences or noisier trajectories.

- **Metric scope.**
Our analysis includes extrinsic curvature, layer-wise deviation, directional angle, and cosine similarity. Additional geometric descriptors (such as **torsion**, **energy**, or **intrinsic curvature tensors**) remain unexplored here but offer promising extensions.
- **Finite sequence effects.** Current analysis is limited to short prompts. Curvature behaviour may vary with longer sequences. The 3-point method may also become unstable with longer sequences.
- **Layer resolution.** With only 26-28 layers, the discrete sampling may miss important dynamics. Further experimentation is required to validate if this scales to models with many more layers (e.g. 80+).

These limitations do not undermine the core findings, but highlight clear directions for future work: broader prompt and model sampling, null calibration, multimodal instrumentation, and behavioural grounding of geometric signals.

7 Future Work

This study introduces a geometric lens on LLM behaviour, grounded in path curvature within native activation space. While the framework rests on principled metric geometry, several promising directions remain:

- **Singularity and embedding structure.**
The insights from section 5.3 suggest new empirical directions. Tokens identified as geometrically singular in E could be tracked across varied prompts to measure how their curvature and salience signatures vary by context. Do some contexts neutralise their irregularity? Do others amplify it? By combining Robinson’s singularity indices with our delta metrics, future work could identify not just which tokens are charged, but when and how they activate that charge. Such studies would further unify pretraining geometry and inference dynamics - and clarify how concern becomes curvature.
- **Interpretive Divergence as Semantic Signal.** One promising interpretive direction emerging from the two-layer geometric perspective outlined in section 5.3 is the role of **contextual-conceptual divergence** as a marker of semantic significance. While this paper focuses on how concern shifts alter the shape of residual trajectories, future work could explore when and why a token’s **contextual meaning trajectory diverges sharply from its latent conceptual position**. Such divergence may signal both **positive behaviours** (e.g. abstraction, creativity, novel synthesis) and **failure modes** (e.g. hallucination, incoherence, misalignment). Tokens used in ways that sharply depart from their pretraining priors may produce high curvature relative to their embedding-based expectation. By measuring this conceptual-contextual gap (using angular divergence, curvature, or alignment metrics) *Curved Inference* could support new forms of semantic anomaly detection and creativity tracing.
- **Force attribution.** While Appendix A frames attention and MLP block layers as “semantic lenses” that induce the observed curvature in the residual stream, we leave a detailed force-alignment study-e.g., correlating per-layer MLP output with discrete accelerations, or ablating high-impact attention heads-to future work.
- **Causal interventions.**
Whether curvature reflects a causal locus of computation remains unclear. Having identified high-curvature layers aligned with concern tokens, future work could test their functional role through patching, ablation, or targeted editing.
- **Scaling and architectural generalisation.**
This study focused on two decoder-only models (1b-3b). Extending the analysis to larger checkpoints and different model families would test the generality of observed curvature behaviours.
- **Intrinsic geometric structure.**
While this work focused on extrinsic curvature in full-dimensional space, it’s possible that token trajectories lie near a lower-dimensional manifold. Exploring intrinsic curvature or geodesic deviation could uncover deeper representational geometry.
- **Alignment and robustness.**
Because curvature often spikes on morally or emotionally charged prompts, it may correlate with semantic

load or alignment risk. Applying curvature analysis to robustness benchmarks or generation pipelines could test its use as a real-time safety signal.

- **Cross-lingual and cross-domain generalisation.**

Curvature signals should be tested across languages, tokenisation regimes, and domains. Prior work shows residual trajectories align across translations [18]; future work could test whether curvature shape generalises similarly, revealing a universal form of concern.

Together, these directions will help test whether curvature is not just diagnostic - but generative, transferable, and actionable.

8 Conclusions

This work introduces *Curved Inference* - a geometry-first framework for analysing how large language models bend their internal representational space in response to latent *concern*. By grounding all measurements in native residual space \mathbb{R}^d and computing path curvature κ_i and salience $S(t)$ using the semantic pullback metric $G = U^\top U$, we eliminate potential distortions introduced by low-dimensional projections and recover meaningful geometric structure.

