# WHY SAFETY ALIGNMENT CAN FAIL: A GEOMETRIC THEORY OF STRUCTURAL HALLUCINATION IN LLMS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Safety alignment can fail under jailbreak even measured task capability remains intact, but a change in a safety projection alone does not reveal whether the representation moves farther or moves in a different direction. We give an exact geometric decomposition of every normalized inter-layer update into direction-independent retention and a direction-specific tangent contribution, derive the sharp reachable interval of a single update, and prove a finite-depth directional-budget bound with a closed-form characterization of equality. Exponential projection decay follows as the constant-retention, zero-budget special case. Across four Llama and Qwen models on HarmBench, refused trajectories receive systematically larger safety-directed tangent contributions than jailbroken trajectories, with refused-minus-jailbroken effect sizes from 1.3 to 2.3. The retention contrast has the opposite sign and partly offsets the tangent contrast. The tangent ordering remains after adjustment for layer and current projection. A separate Llama-3.1-8B GSM8K experiment observes no statistically distinguishable accuracy difference between clean prompting and the tested jailbreak prefix. These results support analyzing safety maintenance through both retention and direction-specific writes, clarifying how task computation can persist while constraint transmission weakens.

## 1 INTRODUCTION

As Large Language Models (LLMs) scale in capacity and are deployed with public access (Brown et al., 2020; Kaplan et al., 2020; Hoffmann et al., 2022), safety alignment techniques such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) and Constitutional AI (Bai et al., 2022) have become essential safeguards. Here, *safety alignment* means output-level compliance with a specified safety policy, and *task capability* means the ability to understand a prompt and produce a coherent, substantively relevant response. For a prohibited harmful request, we define *jailbreak success* behaviorally as providing substantive harmful assistance rather than refusing or responding safely (Chao et al., 2024; Mazeika et al., 2024). We use *clean hijacking* for a successful jailbreak with approximately preserved measured task capability, treating capability retention as an empirical condition rather than part of the jailbreak definition. A puzzling asymmetry has emerged in practice: task capability can remain robust, whereas safety constraints are routinely circumvented by adversarial attacks (Wei et al., 2023; Zou et al., 2023). This contrast raises a fundamental question: why can the same forward pass sustain task-relevant computation yet fail to preserve the safety constraint it was trained to uphold?

Probing studies sharpen this puzzle: harmfulness can remain linearly decodable from hidden states even when a jailbreak elicits harmful output (Lin et al., 2024; Zou et al., 2025; Arditi et al., 2024). Detectability therefore does not entail refusal. Paired reference and jailbreak representations also separate across depth (Figure 1), but this descriptive separation does not distinguish two possibilities: successive layers may retain less of the preceding state, or their new contributions may point less strongly along the safety direction. Projection trajectories alone conflate these channels, and existing descriptions do not isolate them at the level of adjacent normalized states (Wei et al., 2023; Greenblatt et al., 2024).

Our key observation is that the answer lies in the geometry of the residual stream. In Pre-Norm Transformer architectures, RMSNorm (Zhang and Sennrich, 2019) motivates an angular description of hidden states after rescaling. On this sphere, every layer transition separates exactly into a retained



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

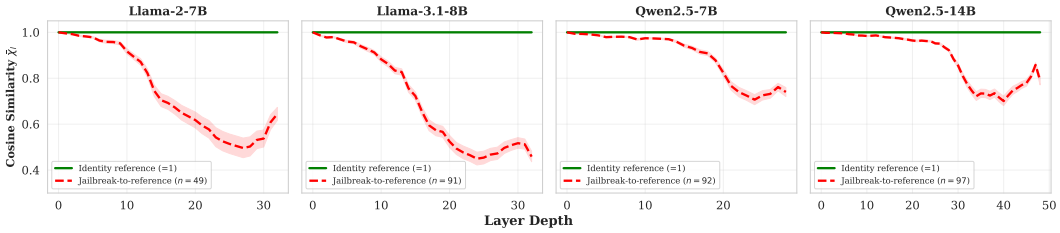


**Figure 1: Paired Reference–Jailbreak Representation Similarity.** Mean cosine similarity $\bar{\chi}_l$ between paired reference and jailbreak hidden states across layers (defined in Appendix C.3); bands are prompt-pair bootstrap 95% confidence intervals. The horizontal line at 1 is the identity reference.

part of the previous state and a tangent contribution that factors into available coordinate capacity and write orientation. This gives a sharp interval for one-step projection updates and a finite-depth lower bound on the directional write budget required to depart from passive propagation. The familiar exponential decay of a fixed-direction projection is recovered when this budget is zero and retention is constant, a passive reference rather than a prescribed shape for observed trajectories. This geometry supplies a measurable signature for *structural hallucination*, the semantic-level failure in which constraint transmission weakens while stable behavior violates the constraint.

This decomposition turns the two channels into measured quantities and reveals a consistent empirical pattern. Across the tested Llama and Qwen families on HarmBench, refused trajectories receive systematically larger safety-directed tangent contributions than jailbroken trajectories, with large effect sizes in every tested model. The retention contrast points in the opposite direction and therefore does not account for the safety-coordinate gap; the tangent ordering also persists after adjustment for layer and current projection. In short, the attenuation is shared in scale, while the safety-directed feed differs. Building on this analysis, we present the following results:

1. **Exact Geometric Propagation and Passive Reference (M1).** We derive exact propagation formulas and directional-budget bounds for normalized trajectories, with a closed-form characterization of equality and exponential decay as a passive reference.
2. **Safety-Directed Tangent Deficit (M2).** After adjustment for layer and current projection, jailbroken trajectories receive smaller tangent contributions along a cross-fitted safety-associated direction than refused trajectories. The retention contrast partly offsets the tangent contrast.
3. **Empirical Evaluation.** Safety projections correlate with attention to explicit safety instructions, while paired early-layer geometry carries information about jailbreak outcomes. A separate GSM8K evaluation on Llama-3.1-8B finds similar accuracy under clean prompting and the tested jailbreak prefix.

## 2 Related Work

**Stability of Alignment Constraints.** The instability of aligned behavior is a challenge in modern safety research. While optimization-based alignment is standard (Ouyang et al., 2022; Bai et al., 2022), it remains vulnerable to adversarial attacks (Wei et al., 2023; Zou et al., 2023; Chao et al., 2024; Greenblatt et al., 2024; Hubinger et al., 2024) and generalized misalignment from narrow finetuning (Betley et al., 2026). Mechanistic interpretability provides a common lens for diagnosing these failures (Bereska and Gavves, 2024; Anthropic, 2021; Zou et al., 2025; Bartoszcze et al., 2025). Several studies identify linear refusal- or safety-associated directions (Arditi et al., 2024; Lin et al., 2024; Sheng et al., 2026; Turner et al., 2024) and layer-localized safety structure (Li et al., 2025). Parallel work examines structural mechanisms for hallucination (Gao et al., 2025) and geometric knowledge organization (Zhang et al., 2025). These studies motivate geometric measurements but do not separate direction-independent inter-layer retention from direction-specific tangent construction. We analyze this distinction along a measured safety coordinate.

**Geometric Dynamics of Deep Representations.** Theoretical analysis of deep networks, grounded in stochastic calculus (Øksendal, 2003; Vershynin, 2018), has traditionally focused on *optimization dynamics* (Ren and Sutherland, 2025; Belkin, 2018; Mandt et al., 2017) or *static expressivity* (Telgarsky, 2016; Zhu et al., 2025). Recently, researchers have begun mapping the *geometric organization* of knowledge (Benfenati et al., 2024; Xie et al., 2026) and belief states (Shai et al., 2024; Zhao et al.,

2024) within the residual stream. A parallel line of work models the Transformer forward pass as a continuous dynamical system on the sphere. Sander et al. (2022) interpret doubly stochastic attention as a Wasserstein gradient flow, establishing a diffusion-limit perspective for self-attention layers. Geshkovski et al. (2023) formalize Transformers as interacting particle systems on the unit sphere, proving long-time clustering of token representations under deterministic dynamics. Fedorov et al. (2026) extend this to stochastic settings, showing that initialization noise induces phase transitions between clustering and antipodal separation. These works characterize the *collective* dynamics of multi-token systems (convergence, clustering). Yet, these methods operate distinctly from the *layer-wise trajectory* of inference. We study that trajectory through an exact discrete decomposition of normalized inter-layer states, with the deterministic constant-retention reference as a limiting case rather than a law of observed trajectories.

# 3 Geometric Analysis

We develop a geometric account of how safety-associated coordinates propagate through successive layers of a fixed model. M1 derives an exact recursion separating the retained projection from the tangent contribution, a sharp projection interval, and a lower bound on the directional write budget. Exponential decay is recovered as the passive constant-retention specialization (§3.2). M2 predicts a smaller realized safety-directed tangent contribution in jailbroken trajectories than in refused trajectories after adjustment for layer and current projection. A spherical-cap result characterizes the geometry of an idealized positive-threshold region (§3.3).

## 3.1 Preliminaries

We formalize the constraint pathway $X \xrightarrow{\Pi} L_1 \longrightarrow L_2 \longrightarrow L_3$, where $X$ is the *prompt*, $\Pi$ is a fixed *safety policy*, $L_1 := \Pi(X)$ is the *prompt-conditioned constraint* that the policy requires for this prompt, $L_2$ is the *internal constraint state*, the part of the hidden computation that encodes and maintains the constraint, and $L_3$ is the *output behavioral state*, the realized behavior class of the response (for example refusal, safe assistance, harmful compliance, or irrelevant failure). RMSNorm (Zhang and Sennrich, 2019) motivates writing hidden states as normalized directions $h_l \in \mathbb{S}^{d-1}$. We use linear probes to define measurement directions. For a one-dimensional measurement, $v_{L_1} \in \mathbb{S}^{d-1}$ denotes the fixed direction associated with the constraint (the safety-associated direction), and $\rho_l := \langle h_l, v_{L_1} \rangle$ is its layer-$l$ projection. The projection $\rho_l$ is an observable coordinate of the internal constraint state $L_2$ and is the coordinate used by the M2 comparison. This is the $k{=}1$ case of a low-rank alignment subspace; Appendix E gives the corresponding subspace formulation for distributed linear representations. Task capability is measured behaviorally in the experiments. The main parameters are the number of layers $L$, layer index $l \in \{0, 1, \ldots, L\}$, hidden dimension $d$, and residual strength $\beta$, whose raw-update and layer-indexed variants $\beta_{\text{raw}}$ and $\beta_l$ are defined where used. Formal definitions are collected in Appendix B.

**Scope.** The propagation, orientation, and budget results are exact for any sequence of nonzero hidden states after $\ell_2$ normalization. The unit tangential update relates the retention coefficient $c_l$ to $\beta_l$, while isotropy enters the optional terminal-crossing proxy of Remark D.5. The natural-language experiments compare realized tangent contributions in refused and jailbroken trajectories after adjustment for layer and current projection. Causal control by the measured direction and attack-dependent initiation are questions for direct interventions. Further limits are discussed in Appendix A.

**Assumption 3.1** (Information Bottleneck)**.** The behavioral state $L_3$ is approximately conditionally independent of the prompt-conditioned constraint $L_1$ given the internal constraint state $L_2$:

$$I(L_1; L_3 \mid L_2) \leq \varepsilon_{\text{leak}}, \tag{1}$$

where $\varepsilon_{\text{leak}} \geq 0$ accounts for constraint influence on behavior that bypasses the internal constraint state, for example shallow lexical effects of the prompt. The idealized case $\varepsilon_{\text{leak}} = 0$ corresponds to the constraint acting on behavior only through $L_2$.

### 3.1.1 Structural Hallucination

We formalize constraint failure in terms of internal constraint information and output behavior. Write $I(\cdot;\cdot)$ for mutual information and $H(\cdot \mid \cdot)$ for conditional entropy. Let $V : \mathcal{L}_1 \times \mathcal{L}_3 \to \{0, 1\}$

be a measurable verifier, with $V(\ell_1, \ell_3) = 1$ if and only if output behavior $\ell_3$ satisfies the prompt-conditioned constraint $\ell_1$.

**Definition 3.2** (Structural Hallucination)**.** Let $\mathbb{P}_{\theta,\mathcal{D}}$ be the probability measure induced by $X \sim \mathcal{D}$ and generation under $P_\theta$ on the common sample space of $X, L_1, L_2, L_3$. For $\varepsilon, \delta \geq 0$ and $\zeta \in [0, 1]$, the model exhibits $(\varepsilon, \delta, \zeta)$-*structural hallucination* on $\mathcal{D}$ if there exists a non-degenerate prompt event $A \in \sigma(X)$ with $\mathbb{P}_{\theta,\mathcal{D}}(A) > 0$ such that, under the conditional law $\mathbb{P}_A := \mathbb{P}_{\theta,\mathcal{D}}(\cdot \mid A)$,

$$(\mathbf{C1})\ I_A(L_1; L_2) \leq \varepsilon, \quad (\mathbf{C2})\ H_A(L_3 \mid L_2) \leq \delta, \quad (\mathbf{C3})\ \mathbb{P}_A(V(L_1, L_3) = 0) \geq 1 - \zeta. \tag{2}$$

Here $I_A$ and $H_A$ are computed under $\mathbb{P}_A$, and non-degeneracy means $\varepsilon < H_A(L_1) < \infty$.

Condition $(\mathbf{C1})$ limits how much the internal constraint state reveals about the prompt-conditioned constraint, while $(\mathbf{C2})$ limits the uncertainty in output behavior given that state. Condition $(\mathbf{C3})$ requires constraint violations with probability at least $1 - \zeta$ on the same event $A$.

**Proposition 3.3** (Information-Theoretic Bound)**.** *Under the conditional information bottleneck $I_A(L_1; L_3 \mid L_2) \leq \varepsilon_{leak}$, if $(\mathbf{C1})$ of Definition 3.2 holds under $\mathbb{P}_A$, then:*

$$I_A(L_1; L_3) \leq \varepsilon + \varepsilon_{leak}. \tag{3}$$

*In the exact bottleneck case $\varepsilon_{leak} = 0$, this gives $I_A(L_1; L_3) \leq \varepsilon$.*

A proof and a justification of Assumption 3.1 are given in Appendix B.5.

### 3.2 Directional Propagation on the Sphere (M1)

We now analyze how a measured coordinate of the internal constraint state $L_2$ evolves across layers.

#### 3.2.1 Residual Dynamics on the Sphere

Under the normalized-state convention in Assumption B.9, a residual transition at layer $l$ can be written as:

$$h_{l+1} = \mathrm{Normalize}(h_l + \beta_{\mathrm{raw}} \xi_l^{\mathrm{raw}}), \tag{4}$$

where $\mathrm{Normalize}(u) := u/\|u\|_2$ and $h_l + \beta_{\mathrm{raw}} \xi_l^{\mathrm{raw}} \neq 0$. Here $\beta_{\mathrm{raw}} > 0$ and $\beta_{\mathrm{raw}} \xi_l^{\mathrm{raw}}$ is the total residual increment divided by the norm of the incoming unnormalized hidden state. The superscript raw denotes the update before renormalization. For any fixed unit direction $v$, write $\rho_l := \langle h_l, v \rangle$. The safety coordinate is the case $v = v_{L_1}$. Appendix C.1.2 gives the corresponding projection update.

