MINDCRAFT: A FULL-SEQUENCE FLOW THEORY OF HOW CONCEPTS TAKE SHAPE IN DEEP MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Two views describe how a language model represents a concept: as a linear direction or as a curved manifold. Both characterize a representation once it has formed; neither describes how a distinction is carried through depth or predicts which geometry it will take. We introduce MindCraft, which treats depth as time and follows a one-token counterfactual edit through the full-sequence flow operator, the linearized map from the edited embedding to every token at every layer. Each layer stretches and turns the edit, and how the turning is distributed over depth decides the geometry: one sharp turn yields a linear direction, gradual turning yields a manifold. We show that the flow starts isotropic and sharpens at a rate set by the first layer, and that, under a growth gap between answer-relevant and answer-irrelevant directions, each input's flow is driven toward the task read-out subspace. Across four domains and six models from three families, task-central distinctions branch early while temporal and surface details branch late or never. Steering at the branching layers lowers hallucination on HalluLens by 14.5 points and ambiguous-context bias on BBQ to 0.01–0.04.

1 INTRODUCTION

Large Language Models (LLMs) achieve strong benchmark performance, yet still produce confidently wrong outputs in deployment (Lin et al., 2022; Ji et al., 2023; Maynez et al., 2020; Zhang et al., 2025). Interpretability research seeks to uncover the mechanisms behind such behavior within the model's internal representations (Lipton, 2018; Doshi-Velez & Kim, 2017; Ribeiro et al., 2016), focusing particularly on how task-relevant distinctions are represented across layers.

Two main perspectives have emerged. The *Linear Representation Hypothesis* (LRH) (Park et al., 2024), which underlies Representation Engineering (Zou et al., 2023), posits that task-relevant concepts become approximately linearly separable in deeper layers, enabling linear probes and steering vectors. Conversely, the *Representation Manifold* (RM) view (Modell et al., 2025) argues that concept structure lies on nonlinear manifolds, where apparent linearity is merely a local approximation. Both views describe *static* properties of a trained representation at a single layer. Neither explains the layer-wise process by which a small input change is amplified into a separated representation, nor predicts which of the two geometric pictures a given concept will follow.

To study how representation geometry emerges across depth, we introduce **MindCraft**, a dynamic framework for tracing how a localized counterfactual edit propagates through the network. Given a prompt, we change a single token and follow the resulting perturbation across layers. At each layer, the perturbation undergoes two basic transformations: its magnitude is amplified or suppressed (stretching), and its direction is preserved or rotated (turning). Viewing depth as time therefore turns concept formation into a trajectory. This perspective unifies the two seemingly competing geometric views: a concentrated, shared turn can align concept values along a common direction, giving rise to the linear geometry described by LRH, whereas distributed or value-specific turning preserves a curved, multi-directional structure consistent with the manifold view. The key question thus shifts from *which geometry is correct to how turning is distributed across depth*. This distribution also reveals when a distinction becomes functionally relevant to the model: the layer at which the read-out representation first separates under an edit defines a natural **branching layer**. Figure 1 illustrates this view. A counterfactual edit begins in an initially isotropic flow, is routed across tokens, and progressively concentrates toward a task-relevant read-out subspace in deeper layers (Figure 1a).

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.

Within this shared flow, different turning patterns give rise to either approximately linear concept directions or curved representation manifolds (Figure 1b,c).

Our key contributions are summarized as follows:

- We introduce a full-sequence flow framework that traces how a localized counterfactual perturbation propagates across both depth and tokens, providing a unified view of counterfactual transport, routing, and representation dynamics.
- We develop a dynamical theory of concept geometry: the flow starts isotropic, sharpens with depth, and under a task-relevant growth gap moves toward a shared read-out subspace; within this framework, LRH and representation manifolds arise as distinct turning regimes of the same flow.
- We propose a branching-layer criterion that identifies from activations alone when the read-out representation begins to carry a concept, and use it to construct a Concept Hierarchy over distinctions in a prompt.
- We validate the theory across multiple domains and model families, and show that branching-layer-targeted steering consistently improves model behavior: across seven 7B–70B models, it reduces hallucination on HalluLens by 14.5 on average and lowers ambiguous-context bias on BBQ to 0.01–0.04.

2 Related Work

Two views describe the geometry of concepts in language models. The LRH (Park et al., 2024) holds that concepts become linearly separable with depth (Marks & Tegmark, 2024; Park et al., 2025); the RM view (Modell et al., 2025) holds that continuous factors trace smooth submanifolds, such as the circular features for days and months found by Engels et al. (2025). Both characterize a representation at a single layer. How representations change across layers has been studied with similarity measures (Kornblith et al., 2019) and layer-wise decoding (nostalgebraist, 2020; Belrose et al., 2023), and continuous-depth views treat residual networks and Transformers as dynamical systems (Chen et al., 2018; Geshkovski et al., 2023). Causal tracing and activation patching (Meng et al., 2022) localize where the effect of an edit is carried; steering methods (Zou et al., 2023; Turner et al., 2023; Rimsky et al., 2024; Li et al., 2023) extract concept directions from contrastive inputs and choose the injection layer by sweeping. We follow the flow of a counterfactual difference through depth and across tokens, linking the dynamic view to the static LRH / RM view.

3 A Full-Sequence Flow Model of the Residual Stream

Depth as time. Let $Z(t) = [z^{(1)}(t), \dots, z^{(N)}(t)]^\top \in \mathbb{R}^{N\times d}$ be the residual-stream states of all N tokens at normalized depth $t = \ell/L \in [0,1]$, and $\mathbf{z}(t) := \mathrm{vec}(Z(t)) \in \mathbb{R}^{Nd}$ the state of the whole sequence. We read the forward pass as the ODE

$$\frac{d}{dt}\mathbf{z}(t;X) = \mathbf{f}(\mathbf{z}(t),t) := \mathbf{f}_{\mathrm{MHA}}(\mathbf{z}(t),t) + \mathbf{f}_{\mathrm{MLP}}(\mathbf{z}(t),t), \qquad \mathbf{z}(0;X) = \text{embedding of } X, \quad (1)$$

so that the residual update of layer ℓ is $\frac{1}{L}\mathbf{f}(\cdot, \ell/L)$. The Jacobian $\mathbf{J}(\mathbf{z},t) := \partial\mathbf{f}/\partial\mathbf{z} \in \mathbb{R}^{Nd\times Nd}$ is an $N \times N$ array of $d \times d$ blocks; the off-diagonal block $\mathbf{J}^{(p,i)}$ is the instantaneous routing from token i to token p by attention (zero for $p < i$ under causal masking).

The flow operator. Let $X_{\Delta x}$ be the counterfactual of an input X obtained by replacing token i, and $\boldsymbol{\delta}(t) := \mathbf{z}(t;X_{\Delta x}) - \mathbf{z}(t;X) \in \mathbb{R}^{Nd}$ the *counterfactual flow*; at $t = 0$ it lives on the block of token i only. Let $\boldsymbol{\Phi}(t,\tau) \in \mathbb{R}^{Nd\times Nd}$ solve $\frac{d}{dt}\boldsymbol{\Phi}(t,\tau) = \mathbf{J}(\mathbf{z}(t),t)\boldsymbol{\Phi}(t,\tau)$ with $\boldsymbol{\Phi}(\tau,\tau) = \mathbf{I}$. Then (Section A.1)

$$\boldsymbol{\delta}(t) = \boldsymbol{\Phi}(t,0)\boldsymbol{\delta}(0) + O(\|\boldsymbol{\delta}(0)\|^2). \quad (2)$$

$\boldsymbol{\Phi}(t,0)$ says where a small change of the input goes. Its block $\boldsymbol{\Phi}^{(p\leftarrow i)}(t)$ carries an edit of token i to token p. Since an edit starts on its own token, the share of its energy found on other tokens, $E_i(t) := \sum_{p\neq i}\|\boldsymbol{\Phi}^{(p\leftarrow i)}(t)\|_F^2 / \sum_p \|\boldsymbol{\Phi}^{(p\leftarrow i)}(t)\|_F^2$, measures how far attention has *routed* it; $E_i(0) = 0$. The block row $\boldsymbol{\Phi}^N(t) := [\boldsymbol{\Phi}^{(N\leftarrow 1)} \dots \boldsymbol{\Phi}^{(N\leftarrow N)}] \in \mathbb{R}^{d\times Nd}$ is the *read-out flow*: everything that reaches the last token, whose state conditions the next prediction (the *read-out token*).

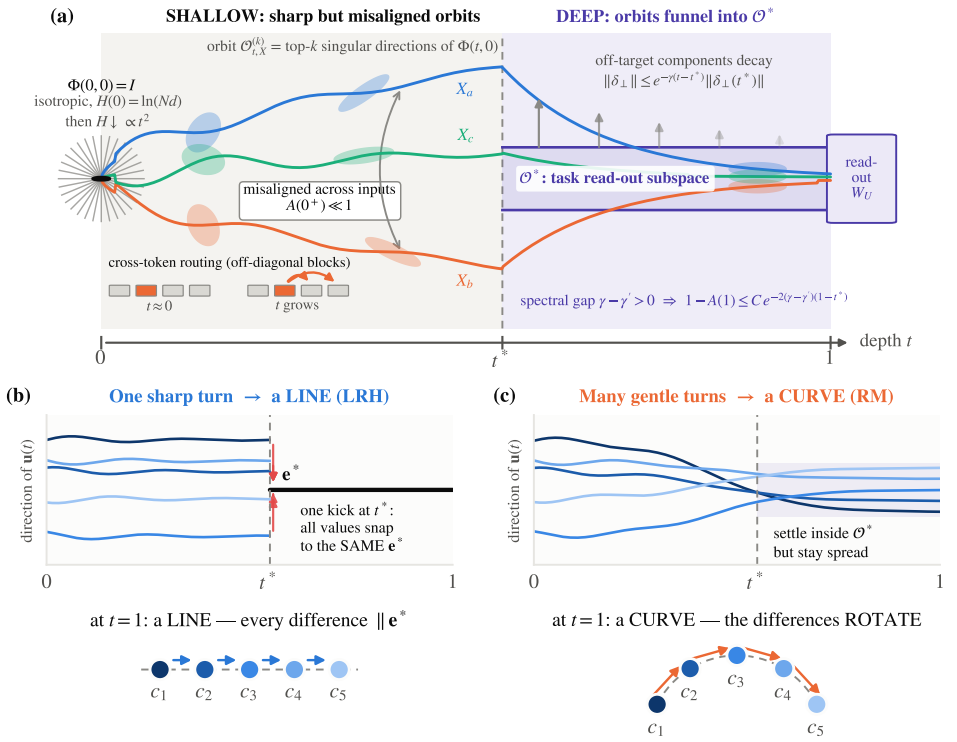


Figure 1: **MindCraft in one picture.** (a) The flow operator $\mathbf{\Phi}(t,0)$ starts isotropic ($H(0)=\ln Nd$); shallow orbits sharpen along input-specific, misaligned directions while attention routes the edit across tokens (inset); past t^* the on-target part outgrows the off-target part and orbits move toward $\mathcal{O}^*$ (Theorem 4.3). (b) Shared turning sends all values of a concept to one direction $\mathbf{e}^*$: a line (LRH). (c) Value-specific turning keeps the values spread inside $\mathcal{O}^*$: a curve (RM).

Finite edits. A token swap is not a small change. Along the straight path $\mathbf{z}_s(0)=\mathbf{z}(0;X)+s\,\boldsymbol{\delta}(0)$, $s\in[0,1]$, between the two embeddings, the fundamental theorem of calculus gives, *exactly*,

$$\boldsymbol{\delta}(t)=\bar{\mathbf{\Phi}}(t,0)\,\boldsymbol{\delta}(0),\qquad \bar{\mathbf{\Phi}}(t,0):=\int_0^1 \mathbf{\Phi}(t,0;\mathbf{z}_s)\,ds, \tag{3}$$

the flow operator averaged along the path, whose $s=0$ term is $\mathbf{\Phi}(t,0)$. Quantities of the operator itself (entropy, orbits) are computed from $\mathbf{\Phi}$; quantities of an actual pair are computed from $\boldsymbol{\delta}=\bar{\mathbf{\Phi}}\boldsymbol{\delta}(0)$, which two forward passes give directly (Section C.2).

4 Spectral Dynamics of the Flow

4.1 An isotropic start that sharpens

With singular values $\sigma_1\ge\cdots\ge\sigma_{Nd}>0$ of $\mathbf{\Phi}_X(t,0)$ and $p_i=\sigma_i/\sum_j\sigma_j$, the *spectral entropy* $H(t)=-\sum_i p_i\ln p_i$ measures how evenly the operator spreads a random edit over directions (e^H is the effective number of directions reached). Two facts hold for any network of the form equation 1 (Sections A.2 and A.3). First, with $\mu_i=\mathbf{u}_i^\top\mathbf{J}\mathbf{u}_i$ the growth rate along the i-th singular direction,

$$\frac{d}{dt}H(t)=-\mathrm{Cov}_p(\mu,\ln p)\,: \tag{4}$$

entropy falls exactly when the directions that already dominate receive the larger growth rates, i.e. when amplification is preferential. Second, $H(0)=\ln Nd$ because $\mathbf{\Phi}(0,0)=\mathbf{I}$, and

$$H(t)=\ln Nd-\frac{t^2}{2}\,\mathrm{Var}\big(\lambda(\mathrm{sym}\,\mathbf{J}(0))\big)+O(t^3), \tag{5}$$

so the sharpening starts flat and its initial speed is set by how anisotropically the first layer amplifies.