Our empirical results confirm that concern-shifted prompts induce measurable and model-specific changes in the internal trajectories of LLMs. In particular:

- **LLaMA3.2-3b** displays clear, statistically significant scaling of curvature and salience in response to increasing concern strength.
- **Gemma3-1b**, while reactive to concern shifts, shows weaker differentiation between moderate and strong variants.

These patterns reflect deeper architectural and geometric dynamics, captured through our proposed distinction between two interacting layers of semantic geometry:

- A **latent conceptual geometry**, encoded in token embeddings (E) and unembedding projections (U), shaped during pretraining.
- A **contextual semantic geometry**, expressed through residual trajectories (x) during inference.

We show that while some tokens occupy irregular or singular regions in embedding space (as described by Robinson et al. [16]), their impact is determined by how context interacts with that potential. Meanwhile, the vec2vec [17] findings reveal that despite architectural divergence, different models often encode semantically compatible latent structures - a finding that our residual-path analysis helps operationalise.

Curved Inference provides a principled method for tracing how **semantic potential becomes semantic movement**. Its native-space metrics expose how meaning is navigated, reoriented, or reinforced within a model’s depth. These geometric insights complement traditional Interpretability methods, and open new paths for alignment monitoring, architectural diagnosis, and semantic probing.

Earlier versions of this work used UMAP-based projections to visualise geometric divergence. These proved inadequate, mischaracterising the timing and scale of curvature effects. This led to the mischaracterisation of LLaMA’s curvature as “late and steep”. Full-space analysis using κ_i and $S(t)$ corrected this, revealing that curvature is strongest **early** in the residual trajectory and often sustained across depth - a correction that highlights the importance of measuring semantic geometry in native activation space.

In mapping internal semantic geometry, *Curved Inference* shifts Interpretability away from static probes and component dissection and toward **dynamic trajectory analysis** - a vantage point that scales with model size and abstraction depth.

Curved Inference is both a **map** and a **diagnostic**:

It reveals what the model finds meaningful - and how it bends to meet it.

References

- 1 - **Yu, Z., et al.** (2023) “Exploring the Residual Stream of Transformers” *arXiv*
- 2 - **Jain, S. & Wallace, B.** (2019) “Attention is not Explanation” *arXiv*
- 3 - **Conneau, A., et al.** (2018) “What you can cram into a single vector: Probing sentence embeddings for linguistic properties” *arXiv*
- 4 - **Olah, C., et al.** (2020) “Zoom In: An Introduction to Circuits” *Distill*
- 5 - **Elhage, N., et al.** (2021) “A Mathematical Framework for Transformer Circuits” *arXiv*
- 6 - **Madsen, A. et al.** (2024) “Interpretability Needs a New Paradigm” *arXiv*
- 7 - **Singh, C. et al.** (2024) “Rethinking Interpretability in the Era of Large Language Models” *arXiv*
- 8 - **Yeo, W. et al.** (2024) “How Interpretable are Reasoning Explanations from Prompting Large Language Models?” *arXiv*
- 9 - **Huang, S. et al.** (2023) “Can Large Language Models Explain Themselves? A Study of LLM-Generated Self-Explanations” *arXiv*
- 10 - **Molina, R.** (2023) “Traveling Words: A Geometric Interpretation of Transformers” *arXiv*
- 11 - **Shai, A. et al.** (2025) “Transformers Represent Belief State Geometry in their Residual Stream” *arXiv*
- 12 - **Manson, R.** (2025) “FRESH (Functionalist & Representationalist Emergent Self Hypothesis)” *Github*
- 13 - **Manson, R.** (2025) “Curved Inference in LLMs - Experiment” *Github*
- 14 - **Google** (2025) “Gemma 3 Model Card” *ai.google.dev*
- 15 - **Meta** (2024) “LLaMA 3.2 Model Card” *Github*
- 16 - **Robinson, M., et al.** (2025) “Token Embeddings Violate the Manifold Hypothesis” *arXiv*
- 17 - **Jha, R., et al.** (2025) “Harnessing the Universal Geometry of Embeddings” *arXiv
- 18 - **Tian, Y., et al.** (2024) “Neural Interlinguae: Layerwise Geometry Aligns Translation Trajectories Across Languages” *arXiv*
- 19 - **Elhage, N., et al.** (2023) “A Privileged Basis for the Transformer Residual Stream” *Transformer Circuits*