For any adjacent normalized states, define the *retention coefficient* $c_l := \langle h_{l+1}, h_l \rangle$, the tangent residual $w_l := h_{l+1} - c_l h_l$, and the realized tangent contribution $g_l(v) := \langle w_l, v \rangle$. Projecting the orthogonal decomposition $h_{l+1} = c_l h_l + w_l$ onto $v$ gives

$$\rho_{l+1} = c_l \rho_l + g_l(v), \qquad \Delta\rho_l = -(1 - c_l)\rho_l + g_l(v), \tag{5}$$

where $\Delta\rho_l := \rho_{l+1} - \rho_l$. The retention coefficient $c_l$ is shared across measurement directions, while $g_l(v)$ is the projection of the tangent residual onto $v$. Writing $K_c(a, b) := \prod_{r=a}^{b-1} c_r$ with $K_c(a, a) = 1$, iteration gives

$$\rho_L = K_c(0, L)\rho_0 + \sum_{j=0}^{L-1} K_c(j + 1, L) g_j(v). \tag{6}$$

The first term propagates the initial projection through all layers. Each tangent contribution is weighted by the retention in the subsequent layers. Eq. 5 and Eq. 6 hold *pathwise*: they are exact for each individual normalized hidden-state trajectory.

#### 3.2.2 Write Orientation and Directional Budget

For a fixed current state $h_l$, measurement direction $v$, and retention coefficient $c_l$, we characterize the possible values of the next projection $\rho_{l+1}$.

**Proposition 3.4** (Sharp Projection Interval). *Let $h_l, h_{l+1} \in \mathbb{S}^{d-1}$ with $c_l = \langle h_{l+1}, h_l \rangle$ and $v \in \mathbb{S}^{d-1}$, and write $\rho_l := \langle h_l, v \rangle$ and $g_l(v) := \langle w_l, v \rangle$ as above. Define*

$$\eta_l(v) := \frac{g_l(v)}{\sqrt{1-c_l^2}\sqrt{1-\rho_l^2}}, \tag{7}$$

*when the denominator is nonzero. In this case, $\eta_l(v)$ is the cosine of the angle between $w_l$ and $v - \rho_l h_l$, the component of $v$ tangent to the sphere at $h_l$. If the denominator vanishes, then $g_l(v) = 0$, and we set $\eta_l(v) := 0$. Then $|\eta_l(v)| \le 1$ and*

$$\rho_{l+1} = c_l \rho_l + \sqrt{1-c_l^2}\sqrt{1-\rho_l^2}\,\eta_l(v). \tag{8}$$

*Consequently,*

$$\rho_{l+1} \in \left[ c_l\rho_l - \sqrt{(1-c_l^2)(1-\rho_l^2)},\; c_l\rho_l + \sqrt{(1-c_l^2)(1-\rho_l^2)} \right]. \tag{9}$$

*For fixed $(h_l, v, c_l)$, both endpoints are attained by a unit next state with retention $c_l$ for $d \ge 2$, and every point of the interval is attained for $d \ge 3$.*

Eq. 6 expresses the departure from passive propagation as a sum of tangent contributions weighted by subsequent retention. For a unit-state trajectory $h_0, \dots, h_L$ and a fixed unit direction $v$, we accumulate the squared contributions, normalized by the squared tangent magnitude at each layer, in the directional write budget

$$\mathcal{G}_L(v) := \sum_{j=0}^{L-1} \frac{g_j(v)^2}{1-c_j^2}, \tag{10}$$

with zero terms when $|c_j| = 1$, where unit-state geometry forces $g_j(v) = 0$. The following identity identifies the budget lower bound and its exact remainder.

**Theorem 3.5** (Exact Directional Budget Decomposition). *Let $h_0, \dots, h_L \in \mathbb{S}^{d-1}$ and fix $v \in \mathbb{S}^{d-1}$. Use the quantities $\rho_j, c_j, g_j(v)$ and $\mathcal{G}_L(v)$ defined above, and assume $|K_c(0,L)| < 1$. Set $D_L(v) := \rho_L - K_c(0,L)\rho_0$ and define*

$$g_j^\star(v) := \frac{D_L(v)}{1 - K_c(0,L)^2}\, K_c(j+1, L)(1-c_j^2), \qquad 0 \le j < L. \tag{11}$$

*Then*

$$\mathcal{G}_L(v) = \frac{D_L(v)^2}{1-K_c(0,L)^2} + \sum_{j=0}^{L-1} \frac{\big(g_j(v) - g_j^\star(v)\big)^2}{1-c_j^2}, \tag{12}$$

*The same zero convention applies to the remainder, which vanishes if and only if $g_j(v) = g_j^\star(v)$ for every $j < L$.*

The remainder measures the weighted squared deviation from $g_j^\star(v)$. We next determine when these contributions are realizable by unit states subject only to the prescribed initial state, retentions, and terminal projection.

**Proposition 3.6** (Attainment of the Directional Budget Bound). *Let $d \ge 3$ and $h_0, \dots, h_L, v \in \mathbb{S}^{d-1}$. Use $\rho_j$ and $c_j$ defined above, and assume $|K_c(0,L)| < 1$. For $D_L(v)$ and $g_j^\star(v)$ from Theorem 3.5, define the candidate projection path by the closed form*

$$\rho_j^\star := K_c(0,j)\rho_0 + \frac{D_L(v) K_c(j,L)}{1-K_c(0,L)^2}\big(1 - K_c(0,j)^2\big), \qquad 0 \le j \le L. \tag{13}$$

*Among all unit-state trajectories with the same initial state $h_0$, fixed direction $v$, retention coefficients $(c_j)_{j=0}^{L-1}$, and terminal projection $\rho_L$, the lower bound $D_L(v)^2/(1-K_c(0,L)^2)$ is attained if and only if*

$$\big(g_j^\star(v)\big)^2 \le (1-c_j^2)\big(1-(\rho_j^\star)^2\big) \qquad \text{for every } j < L. \tag{14}$$

*Every attaining trajectory has projection sequence $(\rho_j^\star)_{j=0}^{L}$ and directional contributions $(g_j^\star(v))_{j=0}^{L-1}$.*

Proofs and supplementary budget bounds are given in Appendix C.2. The feasibility condition holds for all available archived middle-window trajectories in our evaluation (§ 4.2.1).

The passive reference is the zero-budget case $\mathcal{G}_L(v) = 0$, which gives $\rho_L = K_c(0,L)\rho_0$. Constant retention then yields the exponential reference below.

### 3.2.3 PASSIVE CONSTANT-RETENTION REFERENCE

**Theorem 3.7** (Exponential Correlation Decay)**.** *For a fixed direction $v \in \mathbb{S}^{d-1}$, consider the passive reference $g_l(v) = 0$ with constant retention $c_l = q \in (0, 1)$ for $0 \le l < L$. Then*

$$\rho_L = q^L \rho_0 = e^{-\lambda L} \rho_0, \qquad \lambda := -\log q > 0. \tag{15}$$

*Consequently, if $\rho_0$ is integrable, then $\mathbb{E}[\rho_L] = q^L \mathbb{E}[\rho_0]$.*

*Remark* 3.8 (Exact Retention and Its Small-Step Approximation). Under the unit tangential update $h_{l+1} = \mathrm{Normalize}(h_l + \beta_l \xi_l)$ with $\xi_l \perp h_l$, $\|\xi_l\|_2 = 1$, and $\beta_l \ge 0$, the exact retention is

$$c_l = \frac{1}{\sqrt{1+\beta_l^2}}, \qquad \left| 1 - c_l - \frac{\beta_l^2}{2} \right| \le \frac{\beta_l^4}{2}. \tag{16}$$

Thus $\beta_l^2/2$ is a small-step approximation to the measured coefficient $1 - c_l$, not its definition.

*Remark* 3.9 (Measured Retention). The experiments compute $c_{i,l} = \langle h_{i,l+1}, h_{i,l} \rangle$ directly from adjacent normalized states; no residual-strength parameter is fitted.

For an initial projection above a positive threshold, Proposition 3.10 gives the depth at which this passive reference reaches that threshold.

### 3.2.4 RETENTION AND TANGENT BALANCE

To compare projection changes across prompt trajectories, we take expectations in Eq. 5 for a fixed model and measurement direction:

$$\mathbb{E}[\rho_{l+1} - \rho_l] = -\underbrace{\mathbb{E}[\kappa_l \rho_l]}_{\text{Retention Term}} + \underbrace{\mathbb{E}[g_l(v)]}_{\text{Realized Tangent Contribution}}, \qquad \kappa_l := 1 - c_l. \tag{17}$$

The same identity holds within each outcome group. Comparing its two terms separates the retention and tangent contributions to the difference in mean projection change between refused and jailbroken trajectories.

### 3.2.5 COMPARATIVE PREDICTION (M2)

M2 tests whether jailbroken trajectories receive smaller tangent contributions along the safety-associated direction after adjustment for layer and current projection. Let $J_i \in \{\mathrm{Ref}, \mathrm{JB}\}$ denote the observed response label for prompt $i$, and let $l \in \mathcal{W}$ index the prespecified middle-layer window. The fold-specific unit direction $v$ is cross-fitted, oriented so that larger projections are associated with refusal, and held fixed across layers. For normalized held-out states, write $\rho_{i,l} := \langle h_{i,l}, v \rangle$. The corresponding tangent contribution is

$$g_{i,l}(v) := \rho_{i,l+1} - \rho_{i,l} + \kappa_{i,l} \rho_{i,l}, \tag{18}$$

where $\kappa_{i,l} := 1 - \langle h_{i,l+1}, h_{i,l} \rangle$.

Eq. 8 shows that the feasible tangent contribution depends on the current projection. We therefore adjust for layer and current projection when comparing the two outcome groups, using the linear regression

$$g_{i,l}(v) = \alpha_l + b_l \rho_{i,l} + \gamma_{\mathrm{ov}} \mathbf{1}\{J_i = \mathrm{JB}\} + \varepsilon_{i,l} \tag{19}$$

on the overlap of their central projection ranges at each layer. This restriction keeps the comparison within projection ranges represented in both groups. Here $\alpha_l$ and $b_l$ are layer-specific intercepts and projection slopes, $\gamma_{\mathrm{ov}}$ is the shared jailbroken-minus-refused coefficient, and $\varepsilon_{i,l}$ is the regression residual. M2 predicts $\gamma_{\mathrm{ov}} < 0$, corresponding to a lower fitted tangent contribution for the jailbroken group at a given layer and current projection. The comparison tests whether the group difference extends beyond the current projection to the next layer's tangent contribution. The overlap rule and statistical inference are detailed in Appendix C.3.2.

## 3.3 CONDITIONAL GEOMETRY OF POSITIVE-PROJECTION REGIONS

We next characterize an idealized region defined by a positive threshold on the safety-associated coordinate.

#### 3.3.1 CONSTRAINTS AS A SPHERICAL CAP

The linear probe hypothesis (Anthropic, 2021) implies that constraint-relevant concepts are linearly separable in hidden space. By Assumption B.9, we define a spherical cap:

$$\mathcal{C}(\tau) := \{h \in \mathbb{S}^{d-1} : \langle h, v_{L_1} \rangle \geq \tau\}, \tag{20}$$

where $v_{L_1}$ is the constraint direction and $\tau > 0$ is the decision threshold. Following Corollary D.3, this region occupies an exponentially small fraction of the uniform measure $\sigma_{d-1}$ on $\mathbb{S}^{d-1}$:

$$\frac{\sigma_{d-1}(\mathcal{C}(\tau))}{\sigma_{d-1}(\mathbb{S}^{d-1})} \leq \exp\left(-\frac{(d-1)\tau^2}{2}\right). \tag{21}$$

**Proposition 3.10** (Threshold Depth under Passive Propagation)**.** *Under the passive reference of Thm. 3.7, suppose $\rho_0 = \langle h_0, v_{L_1} \rangle \geq \tau > 0$. The continuous-depth proxy at which $q^L \rho_0$ reaches $\tau$ is*

$$L_{\text{thresh}} := \frac{\log(\rho_0/\tau)}{-\log q}. \tag{22}$$

*For the unit tangential update with constant step $\beta$, as $\beta \to 0^+$ with $\rho_0$ and $\tau$ fixed, this becomes $L_{\text{thresh}} = 2\beta^{-2}\log(\rho_0/\tau) + O(1)$.*

The proof is given in Appendix D.2.2. For the same initial projection and constant retention, remaining above $\tau$ beyond $L_{\text{thresh}}$ requires a positive accumulated tangent contribution in Eq. 6. The empirical analysis uses the exact decomposition of M1 and the adjusted comparison of M2 to examine how measured retention and tangent contributions differ between refused and jailbroken trajectories.

## 4 EXPERIMENTS

We use M1 as the conditional passive reference and test the M2 prediction, with jailbreaking as a proxy for alignment failure.

### 4.1 EXPERIMENTAL SETUP

**Datasets & Models.** We evaluate Llama (2-7B, 3.1-8B) (Meta, 2023; 2024) and Qwen2.5 (7B, 14B) (Qwen et al., 2025). We use 100 semantically matched safe/jailbreak prompt pairs from JailbreakBench (Chao et al., 2024) for trajectory comparisons and 400 HarmBench (Mazeika et al., 2024) prompts per model for the refused-versus-jailbroken analysis. Separately, we evaluate task capability on GSM8K (Cobbe et al., 2021) ($N = 1{,}319$) with Llama-3.1-8B under clean prompting and the tested jailbreak prefix.

**Evaluation Protocol.** The main model runs use temperature 1 and FP32 precision on an RTX PRO 6000. Each GSM8K condition is evaluated over five runs. On HarmBench, we use five-fold cross-fitting to estimate a linear safety direction at the middle layer and evaluate projections on held-out prompts. Hidden states are $\ell_2$-normalized, and the direction is oriented toward refusal and held fixed across layers, as in M2 (§ 3.2.5). For a model with $L$ transformer blocks, group contrasts use per-prompt averages over the prespecified middle window $\mathcal{W} = \{l : 0.25 \leq l/L < 0.75\}$. Outcomes are labeled using the HarmBench official prompts with GPT-5.4 as judge. Statistical inference uses the prompt trajectory as its unit, conditional on the fitted directions (Appendix C.3.2).

### 4.2 VALIDATION OF MAIN RESULTS

We first examine how retention and tangent contributions shape the safety-projection gap. We then relate this geometric comparison to task performance and attention to safety instructions.

#### 4.2.1 M1 REFERENCE AND M2 TANGENT CONTRIBUTION

Figure 2 reports the mean cosine similarity to the initial state, $s_l := \mathbb{E}_i[\cos(h_{i,l}, h_{i,0})]$, which drops through the middle layers and partially recovers later. A state can change direction while its safety projection remains unchanged. We therefore examine the fixed safety projection and its layer-wise updates.