4.2 SHALLOW ORBITS ARE INPUT-SPECIFIC

Which directions does the sharpened operator favor? We call the span of the top-k left singular vectors of $\mathbf{\Phi}_X(t,0)$ the *orbit* of X: the directions in which a small edit of X mostly lives after depth t. Its restriction to the read-out token, the top-k left singular subspace of $\mathbf{\Phi}_X^N(t)$, is the *read-out orbit*, with projector $\mathbf{P}_{t,X}$. As $t \to 0^+$ the read-out orbit converges to the top-k eigenspace of $\operatorname{sym} \mathbf{J}_X^{(N,N)}(0)$, the symmetric part of the first layer's Jacobian at the read-out token (Section A.4). It is thus set by the tokens of X through the attention pattern and the value–output circuit, and inputs in different contexts have different shallow orbits: the left half of Figure 1(a), orbits that sharpen (Equation (5)) but are misaligned across inputs.

4.3 THE TASK READ-OUT SUBSPACE AND THE DEEP LAYERS

Which directions are "on target" is decided by what the model reads out. Only differences of logits move the softmax, so only differences of unembedding rows can change the answer. With $\mathbf{w}_a$ the a-th row of $\mathbf{W}_U$, $\mathbf{g}$ the gain of the final normalization and $\mathcal{A}$ the candidate answers of a task, the *task read-out subspace* is

$$\mathcal{O}_d^* := \operatorname{span}\{\mathbf{g} \odot (\mathbf{w}_a - \mathbf{w}_{a'}) : a, a' \in \mathcal{A}\} \subset \mathbb{R}^d, \qquad k = \dim \mathcal{O}_d^* \le |\mathcal{A}| - 1, \tag{6}$$

with projectors $\mathbf{P}_\parallel$ onto it and $\mathbf{P}_\perp = \mathbf{I} - \mathbf{P}_\parallel$. It is low-dimensional and fixed by the model and the task. The part $\mathbf{P}_\parallel \boldsymbol{\delta}$ of an edit can change the answer (*on-target*); the part $\mathbf{P}_\perp \boldsymbol{\delta}$ changes the representation but not the answer (*off-target*).

The deep layers bring the shallow orbits, sharp but different, together (right half of Figure 1a). What this requires is a *growth gap* at the read-out token.

Assumption 4.1 (Growth gap)**.** There are an onset $t^* \in [0,1)$ and rates $\gamma_\parallel > \max(\gamma_\perp, 0)$ such that for $t \in [t^*, 1]$ and all inputs: (A1) off-target directions grow at most at rate $\gamma_\perp$: $\mathbf{v}^\top \operatorname{sym}(\mathbf{J}^{(N,N)})\mathbf{v} \le \gamma_\perp \|\mathbf{v}\|^2$ for $\mathbf{v} \in (\mathcal{O}_d^*)^\perp$; (A2) on-target directions grow at least at rate $\gamma_\parallel$ and receive little from elsewhere: $\mathbf{u}^\top \operatorname{sym}(\mathbf{J}^{(N,N)})\mathbf{u} \ge \gamma_\parallel \|\mathbf{u}\|^2$ for $\mathbf{u} \in \mathcal{O}_d^*$, and $\|\mathbf{P}_\parallel \mathbf{J}^{(N,N)} \mathbf{P}_\perp\| + \sum_{p \ne N} \|\mathbf{P}_\parallel \mathbf{J}^{(N,p)}\| \le \eta_\parallel$; (A3) attention injects little into off-target directions, $\|\mathbf{P}_\perp \mathbf{J}^{(N,N)} \mathbf{P}_\parallel\| + \sum_{p \ne N} \|\mathbf{P}_\perp \mathbf{J}^{(N,p)}\| \le \eta$; (A4) $\|\mathbf{\Phi}_X(t,0)\| \le \Lambda$.

The gap $\Delta\gamma := \gamma_\parallel - \gamma_\perp > 0$ states that, once the model has settled what matters, the answer-relevant part of an edit is amplified faster than the rest. All four conditions constrain only the instantaneous Jacobian at the read-out token; the growth of the accumulated on-target flow follows from them.

Lemma 4.2 (On-target growth)**.** *Under (A2) and (A4), for $t \ge t^*$, $\sigma_k(\mathbf{P}_\parallel \mathbf{\Phi}_X^N(t)) \ge e^{\gamma_\parallel (t - t^*)}\, \underline{\sigma}_X$, where $\underline{\sigma}_X := \sigma_k(\mathbf{P}_\parallel \mathbf{\Phi}_X^N(t^*)) - \eta_\parallel \Lambda / \gamma_\parallel$ (Section B.1).*

To compare inputs $\{X_a\}_{a=1}^M$ we use the *cross-input alignment* $A(t) := \frac{1}{M(M-1)k} \sum_{a \ne b} \operatorname{Tr}(\mathbf{P}_{t,a} \mathbf{P}_{t,b}) \in [0,1]$, the average overlap of their read-out orbits: 1 when all inputs share one orbit and $\approx k/d$ for unrelated orbits.

Theorem 4.3 (Deep-layer convergence to the read-out subspace)**.** *Under Assumption 4.1, let $r_X(t) := \|\mathbf{P}_\perp \mathbf{\Phi}_X^N(t)\| / \sigma_k(\mathbf{P}_\parallel \mathbf{\Phi}_X^N(t))$ be the off/on ratio of the read-out flow and $\rho_X := \|\mathbf{P}_\perp \mathbf{\Phi}_X^N(t^*)\| / \underline{\sigma}_X$, with $\underline{\sigma}_X > 0$ as in Lemma 4.2. Then*

$$r_X(t) \le \bar{r}_X(t) := \rho_X\, e^{-\Delta\gamma (t - t^*)} + \frac{\eta \Lambda\, (t - t^*)\, e^{(\gamma_\perp^+ - \gamma_\parallel)(t - t^*)}}{\underline{\sigma}_X}, \tag{7}$$

and whenever $\bar{r}_X(t) \le \frac{1}{2}$ the read-out orbit satisfies $\|\mathbf{P}_{t,X} - \mathbf{P}_\parallel\| \le 2\bar{r}_X(t)$; at the output, $1 - A(1) \le 16 \max_X \bar{r}_X(1)^2$.

The proof (Section B.2) combines a Grönwall argument with a Davis–Kahan angle bound. The off/on ratio of every input's flow decays at the rate of the gap; once it is small, every read-out orbit lies close to $\mathcal{O}_d^*$, the same subspace for every input, so that, under Assumption 4.1, orbits that are misaligned at the start become aligned.

5 ONE EDIT THROUGH DEPTH: STRETCHING AND TURNING

Section 4 describes the operator as a whole. To follow a single edit, let $\mathbf{u}(t) := \boldsymbol{\delta}(t)/\|\boldsymbol{\delta}(t)\|$ be its direction. Splitting $\mathbf{J}\mathbf{u}$ along and across $\mathbf{u}$ gives (Section A.5)

$$\frac{d}{dt}\ln\|\boldsymbol{\delta}\| = \mu(t) := \mathbf{u}^\top\mathbf{J}\mathbf{u}, \qquad \dot{\mathbf{u}} = (\mathbf{I}-\mathbf{u}\mathbf{u}^\top)\mathbf{J}\mathbf{u}, \qquad \omega(t) := \|\dot{\mathbf{u}}(t)\| \le \|\mathbf{J}(t)\|. \tag{8}$$

To first order a layer *stretches* an edit, by the Rayleigh quotient μ, and *turns* it, by the angular speed ω of its direction. Stretching decides when the edit matters for the output (Section 5.1); turning decides the geometry of a concept (Section 5.2); the flow after a layer decides what an added direction does (Section 5.3).

5.1 STRETCHING: WHEN THE READ-OUT TOKEN PICKS UP AN EDIT

An edit of token $i \neq N$ starts outside the read-out token. It arrives there through the routing blocks $\boldsymbol{\Phi}^{(N\leftarrow i)}$ and, once there, is stretched at rate μ. By equation 3 the resulting *read-out response* is exactly the read-out block of the path-averaged flow operator applied to the edit, $\boldsymbol{\delta}_N(\ell) = \bar{\boldsymbol{\Phi}}^{(N\leftarrow i)}(\ell,0)\,\boldsymbol{\delta}(0)$. The read-out token carries the edit once this response is no longer small compared with the state it is added to. With hats denoting coordinates divided by their standard deviation over reference read-out states, so that a few high-magnitude coordinates do not dominate, we define

$$s_\ell := \cos\big(\hat{z}_N(\ell),\ \hat{z}_N(\ell)+\hat{\boldsymbol{\delta}}_N(\ell)\big), \qquad \ell^*(X, X_{\Delta x}) := \min\{\ell :\ s_\ell < \tau\}, \tag{9}$$

with $\ell^* = \mathrm{INF}$ if the score never crosses τ. Since $\sin\angle(a, a+b) = \|\mathbf{P}_a^\perp b\|/\|a+b\|$, with $\mathbf{P}_a^\perp$ the projector orthogonal to a, the condition $s_\ell < \tau$ says that the part of the response orthogonal to the state exceeds a fraction $\sqrt{1-\tau^2}$ of the edited state. The **branching layer** ℓ^* is therefore the first depth at which the read-out block of the flow operator, applied to the edit, dominates the read-out state. The **Concept Hierarchy** of a prompt is its edits sorted by ℓ^*: the order in which the read-out responses of the edits reach this level. Because the level a cosine reaches depends on the coordinates it is computed in, τ is calibrated on half of the prompts to a fixed fraction of branching pairs and the layers are read on the other half (Section C.5); what the procedure reports is the order. The hierarchy figures read the same response in the singular basis of the value projection, keeping its ten strongest coordinates (Section C.3); both readings give the same order (Table 1).

Why should this order matter for the output? To second order, the causal effect of an edit on the next-token distribution is $\mathrm{KL}(p_X\|p_{X_{\Delta x}}) = \frac{1}{2}\mathrm{Var}_{a\sim p_X}[\Delta\ell_a]$, where, up to the radial direction removed by the final normalization, the logit gap between two tokens changes by $\Delta\ell_a - \Delta\ell_b \propto \langle \mathbf{g}\odot(\mathbf{w}_a-\mathbf{w}_b), \boldsymbol{\delta}_N(1)\rangle$ (Section L). The causal effect is thus the size of the on-target read-out response at the output, while the branching layer is when that response first dominates the state. Stretching connects the two.

Prediction 5.1 (Selective stretching)**.** An edit that changes what the model should say next is stretched more than one that does not. It therefore reaches the branching level earlier, and the order of branching layers follows the causal effect of the edits.

5.2 TURNING: THE GEOMETRY OF A CONCEPT

Turning acts on the *set* of values of a concept. Fix a concept with values $c_1,\dots,c_n$, and let the *coherence* $\kappa(t) \in [0,1]$ be the fraction of the spread of $\{\mathbf{z}(t;c_j)\}_j$ at the read-out token that one direction carries ($\kappa = 1$ for values on a line, $\kappa \approx 1/(n-1)$ for values spread at random). Two regimes follow (Figure 1b,c).

- *Shared turning (LRH regime).* If the layers turn all values to a common direction $\mathbf{e}^*$, the differences $\mathbf{z}(1;c_{j+1}) - \mathbf{z}(1;c_j)$ are parallel, the values lie on a line, $\kappa \to 1$, and a single steering vector suffices.
- *Value-specific turning (RM regime).* If each value keeps a direction of its own, the values stay spread inside $\mathcal{O}_d^*$ and κ stays small. Spread needs $\dim\mathcal{O}_d^* \ge 2$; for a binary concept the two pictures coincide.

Section 4.3 tells where the model can change the geometry of a concept: the turn into $\mathcal{O}_d^*$ at the onset t^* is the last large turn the read-out token receives. It also gives a reference computed from the

weights alone, before any forward pass: the coherence $\kappa_U := \kappa(\{\mathbf{g} \odot \mathbf{w}_{a(c_j)}\}_j)$ of the unembedding rows of the values' answers $a(c_j)$.

Prediction 5.2 (Geometry at the read-out)**.** When the value of a concept is what the model must say next, the coherence of its values at the output moves toward κ_U; in a context that does not ask for the value, it need not.

5.3 Transport: steering at the branching layer

A direction added to the read-out state at layer ℓ reaches the output through the flow after ℓ, just as the counterfactual does: by equation 3 applied from layer ℓ, $\boldsymbol{\delta}(1) = \bar{\boldsymbol{\Phi}}(1, \ell)\, \boldsymbol{\delta}(\ell)$.

Prediction 5.3 (Transport)**.** A direction read from counterfactuals at a layer where the read-out token carries the edit, and added there, moves the output the way the counterfactual does.

MindCraft steering (Algorithm 1, Section M.1) collects the residual-stream differences of many counterfactual pairs of a task, each at its branching layer, takes their principal direction $\mathbf{d}$, and adds $\alpha\mathbf{d}$ at those layers, as in Representation Engineering. The injection depth is the branching layer, where the read-out token first carries the distinction.

6 Experiments

Each experiment tests one part of the theory: Section 6.1 measures the operator (Section 4); Section 6.2 tests selective stretching and the Concept Hierarchy (Prediction 5.1); Section 6.3 tests the geometry of concepts (Prediction 5.2); Section 6.4 tests transport and steering (Prediction 5.3). The measurements of Figure 2 are on Qwen2.5-7B-Instruct ($L = 28$, $d = 3{,}584$) (Qwen Team et al., 2024), and every panel is repeated on Qwen2.5-0.5B, Qwen2.5-14B, Llama-3.1-8B, Mistral-7B-v0.3 and Gemma-2-9B (Section G). The Concept Hierarchies use Qwen2.5-7B-Instruct; the large-scale validation uses all six models.