Appendix A - Semantic Geometry of Transformer Inference

A.1 Overview: Geometry as Trajectory

Transformer inference can be viewed as a geometric process - each token is mapped to a vector in a high-dimensional space, and then pushed through a series of attention and MLP updates. The result is a continuous sequence of transformations forming a trajectory in residual space. This trajectory encodes the evolving semantic state of the model as it processes or generates a sequence. Attention and MLP layers act as dynamic lenses, bending and focusing these trajectories based on contextual and relational signals. This appendix outlines how these trajectories are constructed, how they evolve, and how geometric measurements such as curvature and salience are defined within this process.

Notation Recap:

- E : embedding matrix, maps token IDs to vectors in \mathbb{R}^d
- U : unembedding matrix, maps residual vectors to logit space (often $U = E$)
- x : residual stream vector, the current semantic state
- $l = Ux$: logit vector (dot products between x and each token direction)
- $G = U^\top U$: pullback metric from logit space, defines geometry in residual space

A.2 Token Trajectories and Residual Flow

Tokens are first split by a tokeniser and mapped to unique token IDs. Each token ID is used to look up an embedding vector from the learned embedding matrix $E \in \mathbb{R}^{|V| \times d}$, where $|V|$ is the vocabulary size and d is the model dimension.

The initial residual stream vector $x \in \mathbb{R}^d$ for a token is simply the embedding $E[t]$. In models using Rotary Positional Embedding (RoPE), no position vector is added at this stage. Instead, positional information is injected later during attention via rotation.

At each transformer layer, the residual vector is updated by adding the outputs of the attention and MLP sublayers:

$$x^{(\ell+1)} = x^{(\ell)} + \text{Attention}(x^{(\ell)}) + \text{MLP}(x^{(\ell)})$$

This additive structure means that the residual stream forms a trajectory through \mathbb{R}^d , with each step determined by the semantic influence of the attention and MLP mechanisms.

- The attention layer gathers contextual signals from other positions, modulated by relative position (via RoPE), and contributes a vector update that reflects token-token interaction.
- The MLP layer applies a local, nonlinear transformation that sharpens or redirects the vector - often enhancing its alignment with task-relevant directions in the model.

Together, these updates shape the path taken by each token's representation. We refer to this evolving path as the token's semantic trajectory.

Because each update is added to the previous residual state, it is only in the residual stream that one can observe the full evolution of meaning over depth. Attention and MLP outputs are delta vectors - they cause curvature, but the residual stream *is* the curve.

A.3 Position, RoPE, and Curvature

In models using RoPE, positional information is not embedded additively into the residual vector. Instead, RoPE applies a deterministic, sinusoidal rotation to the query and key vectors used in attention. These rotations encode relative position by angular offset, preserving dot products while modulating attention scores.

Because RoPE does not shift or perturb the initial residual vector, it preserves semantic purity in the early layers. Curvature in the residual stream only arises once RoPE-modulated attention begins to redistribute contextual information across tokens.

This means that curvature is not tied to absolute position, but to *semantic interaction* among tokens that are contextually relevant and positionally adjacent. As a result, the residual trajectory bends not because of where a token is, but because of how it relates to others. RoPE thereby becomes a key generator of semantically aligned curvature.