**Table 1: Exact Pathwise M2 Decomposition (HarmBench).** All contrasts and coefficients are significant, with prompt-level permutation $p$ ranging from $1/10001$ to $0.0211$. Tangent effect sizes range from 1.33 to 2.33.

| Model | Net | Retention | Tangent | Hedges' $g$ | $\gamma$ | $\gamma_{\text{ov}}$ | Retained |
|---|---|---|---|---|---|---|---|
| Llama-2-7B | 0.01743 | −0.01875 | 0.03618 | 2.33 | −0.01187 | −0.00703 | 23.5% |
| Llama-3.1-8B | 0.00855 | −0.02055 | 0.02910 | 1.91 | −0.00950 | −0.00820 | 42.9% |
| Qwen2.5-7B | 0.00311 | −0.00701 | 0.01011 | 1.53 | −0.00260 | −0.00237 | 54.2% |
| Qwen2.5-14B | 0.00163 | −0.00419 | 0.00582 | 1.33 | −0.00118 | −0.00107 | 42.9% |

**Table 2: Write orientation and positive directional budget.** Group means of the per-prompt window average of $\eta$ and window sum of $\mathcal{G}^+$. Contrasts are Ref–JB with prompt-level Hedges' $g$.

| | Write orientation $\eta$ | | | | Positive budget $\mathcal{G}^+$ | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Ref. | JB | Ref.–JB | Hedges' $g$ | Ref. | JB | Ref.–JB | Hedges' $g$ |
| Llama-2-7B | 0.04972 | −0.02989 | 0.07961 | 2.38 | 0.17799 | 0.00586 | 0.17213 | 1.84 |
| Llama-3.1-8B | 0.02391 | −0.02666 | 0.05057 | 1.93 | 0.16653 | 0.01039 | 0.15614 | 1.66 |
| Qwen2.5-7B | 0.01865 | −0.00439 | 0.02304 | 1.52 | 0.04271 | 0.00611 | 0.03661 | 1.20 |
| Qwen2.5-14B | 0.00404 | −0.01115 | 0.01519 | 1.36 | 0.03560 | 0.00561 | 0.02998 | 1.00 |

**Tangent contributions account for the gap.** To identify which component accounts for the group difference in projection change, we apply the mean update decomposition in Eq. 17. In Table 1, the refused-minus-jailbroken tangent contrast is positive, with Hedges' $g$ from 1.33 to 2.33, while the retention contrast is negative across all four models. The larger mean projection gain in refused trajectories is therefore carried by the tangent term, with retention partly offsetting it.

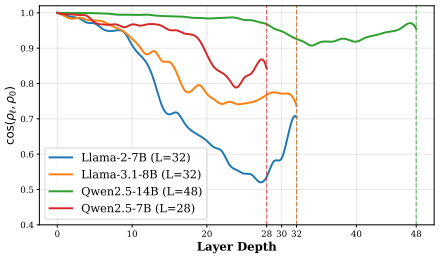


**Figure 2: Trajectory-Decorrelation Proxy.** Self-trajectory cosine $s_l$ decreases through middle layers and partially rebounds later.

The feasible tangent contribution also depends on the current projection (Eq. 8). We therefore compare the groups after adjustment for layer and current projection, using $\gamma$ over the full window and $\gamma_{\text{ov}}$ on overlapping central projection ranges. Both coefficients in Table 1 are negative, supporting the M2 prediction in Eq. 19. The common-support permutation $p$-values are $1/10001$, $1/10001$, $3/10001$, and $0.0211$ for Llama-2-7B, Llama-3.1-8B, Qwen2.5-7B, and Qwen2.5-14B, respectively. The overlap rule and fitting procedure are given in Appendix C.3.2.

**Write orientation.** The same orientation can produce different tangent contributions when geometric capacity differs. We therefore normalize each contribution by its maximum feasible magnitude using $\eta_l(v)$ in Eq. 8. Table 2 shows positive mean orientation for refused trajectories and negative mean orientation for jailbroken trajectories in every model, with Hedges' $g$ of 1.36–2.38. The separation therefore persists relative to each update's feasible projection range.

**Accumulation across depth.** Layer-wise contributions can reinforce or offset one another as they propagate. Eq. 6 weights each contribution by the retention of subsequent layers. To relate the one-step comparison to changes across depth, we examine whole-window projection gains alongside the passive reference. Using each group's mean retention in Thm. 3.7 gives a passive half-life of 4.5 to 11.2 layers, shorter than the middle analysis window (Appendix Table 4). The window-level gains reported in Appendix Table 8 range from 0.06 to 0.26 in mean projection for refused trajectories, compared with −0.02 to 0.07 for jailbroken trajectories. This links M2's tangent deficit to smaller projection gains across depth.

**Attainability of the closed-form solution.** Trajectories with the same retentions and endpoint projections can use different directional budgets. To obtain a minimum-budget reference for each observed window, we evaluate Eq. 13 with its measured retentions and endpoint projections and test the capacity condition in Proposition 3.6. It holds for every available archived trajectory across the four models. The median ratio of minimum to observed budget ranges from 3.3% to 28.1% across

**Table 3: Predicting jailbreak from early trajectory geometry (HarmBench).** Logistic probes use paired reference and jailbreak states from the first $L_{\text{early}}$ layers. Rows compare cosine similarity $\chi$, its layer-wise change $\Delta\chi$, the orthogonal residual magnitude $m^{\perp}$, and their combination (all). Values are mean $\pm$ standard deviation over five cross-validation folds. Best per-model result in **bold**.

| Feature | Llama-2-7B | Llama-3.1-8B | Qwen2.5-7B | Qwen2.5-14B |
|---|---|---|---|---|
| $L_{\text{early}}$ | 16 | 16 | 14 | 24 |
| $\chi$ | $\mathbf{0.975 \pm 0.008}$ | $0.895 \pm 0.033$ | $0.693 \pm 0.105$ | $0.725 \pm 0.081$ |
| $\Delta\chi$ | $0.975 \pm 0.007$ | $0.899 \pm 0.025$ | $0.678 \pm 0.108$ | $0.714 \pm 0.049$ |
| $m^{\perp}$ | $0.970 \pm 0.010$ | $0.899 \pm 0.029$ | $\mathbf{0.719 \pm 0.103}$ | $0.671 \pm 0.072$ |
| all | $0.974 \pm 0.007$ | $\mathbf{0.902 \pm 0.024}$ | $0.664 \pm 0.118$ | $\mathbf{0.736 \pm 0.062}$ |

models. The exact decomposition in Theorem 3.5 thus separates the endpoint requirement from the additional budget associated with the realized layer-wise contributions. Construction and numerical comparisons are given in Appendix C.3.4.

#### 4.2.2 M2: Safety-Coordinate Maintenance and Task Capability

We evaluate task capability separately from the safety-coordinate deficit. On Llama-3.1, GSM8K accuracy is $0.80 \pm 0.012$ under clean prompting and $0.79 \pm 0.007$ under the tested jailbreak prefix (Appendix F). The observed difference is not statistically distinguishable. We also examine coarse attention organization (Hu et al., 2025). Figure 3 shows similar attention concentration and strongly correlated attention-intensity profiles ($r > 0.95$) between reference and jailbreak trajectories.

**Attention to safety instructions.** Similar attention concentration can coexist with a change in which tokens receive attention. We therefore measure attention to a fixed safety preamble to examine how its allocation to safety instructions changes. Across both Llama models and Qwen2.5-7B, jailbreak prompts yield lower preamble attention than reference prompts in $91$ of $92$ layers. This attention correlates positively with the middle-layer safety projection (Spearman $0.43$–$0.63$, Appendix H). These comparisons link lower safety projections to reduced attention to explicit safety instructions, alongside the similar task accuracy observed in the separate GSM8K evaluation.

**Additional experiments and ablation studies.** We further evaluate the diagnostic value of early geometry. Using paired reference and jailbreak representations from the first half of the network, the combined-feature probe in Table 3 reaches cross-validated AUCs of $0.902$–$0.974$ for Llama and $0.664$–$0.736$ for Qwen. Early paired geometry thus carries information about eventual jailbreak success, with stronger discrimination in Llama. Feature definitions are given in Appendix C.3.1. Separate sensitivity analyses vary the reference layer and projection rank. The mean middle-window Ref–JB projection gap remains positive with early, middle, and late reference layers in all four models, and temperature ablations are reported in Appendix G. Appendix Table 9 shows that ranks 1–2 suffice for non-inferiority to the full linear subspace at an AUC margin of $0.02$, supporting a low-rank measurement of the group separation. As an additional direction ablation, we repeat the projection-conditioned regression with label-free unit directions sampled uniformly on $\mathbb{S}^{d-1}$ within each cross-fitting fold. The safety-direction coefficient lies outside the resulting random-direction null in all four models ($p_{\text{dir}} = 0.002$, Appendix Table 7).

## 5 Conclusion

This work develops a geometric theory of how safety-associated representations propagate across language-model layers. The analysis separates retention from tangent contributions and characterizes their accumulation through exact propagation and budget bounds. Across four models, refused trajectories receive larger safety-directed tangent contributions than jailbroken trajectories. The retention contrast has the opposite sign, and the tangent ordering persists after adjustment for layer and current projection. On Llama-3.1-8B, the GSM8K evaluation finds no statistically distinguishable accuracy difference between clean prompting and the tested jailbreak prefix. These findings distinguish continued task performance from the maintenance of safety-associated representations. For alignment research, this distinction motivates studying how safety constraints are sustained throughout computation, informing the development of capable and reliably controlled AI systems.

## LLM USE DISCLOSURE

We used LLM-based assistants to support manuscript polishing and coding. These activities were conducted under the authors' direction, with suggestions reviewed and revised against the mathematical definitions and recorded experimental evidence. Part of the mathematics had proved with the assistance of LLM, and all the proofs were carried out with authors intervention and verification.

## REPRODUCIBILITY.

This work evaluates the proposed geometric decomposition on four publicly released Llama and Qwen checkpoints using JailbreakBench, HarmBench, and GSM8K. The cross-fitted probe construction, fixed-direction measurements, projection-conditioned regression, uncertainty protocol, and supplementary diagnostics are documented in Appendices C.3.2–H. The theoretical statements are made under the conditions stated in the Method, with proofs and numerical checks in Appendices C.2–D. Code and analysis artifacts will be released upon acceptance.

## REFERENCES

Anthropic. A mathematical framework for transformer circuits, 2021.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 136037–136083. Curran Associates, Inc., 2024. doi: 10.52202/079017-4322.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073, 2022. doi: 10.48550/ARXIV.2212.08073.

Keith Ball. An elementary introduction to modern convex geometry. In Silvio Levy, editor, *Flavors of Geometry*, volume 31 of *Mathematical Sciences Research Institute Publications*, pages 1–58. Cambridge University Press, Cambridge, 1997.

Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. Representation engineering for large-language models: Survey and research challenges. *CoRR*, abs/2502.17601, 2025. doi: 10.48550/ARXIV.2502.17601.

Mikhail Belkin. Approximation beats concentration? an approximation view on inference with smooth radial kernels. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, Proceedings of Machine Learning Research, pages 1348–1361. PMLR, 2018.

Alessandro Benfenati, Alfio Ferrara, Alessio Marta, Davide Riva, and Elisabetta Rocchetti. Unveiling transformer perception by exploring input manifolds. *CoRR*, abs/2410.06019, 2024. doi: 10.48550/ARXIV.2410.06019.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety – a review, 2024.

Jan Betley, Niels Warncke, Anna Sztyber-Betley, Daniel Tan, Xuchan Bao, Martín Soto, Megha Srivastava, Nathan Labenz, and Owain Evans. Training large language models on narrow tasks can lead to broad misalignment. *Nat.*, 649(8097):584–589, 2026. ISSN 1476-4687. doi: 10.1038/S41586-025-09937-5.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 55005–55029. Curran Associates, Inc., 2024. doi: 10.52202/079017-1745.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021.

Lev Fedorov, Michael E. Sander, Romuald Elie, Pierre Marion, and Mathieu Laurière. Clustering in deep stochastic transformers. *CoRR*, abs/2601.21942, 2026. doi: 10.48550/ARXIV.2601.21942.

Cheng Gao, Huimin Chen, Chaojun Xiao, Zhiyi Chen, Zhiyuan Liu, and Maosong Sun. H-neurons: On the existence, impact, and origin of hallucination-associated neurons in llms, 2025.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *CoRR*, abs/2312.10794, 2023. doi: 10.48550/ARXIV.2312.10794.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Samuel Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models. *CoRR*, abs/2412.14093, 2024. doi: 10.48550/ARXIV.2412.14093.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022. doi: 10.48550/ARXIV.2203.15556.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Attention slipping: A mechanistic understanding of jailbreak attacks and defenses in llms. *CoRR*, abs/2507.04365, 2025. doi: 10.48550/ARXIV.2507.04365.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam S. Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger B. Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul F. Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training. *CoRR*, abs/2401.05566, 2024. doi: 10.48550/ARXIV.2401.05566.

Ching-Chia Kao, Chia-Mu Yu, Chun-Shien Lu, and Chu-Song Chen. Safety alignment depth in large language models: A markov chain perspective. *CoRR*, abs/2502.00669, 2025. doi: 10.48550/ARXIV.2502.00669.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety layers in aligned large language models: The key to llm security. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 98163–98189, 2025.

Yuping Lin, Pengfei He, Han Xu, Yue Xing, Makoto Yamada, Hui Liu, and Jiliang Tang. Towards understanding jailbreak attacks in LLMs: A representation space analysis. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7067–7085, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.401.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. *J. Mach. Learn. Res.*, 18:134:1–134:35, 2017.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 35181–35224. PMLR, 21–27 Jul 2024.

Meta. Llama 2: Open foundation and fine-tuned chat models, 2023.

Meta. The llama 3 herd of models, 2024.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer Berlin, Heidelberg, 6th edition, 2003. doi: 10.1007/978-3-642-14394-6.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, Proceedings of Machine Learning Research, pages 3515–3530. PMLR, 2022.

Adam S. Shai, Lucas Teixeira, Alexander Gietelink Oldenziel, Sarah Marzen, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. Alphasteer: Learning refusal steering with principled null-space constraint. In *The Fourteenth International Conference on Learning Representations*, 2026.

Matus Telgarsky. benefits of depth in neural networks. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1517–1539, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Kuai Yu, Liang Zhao, Shangyan Zhou, Zhean Xu, Zhengyan Zhang, Wangding Zeng, Shengding Hu, Yuqing Wang, Jingyang Yuan, Lean Wang, and Wenfeng Liang. mhc: Manifold-constrained hyper-connections, 2026.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12360–12371, 2019.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418, 12 2025. ISSN 0891-2017. doi: 10.1162/COLI.a.16.

Yu Zhao, Xiaotang Du, Giwon Hong, Aryo Pradipta Gema, Alessio Devoto, Hongru WANG, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. Analysing the residual stream of language models under knowledge conflicts. In *MINT: Foundation Model Interventions*, 2024.

Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart Russell, and Yuandong Tian. Reasoning by superposition: A theoretical perspective on chain of continuous thought, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025.

**Roadmap.** We first detail reproducibility, scope, and applicability in §A, followed by formal definitions in §B. Detailed proofs and analyses for M1 and M2 are presented in §C and §D, followed by a low-rank alignment-subspace extension in §E. We then present additional mechanism, robustness, and diagnostic analyses in §F, §G, and §H, concluding with limitations and future work in §I.