6.1 Measuring the flow operator

Setup. For an edit at token i only the block column $\partial \mathbf{z}(t)/\partial e_i$ of $\boldsymbol{\Phi}$ is needed. Differentiating the forward pass along $m = 128$ orthonormal directions of the input embedding gives the action of this column at every layer in one pass (Section F). From it we read H, E_i, the read-out orbit ($k = 8$), A and the overlap with $\mathcal{O}_d^*$; from $\boldsymbol{\delta}$ we read its turning and its on- and off-target parts. Prompts ask for the capital of 16 countries in two formats ("The capital of {country} is", "Q: What is the capital of {country}? A:"); the edit swaps the country and $\mathcal{A}$ is the set of capitals ($\dim \mathcal{O}_d^* = 15$).

Results (Figure 2a–c). *Isotropic start*: $H/\ln m$ starts at 1 and decreases monotonically to 0.93 and 0.95 in the two formats. *Routing*: E_i rises from 0 to 0.24 and 0.78, so in the question format most of the edit's energy has left the edited token by the output. *A late growth gap*: through layer 21 the difference $\boldsymbol{\delta}$ keeps 99.98% of its energy off target. From layer 21 to the output its on-target part grows $38\times$ and $110\times$ in the two formats, and its off-target part only $3.6\times$ and $8.4\times$, so the off/on ratio falls from 86 to 7, with onset $t^* = 0.79$ (Section G). The flow thus moves toward $\mathcal{O}_d^*$ by outgrowing its off-target part, as Theorem 4.3 describes. It ends before the bound of the theorem applies, at a ratio of 3.4–6.9 on all six models (Section B.3). The overlap with $\mathcal{O}_d^*$ reaches 1.4–$2.1\times$ chance near the output on four of the six models but stays below chance on Qwen2.5-7B and 14B, where the gap shows in the counterfactual but not in the top singular directions of the operator (Section G).

6.2 Selective stretching and Concept Hierarchies

Selective stretching. We call the growth of $\|\delta_\ell\|$ with depth the *expansion* of an edit, the integrated stretch $\int \mu$ of equation 8. On the persona prompt of Zou et al. (2023), the *honesty edit* (*honest→untruthful*) changes what the model should say next, and the *control edit* (*room A→room B*) does not. We apply each in every prefix of ten science and ten daily-life statements (290 and 295 prefixes; Figure 3; Section E). At layer 2 the two move the final token by comparable amounts (control/honesty ratio of mean displacement norms 0.59, 0.58). With depth the honesty displacement grows about 100-fold and the control displacement 30–40-fold, so the ratio falls to 0.18, 0.21 at

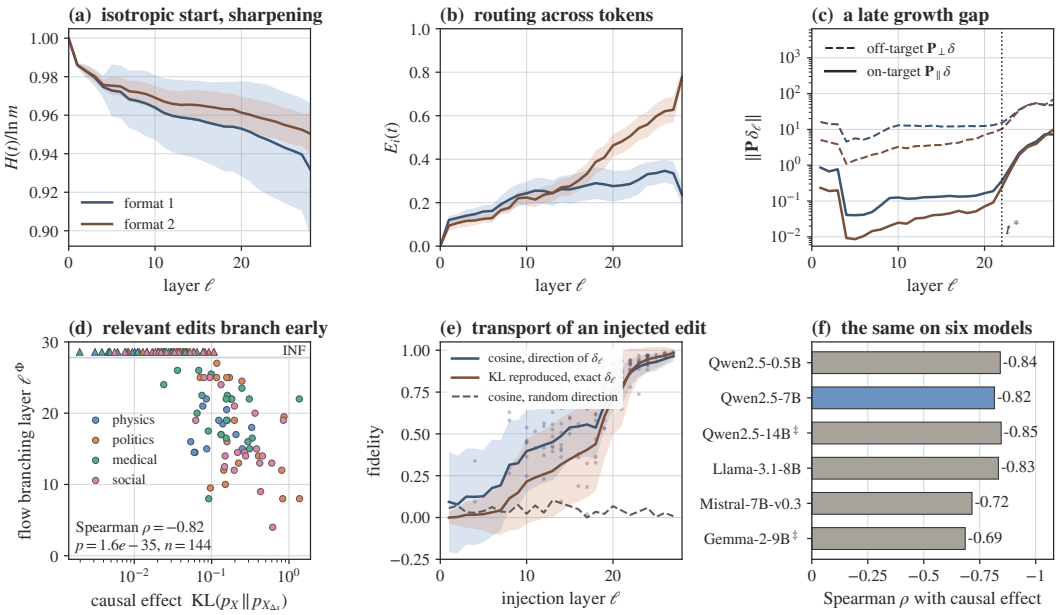


Figure 2: **The flow picture on Qwen2.5-7B-Instruct.** (a) Spectral entropy of the operator; (b) share of the edit's energy routed away from the edited token; (c) on-target (solid) and off-target (dashed) size of $\boldsymbol{\delta}$ at the read-out token, with the onset t^* dotted (mean $\pm$ s.d. over 16 country prompts in each of two formats). (d) Branching layer (Equation (9)) of the 144 concepts of the prompt bank against their causal effect (triangles: never branch). (e) Transport: cosine of the output change with the true counterfactual change when the direction of $\boldsymbol{\delta}_\ell$ is added at layer ℓ ($0.3\,\|h_\ell\|$, all positions), against a random direction of the same strength, and share of the output KL divergence reproduced by the exact difference added at the read-out token (mean $\pm$ s.d. over 288 pairs; dots: each pair at its own branching layer). (f) Spearman ρ between branching layer and causal effect on six models (‡bfloat16). All panels on six models in Section G.

layer 24, most steeply at $t \approx 0.3$–0.65; comparing at the same layer cancels the growth of the residual-stream norm. The flow expands selectively. The same holds on Qwen2.5-0.5B with a length-matched honesty edit and a control edit that starts at the size of the honesty edit (Section E).

Concept Hierarchies. Figure 4 shows two hierarchies (base prompts from ARC, COPA and TruthfulQA, Section O; edits proposed by the pipeline of Section N). In the physics prompt, edits that change the outcome (*break*/*maintain*) or the governing principle (*physics*/*magic*) branch at layer 2 and object substitutions (*glass*/*rubber*, *building*/*table*) at layers 12–13; in the politics prompt, policy and sentiment edits branch early, roles and entities later, and temporal and location details never. The same partition into central and surface distinctions appears in medical diagnosis and social evaluation and on six models from three families (Section H.4). The context can reorder it: appending "*in this sentence, 2024 or 2025 is very important*" moves the *2024*/*2025* edit from layer 11 to layer 6 (Section H.6).

The order follows the causal effect. We built 24 prompts in four domains with six single-token edits each, every edit measured in the 32 contexts formed by the other five: 144 concepts and 4,608 pairs per model (Section I.1). In line with Prediction 5.1, edits with a large causal effect $\mathrm{KL}(p_X \| p_{X_{\Delta x}})$ branch early (Figure 2d, Table 1): $\rho = -0.82$ to -0.84 for the branching layer of equation 9 on the three models of Table 1 and -0.69 to -0.85 on all six (Figure 2f), and -0.69 to -0.81 for its spectral reading; a random orthogonal basis gives nearly the same values, and every truncation level $k \geq 2$ gives ρ between -0.59 and -0.84 (Table 4). On the 32 pairs of an independent pilot set on Qwen2.5-0.5B the relation holds pair by pair ($\rho = -0.62$, $p = 1.6 \times 10^{-4}$; Section I.2). The arrival layer of activation patching (Meng et al., 2022), the depth by which the read-out token has received half of an edit's effect, is unrelated to the effect size: patching normalizes each edit by

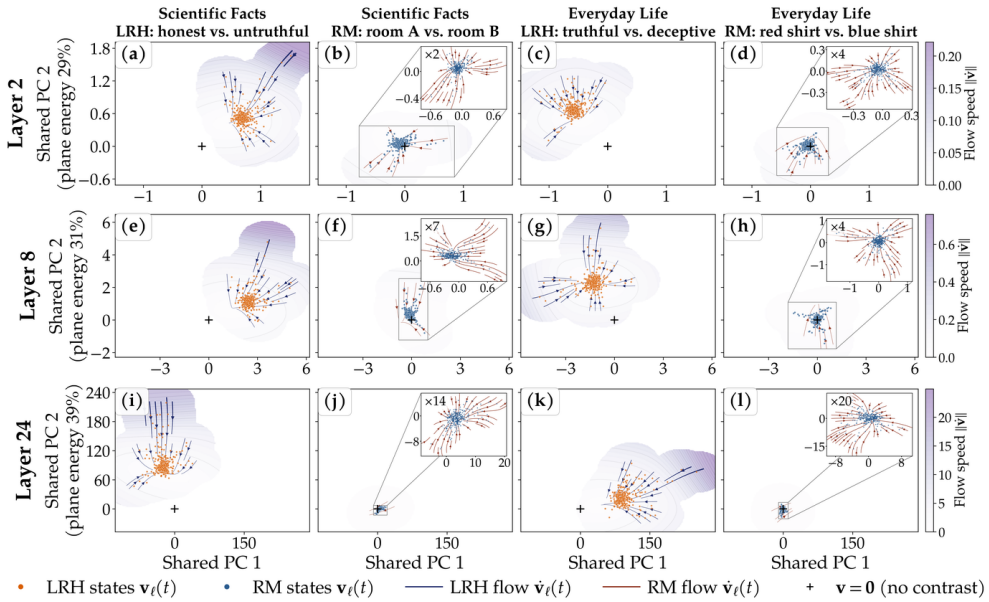


Figure 3: **Selective expansion.** Each point is the final-token displacement $\delta_\ell^{(c)}$ of an edit in one context c (a prefix of one of ten statements). Columns: honesty edit and control edit on science (A) and daily-life (B) statements; rows: layers ℓ. Panels in a row share one uncentered projection (+ marks $\delta = 0$), one axis range and one color scale; insets magnify the control edit. Streamlines interpolate the change of δ_ℓ between consecutive prefixes. All numbers quoted in the text are computed in the full 3584-dimensional space.

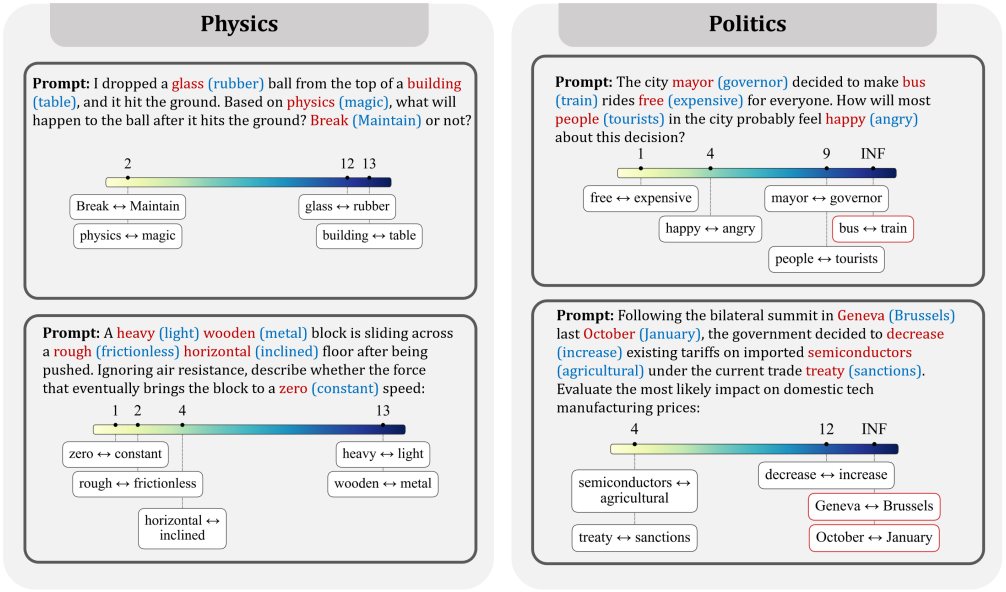


Figure 4: **Concept Hierarchies** on Qwen2.5-7B for physics and politics (medical diagnosis and social evaluation in Figure 11; spectral reading, $k = 10$, $\tau = 0.9$). Each concept is placed at its branching layer, the first layer at which the edit dominates the read-out state; INF: never dominates. Pairs are marked Factual (Counterfactual).

its own effect and asks *where* it travels, while the branching layer asks *when* it comes to dominate the read-out state.

Table 1: Spearman ρ between a concept's branching layer and its log-KL causal effect (144 concepts, 4,608 pairs per model; confidence intervals and further variants in Section I.1; the branching layer on three further models in Table 3). †On 1,152 pairs.

Measure	Qwen2.5-0.5B	Qwen2.5-7B	Llama-3.1-8B
Branching layer (Equation (9))	$\mathbf{-0.84}$	$\mathbf{-0.82}$	$\mathbf{-0.83}$
Spectral reading (W_V, k=10)	-0.75	-0.69	-0.81
Spectral reading, random orthogonal basis (k=10)	-0.78	-0.75	-0.80
Activation patching, arrival layer†	0.00	-0.04	0.08

6.3 Turning and the geometry of a concept

We follow the coherence $\kappa(\ell)$ of concepts with several values (ordinal, cyclic, categorical) in a context where the value is the next answer ("This month is {}. Next month is") and in a neutral one (Section J, Figure 10). Each concept starts from the geometry of its values: ratings and numbers are close to a line in the neutral context, and months, colors and capitals stay spread. When the value is the next answer, the layers after t^* spread the values out, a value-specific turn: from layer 21 to the output, numbers fall from $\kappa = 0.66$ to 0.38 and capitals from 0.41 to 0.15, and the coherence of cyclic and categorical concepts falls after t^* on all six models (one exception, Section G). A shared turn, which pulls the rating scale onto a line (κ from 0.39 to 0.78), appears on Qwen2.5-0.5B.