A.4 Logits and the Pullback Metric

Once the final residual vector is computed for a token, it is projected into logit space by taking a dot product with each row of the unembedding matrix $U \in \mathbb{R}^{|V| \times d}$:

$$l = Ux$$

This yields a logit vector $l \in \mathbb{R}^{|V|}$, where each entry reflects the alignment between x and a possible output token direction u_i . The softmax of l forms a probability distribution over the vocabulary.

Crucially, the matrix U defines a set of semantic directions in residual space. These directions induce a geometry - the pullback metric $G = U^\top U$ redefines how distances and angles are measured in residual space based on the model's output behaviour.

This metric allows curvature to be computed in a way that reflects semantic change. Rather than relying on arbitrary coordinates, we measure changes in the residual trajectory using a geometry aligned with token prediction. This is what gives curvature its interpretability - it reflects changes in internal intent as judged by the model's own output semantics.

A.5 Caching and Contextual Reuse

During inference, transformer models use key-value (KV) caching to avoid recomputing attention outputs for previously processed tokens. Once a token's attention keys and values have been computed, they are stored and reused in subsequent steps. This optimisation ensures that only the *new* token's computations need to be performed at each generation step.

Importantly, this also means that all previous residual vectors are frozen - they are not recomputed or updated. Each prior x forms a fixed semantic anchor. The residual stream for the current token builds on top of these fixed vectors, enabling us to track how each new token evolves in context.

In multi-turn chat settings, the entire chat history is tokenised into a flat prompt. Provided it fits within the context window, only the new portion of the prompt (e.g. user query and assistant response) is recomputed. The rest is reused, including residuals, keys, and values.

A.6 Summary: The Semantic Lens

The transformer can be viewed as a geometric engine. Tokens enter as points in a semantic subspace, pass through layers of contextual and nonlinear modulation, and exit as probability distributions over token space. The residual stream traces the continuous trajectory of each token through this process.

Attention and MLP layers act as semantic lenses. Attention bends trajectories based on relative semantic and positional relevance. MLPs sharpen or redirect them through nonlinear amplification. RoPE enables these transformations to be position-aware without distorting the embedding space directly.

All curvature, salience, and concern arise within the residual stream. It is the only continuous representational path through the model-and the only space in which geometric measurements can meaningfully be made.

Appendix B - Definitions

Meaning

Meaning, in the context of large language models, refers to the implicit content, intent, or conceptual structure represented by the model’s internal activations. It is not a directly observable quantity, but an abstract property inferred from the model’s behaviour and internal geometry.

Conceptual Role

Meaning is what the model *represents* at any point in the forward pass. This could include factual information, sentiment, identity, logical structure, or moral stance. The meaning associated with a given activation depends on the context, the model’s training, and how that activation aligns with downstream predictions.

Meaning becomes accessible through **semantic structure** - the way that internal representations relate to each other and to output tokens. This structure is revealed through geometric properties-such as direction, distance, and curvature-within the residual stream.

Formal Proxy

We do not measure meaning directly. Instead, we study how it **moves** and **changes** through time. This is done by:

- Representing the model’s internal state as a vector $x_t \in \mathbb{R}^d$
- Measuring change (salience) as $\|x_{t+1} - x_t\|_G$
- Measuring reorientation (curvature) as κ_i
- Anchoring this geometry in **semantic space** via the pullback metric:

$$G = U^\top U$$

This metric ensures that the geometry of residual-space movement reflects differences in token-level output probabilities. In this sense, **meaning lives in the structure** of how internal representations flow and bend toward predicted outputs.

Practical Implication

Throughout this work, we treat **meaning** as:

The internal state of the model that gives rise to token predictions and reflects the model’s interpretation of context.

Changes in meaning are inferred from changes in the residual trajectory. High salience means meaning is shifting quickly; high curvature means it is changing direction. Concern identifies directions along which meaning changes matter to the model.

This view of meaning intersects with the idea of superposition - that many abstract features may be simultaneously encoded in overlapping directions within the same residual vector. The geometric structure (e.g. curvature) reflects how these meanings are separated or recombined across layers.