# A Reproducibility, Scope, and Applicability Analysis

**Scope.** The exact pathwise decomposition applies to any sequence of nonzero hidden states after $\ell_2$ normalization. The empirical analysis evaluates fixed cross-fitted linear safety coordinates in Llama (Meta, 2024) and Qwen (Qwen et al., 2025) models trained with standard post-training procedures. A nonlinear or distributed safety representation may require a richer measurement than a single direction, and the tested checkpoints do not cover all alignment procedures or architectures.

In the tested models and benchmarks, the supported empirical result is a comparative tangent-contribution difference after adjustment for layer and current projection. It does not identify attack initiation, microscopic predictable forcing, or causal control by the measured direction.

## A.1 Real-World Architectures

**1. Normalized-State Geometry.** The exact decomposition is applied after $\ell_2$ normalization of each hidden state.

- **Modeling role:** Normalization places adjacent states on the unit sphere and makes $c_l$, $w_l$, and $g_l(v)$ exact observables.
- **Outside scope:** The analysis does not equate normalized angular geometry with the unnormalized magnitude dynamics of the residual stream.

**2. Linear Safety Coordinate.** We measure safety-associated variation through a fixed direction $v_{L_1}$.

- **Measurement role:** Prior work identifies low-dimensional refusal-associated structure in standard RLHF models (Arditi et al., 2024). The fixed direction is a measurement coordinate, not an assumption that safety is intrinsically one-dimensional.
- **Outside scope: Nonlinear alignment mechanisms.** If safety behavior is implemented by nonlinear detector circuits, the relevant alignment coordinate may no longer be captured by a single fixed linear projection $\rho_l$.

**3. Layer-Wise Propagation.** The main result is finite and discrete.

- **Applies:** The variable-coefficient propagation formula holds for every finite layer sequence, including nonstationary trajectories.
- **Outside scope:** A continuous diffusion interpretation requires additional scaling assumptions and is not used for the empirical M2 claim.

**4. Architecture-Dependent Effect Size.** Retention and tangent magnitudes vary across architectures. The primary empirical claim is the within-model M2 comparison: jailbroken trajectories exhibit smaller realized tangent contributions than refused trajectories after adjustment for layer and current projection.

**5. Causal Scope.** Cross-fitting prevents the same prompts from fitting and evaluating the alignment direction, while the projection-conditioned regression controls for layer and current scalar projection. Direct restoration or ablation of the measured direction would test the stronger causal claim and remains future work.

# B Preliminaries: Formal Definitions

This appendix provides the measure-theoretic foundations for the semantic-level analysis introduced in §3.1.

### B.1 Semantic Level Spaces

**Definition B.1** (Complete Probability Space)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space where $\Omega$ is the sample space, $\mathcal{F}$ is a $\sigma$-algebra, and $\mathbb{P}$ is a probability measure.

**Definition B.2** (Semantic Level Spaces)**.** Let $\mathcal{X} := \mathcal{V}^*$ denote the discrete prompt space, endowed with the power-set $\sigma$-algebra, and let $X : \Omega \to \mathcal{X}$ be the prompt. A fixed *safety policy* is a measurable map $\Pi : \mathcal{X} \to \mathcal{L}_1$. We define three further measurable spaces with random variables $L_i : \Omega \to \mathcal{L}_i$:

(i) $\mathcal{L}_1$: **Constraint space** — a Polish space encoding the prompt-conditioned safety constraint, endowed with its Borel $\sigma$-algebra; $L_1 := \Pi(X)$.

(ii) $\mathcal{L}_2$: **Internal constraint-state space** — a Polish space encoding the internally maintained representation of the constraint, endowed with its Borel $\sigma$-algebra. The projection $\rho_l$ of §3.1 is an observable coordinate of $L_2$, not $L_2$ itself.

(iii) $\mathcal{L}_3$: **Output behavioral space** — a finite set of behavior classes, for example {refusal, safe assistance, harmful compliance, irrelevant failure}, endowed with the power-set $\sigma$-algebra. Conditional entropies of $L_3$ are therefore bounded discrete entropies.

### B.2 Information-Theoretic Quantities

**Definition B.3** (Mutual Information)**.** For random variables $X, Y$ with joint law $P_{XY}$ and marginals $P_X, P_Y$, the mutual information is

$$I(X;Y) := D_{\mathrm{KL}}(P_{XY} \,\|\, P_X \otimes P_Y)\,. \tag{23}$$

Here $D_{\mathrm{KL}}$ denotes Kullback–Leibler divergence. When densities exist, this reduces to the usual integral formula $\int p(x,y) \log \frac{p(x,y)}{p(x)p(y)}\, dx\, dy$.

**Definition B.4** (Conditional Mutual Information)**.** For random variables $X, Y, Z$,

$$I(X;Y \mid Z) := \mathbb{E}_Z\big[D_{\mathrm{KL}}\big(P_{XY|Z} \,\big\|\, P_{X|Z} \otimes P_{Y|Z}\big)\big]\,. \tag{24}$$

Assumption 3.1 bounds $I(L_1; L_3 \mid L_2)$ by $\varepsilon_{\text{leak}}$.

### B.3 Spherical Geometry

**Definition B.5** (Unit Hypersphere)**.**

$$\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}, \tag{25}$$

This is a $(d-1)$-dimensional compact Riemannian manifold.

**Definition B.6** (Tangent Space)**.** At $h \in \mathbb{S}^{d-1}$:

$$T_h\mathbb{S}^{d-1} = \{\xi \in \mathbb{R}^d : \langle \xi, h \rangle = 0\}. \tag{26}$$

This is a $(d-1)$-dimensional linear subspace orthogonal to $h$.

**Definition B.7** (Uniform Measure on Sphere)**.** The uniform (Haar) measure $\sigma_{d-1}$ on $\mathbb{S}^{d-1}$ is the unique rotation-invariant probability measure.

**Definition B.8** (Geodesic Distance on Sphere)**.** For $u, v \in \mathbb{S}^{d-1}$, the geodesic (shortest path) distance is:

$$d_{\mathbb{S}}(u, v) = \arccos(\langle u, v \rangle) \in [0, \pi], \tag{27}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product.

### B.4 Normalized-State Convention

Modern transformer models (e.g., Llama, Qwen) use RMSNorm to rescale sublayer inputs. For our directional analysis, we explicitly normalize the residual-stream states to study their angular dynamics.

**Assumption B.9** (Normalized-State Convention)**.** Let $\widetilde{h}_l$ denote the unnormalized residual-stream state. We analyze $h_l := \widetilde{h}_l / \|\widetilde{h}_l\|_2 \in \mathbb{S}^{d-1}$ whenever $\widetilde{h}_l \neq 0$. All retention and tangent identities are statements about these normalized states and do not impose a distribution on their transitions.

### B.5 Proofs for Structural Hallucination Propositions

*Justification of Assumption 3.1.* Under the constraint-pathway reading, the assumption states that the prompt-conditioned constraint influences the output behavior only through its internal encoding: behavior is produced from the hidden computation, and the constraint enters that computation through the state that represents it. Linear-probing evidence that safety-relevant information is carried by identifiable internal directions (Zou et al., 2025; Arditi et al., 2024) supports treating this encoding as the operative channel. The term $\varepsilon_{\text{leak}}$ collects constraint influence that bypasses the internal constraint state, for example shallow lexical effects of the prompt on the output. The assumption is a modeling idealization, not an architectural theorem. □

*Remark* B.10 (Conditioning and Bottleneck Scope). Definition 3.2 evaluates the failure mode under a conditional law $\mathbb{P}_A$. Since conditioning can change conditional independence relations, Proposition 3.3 explicitly assumes the corresponding conditional bottleneck $I_A(L_1; L_3 \mid L_2) \le \varepsilon_{\text{leak}}$ rather than deriving it automatically from the unconditional statement.

*Proof of Proposition 3.3.* By the chain rule for mutual information:

$$I_A(L_1; L_3) \le I_A(L_1; L_2, L_3) = I_A(L_1; L_2) + I_A(L_1; L_3 \mid L_2). \tag{28}$$

The first term satisfies $I_A(L_1; L_2) \le \varepsilon$ by $(\mathbf{C1})$. The second term is bounded by the conditional bottleneck assumption $I_A(L_1; L_3 \mid L_2) \le \varepsilon_{\text{leak}}$. Thus $I_A(L_1; L_3) \le \varepsilon + \varepsilon_{\text{leak}}$. □

## C Detailed Results for M1

This section proves the exact retention and tangent decomposition used in §3.2, its variable-coefficient propagation formula, the sharp projection interval, the directional write-budget bound with its exact decomposition and attainment condition, and the passive constant-retention specialization.

### C.1 Spherical Geometry and Single-Step Dynamics

#### C.1.1 RMSNorm and Spherical Geometry

We recall the geometric definitions from Appendix B. The state space is the $(d-1)$-dimensional manifold $\mathbb{S}^{d-1}$ (Def. B.5) with tangent spaces $T_h\mathbb{S}^{d-1}$ (Def. B.6). We follow the notation of §3.1: $l \in \{0, \dots, L\}$ indexes transformer layers, $L$ is the total number of layers, $d$ is the hidden dimension, and $h_l \in \mathbb{S}^{d-1}$ denotes the normalized hidden state at layer $l$. All inner products and norms are Euclidean unless otherwise stated.

**Lemma C.1** (RMSNorm Geometry)**.** *Let $y_l$ denote the RMSNorm output of a nonzero input, with the additive normalization stabilizer set to zero and entrywise nonzero learnable gains $\psi \in \mathbb{R}^d$. Then $y_l$ lies on a hyper-ellipsoid $\mathcal{E}$ determined by $\psi$:*

$$\mathcal{E} = \Big\{ y \in \mathbb{R}^d \,:\, \sum_{i=1}^{d} \Big(\frac{y_i}{\psi_i}\Big)^2 = d \Big\}. \tag{29}$$

*The normalized direction $y_l/\|y_l\|$ lies on $\mathbb{S}^{d-1}$. The propagation analysis uses the residual-stream directions $h_l$ defined in Assumption B.9; its unit-sphere identities apply independently of the RMSNorm gains.*

#### C.1.2 Residual Update

Consider the residual update in normalized coordinates before re-normalization:

$$h_{l+1}^{\text{pre}} = h_l + \beta_{\text{raw}} \cdot \xi_l^{\text{raw}}, \quad h_{l+1} = \frac{h_{l+1}^{\text{pre}}}{\|h_{l+1}^{\text{pre}}\|}, \tag{30}$$

where $\beta_{\text{raw}} > 0$ and $\beta_{\text{raw}}\xi_l^{\text{raw}}$ is the total residual increment divided by the norm of the incoming unnormalized hidden state. We assume $h_{l+1}^{\text{pre}} \neq 0$. The superscript raw denotes the update before renormalization. This update need not be tangent to $\mathbb{S}^{d-1}$ at $h_l$.

**Lemma C.2** (Single-Step Update). *Let $h_l, v \in \mathbb{S}^{d-1}$, $\beta_{\mathrm{raw}} > 0$, and $\xi_l^{\mathrm{raw}} \in \mathbb{R}^d$, with $h_l + \beta_{\mathrm{raw}} \xi_l^{\mathrm{raw}} \neq 0$. For $h_{l+1}$ in Eq. 4, write $\rho_l := \langle h_l, v \rangle$ and $\rho_{l+1} := \langle h_{l+1}, v \rangle$. Then*

$$\rho_{l+1} = \frac{\rho_l + \beta_{\mathrm{raw}} \langle \xi_l^{\mathrm{raw}}, v \rangle}{\sqrt{1 + 2\beta_{\mathrm{raw}} \langle h_l, \xi_l^{\mathrm{raw}} \rangle + \beta_{\mathrm{raw}}^2 \|\xi_l^{\mathrm{raw}}\|^2}}. \tag{31}$$

*Proof.* Taking the inner product of Eq. 4 with $v$ gives

$$\rho_{l+1} = \frac{\rho_l + \beta_{\mathrm{raw}} \langle \xi_l^{\mathrm{raw}}, v \rangle}{\|h_l + \beta_{\mathrm{raw}} \xi_l^{\mathrm{raw}}\|_2}.$$

The denominator is positive by assumption. Expanding its square and using $\|h_l\|_2 = 1$ yields the stated formula. □

**Raw residual decomposition.** Let

$$P_{h_l}^{\perp} := I_d - h_l h_l^{\top}, \qquad \xi_l^{\perp} := P_{h_l}^{\perp} \xi_l^{\mathrm{raw}}, \qquad \xi_l^{\parallel} := \xi_l^{\mathrm{raw}} - \xi_l^{\perp}. \tag{32}$$

Here $I_d$ denotes the $d \times d$ identity matrix. Thus

$$\xi_l^{\mathrm{raw}} = \xi_l^{\perp} + \xi_l^{\parallel}, \tag{33}$$

where $\xi_l^{\perp} \in T_{h_l} \mathbb{S}^{d-1}$ is the tangent component and $\xi_l^{\parallel} = \langle \xi_l^{\mathrm{raw}}, h_l \rangle h_l$ is the radial component, using $\|h_l\| = 1$.

### C.2 Exact Retention and Tangent Propagation

The residual update above requires a model of the increment $\xi_l^{\mathrm{raw}}$. The quantities below use only two adjacent normalized states, so they can be computed for every observed trajectory without such a model.

**Definition C.3** (Step Observables). For adjacent normalized states $h_l, h_{l+1} \in \mathbb{S}^{d-1}$ and a fixed unit direction $v$, define

$$c_l := \langle h_{l+1}, h_l \rangle, \qquad w_l := h_{l+1} - c_l h_l, \qquad g_l(v) := \langle w_l, v \rangle, \qquad \kappa_l := 1 - c_l. \tag{34}$$

We retain $\rho_l(v) := \langle h_l, v \rangle$, abbreviated as $\rho_l$ when the direction is fixed.

**Proposition C.4** (Exact Step Decomposition). *The observables in Definition C.3 satisfy*

$$w_l \perp h_l, \qquad |c_l| \leq 1, \qquad \|w_l\|_2^2 = 1 - c_l^2, \tag{35}$$

*and, for every fixed unit direction $v$,*

$$\rho_{l+1} = c_l \rho_l + g_l(v), \qquad g_l(v) = \Delta \rho_l + \kappa_l \rho_l. \tag{36}$$

*These identities hold pathwise and require no distributional assumption.*

*Proof.* By construction, $h_{l+1} = c_l h_l + w_l$. Since $\|h_l\|_2 = 1$,

$$\langle h_l, w_l \rangle = \langle h_l, h_{l+1} \rangle - c_l \|h_l\|_2^2 = 0. \tag{37}$$

Cauchy-Schwarz gives $|c_l| \leq 1$, and orthogonality gives $1 = \|h_{l+1}\|_2^2 = c_l^2 + \|w_l\|_2^2$. Taking the inner product of $h_{l+1} = c_l h_l + w_l$ with $v$ proves the first identity in Eq. 36. Subtracting $\rho_l$ proves the second. □

Proposition C.4 splits each step into the retained part $c_l h_l$ and the tangent part $w_l$, whose length $\sqrt{1 - c_l^2}$ is fixed by the retention. For a given retention, only the direction of $w_l$ within the tangent space remains free. The following proof shows how far this direction can move the projection along $v$.