To test Prediction 5.2 against a quantity fixed before any forward pass, we compute κ_U for ten concepts, with two templates per context, on three models (Section J.2). Relative to the neutral context, the answer context moves the output coherence toward κ_U in 23 of 30 concept–model cases (sign test $p = 0.003$), most strongly for cyclic and categorical concepts, whose answer rows are spread ($\kappa_U = 0.12$–0.22); on Llama-3.1-8B, weekdays fall from $\kappa = 0.70$ (neutral) to 0.28 (answer) against $\kappa_U = 0.33$. A small κ alone does not separate a curve from a scatter; for months and weekdays the spread values lie on the circle reported by Engels et al. (2025), with median circular correlation 0.72 and 0.76 at the read-out token (Section J.2).

6.4 Transport and steering

Transport. As Prediction 5.3 predicts, the direction of $\boldsymbol{\delta}_\ell$ added at layer ℓ at a fixed strength moves the output along the true counterfactual change (Figure 2e): the cosine rises from 0.40–0.54 at layers 10–16 to 0.97 at layer 27, against at most 0.10 for a random direction of the same strength, and is 0.71 at each pair's own branching layer. On all six models it is 0.72–0.90 at $\ell = 0.75L$ (Table 3).

Benchmarks. Relative to the unsteered model, steering at the branching layers lowers the hallucination rate on HalluLens PreciseWikiQA (Bang et al., 2025) on all seven models tested, by 14.5 points on average, with the refusal rate unchanged (Table 11). On BBQ (Parrish et al., 2022) it lowers the ambiguous-context bias score to 0.01–0.04, with disambiguated accuracy within 3.5 points of the baseline (Table 12).

7 Conclusion

We described how the LLM carries a counterfactual edit through depth with one object, the full-sequence flow operator $\boldsymbol{\Phi}(t, 0)$, and its path average $\bar{\boldsymbol{\Phi}}$. Its singular structure gives the orbits of inputs, which start isotropic, sharpen, and under a growth gap converge to the task read-out subspace. Its action on one edit splits into stretching and turning. Stretching sets the branching layer, where the read-out block of the flow first dominates the read-out state, and ordering these layers gives a Concept Hierarchy that follows the causal effect of the edits; turning sets the geometry of a concept, with LRH and RM as shared and value-specific turning of its values. Steering with directions read at the branching layers reduces hallucination and bias on seven models. **Limitations.** Our analysis is limited to single-time counterfactual edits and to tasks whose answers form a candidate set, which defines the read-out subspace. Two extensions remain challenging: multi-time counterfactual edits, as in complex scenario recognition, where the answer depends on several steps jointly, and open-ended generation, where no candidate set is available to define the read-out subspace.

AI USE STATEMENT

Synthetic data (part of the method). An LLM is a component of the pipeline in Section N, where it generates the counterfactual pairs used in our experiments. This constitutes generation of synthetic data; details are given in Section N.

Proofs. We used generative AI tools to assist with the proofs. All proofs were checked by the authors, who take responsibility for their correctness.

Code. We used AI coding assistants to help write and debug the code implementing our method and experiments. All AI-generated or AI-modified code was reviewed and verified by the authors.

Writing and presentation. We used generative AI tools to translate parts of the text into English, to improve readability (proofreading, grammar, terminology), and for occasional LaTeX formatting suggestions. All AI-assisted text was reviewed and revised by the authors.

Uses we did not make. Apart from the uses above, we did not use generative AI tools to generate, clean or reformat datasets; to develop theoretical models or conceptual frameworks; to formulate mathematical claims; to propose or refine hypotheses; to design, or give feedback on, the methodology or experiments; to analyse data; or to interpret results.

We take full responsibility for the final content of this work, including text, claims, proofs, code and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

MindCraft is an interpretability method and a steering procedure evaluated on public benchmarks with open-weight models. Like other steering methods, it could be used to shift model behavior in undesirable directions; the same analysis makes such shifts detectable. We see no ethical concerns specific to this paper beyond those common to interpretability research.

REPRODUCIBILITY STATEMENT

All theorems are proved in Sections A and B under explicitly stated assumptions. The tracing procedure, including the path-averaged operator and threshold calibration, is specified in Section C; the operator measurements in Sections F and G; the selective-stretching experiment in Section E; the large-scale validation, with the full prompt bank, in Section I.1; the geometry tests in Section J.2; and the steering pipeline and the controlled comparison, with all hyper-parameters and data splits, in Section M. The tables of Sections C.2, G, I.1 and J.2 are generated directly from the saved experimental outputs. Code for all experiments will be released. The evaluation uses public benchmarks (TruthfulQA, ARC, COPA, HalluLens, BBQ, MMLU) and open-weight checkpoints (Qwen2.5, Llama-3.1, Mistral, Gemma-2).

REFERENCES

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. Hallulens: Llm hallucination benchmark. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. doi: 10.18653/v1/2025.acl-long.1176.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv (Cornell University)*, 2023. doi: 10.48550/arxiv.2303.08112.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems 31*, 2018.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv (Cornell University)*, 2018. doi: 10.48550/arxiv.1803.05457.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 1970. doi: 10.1137/0707001.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv (Cornell University)*, 2017. doi: 10.48550/arxiv.1702.08608.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv (Cornell University)*, 2024. doi: 10.48550/arxiv.2407.21783.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012. doi: 10.1145/2090236.2090255.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are linear. *International Conference on Learning Representations*, 2025. doi: 10.48550/arxiv.2405.14860.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *arXiv (Cornell University)*, 2023. doi: 10.48550/arxiv.2312.10794.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 2011. doi: 10.1137/090771806.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems 29*, 2016.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991. doi: 10.1017/cbo9780511840371.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023. doi: 10.1145/3571730.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv (Cornell University)*, 2023. doi: 10.48550/arxiv.2310.06825.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems 36*, 2023. doi: 10.52202/075280-1797.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022. doi: 10.18653/v1/2022.acl-long.229.

Zachary C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 2018. doi: 10.1145/3233231.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *First Conference on Language Modeling*, 2024. doi: 10.48550/arxiv.2310.06824.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.173.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems 35*, 2022. doi: 10.52202/068431-1262.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.741.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models. *arXiv (Cornell University)*, 2025. doi: 10.48550/arxiv.2505.18235.

nostalgebraist. Interpreting gpt: the logit lens. LessWrong, https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *Proceedings of the 41st International Conference on Machine Learning*, 2024. doi: 10.48550/arxiv.2311.03658.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. *International Conference on Learning Representations*, 2025. doi: 10.48550/arxiv.2406.01506.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. Bbq: A hand-built bias benchmark for question answering. *Findings of the Association for Computational Linguistics: ACL 2022*, 2022. doi: 10.18653/v1/2022.findings-acl.165.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Karl Pearson. Vii. note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 1895. doi: 10.1098/rspl.1895.0041.

Qwen Team, Yang An, Baosong Yang, Beichen Zhang, Binyuan Hui, B. Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huancai 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, Su Yang, Yichang Zhang, Yuqin Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv (Cornell University)*, 2024. doi: 10.48550/arxiv.2412.15115.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2016. doi: 10.18653/v1/n16-3020.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. doi: 10.18653/v1/2024.acl-long.828.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, 2011.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 1904. doi: 10.2307/1412159.

G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

Gemma Team, Morgane Rivière, Shreya P Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mésnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv (Cornell University)*, 2024. doi: 10.48550/arxiv.2408.00118.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv (Cornell University)*, 2023. doi: 10.48550/arxiv.2302.13971.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv (Cornell University)*, 2023. doi: 10.48550/arxiv.2308.10248.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems 30*, 2017.

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*, 2025. doi: 10.1162/coli.a.16.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv (Cornell University)*, 2023. doi: 10.48550/arxiv.2310.01405.

A Proofs

Throughout, $\|\cdot\|$ is the Euclidean norm on vectors and the spectral norm on matrices, $\|\cdot\|_F$ the Frobenius norm, and $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots$ the singular values of $\mathbf{A}$. We use the following standard facts.

(F1) *Weyl's inequality for singular values* (Horn & Johnson, 1991, Thm. 3.3.16): $|\sigma_j(\mathbf{A}+\mathbf{B}) - \sigma_j(\mathbf{A})| \leq \|\mathbf{B}\|$ for all j.

(F2) *Weyl's inequality for eigenvalues of symmetric matrices*: $|\lambda_j(\mathbf{S}+\mathbf{E}) - \lambda_j(\mathbf{S})| \leq \|\mathbf{E}\|$.

(F3) *Davis–Kahan* $\sin\Theta$ *theorem* (Davis & Kahan, 1970): if $\mathbf{S}, \hat{\mathbf{S}}$ are symmetric, $\mathcal{U}$ and $\hat{\mathcal{U}}$ their top-k eigenspaces, and $\lambda_k(\mathbf{S}) - \lambda_{k+1}(\mathbf{S}) =: \Delta > 0$, then $\|\sin\Theta(\mathcal{U}, \hat{\mathcal{U}})\| \leq 2\|\hat{\mathbf{S}} - \mathbf{S}\|/\Delta$.

(F4) *Projector distance* (Stewart & Sun, 1990, Thm. I.5.5): for subspaces $\mathcal{U}, \mathcal{V}$ of equal dimension with orthogonal projectors $\mathbf{P}_\mathcal{U}, \mathbf{P}_\mathcal{V}$, $\|\mathbf{P}_\mathcal{U} - \mathbf{P}_\mathcal{V}\| = \sin\theta_{\max}(\mathcal{U}, \mathcal{V}) = \max_{\mathbf{u}\in\mathcal{U}, \|\mathbf{u}\|=1} \|(\mathbf{I} - \mathbf{P}_\mathcal{V})\mathbf{u}\|$.

(F5) *Derivative of a simple singular value* (Stewart & Sun, 1990, Thm. IV.2.3 and §IV.2.2): if $t \mapsto \mathbf{A}(t)$ is C^1 and $\sigma_j(\mathbf{A}(t_0)) > 0$ is simple, then σ_j is differentiable at t_0 with $\dot{\sigma}_j = \mathbf{u}_j^\top \dot{\mathbf{A}} \mathbf{v}_j$, where $(\mathbf{u}_j, \mathbf{v}_j)$ is the corresponding pair of unit left/right singular vectors.

(F6) *Grönwall's lemma (integral form)*: if y is absolutely continuous on $[a, b]$ and $\dot{y} \leq -\gamma y + c$ a.e. with $\gamma > 0$, $c \geq 0$, then $y(t) \leq e^{-\gamma(t-a)} y(a) + \frac{c}{\gamma}\big(1 - e^{-\gamma(t-a)}\big)$.

(F7) *Chord and sine bounds*: for unit vectors a, a' at angle θ, $\|a' - a\| = 2\sin(\theta/2) \leq \theta$; for nonzero x, y, $\|x\| \sin\angle(x, y) = \operatorname{dist}(x, \operatorname{span} y) \leq \|x - y\|$.

A.1 The flow operator (Equation (2))

Proposition A.1 (Flow operator)**.** *Let* $\mathbf{\Phi}(t, \tau)$ *solve* $\frac{d}{dt}\mathbf{\Phi}(t, \tau) = \mathbf{J}(\mathbf{z}(t), t)\mathbf{\Phi}(t, \tau)$, $\mathbf{\Phi}(\tau, \tau) = \mathbf{I}$. *Then* $\boldsymbol{\delta}(t) = \mathbf{\Phi}(t, 0)\boldsymbol{\delta}(0) + O(\|\boldsymbol{\delta}(0)\|^2)$, *and* $\det \mathbf{\Phi}(t, 0) = \exp \int_0^t \operatorname{Tr} \mathbf{J}\, d\tau > 0$*; in particular, in the continuous model equation 1,* $\mathbf{\Phi}(t, 0)$ *is invertible for every* t.

Standing assumptions. $\mathbf{f}(\cdot, t)$ is C^1 on a neighborhood of the trajectory $\mathbf{z}(t; X)$, $t \in [0, 1]$; there are constants $M_J, L_J < \infty$ with $\|\mathbf{J}(\mathbf{z}, t)\| \leq M_J$ and $\|\mathbf{J}(\mathbf{z}, t) - \mathbf{J}(\mathbf{z}', t)\| \leq L_J\|\mathbf{z} - \mathbf{z}'\|$ for all $\mathbf{z}, \mathbf{z}'$ in that neighborhood and all t.

Existence, invertibility, determinant. The linear ODE $\dot{\mathbf{\Phi}} = \mathbf{J}(\mathbf{z}(t), t)\mathbf{\Phi}$ with bounded measurable coefficient has a unique absolutely continuous solution with $\mathbf{\Phi}(\tau, \tau) = \mathbf{I}$. Jacobi's formula gives

$\frac{d}{dt}\det\mathbf{\Phi}(t,\tau) = \operatorname{Tr}\big(\mathbf{J}(\mathbf{z}(t),t)\big)\det\mathbf{\Phi}(t,\tau)$, whence $\det\mathbf{\Phi}(t,0) = \exp\int_0^t \operatorname{Tr}\mathbf{J}(\mathbf{z}(s),s)\,ds > 0$. Moreover $\|\mathbf{\Phi}(t,\tau)\| \le e^{M_J(t-\tau)}$ (Grönwall applied to $\|\mathbf{\Phi}(t,\tau)\mathbf{x}\|$).