Semantic Space

Semantic space refers to the internal vector space in which a model encodes and manipulates meaning. It is the geometric arena where representations of language, context, and concepts take shape and evolve during inference. In

transformer-based LLMs, this space is typically identified with the **residual stream** - but only when measured under a **meaning-preserving metric**.

Formal Definition

We define semantic space as the residual space \mathbb{R}^d , equipped with a metric derived from the model's output behaviour:

$$G = U^\top U$$

where $U \in \mathbb{R}^{V \times d}$ is the unembedding matrix that projects residual activations $x_t \in \mathbb{R}^d$ to logits over the vocabulary. The inner product and norm induced by G give rise to a geometry in which:

- **Distances** correspond to shifts in output token probabilities
- **Directions** correspond to latent semantic operations
- **Curves** correspond to evolving meaning across layers

This pullback metric transforms residual space into a **logit-aligned semantic space**.

Distinctions and Relationships

Table 6 Comparison of Spaces

Space	Contents	Function
Embedding space	Token embeddings e_i	Stores static lexical representations
Logit space	Output predictions $\ell = Ux$	Determines token-level output
Residual space	Internal state x_t	Active inference trajectory
Semantic space	Residual space + G metric	Geometry aligned with meaning and output

Semantic space is not defined purely by coordinate axes-it emerges from the **functional role** of directions and distances under the model's output logic. That is, it reflects how the model internally represents and differentiates concepts, rather than any superficial arrangement of neurons.

Practical Use in This Work

All geometric quantities in this paper (salience, curvature, directional shifts), are computed in **semantic space**, using the pullback metric G . This ensures that our analysis reflects what the model *does* with its internal states, not just how they appear numerically.

Concern

In this paper, we study concern **extrinsically** by introducing controlled shifts in prompt semantics - what we call **concern-shifted prompts**. These allow us to probe whether and how the model reorients its internal trajectory in response to targeted emotional, moral, or logical pressure. While this is an external perturbation, we interpret the resulting geometric responses as signals of **internal concern sensitivity**.

Operational Definition (This Paper)

In this study, concern is defined **extrinsically** through a controlled prompt-design strategy. Each scaffold contains a **neutral base** plus **concern-shifted prompts**, chosen to introduce a targeted form of concern (e.g. moral framing, emotional tone, identity cue). The resulting **geometric divergence** (measured via residual stream metrics such as path curvature κ_i), is interpreted as a behavioural signal of the model's sensitivity to that shift.

This approach allows us to identify **concern-induced curvature**, where the model's internal trajectory bends or accelerates in response to semantic pressure, even if token-level differences are minimal.

Conceptual Distinction

Concern is orthogonal to salience and curvature.

Table 7 Comparison of 3 Primary Concepts

Term	Type	What it captures	Example
Salience	First-order ($\ x_{t+1} - x_t\ $)	How fast the model's state is changing	Rapid progression through a narrative
Curvature	Second-order (κ_i)	Whether the trajectory is turning	Reorienting after a moral twist
Concern	Priority weighting	Whether the model treats this direction as significant	Responding to identity or moral cues

Potential Future (Intrinsic) Formulations

Longer-term, we aim to explore replacing the extrinsic scaffold design with **intrinsic concern measures** derived from model-internal behaviour:

- **Gradient-based sensitivity:**

$$\text{Concern}(x_t) = \left\| \frac{\partial \ell}{\partial x_t} \right\|_G$$

Captures how sensitive the output is to changes at layer t , measured in the pullback metric.

- **Subspace projection:**

$$\text{Concern}(x_t) = \|P_C x_t\|_G$$

Where \mathcal{C} is a concern-relevant subspace extracted via PCA, CCA, or probing.

- **Behavioural perturbation:** KL divergence between softmax outputs after small displacements in specific directions.

Each of these aims to isolate *what the model finds meaningful*, as inferred from its own structure and behaviour.