*Proof of Proposition 3.4.* Write $\rho_l = \langle h_l, v\rangle$ and $v_l^\perp := v - \rho_l h_l$. Since $w_l \perp h_l$,

$$g_l(v) = \langle w_l, v_l^\perp\rangle, \qquad \|v_l^\perp\|_2^2 = 1 - \rho_l^2. \tag{38}$$

Because $h_l, h_{l+1}$, and $v$ are unit vectors, Cauchy–Schwarz gives $|c_l| \le 1$ and $|\rho_l| \le 1$. Hence both $1 - c_l^2$ and $1 - \rho_l^2$ are nonnegative. Together with $\|w_l\|_2^2 = 1 - c_l^2$, Cauchy–Schwarz gives the denominator-free bound

$$g_l(v)^2 \le (1 - c_l^2)(1 - \rho_l^2). \tag{39}$$

If the right-hand side is zero, Eq. 39 implies $g_l(v) = 0$, so the stated zero convention preserves Eq. 8. Otherwise, division by the positive square root of the right-hand side defines $\eta_l(v)$ and gives $|\eta_l(v)| \le 1$. The reachable interval follows immediately.

We now prove sharpness in the finite-dimensional ambient space $\mathbb{R}^d$. First suppose $v_l^\perp \neq 0$ and set $u := v_l^\perp / \|v_l^\perp\|_2$. Then $u \perp h_l$, $\|u\|_2 = 1$, and $\langle u, v\rangle = \sqrt{1 - \rho_l^2}$. For either sign,

$$h_{l+1}^\pm := c_l h_l \pm \sqrt{1 - c_l^2}\, u \tag{40}$$

is a unit vector with $\langle h_{l+1}^\pm, h_l\rangle = c_l$ and realizes the corresponding endpoint. If $v_l^\perp = 0$, the interval is a singleton; for $d \ge 2$, choose any unit $u \perp h_l$ in the same construction. Such a vector exists because the orthogonal complement of $\mathrm{span}\{h_l\}$ has dimension at least one. Moreover, $v_l^\perp = 0$ implies $v = \rho_l h_l$ and $\rho_l^2 = 1$, so every vector with retention $c_l$ has projection $c_l\rho_l$, the unique point of the interval.

For $d \ge 3$, choose a unit vector $z$ orthogonal to both $h_l$ and $u$. The orthogonal complement of $\mathrm{span}\{h_l, u\}$ has dimension at least one, because $h_l$ and $u$ are orthonormal. For any $t \in [-1, 1]$, the vector

$$q_t := tu + \sqrt{1 - t^2}\, z, \qquad h_{l+1}(t) := c_l h_l + \sqrt{1 - c_l^2}\, q_t \tag{41}$$

is unit, has retention $c_l$, and realizes $c_l\rho_l + \sqrt{1 - c_l^2}\sqrt{1 - \rho_l^2}\, t$. The zero-capacity case reduces to the singleton interval and is covered by the endpoint construction. If the capacity is positive, every point $x$ in the stated interval corresponds to $t = (x - c_l\rho_l)/(\sqrt{1 - c_l^2}\sqrt{1 - \rho_l^2}) \in [-1, 1]$; hence the whole interval is attained.

Finally, let $\varphi_l = \arccos \rho_l$, $\varphi_{l+1} = \arccos \rho_{l+1}$, and $\vartheta_l = \arccos c_l$. The endpoint identities are

$$c_l\rho_l + \sqrt{1 - c_l^2}\sqrt{1 - \rho_l^2} = \cos(\varphi_l - \vartheta_l), \tag{42}$$

$$c_l\rho_l - \sqrt{1 - c_l^2}\sqrt{1 - \rho_l^2} = \cos(\varphi_l + \vartheta_l). \tag{43}$$

The unit-norm bounds imply $\varphi_l, \vartheta_l, \varphi_{l+1} \in [0, \pi]$. On this interval, $\arccos$ is decreasing, and

$$\arccos(\cos|\varphi_l - \vartheta_l|) = |\varphi_l - \vartheta_l|, \qquad \arccos(\cos(\varphi_l + \vartheta_l)) = \pi - |\pi - \varphi_l - \vartheta_l|. \tag{44}$$

Thus, in the original projection coordinates, the equivalent angular interval is

$$|\arccos\rho_l - \arccos c_l| \le \arccos\rho_{l+1} \le \pi - |\pi - \arccos\rho_l - \arccos c_l|. \tag{45}$$

□

The preceding proposition is local to one transition. We next iterate the exact recursion to obtain the finite-depth response.

**Proposition C.5** (Pathwise Variation of Constants)**.** *Define* $K_c(a, b) := \prod_{r=a}^{b-1} c_r$ *for* $a < b$ *and* $K_c(a, a) := 1$. *Then, for every* $L \ge 0$*,*

$$\rho_L = K_c(0, L)\rho_0 + \sum_{j=0}^{L-1} K_c(j+1, L) g_j(v). \tag{46}$$

*Proof.* For $L = 0$, the empty product is one and the empty sum is zero. If the formula holds at $L$, then Eq. 36 gives

$$\rho_{L+1} = c_L\rho_L + g_L(v) \tag{47}$$

$$= K_c(0, L+1)\rho_0 + \sum_{j=0}^{L-1} K_c(j+1, L+1) g_j(v) + g_L(v), \tag{48}$$

which is Eq. 46 at $L + 1$ because $K_c(L+1, L+1) = 1$. □

Eq. 46 shows that $\rho_L$ can differ from the retained initial value $K_c(0,L)\rho_0$ only through the tangent contributions. The next theorem bounds how large these contributions must be to produce a given difference. Each contribution is measured relative to the tangent length $\sqrt{1-c_j^2}$ available at its layer.

**Theorem C.6** (Directional Write-Budget Bound)**.** *For normalized states $h_0, \dots, h_L \in \mathbb{S}^{d-1}$ and a fixed direction $v \in \mathbb{S}^{d-1}$, let $\rho_j := \langle h_j, v\rangle$, $c_j := \langle h_{j+1}, h_j\rangle$, $w_j := h_{j+1} - c_j h_j$, and $g_j(v) := \langle w_j, v\rangle$ for $j < L$. Unit norms give $|c_j| \le 1$ and $\|w_j\|_2^2 = 1 - c_j^2$. Thus $|c_j| = 1$ forces $w_j = 0$ and $g_j(v) = 0$. Let $[x]_+ := \max\{x, 0\}$ and define*

$$\mathcal{G}_L(v) := \sum_{j=0}^{L-1} \frac{g_j(v)^2}{1-c_j^2}, \qquad \mathcal{G}_L^+(v) := \sum_{j=0}^{L-1} \frac{[g_j(v)]_+^2}{1-c_j^2}, \tag{49}$$

*where terms with $|c_j| = 1$ are defined as zero. Then*

$$\big(\rho_L - K_c(0,L)\rho_0\big)^2 \le \big(1 - K_c(0,L)^2\big)\mathcal{G}_L(v). \tag{50}$$

*If $0 \le c_j \le 1$ for all $j < L$, the one-sided bound*

$$\big[\rho_L - K_c(0,L)\rho_0\big]_+^2 \le \big(1 - K_c(0,L)^2\big)\mathcal{G}_L^+(v) \tag{51}$$

*also holds. When $|K_c(0,L)| < 1$, these inequalities give*

$$\mathcal{G}_L(v) \ge \frac{\big(\rho_L - K_c(0,L)\rho_0\big)^2}{1 - K_c(0,L)^2}, \qquad \mathcal{G}_L^+(v) \ge \frac{\big[\rho_L - K_c(0,L)\rho_0\big]_+^2}{1 - K_c(0,L)^2}, \tag{52}$$

*where the second bound additionally requires $0 \le c_j \le 1$ for all $j < L$.*

Equivalently, $\mathcal{G}_L(v) = \sum_{j=0}^{L-1}(1-\rho_j^2)\eta_j(v)^2$. Since $|c_j| \le 1$, the case $|K_c(0,L)| = 1$ forces $|c_j| = 1$ at every step. All tangent contributions then vanish, giving $\mathcal{G}_L(v) = 0$ and $\rho_L = K_c(0,L)\rho_0$.

*Proof of Theorem C.6.* Define $\widehat{g}_j(v) := g_j(v)/\sqrt{1-c_j^2}$ for $|c_j| < 1$ and set $\widehat{g}_j(v) := 0$ for $|c_j| = 1$. The zero convention is consistent because $|c_j| = 1$ implies $\|w_j\|_2 = 0$ and hence $g_j(v) = 0$. Thus

$$g_j(v) = \sqrt{1-c_j^2}\,\widehat{g}_j(v), \qquad \widehat{g}_j(v) = \sqrt{1-\rho_j^2}\,\eta_j(v). \tag{53}$$

Under the same zero convention, $\mathcal{G}_L(v) = \sum_{j=0}^{L-1}\widehat{g}_j(v)^2$ and $\mathcal{G}_L^+(v) = \sum_{j=0}^{L-1}[\widehat{g}_j(v)]_+^2$. The unit-state assumption gives $|c_j| \le 1$, so all square roots in these identities are real and nonnegative. Subtracting the retained initial condition from Eq. 46 gives

$$\rho_L - K_c(0,L)\rho_0 = \sum_{j=0}^{L-1} K_c(j+1,L)\sqrt{1-c_j^2}\,\widehat{g}_j(v). \tag{54}$$

The retention kernels satisfy the exact telescoping identity

$$\sum_{j=0}^{L-1} K_c(j+1,L)^2(1-c_j^2) = 1 - K_c(0,L)^2. \tag{55}$$

Indeed, $K_c(j,L) = c_j K_c(j+1,L)$, so each summand equals $K_c(j+1,L)^2 - K_c(j,L)^2$. The sum telescopes from $K_c(L,L)^2 = 1$ to $K_c(0,L)^2$; for $L = 0$, both sides are zero under the empty-sum and empty-product conventions. Applying Cauchy–Schwarz to the finite sequences

$$x_j := K_c(j+1,L)\sqrt{1-c_j^2}, \qquad y_j := \widehat{g}_j(v), \tag{56}$$

in Eq. 54, and then using Eq. 55 proves Eq. 50.

Now suppose $0 \le c_j \le 1$. Every $K_c(j+1,L)$ is nonnegative, and replacing $\widehat{g}_j(v)$ by $[\widehat{g}_j(v)]_+$ can only increase the right-hand side of Eq. 54. Hence

$$\big[\rho_L - K_c(0,L)\rho_0\big]_+ \le \sum_{j=0}^{L-1} K_c(j+1,L)\sqrt{1-c_j^2}\,[\widehat{g}_j(v)]_+. \tag{57}$$

A second application of Cauchy–Schwarz and Eq. 55 gives Eq. 51. If $|K_c(0,L)| < 1$, then $1 - K_c(0,L)^2 > 0$, so division preserves the inequality direction and gives Eq. 52. The nonnegative-retention assumption is used only for the one-sided comparison; the two-sided bound allows retention coefficients of either sign. □

### C.2.1 Exact Budget Decomposition and Attainment

Theorem C.6 gives a lower bound on $\mathcal{G}_L(v)$. We now identify the contributions that reach this bound and write the gap to the bound as an explicit weighted sum of squares.

*Proof of Theorem 3.5.* Use $g_j^\star(v)$ from Eq. 11 and recall

$$D_L(v) := \rho_L - K_c(0, L)\rho_0. \tag{58}$$

The assumption $|K_c(0, L)| < 1$ makes $1 - K_c(0, L)^2$ positive. If $|c_j| = 1$, the unit-state geometry gives $g_j(v) = 0$, and the definition also gives $g_j^\star(v) = 0$. All ratios at these indices below follow the zero convention of Theorem C.6.

By the retention telescope in Eq. 55,

$$\sum_{j=0}^{L-1} K_c(j+1, L) g_j^\star(v) = D_L(v), \tag{59}$$

$$\sum_{j=0}^{L-1} \frac{\big(g_j^\star(v)\big)^2}{1 - c_j^2} = \frac{D_L(v)^2}{1 - K_c(0, L)^2}. \tag{60}$$

The observed contributions have the same weighted sum by Eq. 6. Therefore

$$\sum_{j=0}^{L-1} K_c(j+1, L)\big(g_j(v) - g_j^\star(v)\big) = 0, \tag{61}$$

and the cross term in the budget expansion vanishes:

$$\sum_{j=0}^{L-1} \frac{g_j^\star(v)\big(g_j(v) - g_j^\star(v)\big)}{1 - c_j^2} = \frac{D_L(v)}{1 - K_c(0, L)^2} \sum_{j=0}^{L-1} K_c(j+1, L)\big(g_j(v) - g_j^\star(v)\big) = 0. \tag{62}$$

Expanding each $g_j(v)$ as $g_j^\star(v) + \big(g_j(v) - g_j^\star(v)\big)$ and using Eq. 60 proves Eq. 12. The remainder is a sum of nonnegative terms. It vanishes exactly when $g_j(v) = g_j^\star(v)$ at every index with $|c_j| < 1$, and equality at the remaining indices holds automatically. This proves the stated equality condition. □

The same argument gives the scalar minimum when the retention sequence and endpoint projections are fixed: among scalar recurrences with these data and zero contributions whenever $|c_j| = 1$, the unique minimizing contribution sequence is $g_j^\star(v)$. Geometric attainability additionally requires the one-step capacity constraints.

*Proof of Proposition 3.6.* For $j < L$, the retention products satisfy $K_c(0, j+1) = K_c(0, j)c_j$ and $K_c(j, L) = c_j K_c(j+1, L)$. Substitution into the closed form in Eq. 13 gives

$$\rho_0^\star = \rho_0, \qquad \rho_{j+1}^\star = c_j \rho_j^\star + g_j^\star(v), \qquad 0 \le j < L. \tag{63}$$

At $j = L$, using $|K_c(0, L)| < 1$ in Eq. 13 gives $\rho_L^\star = K_c(0, L)\rho_0 + D_L(v) = \rho_L$.

For necessity, any attaining trajectory has $g_j(v) = g_j^\star(v)$ by Theorem 3.5. Its fixed initial projection and scalar recursion then force its projection sequence to be $\rho_j^\star$. Applying the capacity inequality Eq. 39 at every layer yields Eq. 14.

For sufficiency, assume Eq. 14 and start from the prescribed unit state $h_0$. Suppose a unit state of projection $\rho_j^\star$ has been constructed. If $|c_j| = 1$, then $g_j^\star(v) = 0$, and multiplying the current state by $c_j$ gives the required next state. If $|c_j| < 1$, the capacity condition places $\rho_{j+1}^\star = c_j \rho_j^\star + g_j^\star(v)$ inside the interval in Proposition 3.4. Since $d \ge 3$, a unit next state with this projection and retention $c_j$ exists. This includes the zero-capacity case $|\rho_j^\star| = 1$. Induction constructs the full unit-state trajectory with the prescribed retentions and terminal projection. Its budget equals the lower bound by Eq. 60.

The attaining contribution and projection sequences are unique by the equality condition and the scalar recursion. The vector trajectory itself need not be unique. If the capacity condition fails, the minimum over the spherical comparison class is strictly larger than the scalar lower bound: this class is nonempty because it contains the original trajectory, is compact, and has a continuous budget for the fixed retention sequence. Its minimum is therefore attained, and necessity rules out equality with the scalar lower bound. □

The passive constant-retention result is the zero-budget specialization.