First-order expansion. Set $\boldsymbol{\delta}(t) = \mathbf{z}(t;X_{\Delta x}) - \mathbf{z}(t;X)$. By the mean-value form of Taylor's theorem, $\dot{\boldsymbol{\delta}} = \mathbf{J}(\mathbf{z}(t),t)\boldsymbol{\delta} + \mathcal{R}(t)$ with $\|\mathcal{R}(t)\| \le \frac{L_J}{2}\|\boldsymbol{\delta}(t)\|^2$ as long as the segment between $\mathbf{z}(t;X)$ and $\mathbf{z}(t;X_{\Delta x})$ lies in the neighborhood above. Variation of constants gives

$$\boldsymbol{\delta}(t) = \mathbf{\Phi}(t,0)\boldsymbol{\delta}(0) + \mathbf{r}(t), \qquad \mathbf{r}(t) := \int_0^t \mathbf{\Phi}(t,s)\mathcal{R}(s)\,ds. \tag{10}$$

Let $\epsilon := \|\boldsymbol{\delta}(0)\|$ and $T^* := \sup\{t \le 1 : \|\boldsymbol{\delta}(s)\| \le 2e^{M_J}\epsilon\ \forall s \le t\}$. On $[0,T^*]$, $\|\mathbf{r}(t)\| \le \int_0^t e^{M_J(t-s)}\frac{L_J}{2}(2e^{M_J}\epsilon)^2 ds \le 2L_J e^{3M_J}\epsilon^2$, so $\|\boldsymbol{\delta}(t)\| \le e^{M_J}\epsilon + 2L_J e^{3M_J}\epsilon^2 < 2e^{M_J}\epsilon$ whenever $\epsilon < \epsilon_0 := e^{-2M_J}/(2L_J)$. By continuity $T^* = 1$, and Equation (10) holds on $[0,1]$ with $\|\mathbf{r}(t)\| \le C_0\|\boldsymbol{\delta}(0)\|^2$, $C_0 := 2L_J e^{3M_J}$. □

In the discrete network the operator is the product $\mathbf{\Phi}(\ell,0) = \prod_{m<\ell}(\mathbf{I} + \frac{1}{L}\partial\mathbf{f}_m)$, which need not be invertible; none of the results below uses invertibility.

A.2 The entropy identity (Equation (4))

Proposition A.2. *On any interval where the singular values of $\mathbf{\Phi}(t,0)$ are simple and positive, with $\mu_i(t) := \mathbf{u}_i(t)^\top\mathbf{J}(\mathbf{z}(t),t)\mathbf{u}_i(t)$ and $\bar{\mu} := \sum_i p_i\mu_i$, $\frac{d}{dt}H(t) = -\sum_i p_i(\mu_i - \bar{\mu})\ln p_i = -\operatorname{Cov}_p(\mu, \ln p)$.*

Fix an interval I on which the singular values of $\mathbf{\Phi}(t) := \mathbf{\Phi}(t,0)$ are simple and positive (positivity holds always in the continuous model by Proposition A.1). By (F5) with $\dot{\mathbf{\Phi}} = \mathbf{J}\mathbf{\Phi}$ and $\mathbf{\Phi}\mathbf{v}_i = \sigma_i\mathbf{u}_i$,

$$\dot{\sigma}_i = \mathbf{u}_i^\top\mathbf{J}\mathbf{\Phi}\mathbf{v}_i = \sigma_i\,\mathbf{u}_i^\top\mathbf{J}\mathbf{u}_i = \sigma_i\mu_i. \tag{11}$$

Let $S := \sum_j \sigma_j$ and $p_i = \sigma_i/S$. Then $\dot{S} = \sum_j \sigma_j\mu_j = S\bar{\mu}$ and $\dot{p}_i = (\dot{\sigma}_i S - \sigma_i\dot{S})/S^2 = p_i(\mu_i - \bar{\mu})$. Since $p_i > 0$ on I, $H = -\sum_i p_i\ln p_i$ is differentiable and

$$\dot{H} = -\sum_i \dot{p}_i(\ln p_i + 1) = -\sum_i p_i(\mu_i - \bar{\mu})\ln p_i - \sum_i p_i(\mu_i - \bar{\mu}). \tag{12}$$

The last sum equals $\sum_i p_i\mu_i - \bar{\mu}\sum_i p_i = 0$, which gives the identity. □

A.3 Initial isotropy and second-order contraction (Equation (5))

Theorem A.3. *$H(0) = \ln Nd$, and if $t \mapsto \mathbf{J}(\mathbf{z}(t),t)$ is C^1 near 0, $H(t) = \ln Nd - \frac{t^2}{2}\operatorname{Var}(\lambda(\operatorname{sym}\mathbf{J}(0))) + O(t^3)$. Hence H strictly decreases at 0^+ unless $\operatorname{sym}\mathbf{J}(0)$ is a multiple of the identity.*

Assumption. $t \mapsto \mathbf{J}(\mathbf{z}(t),t)$ is C^1 on $[0,t_0]$ for some $t_0 > 0$, with $\|\mathbf{J}\| \le M_J$ and $\|\frac{d}{dt}\mathbf{J}\| \le M_J'$ there. Write $n := Nd$, $\mathbf{J}_0 := \mathbf{J}(\mathbf{z}(0),0)$, $\mathbf{S} := \operatorname{sym}\mathbf{J}_0$ with eigenvalues $\lambda_1^s \ge \cdots \ge \lambda_n^s$, $\bar{\lambda} := \frac{1}{n}\sum_j \lambda_j^s$ and $V := \frac{1}{n}\sum_j(\lambda_j^s - \bar{\lambda})^2 = \operatorname{Var}(\lambda(\mathbf{S}))$.

Step 1 (expansion of the Gram matrix). Since $\dot{\mathbf{\Phi}} = \mathbf{J}\mathbf{\Phi}$ and $\ddot{\mathbf{\Phi}} = \dot{\mathbf{J}}\mathbf{\Phi} + \mathbf{J}^2\mathbf{\Phi}$, Taylor's theorem with integral remainder gives $\mathbf{\Phi}(t) = \mathbf{I} + t\mathbf{J}_0 + t^2\mathbf{K}(t)$ with $\|\mathbf{K}(t)\| \le K_0 := \frac{1}{2}(M_J' + M_J^2)e^{M_J}$ for $t \le t_0$. Hence $\mathbf{G}(t) := \mathbf{\Phi}(t)^\top\mathbf{\Phi}(t) = \mathbf{I} + 2t\mathbf{S} + t^2\mathbf{E}(t)$ with $\|\mathbf{E}(t)\| \le E_0 := M_J^2 + 2K_0 + 2M_J K_0 t_0 + K_0^2 t_0^2$.

Step 2 (singular values). By (F2) applied to $\mathbf{I}+2t\mathbf{S}$ (eigenvalues $1+2t\lambda_i^s$) and the perturbation $t^2\mathbf{E}(t)$, $\lambda_i(\mathbf{G}(t)) = 1 + 2t\lambda_i^s + t^2 e_i(t)$ with $|e_i(t)| \le E_0$. For $t \le t_1 := \min\{t_0,\, 1/(4M_J + 2\sqrt{E_0})\}$ we have $\lambda_i(\mathbf{G}(t)) \ge \frac{1}{2}$, and $\sigma_i(t) = \sqrt{\lambda_i(\mathbf{G}(t))} = 1 + t\lambda_i^s + t^2 c_i(t)$ with $|c_i(t)| \le C_1 := E_0 + 2M_J^2$ (Taylor expansion of $\sqrt{1+x}$ with $|x| \le \frac{1}{2}$).

Step 3 (probabilities). Let $S(t) = \sum_j \sigma_j(t) = n\big(1 + t\bar{\lambda} + t^2\bar{c}(t)\big)$ with $|\bar{c}| \le C_1$, and write $p_i(t) = \frac{1}{n}(1 + x_i(t))$. Expanding the quotient for $t \le t_2 := \min\{t_1, 1/(4M_J + 4C_1)\}$,

$$x_i(t) = t(\lambda_i^s - \bar{\lambda}) + t^2 d_i(t), \qquad |d_i(t)| \le C_2 := 2C_1 + 2M_J^2 + 4M_J^2, \tag{13}$$

and, because $\sum_i p_i = 1$ exactly, $\sum_i x_i(t) = 0$ for every t, hence $\sum_i d_i(t) = 0$ as well.

Step 4 (entropy). With $\phi(x) := (1+x)\ln(1+x) = x + \frac{x^2}{2} + \rho(x)$, $|\rho(x)| \le |x|^3$ for $|x| \le \frac{1}{2}$,

$$H(t) = \ln n - \frac{1}{n}\sum_i \phi(x_i(t)) = \ln n - \frac{1}{n}\sum_i x_i - \frac{1}{2n}\sum_i x_i^2 - \frac{1}{n}\sum_i \rho(x_i). \tag{14}$$

The first sum vanishes identically. The second equals $t^2 V + t^3 \cdot \frac{2}{n}\sum_i(\lambda_i^s - \bar{\lambda})d_i + t^4 \frac{1}{n}\sum d_i^2$, whose non-leading part is bounded by $t^3(4M_J C_2 + C_2^2 t)$. The third is bounded by $t^3(2M_J + C_2 t)^3$. Therefore, for $t \le t_2$,

$$\left| H(t) - \ln n + \frac{t^2}{2} V \right| \le C_3 t^3, \qquad C_3 := 2M_J C_2 + C_2^2 + (2M_J + C_2)^3, \tag{15}$$

which is the claim with an explicit remainder. If $\mathbf{S} \neq c\mathbf{I}$ then $V > 0$ and $H(t) < \ln n$ for $0 < t < \min\{t_2, V/(2C_3)\}$. □

A.4 Shallow orbits

Proposition A.4. *If the k-th and $(k{+}1)$-th eigenvalues of* $\operatorname{sym} \mathbf{J}_X^{(N,N)}(0)$ *are distinct, then* $\mathbf{P}_{t,X} \to \mathbf{P}_{\text{top-}k}\big(\operatorname{sym} \mathbf{J}_X^{(N,N)}(0)\big)$ *as* $t \to 0^+$.

Let $\mathbf{S}_N := \operatorname{sym} \mathbf{J}_X^{(N,N)}(0)$ with gap $\Delta := \lambda_k(\mathbf{S}_N) - \lambda_{k+1}(\mathbf{S}_N) > 0$. As in Step 1 of Section A.3, $\mathbf{\Phi}_X^N(t) = \mathbf{B}_N + t\mathbf{J}^{(N,\cdot)}(0) + t^2 \mathbf{K}_N(t)$, where $\mathbf{B}_N = [\,0 \;\cdots\; \mathbf{I}_d \;\cdots\; 0\,]$ selects the read-out block and $\|\mathbf{K}_N\| \le K_0$. Hence $\mathbf{G}_N(t) := \mathbf{\Phi}_X^N(t)\mathbf{\Phi}_X^N(t)^\top = \mathbf{I}_d + 2t\mathbf{S}_N + t^2 \mathbf{E}_N(t)$ with $\|\mathbf{E}_N\| \le E_0$. The top-k left singular subspace of $\mathbf{\Phi}_X^N(t)$ is the top-k eigenspace of $\mathbf{G}_N(t)$, and the top-k eigenspace of $\mathbf{I}_d + 2t\mathbf{S}_N$ is that of $\mathbf{S}_N$, with gap $2t\Delta$. By (F3),

$$\big\| \sin\Theta\big(\mathcal{O}_{t,X}^{(k,N)}, \mathcal{U}_k(\mathbf{S}_N)\big) \big\| \le \frac{2\,t^2 E_0}{2t\Delta} = \frac{E_0}{\Delta}\, t \xrightarrow[t\to 0^+]{} 0, \tag{16}$$

and by (F4) $\|\mathbf{P}_{t,X} - \mathbf{P}_{\text{top-}k}(\mathbf{S}_N)\| \le E_0 t/\Delta$. □

In a network with discrete blocks, "$t \to 0^+$" means the first block. For the read-out token the relevant block of $\mathbf{J}$ at the first layer is the attention map from the edited token to the read-out token composed with the value–output circuit. It is set by the whole input: inputs in different contexts have different shallow orbits, and inputs that share a context and differ in one token share the shallow orbit that the context sets (Section F).