Salience

Salience quantifies how rapidly a model's internal state is changing as it processes a prompt. In geometric terms, it is the **first-order velocity** of the residual stream trajectory - how far the model moves in semantic space from one layer to the next. High salience indicates a rapid update in the model's internal representation, even if that movement follows a straight path.

Operational Definition

For a residual stream trajectory $x_0, x_1, \dots, x_L \subset \mathbb{R}^d$, the **layer-wise salience** at layer t is defined as:

$$\text{Salience}(t) = \|x_{t+1} - x_t\|_G$$

where $\|\cdot\|_G$ denotes the norm induced by the **semantic metric**:

$$G = U^\top U$$

Here, U is the unembedding matrix that maps residual states to logits, and the pullback metric G aligns geometric measurements in residual space with token-level semantic structure.

Semantic Interpretation

Salience tracks the **rate of change of internal meaning**, where “meaning” is defined by how the residual vector projects into logit space. It captures *how much* the model updates its belief or understanding at each layer - irrespective of direction.

A model may have:

- **High salience, low curvature** → confidently elaborating or reinforcing an idea
- **Low salience, high curvature** → making a subtle but meaningful reorientation
- **Low salience, low curvature** → continuing steadily with no shift in interpretation

Aggregation

The total salience over a trajectory can be defined as cumulative arc length:

$$S = \sum_{t=0}^{L-1} \|x_{t+1} - x_t\|_G$$

This is used in later analysis (e.g., for arc-length normalisation in curvature metrics).

Curvature

Curvature captures how sharply the model’s internal representation is **changing direction** as it processes a prompt. In geometric terms, it is the **second-order property** of the residual stream trajectory - the rate at which the model’s semantic path bends, rather than continues in a straight line.

Operational Definition

Let the residual stream activations across layers be denoted:

$$x_0, x_1, \dots, x_L \subset \mathbb{R}^d$$

To estimate curvature, we apply a discrete 3-point finite-difference scheme to the sequence of residual stream vectors. For each interior point i , we compute the first and second derivatives using a discrete 3-point central difference method that accounts for unequal step sizes, then apply the standard extrinsic curvature formula.

The **extrinsic curvature** at index i is defined as:

$$\kappa_i = \frac{\sqrt{\|a_i\|_G^2 \cdot \|v_i\|_G^2 - \langle a_i, v_i \rangle_G^2}}{\|v_i\|_G^3}$$

where:

- v_i is the first derivative (velocity)
- a_i is the second derivative (acceleration)
- $\langle \cdot, \cdot \rangle_G$ and $|\cdot|_G$ denote the inner product and norm under the pullback metric $G = U^\top U$

This is the standard formula for curvature in Euclidean space, extended here to a semantically aligned geometry via the metric G . It is invariant to orthogonal coordinate transformations and reflects intrinsic trajectory shape rather than coordinate artefacts.

Semantic Interpretation

Whereas **salience** measures how far the model moves between steps, **curvature** measures how much it **reorients** e.g. whether the model continues in a consistent direction or turns sharply at some layer. High curvature indicates a structural shift in internal representation, such as a reinterpretation, contradiction, or redirection in meaning.

Examples:

- A strong moral reversal → high curvature
- Steady elaboration of a factual detail → low curvature

Aggregation and Summary Statistics

From the full curvature series κ_i , we derive summary metrics per prompt variant:

- **Mean curvature:**

$$\bar{\kappa} = \frac{1}{L-1} \sum_{i=1}^{L-1} \kappa_i$$

- **Maximum curvature:**

$$\kappa_{\max} = \max_i \kappa_i$$

- **Layer of maximum curvature:**

$$i^* = \arg \max_i \kappa_i$$

Curvature is only defined at interior indices i , where a discrete 3-point central difference can be used to estimate derivatives. Boundary positions $i = 0$ and $i = L$ are excluded because symmetric differencing is not possible.

Comparison

The following table summarises the key interpretive and mathematical roles of the five core terms used throughout this paper.