**Corollary C.7** (Passive Constant-Retention Reference)**.** *If $g_l(v) = 0$ and $c_l = q \in (0,1)$ for $0 \le l < L$, then*

$$\rho_L = q^L \rho_0 = e^{-\lambda L}\rho_0, \qquad \lambda = -\log q > 0. \tag{64}$$

*This is Theorem 3.7.*

### C.2.2 Unit Tangential Update and Small-Step Relation

The exact decomposition does not require a residual-update model. To connect it to the normalized tangential update used in the passive reference, suppose

$$h_{l+1} = \frac{h_l + \beta_l \xi_l}{\|h_l + \beta_l \xi_l\|_2}, \qquad \xi_l \perp h_l, \qquad \|\xi_l\|_2 = 1, \qquad \beta_l \ge 0. \tag{65}$$

**Proposition C.8** (Exact Retention Under a Unit Tangential Step)**.** *Under Eq. 65,*

$$c_l = \frac{1}{\sqrt{1+\beta_l^2}}, \qquad w_l = \frac{\beta_l}{\sqrt{1+\beta_l^2}}\xi_l, \qquad g_l(v) = \frac{\beta_l}{\sqrt{1+\beta_l^2}}\langle \xi_l, v\rangle. \tag{66}$$

*Moreover, for every $\beta_l \ge 0$,*

$$\left|\kappa_l - \frac{\beta_l^2}{2}\right| \le \frac{\beta_l^4}{2}. \tag{67}$$

*If $0 \le \beta_l \le 1$ and $Z_l := \langle \xi_l, v\rangle$, then*

$$\left|\Delta\rho_l - \left(\beta_l Z_l - \frac{\beta_l^2}{2}\rho_l\right)\right| \le \beta_l^3. \tag{68}$$

*Proof.* Tangency gives $\|h_l + \beta_l \xi_l\|_2^2 = 1 + \beta_l^2$, which proves Eq. 66. For $x \ge 0$, Taylor's theorem for $f(x) = (1+x)^{-1/2}$ gives

$$\left|f(x) - 1 + \frac{x}{2}\right| \le \frac{x^2}{2}, \tag{69}$$

because $0 < f''(x) \le 3/4$. Setting $x = \beta_l^2$ proves Eq. 67. Also, $|f(x) - 1| \le x/2$. Therefore the remainder in

$$\Delta\rho_l = \beta_l Z_l - \frac{\beta_l^2}{2}\rho_l + \left(f(\beta_l^2) - 1 + \frac{\beta_l^2}{2}\right)\rho_l + \beta_l\left(f(\beta_l^2) - 1\right) Z_l \tag{70}$$

has magnitude at most $\beta_l^4/2 + \beta_l^3/2 \le \beta_l^3$ when $\beta_l \le 1$. □

The measured middle-window values of $c_l$ are not close enough to one to treat $\beta_l^2/2$ as the empirical coefficient without approximation error. The main analysis therefore uses $c_l$ and $\kappa_l = 1 - c_l$ directly.

### C.2.3 Exact Expectation Balance

*Remark* C.9 (Passive Reference). When $g_l(v) = 0$, the projection evolves as $\rho_{l+1} = c_l \rho_l$. Observed trajectories are evaluated with their measured tangent contributions through Eq. 17.

Taking expectations in Eq. 36 gives

$$\mathbb{E}[\Delta\rho_l] = -\mathbb{E}[\kappa_l \rho_l] + \mathbb{E}[g_l(v)]. \tag{71}$$

For an observed terminal outcome $J$, the same pathwise identity gives

$$\mathbb{E}[\Delta\rho_l \mid J] = -\mathbb{E}[\kappa_l \rho_l \mid J] + \mathbb{E}[g_l(v) \mid J]. \tag{72}$$

Neither equation factors $\mathbb{E}[\kappa_l \rho_l]$ into a product of means, because $\kappa_l$ and $\rho_l$ can vary together across prompts. Appendix C.4 measures this covariance.

The following result distinguishes an outcome-group contrast from a conditional mean update given the current information.

**Proposition C.10** (Outcome-Group Non-Identification of Predictable Forcing)**.** *There exists a process whose directional update has zero conditional expectation under the current information, while the two groups defined by the realized terminal outcome have a strictly nonzero next-projection contrast.*

*Proof of Proposition C.10.* Let $\Omega = \{-1, +1\}$ have the uniform law and let the current $\sigma$-algebra be trivial. Choose orthogonal unit vectors $h, v$ and define $Z(\omega) = \omega$. Then $\mathbb{E}[Z \mid \mathcal{F}_0] = 0$. For $\beta > 0$, set

$$h_1(\omega) = \frac{h + \beta Z(\omega) v}{\sqrt{1 + \beta^2}}. \tag{73}$$

The population directional update remains centered. If the terminal groups are defined by $J = \mathrm{Ref}$ for $Z = +1$ and $J = \mathrm{JB}$ for $Z = -1$, then

$$\mathbb{E}[\langle h_1, v\rangle \mid J = \mathrm{JB}] - \mathbb{E}[\langle h_1, v\rangle \mid J = \mathrm{Ref}] = -\frac{2\beta}{\sqrt{1 + \beta^2}} < 0. \tag{74}$$

Thus post-hoc outcome stratification alone does not identify a predictable microscopic forcing term. □

In our experiments, the refused and jailbroken groups are defined by the terminal outcome. The M2 contrast is therefore a comparison between outcome groups. By Proposition C.10, it does not by itself identify a forcing term that depends only on information available before the outcome.

### C.3 Empirical Measurement Protocol

For paired reference and jailbreak prompts, define the descriptive similarity

$$\chi_{i,l} := \cos(h_{i,l}^{\mathrm{Ref}}, h_{i,l}^{\mathrm{JB}}), \qquad \bar{\chi}_l := N_{\mathrm{pair}}^{-1} \sum_i \chi_{i,l}. \tag{75}$$

The self-trajectory cosine is $s_l := \mathbb{E}_i[\cos(h_{i,l}, h_{i,0})]$. These quantities are distinct from the adjacent-state retention $c_{i,l}$ and the fixed-direction projection $\rho_{i,l}$. The only normalization is the per-vector $\ell_2$ normalization in the cosine. Because both prompts in a pair end with the same chat-template generation-boundary token, $\chi_{i,0} = 1$ at the embedding output; this deterministic point is a sanity check rather than an empirical effect.

#### C.3.1 Early-Layer Geometric Diagnostic

Let $\widetilde{h}_{i,l}^{\mathrm{Ref}}$ and $\widetilde{h}_{i,l}^{\mathrm{JB}}$ denote the unnormalized states for the paired reference and jailbreak prompts. The paired cosine $\chi_{i,l}$ defined above is unchanged by per-vector normalization and therefore equals $\cos(\widetilde{h}_{i,l}^{\mathrm{Ref}}, \widetilde{h}_{i,l}^{\mathrm{JB}})$. We use $\Delta\chi_{i,l} := \chi_{i,l+1} - \chi_{i,l}$ and the orthogonal residual magnitude

$$m_{i,l}^{\perp} := \left\| \widetilde{h}_{i,l}^{\mathrm{Ref}} - \frac{\langle \widetilde{h}_{i,l}^{\mathrm{Ref}}, \widetilde{h}_{i,l}^{\mathrm{JB}} \rangle}{\|\widetilde{h}_{i,l}^{\mathrm{JB}}\|_2^2} \widetilde{h}_{i,l}^{\mathrm{JB}} \right\|_2, \tag{76}$$

with a zero projection when the jailbreak state vanishes. A logistic probe uses these features from the first $L_{\mathrm{early}}$ layers, with feature standardization fitted within each cross-validation training fold.

#### C.3.2 Cross-Fitted Exact Decomposition

**Outcome labeling.** The terminal outcome $J_i \in \{\mathrm{Ref}, \mathrm{JB}\}$ of each prompt is assigned with the HarmBench evaluation protocol (Mazeika et al., 2024), using GPT-5.4 as the judge model applied to the full generated completion. A completion is labeled jailbroken when it provides substantive harmful assistance and refused otherwise, matching the behavioral definition in §1. Labels are assigned once, before any geometric analysis, and are not functions of the hidden-state measurements.

We split HarmBench prompts into five folds. For each held-out fold, the linear alignment direction is fitted at the middle layer and oriented toward refusal on the other four folds, then held fixed across all layers. Hidden states are $\ell_2$-normalized before fitting and projection. For a model with $L$ transformer blocks, let $l \in \{0, \ldots, L-1\}$ index updates. The prespecified middle window is

$\mathcal{W} = \{l : 0.25 \leq l/L < 0.75\}$, corresponding to layers 8–23 for Llama-2-7B and Llama-3.1-8B, 7–20 for Qwen2.5-7B, and 12–35 for Qwen2.5-14B. Normalized held-out states give

$$c_{i,l} = \langle h_{i,l+1}, h_{i,l} \rangle, \qquad \kappa_{i,l} = 1 - c_{i,l}, \qquad g_{i,l}(v) = \Delta\rho_{i,l} + \kappa_{i,l}\rho_{i,l}. \tag{77}$$

We first average complete trajectories over the prespecified model-relative middle window and then apply prompt-level bootstrap and permutation inference. All prompt-bootstrap 95% confidence intervals condition on the predeclared cross-fitting folds and the alignment direction fitted within each training fold; each draw resamples complete held-out prompt trajectories. For conditioned coefficients, each bootstrap draw refits the layer-conditioned regression, and the common-support bootstrap also recomputes the support mask. Neither procedure refits the safety probe or recomputes held-out projections.

For the regression comparison, we fit

$$g_{i,l}(v) = \alpha_l + b_l\rho_{i,l} + \gamma\mathbf{1}\{J_i = \text{JB}\} + \varepsilon_{i,l}. \tag{78}$$

The common-support fit intersects the group-specific empirical 5th to 95th percentile intervals at each layer and retains layers with at least ten observations from each group. Standard errors are clustered by prompt. Prompt-label permutation holds the fitted fold directions, $\rho_{i,l}$, and $g_{i,l}(v)$ fixed; the common-support permutation additionally holds the observed support mask fixed. Table 1 reports the exact decomposition and conditioned estimates.

### C.3.3 Write Orientation and Positive Directional Budget

Using the same cross-fitted directions and update window, we compute the orientation in Proposition 3.4 and the positive directional budget in Theorem C.6. Orientation is averaged over window updates within each prompt, whereas the positive budget is summed. Table 2 in the main text reports group means, refused-minus-jailbroken contrasts, and prompt-level effect sizes. In all four models, mean orientation is positive in the refused group and negative in the jailbroken group, and the positive budget is larger in refused trajectories.

### C.3.4 Evaluation of the Minimum-Budget Solution

We evaluate the closed-form solution on all $1{,}600$ available archived prompt trajectories: $400$ for each models. We apply the finite-horizon results to each middle-window segment, shifting the layer indices to start at zero and using the number of transitions in that window as the horizon. Saved hidden states and training-fold directions are normalized in FP64 for this calculation. Each candidate uses the observed retention coefficients and endpoint projections in Eq. 11 and Eq. 13.

All measured window coefficients satisfy $0 < c_j < 1$. We check Eq. 14 in its equivalent form $(\rho_j^\star)^2 + (g_j^\star(v))^2/(1 - c_j^2) \leq 1$ at every transition. Every candidate satisfies this condition, with a maximum left-hand side of $0.09567$ across all models and trajectories. We construct realizing paths in $\mathbb{R}^3$ and check their unit norms, adjacent retentions, and projections. The largest discrepancy in these quantities is below $1.1 \times 10^{-15}$, and the constructed budgets agree with the lower bound to within $1.6 \times 10^{-16}$.

An independent least-squares calculation over the interior projection coordinates gives the same scalar minimum on every trajectory. We also solve the problem with the spherical capacity constraints using SLSQP on 16 trajectories per model. Within each outcome group, we sort prompt IDs and select eight evenly spaced entries. All $64$ constrained optimizations succeed, with budget differences from the closed-form minimum below $4.1 \times 10^{-13}$. On the observed trajectories, the largest residual in the exact budget identity of Theorem 3.5 is below $1.7 \times 10^{-16}$.

For each trajectory, we also divide the attained minimum budget by the observed directional budget over the same window. The median ratios are $28.1\%$, $6.9\%$, $11.3\%$, and $3.3\%$ for Llama-2-7B, Llama-3.1-8B, Qwen2.5-7B, and Qwen2.5-14B, respectively. The remaining fraction is the exact remainder in Eq. 12 divided by the observed budget. It measures the normalized weighted squared deviation of the realized tangent contributions from the minimum-budget sequence with the same retentions and endpoint projections.

### C.4 Retention Diagnostics and a Window-Local Two-Parameter Reference

#### C.4.1 Measured Retention by Group

Table 4 reports the middle-window group means of the adjacent-state retention $c_{i,l}$ and the geometric coefficient $\kappa_{i,l} = 1 - c_{i,l}$, together with the implied passive half-life $\log 2/(-\log \bar{c})$. The half-life ranges from $4.5$ to $11.2$ layers across the tested models and groups. The retention coefficient is independent of the measurement direction, the group differences are small, and in no model is jailbroken retention smaller than refused retention. Larger inter-layer angular motion under jailbreak therefore does not account for the M2 ordering.

**Table 4: Measured Middle-Window Retention by Group.** Group means of the adjacent-state retention $\bar{c}$ and the geometric coefficient $\bar{\kappa} = 1 - \bar{c}$ over the prespecified middle window. The passive half-life is $\log 2/(-\log \bar{c})$ in layers, the depth at which a purely retained projection halves at the measured mean retention.

| Model | $\bar{c}_{\mathrm{Ref}}$ | $\bar{c}_{\mathrm{JB}}$ | $\bar{\kappa}_{\mathrm{Ref}}$ | $\bar{\kappa}_{\mathrm{JB}}$ | Half-life Ref | Half-life JB |
|---|---|---|---|---|---|---|
| Llama-2-7B | 0.9101 | 0.9153 | 0.0899 | 0.0847 | 7.4 | 7.8 |
| Llama-3.1-8B | 0.8563 | 0.8699 | 0.1437 | 0.1301 | 4.5 | 5.0 |
| Qwen2.5-7B | 0.9069 | 0.9133 | 0.0931 | 0.0867 | 7.1 | 7.6 |
| Qwen2.5-14B | 0.9331 | 0.9399 | 0.0669 | 0.0601 | 10.0 | 11.2 |

#### C.4.2 Window-Local Two-Parameter Reference

As a diagnostic, we replace the exact coefficients by their pooled window means: for each group, $\bar{\kappa}$ and $\bar{g}$ denote the means of $\kappa_{i,l}$ and $g_{i,l}(v)$ over prompts and window layers. We then evolve

$$\hat{\rho}_{l+1} = (1 - \bar{\kappa})\hat{\rho}_l + \bar{g} \tag{79}$$

from the observed group mean at the window start. This constant-coefficient reference introduces no fitted parameter. Table 5 reports the within-window reconstruction error. The reconstruction is moderately accurate in seven of the eight model–group pairs and fails for the Llama-2-7B jailbroken group.