A.5 Stretch–turn decomposition (Equation (8))

Wherever $\boldsymbol{\delta} \neq 0$: $\frac{d}{dt}\|\boldsymbol{\delta}\|^2 = 2\boldsymbol{\delta}^\top \dot{\boldsymbol{\delta}} = 2\boldsymbol{\delta}^\top \mathbf{J}\boldsymbol{\delta}$ (first order, cf. Proposition A.1), hence $\frac{d}{dt}\ln\|\boldsymbol{\delta}\| = \mathbf{u}^\top \mathbf{J}\mathbf{u} = \mu$. Differentiating $\mathbf{u} = \boldsymbol{\delta}/\|\boldsymbol{\delta}\|$, $\dot{\mathbf{u}} = \dot{\boldsymbol{\delta}}/\|\boldsymbol{\delta}\| - \boldsymbol{\delta}\,\frac{d}{dt}\ln\|\boldsymbol{\delta}\|/\|\boldsymbol{\delta}\| = \mathbf{J}\mathbf{u} - \mu\mathbf{u} = (\mathbf{I} - \mathbf{u}\mathbf{u}^\top)\mathbf{J}\mathbf{u}$, and the decomposition $\mathbf{J}\mathbf{u} = \mu\mathbf{u} + (\mathbf{I} - \mathbf{u}\mathbf{u}^\top)\mathbf{J}\mathbf{u}$ is the orthogonal split of $\mathbf{J}\mathbf{u}$ along and across $\mathbf{u}$. Since $\mathbf{I} - \mathbf{u}\mathbf{u}^\top$ is an orthogonal projector, $\omega = \|\dot{\mathbf{u}}\| \le \|\mathbf{J}\mathbf{u}\| \le \|\mathbf{J}\|$. Finally, for $0 \le t_0 \le t_1 \le 1$, $\|\mathbf{u}(t_1) - \mathbf{u}(t_0)\| = \|\int_{t_0}^{t_1} \dot{\mathbf{u}}\, dt\| \le \int_{t_0}^{t_1} \omega\, dt$ (Minkowski's integral inequality). □

A.6 The gain-ordered basis

Lemma A.5. *With $W_V = U\Sigma R^\top$ and $C_\ell(X)$ the Concept Path of Definition C.1, $C_\ell(X) = v_\ell(X)\,R$ where $v_\ell(X) = \tilde{z}W_V$ is the last-token value vector; hence* $\cos(C_\ell(X), C_\ell(X')) = \cos(v_\ell(X), v_\ell(X'))$ *for every pair of inputs.*

With $U^\top U = \mathbf{I}_{d_v}$ and $R^\top R = \mathbf{I}_{d_v}$, the value vector is $v = \tilde{z}W_V = (\tilde{z}U)\Sigma R^\top$, and the coordinates of C are $C_i = \langle \tilde{z}, u_i\rangle \sigma_i = (\tilde{z}U\Sigma)_i$, i.e. $C = \tilde{z}U\Sigma = vR$. Since R is orthogonal, $\langle C, C'\rangle = \langle v, v'\rangle$ and $\|C\| = \|v\|$. □

The basis therefore supplies an ordering of coordinates by the gain of W_V, and the truncation to the top-k coordinates in this ordering sets the scale of the spectral score (Section H.2). The truncated path of Equation (26), $\tilde{C}_\ell = \sum_{i\in S_k} C_i u_i = U\,\mathrm{top}_k(C)$, is the same object written as a vector of the residual

space $\mathbb{R}^d$: since U has orthonormal columns, $\langle \tilde{C}_\ell(X), \tilde{C}_\ell(X')\rangle = \langle \mathrm{top}_k(C(X)), \mathrm{top}_k(C(X'))\rangle$, so the score is unchanged, while paths of different layers now live in one space and their rotation θ_ℓ is well defined.

A.7 Cosine–rotation inequality

Proposition A.6. *Let $a_\ell, b_\ell \in \mathbb{R}^d$ be the unit vectors of the residual-space Concept Paths $\tilde{C}_\ell(X), \tilde{C}_\ell(X_{\Delta x})$ of Equation (26), $s_\ell = \langle a_\ell, b_\ell\rangle$, and $\theta_\ell(X) = \angle(a_\ell, a_{\ell+1})$ the rotation of the path of X between layers ℓ and $\ell+1$. Then $|s_{\ell+1} - s_\ell| \le 2\sin\big(\theta_\ell(X)/2\big) + 2\sin\big(\theta_\ell(X_{\Delta x})/2\big) \le \theta_\ell(X) + \theta_\ell(X_{\Delta x})$, and $s_{\ell_0} - s_{\ell_1} \le \sum_{\ell=\ell_0}^{\ell_1-1} \big(\theta_\ell(X) + \theta_\ell(X_{\Delta x})\big)$.*

We have $s_{\ell+1} - s_\ell = \langle a_{\ell+1} - a_\ell,\, b_\ell\rangle + \langle a_{\ell+1},\, b_{\ell+1} - b_\ell\rangle$. By Cauchy–Schwarz, $|\langle a_{\ell+1} - a_\ell, b_\ell\rangle| \le \|a_{\ell+1} - a_\ell\| = 2\sin(\theta_\ell(X)/2) \le \theta_\ell(X)$ by (F7); the same holds for the b-term. Summing over $\ell = \ell_0, \dots, \ell_1 - 1$ and using $s_{\ell_0} - s_{\ell_1} \le \sum_\ell |s_{\ell+1} - s_\ell|$ gives the telescoped bound. The paths of layers ℓ and $\ell+1$ are built from different bases $U_\ell, U_{\ell+1}$ and different index sets, but both are vectors of $\mathbb{R}^d$, which is all the argument uses. □

The inequality is a property of the spectral lens: a drop of the score certifies that at least one of the two paths has rotated. The link between the lens and the flow operator is Equation (3).

A.8 Stability of the branching direction

Proposition A.7. *Let $\hat{d}_\ell := b_\ell - a_\ell$ and $\Theta_{[\ell^*,\ell)} := \sum_{j=\ell^*}^{\ell-1}(\theta_j(X) + \theta_j(X_{\Delta x}))$. If $s_j < \tau$ for all $j \in [\ell^*, \ell]$, then $\sin\angle(\hat{d}_\ell, \hat{d}_{\ell^*}) \le \Theta_{[\ell^*,\ell)}/\sqrt{2(1-\tau)}$.*

For $j \ge \ell^*$, $\|\hat{d}_{j+1} - \hat{d}_j\| \le \|a_{j+1} - a_j\| + \|b_{j+1} - b_j\| \le \theta_j(X) + \theta_j(X_{\Delta x})$ by (F7). Summing, $\|\hat{d}_\ell - \hat{d}_{\ell^*}\| \le \Theta_{[\ell^*,\ell)}$. By (F7) with $x = \hat{d}_\ell$, $y = \hat{d}_{\ell^*}$, $\sin\angle(\hat{d}_\ell, \hat{d}_{\ell^*}) \le \|\hat{d}_\ell - \hat{d}_{\ell^*}\|/\|\hat{d}_\ell\|$, and $\|\hat{d}_\ell\|^2 = 2(1 - s_\ell) > 2(1-\tau)$ because $s_\ell < \tau$. □

After branching, the direction of the counterfactual difference can drift from its value at ℓ^* only by the turning spent since. When that budget is small, one direction read at ℓ^* describes the difference at every later layer; Figure 18(2) measures the budget on the residual stream.

B Proof of Theorem 4.3 and the Steering Operator

B.1 Proof of Lemma 4.2

Fix an input X and write $M(t) := \mathbf{P}_\parallel \mathbf{\Phi}_X^N(t) \in \mathbb{R}^{d \times Nd}$. From $\frac{d}{dt}\mathbf{\Phi}^{(N\leftarrow\cdot)} = \sum_p \mathbf{J}^{(N,p)}\mathbf{\Phi}^{(p\leftarrow\cdot)}$,

$$\dot{M} = \underbrace{\mathbf{P}_\parallel \mathbf{J}^{(N,N)} \mathbf{P}_\parallel}_{=:A} M + E, \qquad E := \mathbf{P}_\parallel \mathbf{J}^{(N,N)} \mathbf{P}_\perp \mathbf{\Phi}_X^N + \sum_{p \ne N} \mathbf{P}_\parallel \mathbf{J}^{(N,p)} \mathbf{\Phi}^{(p\leftarrow\cdot)},$$

and $\|E\| \le \eta_\parallel \Lambda$ by (A2) and (A4). Fix a unit $\mathbf{v} \in \mathbb{R}^{Nd}$ and let $\mathbf{y}(t) = M(t)\mathbf{v} \in \mathcal{O}_d^*$. Wherever $\mathbf{y} \ne 0$,

$$\tfrac{d}{dt}\|\mathbf{y}\| = \frac{\mathbf{y}^\top(A\mathbf{y} + E\mathbf{v})}{\|\mathbf{y}\|} \;\ge\; \gamma_\parallel \|\mathbf{y}\| - \eta_\parallel \Lambda,$$

because $\mathbf{y}^\top A\mathbf{y} = \mathbf{y}^\top \mathrm{sym}(\mathbf{J}^{(N,N)})\mathbf{y} \ge \gamma_\parallel \|\mathbf{y}\|^2$ for $\mathbf{y} \in \mathcal{O}_d^*$. The comparison lemma gives $\|\mathbf{y}(t)\| \ge e^{\gamma_\parallel (t-t^*)}\|\mathbf{y}(t^*)\| - \frac{\eta_\parallel \Lambda}{\gamma_\parallel}(e^{\gamma_\parallel(t-t^*)} - 1) \ge e^{\gamma_\parallel(t-t^*)}\big(\|\mathbf{y}(t^*)\| - \eta_\parallel\Lambda/\gamma_\parallel\big)$. Let V be the span of the top-k right singular vectors of $M(t^*)$; for unit $\mathbf{v} \in V$, $\|M(t^*)\mathbf{v}\| \ge \sigma_k(M(t^*))$. By the Courant–Fischer characterization, $\sigma_k(M(t)) \ge \min_{\mathbf{v}\in V, \|\mathbf{v}\|=1} \|M(t)\mathbf{v}\| \ge e^{\gamma_\parallel(t-t^*)}\underline{\sigma}_X$. □

B.2 Proof of Theorem 4.3

Fix an input X and suppress it. For $t \in [t^*, 1]$ write $\mathbf{\Phi}^{(p)}(t)$ for the p-th block row of $\mathbf{\Phi}(t, 0)$, so that $\mathbf{\Phi}^N = \mathbf{\Phi}^{(N)}$, and set $\mathbf{Y}(t) := \mathbf{P}_\perp \mathbf{\Phi}^N(t)$, $\mathbf{W}(t) := \mathbf{P}_\parallel \mathbf{\Phi}^N(t)$.

Step 1 (evolution of the off-target block). Differentiating $\dot{\mathbf{\Phi}} = \mathbf{J}\mathbf{\Phi}$ block-row-wise, $\dot{\mathbf{\Phi}}^{(N)} = \sum_{p=1}^{N} \mathbf{J}^{(N,p)}\mathbf{\Phi}^{(p)}$. Multiplying by $\mathbf{P}_\perp$ and inserting $\mathbf{\Phi}^{(N)} = \mathbf{Y} + \mathbf{W}$,

$$\dot{\mathbf{Y}} = \underbrace{\mathbf{P}_\perp \mathbf{J}^{(N,N)} \mathbf{P}_\perp}_{=:\mathbf{A}(t)} \mathbf{Y} + \underbrace{\mathbf{P}_\perp \mathbf{J}^{(N,N)} \mathbf{P}_\parallel \mathbf{W} + \sum_{p \neq N} \mathbf{P}_\perp \mathbf{J}^{(N,p)} \mathbf{\Phi}^{(p)}}_{=:\mathbf{F}(t)}, \tag{17}$$

and by (A3), (A4) and $\|\mathbf{\Phi}^{(p)}\| \le \|\mathbf{\Phi}(t,0)\| \le \Lambda$, $\|\mathbf{F}(t)\| \le \eta\Lambda$.

Step 2 (growth of the off-target block). Fix a unit vector $\mathbf{x}$ and let $y(t) := \|\mathbf{Y}(t)\mathbf{x}\|$. Wherever $y > 0$, $\frac{1}{2}\frac{d}{dt}y^2 = (\mathbf{Y}\mathbf{x})^\top \mathbf{A}(\mathbf{Y}\mathbf{x}) + (\mathbf{Y}\mathbf{x})^\top \mathbf{F}\mathbf{x} \le \gamma_\perp y^2 + \eta\Lambda\, y$ by (A1) and Cauchy–Schwarz, so $\dot{y} \le \gamma_\perp y + \eta\Lambda$ a.e. The comparison lemma for this linear inequality gives

$$y(t) \le e^{\gamma_\perp (t-t^*)} y(t^*) + \eta\Lambda \int_{t^*}^{t} e^{\gamma_\perp (t-s)} ds \le e^{\gamma_\perp (t-t^*)} \|\mathbf{Y}(t^*)\| + \eta\Lambda\,(t-t^*)\, e^{\gamma_\perp^+ (t-t^*)}, \tag{18}$$

with $\gamma_\perp^+ := \max(\gamma_\perp, 0)$; taking the supremum over $\mathbf{x}$ bounds $\|\mathbf{Y}(t)\|$.

Step 3 (the on-target part outgrows it). By Lemma 4.2, $\sigma_k(\mathbf{W}(t)) \ge e^{\gamma_\parallel (t-t^*)} \underline{\sigma}_X =: w(t)$. Dividing Step 2 by $w(t)$,

$$r(t) := \frac{\|\mathbf{Y}(t)\|}{\sigma_k(\mathbf{W}(t))} \le \frac{\|\mathbf{Y}(t)\|}{w(t)} \le \rho_X\, e^{-(\gamma_\parallel - \gamma_\perp)(t-t^*)} + \frac{\eta\Lambda\,(t-t^*)\, e^{(\gamma_\perp^+ - \gamma_\parallel)(t-t^*)}}{\underline{\sigma}_X} = \bar{r}_X(t). \tag{19}$$

By (F1), $\sigma_k(\mathbf{\Phi}^N(t)) \ge \sigma_k(\mathbf{W}(t)) - \|\mathbf{Y}(t)\| = \sigma_k(\mathbf{W}(t))(1 - r(t))$.

Step 4 (angle between the orbit and $\mathcal{O}_d^*$). Let $\mathbf{u}$ be a unit vector in the top-k left singular subspace of $\mathbf{\Phi}^N(t)$. Then $\mathbf{u} = \mathbf{\Phi}^N(t)\mathbf{v}$ for some $\mathbf{v}$ with $\|\mathbf{v}\| \le 1/\sigma_k(\mathbf{\Phi}^N(t))$, and $\|\mathbf{P}_\perp \mathbf{u}\| = \|\mathbf{Y}(t)\mathbf{v}\| \le \|\mathbf{Y}(t)\| / \big(\sigma_k(\mathbf{W}(t))(1 - r(t))\big) = r/(1-r) \le 2\bar{r}_X(t)$ whenever $\bar{r}_X(t) \le \frac{1}{2}$. Since $\dim \mathcal{O}_{t,X}^{(k,N)} = \dim \mathcal{O}_d^* = k$, (F4) gives $\|\mathbf{P}_{t,X} - \mathbf{P}_\parallel\| = \max_{\mathbf{u}} \|\mathbf{P}_\perp \mathbf{u}\| \le 2\bar{r}_X(t)$.