Table 8 Comparison of Interpretive and Mathematical Roles of Key Terms

Term	Type / Role	What It Captures	Operational Form
Meaning	Latent content	The internal representation of context, concepts, and intent	Inferred from geometry and output alignment
Semantic space	Geometric setting	The metric space in which meaning is encoded and compared	Residual space with pullback metric $G = U^\top U$
Salience	First-order (velocity)	How rapidly the model's internal state is changing	$\ x_{t+1} - x_t\ _G$
Curvature	Second-order (reorientation)	How sharply the model changes direction in semantic space	κ_i from discrete 3-point finite differences
Concern	Directional importance	Whether a direction is semantically or behaviourally significant	Currently via prompt-class variation - future via gradients or projections

This taxonomy supports a structured analysis of model behaviour in geometric terms - *salience* describes motion, *curvature* describes trajectory shape, and *concern* identifies which directions matter. All are grounded in *semantic space*, which gives these geometric properties functional meaning in terms of the model's outputs.

Supporting Definitions

Semantic Metric

We define the pullback metric ($G = U^\top U$) on residual space, where (U) is the unembedding matrix from residual space to logits. This equips residual activations with a geometry that respects the model's token-level semantics. All distances, angles, and curvatures are computed under this metric.

This pullback metric approach reflects a broader insight that residual directions are not functionally uniform - some carry disproportionate semantic weight as shown in Elhage et al. [19] through their analysis of privileged basis vectors.

Notation

- $(x_t \in \mathbb{R}^d)$: residual stream activation at layer t
- (x_i) : residual stream activation at layer i , forming a discrete trajectory in \mathbb{R}^d
- (κ_i) : discrete curvature at interior index i , computed from local 3-point finite differences
- (v_i) : estimated velocity at index i (first derivative of x)
- (a_i) : estimated acceleration at index i (second derivative of x)
- $(G = U^\top U)$: semantic (pullback) metric

Appendix C - Discrete Curvature: Geometric Criterion for Semantic Divergence

This appendix provides a minimal geometric condition under which a concern-shifted prompt must exhibit curvature in the model’s residual stream trajectory.

C.1 Core Result

Let (x_0, x_1, \dots, x_L) be a residual stream trajectory, and define inter-layer steps:

$$v_\ell = x_{\ell+1} - x_\ell \quad \text{for } \ell = 0, \dots, L-1$$

Let v_ℓ^{ctrl} be the step at layer ℓ for the control prompt, and define the shift-induced difference:

$$\Delta v_\ell = v_\ell^{\text{shift}} - v_\ell^{\text{ctrl}}$$

If there exists any layer ℓ^* such that Δv_{ℓ^*} is **not colinear** with $v_{\ell^*}^{\text{ctrl}}$, then the concern-shifted trajectory exhibits strictly positive curvature at that layer. That is, the path bends in a two-dimensional plane spanned by $\{v_{\ell^*}^{\text{ctrl}}, \Delta v_{\ell^*}\}$.

Conversely, if $\Delta v_\ell \parallel v_\ell^{\text{ctrl}}$ for all ℓ , then the shifted trajectory lies in a single affine line and all curvature vanishes.

C.2 Interpretation

This condition ensures that a geometric bend arises when the shift-induced step differs in direction-not merely in magnitude-from the control. It provides a discrete diagnostic: **directional deviation implies curvature**.

C.3 Coordinate Considerations

All measurements are taken in the residual space \mathbb{R}^d , using the semantic (pullback) metric $G = U^\top U$ derived from the model’s unembedding matrix U .

The curvature definition used in this study is invariant to **orthogonal changes of basis** in \mathbb{R}^d -that is, coordinate rotations do not affect the result. While such transformations are not typical within transformer operations, this mathematical invariance ensures that curvature is a property of the trajectory’s shape, not of any arbitrary axis labelling.

This result confirms that discrete, layerwise divergences can induce true geometric bending-so long as they are directionally expressive. It supports the use of κ_i as a valid and semantically aligned curvature estimate.