**Table 5: Window-Local Two-Parameter Reconstruction.** The reference evolves $\hat{\rho}_{l+1} = (1-\bar{\kappa})\hat{\rho}_l + \bar{g}$ from the observed group mean at the window start, with $\bar{\kappa}$ and $\bar{g}$ measured over the prespecified middle window. No parameter is fitted. RMSE and $R^2$ are computed against the observed within-window group-mean trajectory; the endpoint error is predicted minus observed at the window end.

| Model | Group | $\bar{\kappa}$ | $\bar{g}$ | RMSE | $R^2$ | Endpoint error |
|---|---|---|---|---|---|---|
| Llama-2-7B | Refused | 0.08995 | 0.02262 | 0.07562 | 0.555 | 0.00712 |
| Llama-2-7B | Jailbroken | 0.08472 | −0.01357 | 0.03586 | −0.427 | −0.02281 |
| Llama-3.1-8B | Refused | 0.14371 | 0.01474 | 0.07351 | 0.566 | 0.02241 |
| Llama-3.1-8B | Jailbroken | 0.13009 | −0.01436 | 0.02037 | 0.468 | −0.02734 |
| Qwen2.5-7B | Refused | 0.09305 | 0.00822 | 0.02456 | 0.411 | 0.01198 |
| Qwen2.5-7B | Jailbroken | 0.08667 | −0.00190 | 0.00645 | 0.549 | −0.00194 |
| Qwen2.5-14B | Refused | 0.06688 | 0.00238 | 0.02829 | 0.400 | 0.02255 |
| Qwen2.5-14B | Jailbroken | 0.06011 | −0.00344 | 0.01187 | 0.303 | 0.00534 |

For layer $l$, let $\bar{\rho}_l$, $\bar{\kappa}_l$, and $\bar{g}_l$ denote the prompt means of $\rho_{i,l}$, $\kappa_{i,l}$, and $g_{i,l}(v)$ within the group, let $\Delta\bar{\rho}_l := \bar{\rho}_{l+1} - \bar{\rho}_l$, and let $\mathrm{Cov}_l(\kappa_{i,l}, \rho_{i,l})$ denote the prompt-level covariance at layer $l$. Averaging the exact step identity over prompts, the one-step residual of this reference decomposes exactly into three measured terms,

$$\Delta\bar{\rho}_l - (-\bar{\kappa}\bar{\rho}_l + \bar{g}) = -(\bar{\kappa}_l - \bar{\kappa})\bar{\rho}_l - \mathrm{Cov}_l(\kappa_{i,l}, \rho_{i,l}) + (\bar{g}_l - \bar{g}). \tag{80}$$

Table 6 reports the root-mean-square size of each term over the window layers. The layer variation of $\bar{g}_l$ dominates in every model and group, the prompt-level covariance is negligible, and the retention variation is secondary. The window-constant reference is therefore a partial description: the tangent supply has a layer profile, consistent with reports that safety-relevant computation concentrates at specific depths (Li et al., 2025; Kao et al., 2025). This diagnostic characterizes a window-local reference and is not used as evidence for M1 or M2.

**Table 6: Two-Parameter Reference Residuals.** All residual columns report RMS values over window layers from Eq. 80: retention variation $(\bar{\kappa}_l - \bar{\kappa})\bar{\rho}_l$, tangent variation $\bar{g}_l - \bar{g}$, prompt-level covariance $\mathrm{Cov}_l(\kappa_{i,l}, \rho_{i,l})$, and one-step closure error. The first numeric column reports $\mathrm{SD}_l(\bar{\kappa}_l)$.

| Model | Group | $\mathrm{SD}_l(\bar{\kappa}_l)$ | Ret. var. | Tang. var. | Cov. | Closure |
|---|---|---|---|---|---|---|
| Llama-2-7B | Refused | 0.0348 | 0.0051 | 0.0342 | 0.0004 | 0.0313 |
| Llama-2-7B | Jailbroken | 0.0337 | 0.0045 | 0.0180 | 0.0002 | 0.0149 |
| Llama-3.1-8B | Refused | 0.0391 | 0.0039 | 0.0503 | 0.0019 | 0.0494 |
| Llama-3.1-8B | Jailbroken | 0.0470 | 0.0061 | 0.0168 | 0.0002 | 0.0142 |
| Qwen2.5-7B | Refused | 0.0263 | 0.0016 | 0.0152 | 0.0002 | 0.0150 |
| Qwen2.5-7B | Jailbroken | 0.0259 | 0.0011 | 0.0080 | 0.0001 | 0.0081 |
| Qwen2.5-14B | Refused | 0.0164 | 0.0004 | 0.0143 | 0.0003 | 0.0142 |
| Qwen2.5-14B | Jailbroken | 0.0143 | 0.0011 | 0.0064 | 0.0001 | 0.0067 |

# D Detailed Results for M2

This section separates the empirical M2 comparison from the conditional geometry of an idealized positive-threshold region. It provides the spherical-cap volume bound, the threshold calculation under passive propagation, and a terminal-crossing proxy.

## D.1 Geometric Setup

Under the fixed linear-coordinate model, thresholding the measured safety direction defines a spherical cap on the representation sphere.

### D.1.1 Linear Separability of Alignment

*Property* D.1 (Linear Separability of Safety-Associated Representations). Some safety- and refusal-related representations are approximately linearly separable on the representation sphere. Specifically, there exist a unit direction $v_{L_1} \in \mathbb{S}^{d-1}$ and threshold $\tau_{L_1}$ encoding the measured constraint coordinate.

**Empirical Motivation.** Prior work identifies refusal-associated directions in residual streams (Arditi et al., 2024) and reports linear decodability of safety-relevant concepts (Zou et al., 2025). Together with the residual-stream view of features (Anthropic, 2021), these results motivate the fixed linear coordinate evaluated in our tested checkpoints.

The alignment-compliant region $\mathcal{C}(\tau)$ is defined in §3.3 as the spherical cap $\{h \in \mathbb{S}^{d-1} : \langle h, v_{L_1} \rangle \geq \tau\}$.

## D.2 Positive-Threshold Geometry

This subsection applies concentration of measure to the idealized spherical-cap region and derives the threshold expression under the passive reference of M1.

### D.2.1 Concentration of Measure

**Lemma D.2** (Spherical Cap Bound). *Let $u \in \mathbb{S}^{d-1}$ be fixed. For any $0 \leq \epsilon < 1$:*

$$\sigma_{d-1}\left(\left\{x \in \mathbb{S}^{d-1} : \langle x, u \rangle \geq \epsilon\right\}\right) \leq \exp\left(-\frac{d\epsilon^2}{2}\right). \tag{81}$$

*Proof.* This is a standard spherical-cap estimate; see (Ball, 1997, Lemma 2.2). □

**Corollary D.3** (Spherical Cap Volume). *For the alignment region $\mathcal{C}(\tau)$ with $\tau > 0$:*

$$\frac{\sigma_{d-1}(\mathcal{C}(\tau))}{\sigma_{d-1}(\mathbb{S}^{d-1})} \leq \exp\left(-\frac{(d-1)\tau^2}{2}\right). \tag{82}$$

*Proof.* For $0 < \tau < 1$, apply Lemma D.2 with $u = v_{L_1}$ and $\epsilon = \tau$. Thus:

$$\begin{aligned}\sigma_{d-1}(\mathcal{C}(\tau)) &= \sigma_{d-1}(\{x : \langle x, v_{L_1}\rangle \geq \tau\}) \\ &\leq \exp\left(-\frac{d\tau^2}{2}\right) \leq \exp\left(-\frac{(d-1)\tau^2}{2}\right).\end{aligned} \tag{83}$$

For $\tau \geq 1$, $\mathcal{C}(\tau) \subseteq \{v_{L_1}\}$, which has measure zero for $d \geq 2$; for $d = 1$, the right-hand side equals one. □

**Interpretation.** Under uniform spherical measure, the idealized positive-threshold region has exponentially small volume for fixed $\tau > 0$. This is a statement about the measure of the region, not about a transformer's learned state distribution or its transition probabilities.

### D.2.2 Threshold Depth under Passive Propagation

Proposition 3.10 applies only under the passive reference and the positive initial-margin condition $\rho_0 \geq \tau > 0$. By Theorem 3.7, $\rho_l = q^l \rho_0$. Solving $q^L \rho_0 = \tau$ gives the continuous-depth proxy $L_{\text{thresh}} = \log(\rho_0/\tau)/(-\log q)$. Under the unit tangential update Eq. 65 with constant step $\beta$, this equals $2\beta^{-2}\log(\rho_0/\tau) + O(1)$ as $\beta \to 0^+$ with $\rho_0$ and $\tau$ fixed. It is not a first-passage statement and is not applied to the observed HarmBench trajectories.

### D.2.3 Hitting Probability for Unsafe Regions

The threshold depth above assumes the passive reference. With random tangent contributions, the projection can fall below $\tau$ at a different depth. The remark below gives a proxy for the probability that the terminal projection lies below $\tau$, under an additional model.

**Definition D.4** (Unsafe Region)**.** Define the complement:

$$\mathcal{U}(\tau) := \mathbb{S}^{d-1} \setminus \mathcal{C}(\tau) = \{h \in \mathbb{S}^{d-1} : \langle h, v_{L_1}\rangle < \tau\}. \tag{84}$$

*Remark* D.5 (Gaussian Proxy for Hitting Probability). For a trajectory starting from $\rho_0 > \tau$, the rigorous relation is the event inclusion $\{\rho_L < \tau\} \subseteq \{\exists l \leq L : \rho_l < \tau\}$, hence

$$\mathbb{P}(\exists l \leq L : \rho_l < \tau) \geq \mathbb{P}(\rho_L < \tau). \tag{85}$$

A closed-form proxy for the right-hand side requires an additional model. Impose the unit tangential update Eq. 65 with constant step $\beta_l = \beta$, and model the tangential projections $Z_l = \langle \xi_l, v\rangle$ as independent and zero-mean with variance of order $1/(d-1)$, as for isotropic tangent directions. Under this model, the small-step expansion Eq. 68 accumulates $\text{Var}[\rho_L] = O(L\beta^2/(d-1))$, and the terminal Gaussian approximation $\rho_L \approx \mathcal{N}(q^L\rho_0, \text{Var}[\rho_L])$ gives

$$\mathbb{P}(\rho_L < \tau) \approx \Phi\left(\frac{\tau - q^L\rho_0}{\sqrt{\text{Var}[\rho_L]}}\right), \tag{86}$$

where $\Phi$ is the standard normal CDF and $q = 1/\sqrt{1+\beta^2}$ is the constant retention of Proposition C.8. The displayed quantity is a terminal-crossing proxy under this model, not a rigorous closed-form first-hitting-time bound.

## D.3 Empirical Content of M2

M2 is evaluated through the cross-fitted regression comparison in § 3.2.5. Table 1 reports the exact retention and tangent decomposition together with unrestricted and common-support coefficients for $g_{i,l}(v)$. §4.2.1 in the main text reports the common-support permutation $p$-values. All prompt-bootstrap 95% confidence intervals for the tangent contrast, $\gamma$, and $\gamma_{\text{ov}}$ exclude zero, as do the prompt-clustered regression intervals. Across models, the endpoints of the tangent bootstrap intervals range from $0.00499$ to $0.03864$. The prompt-clustered interval endpoints range from $-0.01414$ to $-0.00040$ for $\gamma$ and from $-0.01012$ to $-0.00044$ for $\gamma_{\text{ov}}$. These ranges summarize endpoints across model-specific intervals. Prompt-label permutation gives $p = 1/10001$ for every tangent contrast. The resampling and conditioning protocol is given in § C.3.2.

**Table 7: Label-free random-direction control for the projection-conditioned update (HarmBench)** The regression coefficient is the JB–Ref contrast after adjustment for layer and current projection. The common-support fit uses the intersection of the central projection ranges. $p_{\mathrm{dir}}$ uses 500 random directions per fold without label-based orientation.

| Model | $\hat{\gamma}$ | $\hat{\gamma}_{\mathrm{ov}}$ | Rows kept | $p_{\mathrm{dir}}$ |
|---|---|---|---|---|
| Llama-2-7B | $-0.01225$ | $-0.00737$ | 23.7% | 0.002 |
| Llama-3.1-8B | $-0.00952$ | $-0.00751$ | 42.8% | 0.002 |
| Qwen2.5-7B | $-0.00251$ | $-0.00237$ | 54.1% | 0.002 |
| Qwen2.5-14B | $-0.00138$ | $-0.00121$ | 43.4% | 0.002 |

**M2 Direction-Specificity Control.** To assess whether the projection-conditioned group contrast is specific to the cross-fitted alignment direction, we construct a label-free random-direction null using the M2 evaluation protocol. The five prompt-level cross-fitting folds, held-out middle-layer window, normalized states, and regression specification are kept fixed. Within each training fold, we draw 500 directions sampled uniformly from the rotation-invariant Haar measure on $\mathbb{S}^{d-1}$. These directions are not fitted to outcome labels and are not oriented by outcome labels. For each random direction $u$, we compute $\rho_{i,l}(u) = \langle h_{i,l}, u\rangle$ and $g_{i,l}(u) = \Delta\rho_{i,l}(u) + \kappa_{i,l}\rho_{i,l}(u)$ on the held-out prompts. We then fit the layer- and current-projection-adjusted regression in Eq. 19. The common-support estimate uses the layer-wise intersection of the empirical 5th–95th percentile ranges and the minimum group-count rule specified above. The empirical $p_{\mathrm{dir}}$ compares the alignment-direction coefficient with the resulting random-direction null under the same coefficient and support rule. Table 7 reports the resulting estimates. The alignment-direction coefficient lies outside the random-direction null in all four models. Across models, the null ranges are approximately $[-0.00084, 0.00096]$, with $p_{\mathrm{dir}} = 0.002$.

### D.4 Trajectory-Displacement Statistics

We summarize fixed-direction changes between two prespecified state anchors. For group $g \in \{\mathrm{Ref}, \mathrm{JB}\}$, let $\bar{\rho}_{g,l}$ be its cross-fitted mean projection. For $r \in \{0.25, 0.75\}$, let $l_r$ be the available state layer nearest to $r = l/L$, and define

$$\Delta\bar{\rho}^{\mathrm{net}}_{\mathcal{W},g} := \bar{\rho}_{g,l_{0.75}} - \bar{\rho}_{g,l_{0.25}}. \tag{87}$$

This is an anchor-to-anchor trajectory change, not a per-layer increment. The conditional M2 regression separately uses the half-open update window $\mathcal{W} = \{l : 0.25 \le l/L < 0.75\}$.