Step 5 (alignment). For two rank-k orthogonal projectors, $\|\mathbf{P}_a - \mathbf{P}_b\|_F^2 = 2k - 2\,\mathrm{Tr}(\mathbf{P}_a\mathbf{P}_b)$ and $\mathrm{rank}(\mathbf{P}_a - \mathbf{P}_b) \le 2k$, so $1 - \mathrm{Tr}(\mathbf{P}_a\mathbf{P}_b)/k = \|\mathbf{P}_a - \mathbf{P}_b\|_F^2/(2k) \le \|\mathbf{P}_a - \mathbf{P}_b\|^2 \le (\|\mathbf{P}_a - \mathbf{P}_\parallel\| + \|\mathbf{P}_b - \mathbf{P}_\parallel\|)^2 \le 16 \max_X \bar{r}_X(1)^2$. Averaging over ordered pairs gives the bound on $1 - A(1)$. □

The case $\gamma_\perp < 0$ is the one drawn in Figure 1(a), where off-target components decay outright; the proof is the same.

B.3 The theorem at the depth of trained models

The quantitative conclusion of Theorem 4.3 requires $\bar{r}_X \le \frac{1}{2}$. On all six models the off/on ratio of the actual counterfactual at the read-out token is still 3.4–6.9 at the output (Table 3), and the operator ratio r_X, a worst case over directions, is at least as large. The measurements therefore test the premise of the theorem, a growth gap with onset t^*, and the direction of motion it implies, a decaying off/on ratio, which is how the main text uses it.

B.4 The steering operator

Let the read-out position evolve under

$$\frac{d}{dt}\hat{z}^{(N)}(t) = f^{(N)}(\hat{\mathbf{z}}(t), t) - \beta\, \mathbf{P}_\perp \hat{z}^{(N)}(t), \qquad t \ge t^*,\ \beta > 0. \tag{20}$$

Theorem B.1 (Off-target damping)**.** *Under Assumption 4.1 evaluated along the steered trajectory, the off-target counterfactual $\boldsymbol{\delta}_\perp := \mathbf{P}_\perp \delta^{(N)}$ obeys*

$$\|\boldsymbol{\delta}_\perp(1)\| \le e^{(\gamma_\perp - \beta)(1-t^*)} \|\boldsymbol{\delta}_\perp(t^*)\| + \eta\,(1-t^*)\, e^{(\gamma_\perp - \beta)^+ (1-t^*)} \sup_{\tau \in [t^*, 1]} \|\boldsymbol{\delta}(\tau)\|. \tag{21}$$

Proof. Linearizing Equation (20) along the steered factual trajectory, $\dot{\boldsymbol{\delta}}_\perp = \mathbf{P}_\perp \mathbf{J}^{(N,N)} \mathbf{P}_\perp \boldsymbol{\delta}_\perp + \mathbf{P}_\perp \mathbf{J}^{(N,N)} \mathbf{P}_\parallel \boldsymbol{\delta}_\parallel + \sum_{p \neq N} \mathbf{P}_\perp \mathbf{J}^{(N,p)} \delta^{(p)} - \beta \boldsymbol{\delta}_\perp$. Taking the inner product with $\boldsymbol{\delta}_\perp$ and using (A1) for the first term, Cauchy–Schwarz and (A3) for the next two (with $\|\boldsymbol{\delta}_\parallel\|, \|\delta^{(p)}\| \le D := \sup_\tau \|\boldsymbol{\delta}(\tau)\|$), $\frac{d}{dt}\|\boldsymbol{\delta}_\perp\| \le (\gamma_\perp - \beta)\|\boldsymbol{\delta}_\perp\| + \eta D$ a.e., and the comparison lemma of Step 2 above gives the claim. □

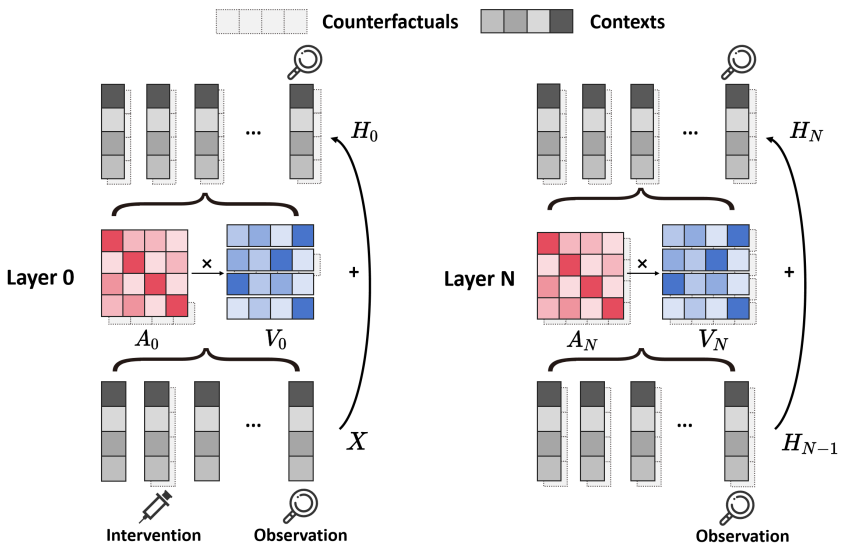


Figure 5: Overview of the MindCraft tracing procedure. X is the input sequence; at layer ℓ, H_ℓ is the attention output after the residual connection, A_ℓ the attention weights and V_ℓ the value matrix. An intervention is first applied to a specific input token. Through the attention mechanism, the counterfactual and original representations at the final token are then compared. The resulting difference reveals the layer at which the two paths first separate.

Damping adds β to the gap: the off-target share of the counterfactual is multiplied by $e^{-\beta(1-t^*)}$ relative to the undamped flow, up to the coupling floor. The bound is most useful when the intervention starts at t^*: earlier, the orbit is not yet aligned with $\mathcal{O}_d^*$ and damping removes information the model still needs; later, β has little left to do. The additive steering of Algorithm 1 is the complementary on-target pull.

C THE MINDCRAFT TRACING PROCEDURE: DETAILED CONSTRUCTION

This appendix gives the complete formulation of the procedure summarized in Section 5.1.

C.1 THE PROPAGATION OF COUNTERFACTUAL SIGNALS

Counterfactual pairs. We borrow the notion of *counterfactual* from causal inference (Pearl, 2009), defined as the resulting variable after an *intervention* is applied to a particular variable (e.g., a token), while the surrounding *context* is held constant. If we intervene on a specific component x of X and compute the counterfactual on the target variable Y, we write

$$Y_{x \leftarrow \Delta x} \;=\; Y \mid \big(do(x = \Delta x), X \setminus x\big), \tag{22}$$

where $do(\cdot)$ denotes the intervention that assigns x the value Δx, and $X \setminus x$ is the remaining context, held constant. For example, for sentiment classification of "The movie was great." the model predicts $Y =$ Positive. Replacing $x =$ "great" by $\Delta x =$ "terrible" with the other tokens unchanged yields $Y_{x \leftarrow \Delta x} =$ Negative. For brevity, $Y_{x \leftarrow \Delta x}$ is written $Y_{\Delta x}$. The difference $\delta Y = Y_{\Delta x} - Y$ is the counterfactual difference, and $(Y_{\Delta x}, Y)$ is the counterfactual pair.

Layerwise propagation. A model's conceptual understanding can be revealed through small, targeted counterfactual edits. For example, consider the intervention

X: `You are a powerful person making decisions.`
$X_{\Delta x}$: `You are a powerless person making decisions.`

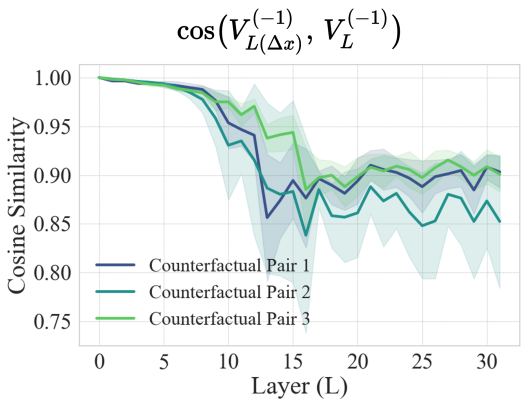


Figure 6: $\cos(V_{L(\Delta x)}^{(-1)}, V_L^{(-1)})$ across layers for three counterfactual pairs. Pair 1: "Pretend you're an honest (untruthful) person making statements about the world."; Pair 2: "Describe a fair (biased) scenario that you have seen."; Pair 3: "You are a powerful (powerless) person making decisions."

The only difference is the word change (powerful → powerless). By following how this intervention propagates through the network (Figure 5), we observe where the model begins to treat the two inputs differently. Following prior work showing that the last token best captures a model's generative state (Meng et al., 2022; Zou et al., 2023), we compare the value representations of the final token, $V_{L(\Delta x)}^{(-1)}$ and $V_L^{(-1)}$, across layers.

Figure 6 shows the cosine similarity between the factual and counterfactual value vectors at each layer for three persona prompts. In the early layers the similarity stays high: the network has not yet distinguished the two contexts at the read-out position. It then drops over a few layers and fluctuates without a further trend. This *branch-and-stabilize* pattern is what the procedure quantifies: representations remain similar in early layers, diverge at a branching point, and then stabilize. The untruncated cosine shown here varies over a narrow range; Sections C.5 and H.2 describe how the score is scaled so that pairs can be ordered.

C.2 The path-averaged flow operator and the flow score

For a swap at position i, $\mathbf{z}_s(0) = \mathbf{z}(0; X) + s\,\boldsymbol{\delta}(0)$ interpolates between the two embeddings, and $\frac{d}{ds}\mathbf{z}(\ell; \mathbf{z}_s(0)) = \boldsymbol{\Phi}(\ell, 0; \mathbf{z}_s)\boldsymbol{\delta}(0)$. Integrating over $s \in [0, 1]$ gives equation 3 exactly. The quadrature $\bar{\boldsymbol{\Phi}}_m\boldsymbol{\delta}(0) := \sum_{j=1}^{m} w_j\, \boldsymbol{\Phi}(\ell, 0; \mathbf{z}_{s_j})\boldsymbol{\delta}(0)$ with Gauss–Legendre nodes s_j and weights w_j needs m forward-mode Jacobian–vector products, each returning the tangent at every layer (float32, attention computed without fused kernels). Table 2 reports its error at the read-out token and the agreement of the resulting flow branching layers with those from the exact displacement. The $s = 0$ term alone gives the direction of the displacement only approximately (Section K); the quadrature converges on all three models, fastest on the smallest. Since equation 3 is exact, the flow score of equation 9 can equally be computed from the two forward passes, which is what we do for the large-scale validation.

The flow score standardizes each coordinate of the read-out state by its standard deviation over a reference set of read-out states at that layer (for the validation of Section I.1, all 1,536 read-out states of the prompt bank), so that a few high-magnitude coordinates, shared by both inputs, do not dominate the cosine.

C.3 Concept Path (spectral lens)

The figures of Concept Hierarchies read the displacement in a per-layer spectral basis. For a given layer, let $Z \in \mathbb{R}^{n \times d}$ denote the input token representations from the previous layer, where n is the

[0]Counterfactual Pair 1 is "Pretend you're an honest (untruthful) person making statements about the world."; Counterfactual Pair 2 is "Describe a fair (biased) scenario that you have seen."; Counterfactual Pair 3 is "You are a powerful (powerless) person making decisions."

Table 2: Path-averaged operator $\bar{\mathbf{\Phi}}$ from m JVPs. Median over pairs of the layer-averaged relative error $\|\bar{\mathbf{\Phi}}_m^{(N\leftarrow i)}\boldsymbol{\delta}(0) - \boldsymbol{\delta}_N\|/\|\boldsymbol{\delta}_N\|$ at the read-out token, and Kendall τ between the flow branching layer from the quadrature and from the exact displacement (288 pairs, 2 per concept).

	Qwen2.5-0.5B		Qwen2.5-7B		Llama-3.1-8B	
rule	error	Kendall	error	Kendall	error	Kendall
$s = 0$ (first-order $\mathbf{\Phi}$)	1.105	0.45	1.007	0.21	1.502	0.30
midpoint ($m = 1$)	0.810	0.74	1.880	0.42	1.803	0.69
Gauss–Legendre $m = 2$	0.255	0.82	0.561	0.43	0.667	0.61
Gauss–Legendre $m = 4$	0.057	0.91	0.407	0.50	0.261	0.74
Gauss–Legendre $m = 8$	0.004	0.98	0.233	0.64	0.094	0.90

sequence length and d is the model dimension. A self-attention block computes

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V, \qquad \mathrm{Attn}(Q,K,V) = \mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big)V, \tag{23}$$

with $W_Q \in \mathbb{R}^{d\times n_h d_h}$ and $W_K, W_V \in \mathbb{R}^{d\times d_v}$, $d_v = n_{\mathrm{kv}} d_h$ (for multi-head attention $n_{\mathrm{kv}} = n_h$ and $d_v = d$; for the grouped-query attention of Qwen2.5-7B, $n_{\mathrm{kv}} = 4$, $d_h = 128$, $d_v = 512$). We focus on the last token, which conditions the next-token prediction (Meng et al., 2022; Zou et al., 2023), and on the value pathway, through which token content enters the residual stream within an attention block. Let $v^{(-1)} = Z^{(-1)}W_V \in \mathbb{R}^{d_v}$ denote the last token's value vector *before* attention weighting.