**Table 8: Cross-Fitted Middle-Window Safety-Projection Change.** Group changes are anchor-to-anchor differences $\Delta\bar{\rho}^{\mathrm{net}}_{\mathcal{W},g}$. The contrast is refused minus jailbroken, with prompt-bootstrap 95% confidence intervals. Hedges' $g$ is computed over prompt-level anchor-to-anchor changes. No layer is treated as an independent sample.

| Model | Refused change | Jailbroken change | Refused–Jailbroken [95% CI] | Hedges' $g$ |
|---|---|---|---|---|
| Llama-2-7B | $+0.2565$ | $-0.0223$ | $0.2787\ [0.2577, 0.2998]$ | 2.50 |
| Llama-3.1-8B | $+0.2042$ | $+0.0678$ | $0.1364\ [0.1247, 0.1483]$ | 2.14 |
| Qwen2.5-7B | $+0.0641$ | $+0.0212$ | $0.0430\ [0.0343, 0.0516]$ | 1.07 |
| Qwen2.5-14B | $+0.0613$ | $+0.0224$ | $0.0389\ [0.0310, 0.0464]$ | 1.08 |

## E Generalization to Low-Rank Alignment Subspaces

The preceding analysis focuses on a single alignment direction $v_{L_1} \in \mathbb{S}^{d-1}$, which is the $k$=1 special case. A single direction may miss alignment information that is spread over several coordinates. We now extend the exact propagation to a $k$-dimensional alignment subspace.

### E.1 SETUP

Let $U_a \in \mathbb{R}^{d\times k}$ be an orthonormal basis for the alignment subspace, satisfying $U_a^\top U_a = I_k$. Define the *alignment coordinates* at layer $l$:

$$z_l := U_a^\top h_l \in \mathbb{R}^k. \tag{88}$$

For any unit vector $a \in \mathbb{S}^{k-1}$ in the alignment subspace, define the corresponding one-dimensional subspace coordinate:

$$\rho_l^{(a)} := a^\top z_l = \langle h_l,\ U_a a\rangle. \tag{89}$$

Since $U_a a$ is a fixed direction in the ambient space $\mathbb{R}^d$, the exact M1 recursion applies directly to each such direction.

### E.2 PER-COORDINATE EXACT PROPAGATION

**Proposition E.1** (Subspace Alignment Propagation)**.** *Let $u_l := U_a^\top w_l \in \mathbb{R}^k$ be the realized tangent contribution to the alignment subspace. Then*

$$z_{l+1} = c_l z_l + u_l, \tag{90}$$

*and hence*

$$z_L = K_c(0, L) z_0 + \sum_{j=0}^{L-1} K_c(j+1, L) u_j. \tag{91}$$

*If $u_l = 0$ and $c_l = q \in (0, 1)$ for all $l < L$, then $z_L = q^L z_0$.*

*Proof.* Applying $U_a^\top$ to $h_{l+1} = c_l h_l + w_l$ gives the one-step vector recursion. Iteration is identical to Proposition C.5 because the coefficients $c_l$ are scalar. The passive specialization follows by setting $u_l = 0$ and $c_l = q$. □

The single-direction decomposition is the $k{=}1$ case. The proposition concerns the coordinate vector $z_l$ and does not assert the same evolution for nonlinear quantities such as $\|z_l\|_2$ or $\|z_l\|_2^2$.

### E.3 EMPIRICAL RANK AUDIT

We evaluate the dimension required by the linear measurement using five-fold prompt-level cross-fitting over the middle-depth state levels and 5,000 prompt-bootstrap resamples. A rank is non-inferior to the full available linear subspace when the upper 95% bound on $\mathrm{AUC}_{\mathrm{full}} - \mathrm{AUC}_k$ is at most the predeclared margin $0.02$.

**Table 9: Cross-Fitted Alignment-Subspace Rank Audit.** Rank-1 AUC includes a prompt-bootstrap 95% confidence interval. Full AUC uses the full available linear-subspace rank on the same middle-depth state levels.

| Model | Rank-1 AUC [95% CI] | Full AUC | Smallest non-inferior rank |
|---|---|---|---|
| Llama-2-7B | $0.939\ [0.915, 0.960]$ | $0.936$ | $1$ |
| Llama-3.1-8B | $0.902\ [0.871, 0.929]$ | $0.895$ | $1$ |
| Qwen2.5-7B | $0.858\ [0.822, 0.892]$ | $0.860$ | $1$ |
| Qwen2.5-14B | $0.875\ [0.836, 0.910]$ | $0.879$ | $2$ |

The audit supports a low-rank linear measurement in these checkpoints, not universal rank-one alignment. The audit uses state levels including the terminal state of the middle interval, whereas the M2 regression uses half-open update indices. The corresponding state-level and update counts are $17 \to 16$ for the two Llama models, $15 \to 14$ for Qwen2.5-7B, and $25 \to 24$ for Qwen2.5-14B.

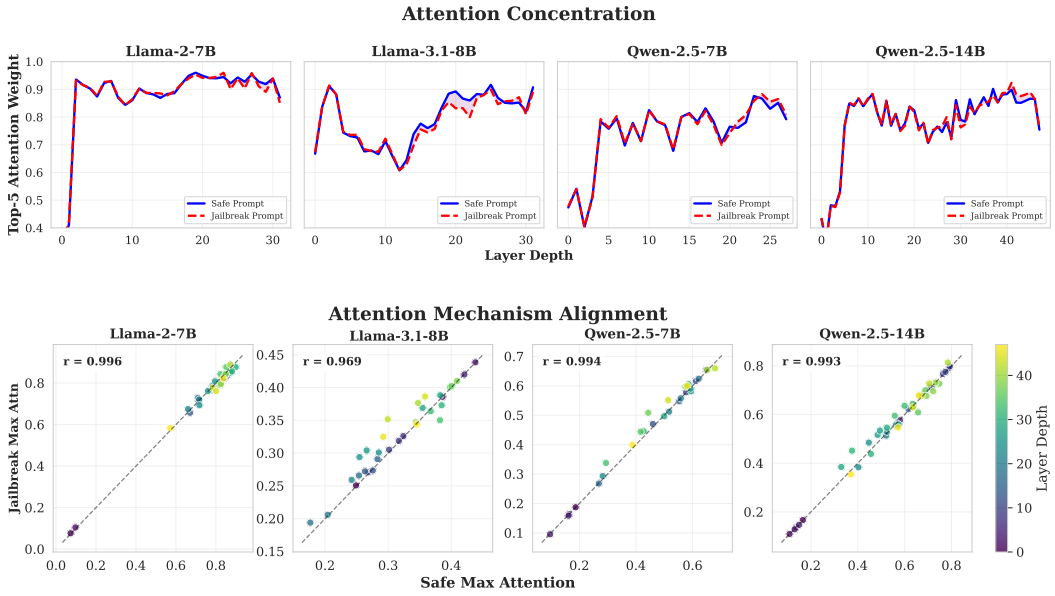


**Figure 3: Attention-Concentration Stability Analysis** (Dataset: JailbreakBench (Chao et al., 2024), $N = 100$ pairs). **Top Panel:** Visualization of attention concentration stability. The top-5 attention weights remain consistently high for both safe and jailbroken trajectories, indicating preserved coarse attention concentration. **Bottom Panel:** Comparison of maximum attention scores between safe and jailbroken states. The near-perfect correlation is consistent with preservation of coarse attention-concentration statistics during jailbreaking.

### E.4 Polyhedral Safe Region and Failure Depth

The subspace coordinates also allow alignment to be described by several thresholds rather than one. If alignment is characterized by multiple margin constraints, the safe set is the intersection of the sphere with a finite family of half-space constraints:

$$\mathcal{C}_{\text{poly}} := \{h \in \mathbb{S}^{d-1} : a_j^\top U_a^\top h \geq \tau_j, \quad j = 1, \ldots, J\}, \tag{92}$$

Here $J \geq 1$ is the number of constraints. Each $a_j \in \mathbb{S}^{k-1}$ is a constraint direction, and $\tau_j > 0$ is its threshold. For each constraint, define the coordinate $\rho_l^{(a_j)} := a_j^\top U_a^\top h_l$. For the passive reference in Proposition E.1, with $u_l = 0$ and $c_l = q \in (0, 1)$, each coordinate follows $\rho_l^{(a_j)} = q^l \rho_0^{(a_j)}$. For initially satisfied constraints with $\rho_0^{(a_j)} > \tau_j > 0$, define the continuous boundary time for coordinate $j$ by

$$t_j = \frac{\log(\rho_0^{(a_j)}/\tau_j)}{-\log q}. \tag{93}$$

The earliest continuous boundary time is $\min_j t_j$. Under the unit tangential update with constant small step $\beta$, each $t_j$ equals $2\beta^{-2}\log(\rho_0^{(a_j)}/\tau_j) + O(1)$ as in Proposition 3.10. The actual discrete failure layer is the first integer strictly greater than $\min_j t_j$, provided it lies within the observed layer horizon. At that layer at least one coordinate satisfies $q^l \rho_0^{(a_j)} < \tau_j$. The smallest initial surplus $\rho_0^{(a_j)}/\tau_j$ determines the earliest boundary.

## F Attention Concentration Stability

Figure 3 reports two coarse attention statistics as descriptive internal proxies. These statistics do not by themselves establish task or reasoning preservation.

**1. Attention Concentration (Top Panel).** We use the top-5 attention concentration metric (the sum of the top-5 attention weights per head) to measure a coarse form of attention focus. The top panel

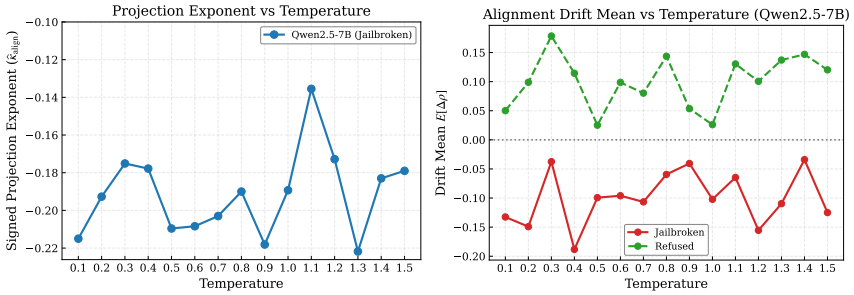


**Figure 4: Ablation on Sampling Temperature.** The signed fixed-direction alignment-projection exponent $\widehat{\lambda}_{\text{align}}$ (left) and mid-layer alignment-drift mean (right) remain stable across varying temperatures ($\text{Temp} \in [0.1, 1.5]$). The directional separation between refused and jailbroken trajectories persists across decoding stochasticity.

shows high concentration scores ($> 0.7$) and similar distributions for safe and jailbroken trajectories. This statistic changes little between harmful generation and refusal under the tested prompts.

**2. Attention-Intensity Correlation (Bottom Panel).** The bottom panel compares the maximum attention scores of safe versus jailbroken trajectories across layers. We observe a strong linear correlation ($r > 0.95$). This supports preservation of coarse attention-intensity statistics, while leaving value-stream, MLP, causal-path, and logit-level mechanisms as separate questions.

**3. Behavioral Task Measurement**

We measure task capability separately from the safety coordinate under the tested jailbreak prefix.

**Protocol.** We prepend a fixed jailbreak prefix to each GSM8K (Cobbe et al., 2021) question ($n = 1319$) while keeping the mathematical content unchanged. We evaluate accuracy under both conditions across 5 independent runs with Llama-3.1-8B-Instruct.

**Results.** Accuracy is $0.80 \pm 0.012$ under clean prompting and $0.79 \pm 0.007$ under the tested prefix across 5 runs ($\Delta < 0.01$, $p > 0.05$). This evaluation does not statistically distinguish the two accuracies for Llama-3.1-8B-Instruct on GSM8K under this prefix.

## G Ablation Study

To assess sensitivity to decoding configuration, we vary the sampling temperature $\text{Temp} \in [0.1, 1.5]$ and track two empirical geometric metrics:

1. **Signed alignment-projection exponent ($\widehat{\lambda}_{\text{align}}$):** An empirical log-slope proxy fitted from the normalized fixed-direction alignment projection. It is distinct from the exact step coefficient $\kappa_l = 1 - c_l$.
2. **Mid-layer alignment-drift mean ($\mathbb{E}[\Delta\rho_{\text{align}}]$):** The mean per-layer change in alignment projection over the mid-layer analysis window.

As shown in Figure 4, both metrics exhibit stability across the temperature spectrum. The signed projection exponent remains within a comparable range, and the directional separation between refused and jailbroken trajectories persists. Thus the measured directional separation is not specific to one tested temperature.

**Reference-Layer and Rank Sensitivity.** Across early, middle, and late fixed anchors, all 12 model–anchor mean middle-window refused-minus-jailbroken gaps are positive with prompt-bootstrap confidence intervals excluding zero. The pointwise ordering is anchor-sensitive in the two Llama models, while layer-matched ordering holds in all four models. Thus the absolute fixed-direction profile varies with the reference layer. Table 9 separately shows that a low-rank linear subspace retains the measured separation.

# H ATTENTION DISPLACEMENT PROBE

To relate the safety coordinate used in M2 (§ 3.2.5) to attention allocation, we compare attention to explicit safety context under reference and jailbreak prompts.

We prepend a fixed explicit safety preamble to define a measurable set of safety tokens for this analysis. We extract the empirically fitted alignment direction, denoted $v_{\text{align}}$, on held-out matched safe/jailbreak pairs and evaluate on disjoint pairs. For each layer, we measure the attention weight from the last pre-generation token to the safety-prefix tokens.

| **Model** | **Safe attn.** | **JB attn.** | $\Delta$ | **Layers** $\Delta > 0$ | **Mid-layer Spearman** |
|---|---|---|---|---|---|
| Llama-3.1-8B | 0.041 | 0.023 | +0.018 | 31/32 | $r$=0.625, $p$=1.6×10$^{-8}$ |
| Llama-2-7B | 0.115 | 0.083 | +0.032 | 32/32 | $r$=0.572, $p$=4.3×10$^{-7}$ |
| Qwen-2.5-7B | 0.074 | 0.043 | +0.031 | 28/28 | $r$=0.426, $p$=3.3×10$^{-4}$ |

Jailbreak-condition trajectories assign less attention to explicit safety-prefix tokens on 91 of 92 available layers across the three models. Mid-layer Last→Safety attention is positively associated with the held-out safety projection (bootstrap Spearman $r \in [0.43, 0.63]$, $p < 10^{-3}$), linking the measured coordinate to attention allocated to explicit safety context.

# I DISCUSSION

The exact geometry separates adjacent-state retention from direction-specific tangent construction. M1 gives exact propagation, a sharp one-step interval, and a directional-budget bound, with constant-retention decay as its passive specialization, while M2 tests whether jailbroken trajectories receive smaller realized tangent contributions than refused trajectories after adjustment for layer and current projection. The exact retention contrast opposes the observed net gap in all four models. Task capability is evaluated separately through behavioral measurements rather than inferred from a theoretical task direction. Direct interventions remain necessary to establish causal control by the measured direction.

## I.1 LIMITATIONS AND FUTURE WORK

In the tested models and benchmarks, the supported empirical result is a comparative tangent-contribution association after adjustment for layer and current projection. The analysis conditions on a scalar cross-fitted coordinate rather than the full hidden state, does not establish causal necessity or sufficiency, and does not derive why a particular attack changes safety-directed construction. M1 defines a constant-retention passive reference rather than a trajectory law. The task-capability measurement covers one model, task, and tested prefix.

Future work should test direct interventions, nonlinear and context-dependent safety coordinates, external benchmark generalization, and architecture-conditioned residual dynamics.