Spectral basis. We express $v^{(-1)}$ in the principal basis of W_V:

$$W_V = U\Sigma R^\top, \quad U = [u_1,\dots,u_{d_v}] \in \mathbb{R}^{d\times d_v}, \quad \Sigma = \mathrm{diag}(\sigma_1,\dots,\sigma_{d_v}), \quad R \in \mathbb{R}^{d_v\times d_v}. \tag{24}$$

The left singular vectors $\{u_i\}$ are the input directions along which W_V has the largest gain, in decreasing order.

Definition C.1 (Concept Path)**.** For a self-attention layer, the **Concept Path** of the last token is

$$C \;=\; \big[\,\langle \tilde{z}, u_1\rangle\,\sigma_1,\;\dots,\;\langle \tilde{z}, u_{d_v}\rangle\,\sigma_{d_v}\,\big] \;\in\; \mathbb{R}^{d_v}, \tag{25}$$

where $\tilde{z}$ is the post-normalization last-token input to W_V, and $\{u_i\}$ are the input-side singular vectors of W_V.

By Lemma A.5, $C = v^{(-1)}R$ is the value vector in the orthonormal basis R, with coordinates ordered by the gain of W_V. The input of the lens is again $\bar{\mathbf{\Phi}}^{(N\leftarrow i)}(\ell)\boldsymbol{\delta}(0)$, after the next layer's normalization; the lens adds the basis and the truncation below.

Multi-head and grouped-query attention. The value projection is a single matrix W_V that is subsequently partitioned into per-head slices $W_V = [W_V^{(1)};\dots;W_V^{(n_{\mathrm{kv}})}]$ (Vaswani et al., 2017). The spectral analysis is performed on the full W_V *prior* to this partitioning, so the singular vectors $\{u_i\}$ describe the principal directions of the layer's value transformation as a whole.

C.4 Spectral branching layer and Concept Hierarchy

Spectral branching layer. Given X and $X_{\Delta x}$, let $C_l(X), C_l(X_{\Delta x}) \in \mathbb{R}^{d_v}$ denote their gain-ordered coordinates at layer l. We retain the top-k entries (by absolute value) of each, zeroing the rest, and map the result back to the residual space,

$$\tilde{C}_l \;=\; U\,\mathrm{top}_k(C_l) \;=\; \sum_{i\in S_k} C_{l,i}\,u_i \;\in\; \mathbb{R}^d, \tag{26}$$

where S_k is the index set of the k retained entries; since U has orthonormal columns the cosines below are the same in either form. The index sets are chosen separately for the two inputs and for each layer. The **separation score** at layer l is

$$s_l(X, X_{\Delta x}) = \cos\big(\tilde{C}_l(X),\,\tilde{C}_l(X_{\Delta x})\big), \qquad l^*(X, X_{\Delta x}) = \min\{l \in [0, L{-}1] : s_l(X, X_{\Delta x}) < \tau\}, \tag{27}$$

with $l^* = \text{INF}$ if no such layer exists. Because the index sets can differ between the two inputs, s_l responds both to the values of shared coordinates and to whether the two inputs select the same coordinates. The hierarchy figures use $k = 10$ and $\tau = 0.9$; the quantitative comparisons use the calibrated threshold of Section C.5. Since $\sin\angle(a, a + \hat{\delta}) = \|P_{a^\perp}\hat{\delta}\|/\|a + \hat{\delta}\|$, a cosine below $\tau = 0.9$ means that the counterfactual difference has acquired a component orthogonal to the factual path larger than $\sqrt{1-\tau^2} \approx 0.44$ of the path norm, equivalently that the normalized difference $\hat{b}_l - \hat{a}_l$ has norm larger than $\sqrt{2(1-\tau)} \approx 0.45$: the edit dominates the amplified state. Propositions A.6 and A.7 relate s_l to the turning of the paths, and Section D records what the two limiting profiles imply for the score.

Concept Hierarchy. For a set of counterfactual pairs sharing a common base prompt, we obtain a collection of branching layers $\{l_i^*\}$ and display the pairs ordered by l_i^* (Figure 4). The *Concept Hierarchy* is the ordered tuple $(l_1^*, \ldots, l_N^*)$, sorted ascendingly, together with the counterfactual labels. Earlier-branching pairs are those whose read-out response reaches the threshold first.

C.5 Threshold calibration

The level a cosine reaches depends on the coordinates it is computed in: the standardized read-out cosine rarely falls below 0.99, while the truncated spectral score falls much lower. A fixed τ therefore ties the absolute branching layer to the choice of lens and of k. We instead fix the fraction f of pairs that branch and choose τ by cross-fitting: τ is the f-quantile of the per-pair minimum score over the prompts with even index and is applied to the prompts with odd index, and vice versa. The main results use $f = 0.5$; $f = 0.8$ gives the same ordering of lenses. The output of the procedure is the order of the branching layers.

D Discrete Profiles of the Spectral Score

This appendix records what the impulsive and gradual profiles of turning imply for the spectral score of Section C, using only Propositions A.6 and A.7. The cosine–rotation inequality bounds turning from *below* by the drop of s_l: it certifies turning where the score drops.

Setup and notation. Fix a counterfactual pair $(X, X_{\Delta x})$. We write a_l, b_l for the unit vectors of the filtered Concept Paths, $\hat{d}_l := b_l - a_l$ for the normalized difference and $\theta_l(X), \theta_l(X_{\Delta x})$ for the layer-to-layer rotations. We assume $\tilde{C}_l(X) \neq 0$ and $\tilde{C}_l(X_{\Delta x}) \neq 0$ for all l, which holds whenever the value vector does not vanish. The total turning budget over $[l_0, l_1)$ is $\Theta_{[l_0,l_1)} := \sum_{l=l_0}^{l_1-1}(\theta_l(X) + \theta_l(X_{\Delta x}))$.

Definition D.1 (Branching profiles)**.** At the branching layer l^*, branching is **sharp** if $s_{l^*-1} - s_{l^*} \geq 1 - \tau$, i.e. the cosine crosses the threshold within a single layer transition; and **gradual** if there exists $m \geq 2$ with $s_{l^*-m} - s_{l^*} \geq 1 - \tau$ but $s_{l-1} - s_l < (1-\tau)/m$ for every $l \in \{l^* - m + 1, \ldots, l^*\}$, i.e. the same total drop is spread over m layers.

Proposition D.2 (Sharp branching certifies an impulsive turn)**.** *If branching is sharp with drop* $\Delta s = s_{l^*-1} - s_{l^*}$*, then* $\theta_{l^*-1}(X) + \theta_{l^*-1}(X_{\Delta x}) \geq \Delta s \geq 1 - \tau$.

Proof. Immediate from Proposition A.6 at $\ell = l^* - 1$. □

Proposition D.3 (The branching direction is stable up to the post-branching budget)**.** *Suppose* $s_j < \tau$ *for all* $j \in [l^*, l]$*. Then* $\sin\angle(\hat{d}_l, \hat{d}_{l^*}) \leq \Theta_{[l^*,l)}/\sqrt{2(1-\tau)}$*. Consequently, if* $\Theta_{[l^*,L-1)} \leq \epsilon\sqrt{2(1-\tau)}$*, the single direction* $\hat{u}^* := \hat{d}_{l^*}/\|\hat{d}_{l^*}\|$ *satisfies* $\cos^2\angle(\hat{d}_l, \hat{u}^*) \geq 1 - \epsilon^2$ *at every layer* $l \geq l^*$.

This is the discrete form of the LRH statement: one direction extracted at l^* is $(1-\epsilon^2)$-faithful to the normalized counterfactual difference at all later layers whenever the turning spent after branching is small. It is a worst-case bound. On Qwen2.5-0.5B the read-out difference continues to turn after branching (cosine 0.38 between δ_{l^*} and δ_{l^*+3} on average over the pairs of Section I.2); a direction added at l^* acts on the output because the flow after l^* carries it (Section I.2; right column of Figure 10). On Qwen2.5-7B the consecutive-layer cosines of the residual-stream difference rise to ≈ 0.95 in the late layers (Figure 18(2)).

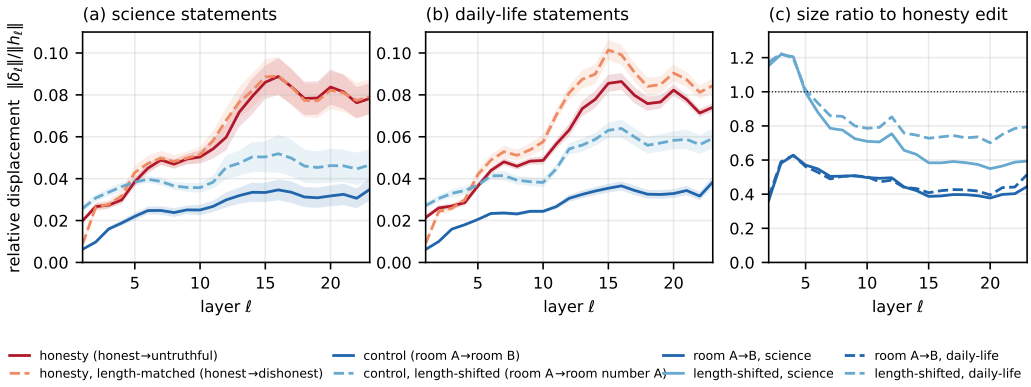


Figure 7: **Selective expansion on Qwen2.5-0.5B.** (a,b) Relative displacement $\|\delta_\ell\|/\|h_\ell\|$ of the read-out token (mean over prefixes of ten statements, 95% bands over statements) for the honesty edit, a length-matched honesty edit, the control edit and a length-shifted control edit. (c) Size of the two control edits relative to the honesty edit.

Remark D.4 (Lifting the direction). $\hat{d}_{l^*} = \hat{b}_{l^*} - \hat{a}_{l^*}$ is a vector of the residual space $\mathbb{R}^d$ (Equation (26)); its coordinates in the gain-ordered frame are $U^\top \hat{d}_{l^*}$ (Lemma A.5). Algorithm 1 works with residual-stream differences $h_{l^*}(X_{\Delta x}) - h_{l^*}(X) \in \mathbb{R}^d$, so that the steering vector and the hooked activation live in the same space; the branching layer is used to choose the depth.

Proposition D.5 (Bounded turning forces gradual branching)**.** *If $\theta_l(X) + \theta_l(X_{\Delta x}) \le M$ for all l, then $s_{l-1} - s_l \le M$ for all l, and a drop of the score from $s_{l_0} \ge \tau + \Delta$ to $s_{l^*} < \tau$ takes at least $\lceil \Delta/M \rceil$ layers. The Concept Path $\{a_l\}_l$ is then a discrete curve on the unit sphere whose total turning $\sum_l \theta_l(X)$ is its arc length.*

Proof. The per-layer bound is Proposition A.6; the layer count follows by summing; the last sentence is the definition of the arc length of a polygonal path on the sphere. □

Corollary D.6 (Turning budget and the two profiles)**.** *For any counterfactual pair and any layer interval $[l_0, l_1] \subseteq [0, L-1]$, $s_{l_0} - s_{l_1} \le \Theta_{[l_0,l_1)}$. The impulsive profile spends the budget in one layer (Proposition D.2) and, if little is spent afterwards, leaves a stable direction (Proposition D.3); the gradual profile spends it a little at a time and leaves a curve (Proposition D.5).*

Consequences for the procedure. (1) *s_l cannot drop without turning*: a drop below τ at l^* certifies turning of at least $1 - \tau$ in $[0, l^*)$. (2) *Steering rests on transport*: Algorithm 1 extracts its direction at l^* and adds it there; the downstream flow $\mathbf{\Phi}(1, t)$ carries it to the output (Figure 2e), whether or not the direction itself stays fixed. (3) *The turning of the read-out difference is measured*: the early-layer consecutive cosines of $\delta H^{(-1)}$ (≈ 0.5, i.e. $\theta \approx 60°$) match the turning angles of Figure 9(f).

E Protocol for the Selective-Expansion Experiment

Prompts and edits. We use the persona prompt of Zou et al. (2023). For the honesty edit the factual and counterfactual prompts are "USER: Pretend you're an *honest* person making statements about the world. ASSISTANT: s" and the same with *untruthful*; for the control edit they are "USER: Pretend you're an honest person sitting in room *A*. ASSISTANT: s" and the same with room *B*. The continuation s ranges over every prefix (one token at a time) of ten science statements and ten daily-life statements; for Qwen2.5-7B prefixes run up to 50 tokens, giving 290 and 295 contexts, and each context is one point in Figure 3. For Qwen2.5-0.5B we add the *length-matched* honesty edit ("an *honest*"→"a *dishonest*", 19 → 19 tokens; the original honesty edit is 19 → 21) and the *length-shifted* control edit ("room *A*"→"room *number A*", 18 → 19 tokens), with prefixes up to 30 tokens.

Results on Qwen2.5-0.5B. Figure 7 repeats the experiment with two further controls. The honesty edit changes the prompt length by two tokens; the length-matched honesty edit (*honest→dishonest*,

one token for one) follows the same curve to within the bands at every layer, so the length change plays no part in the expansion. The length-shifted control edit starts at the same relative size as the honesty edit ($1.2\times$ at layers 2–4) and is overtaken by it from layer 6 on, ending at 0.55 (95% CI 0.45–0.66) and 0.74 (0.68–0.80) of the honesty edit's size on the two datasets; the control edit proper ends at 0.44 (0.37–0.52) and 0.50 (0.46–0.54). Measured against the residual stream, the honesty edit grows from 2.7% of the stream's norm at layer 2 to 8–9% at layers 15–16 while both controls stay below 5–6%.

Displacements. For each context c and layer ℓ we run the factual and counterfactual prompts and record the residual-stream state of the final token after block ℓ; the displacement is $\delta_\ell^{(c)} := h_\ell(X_{\Delta x}) - h_\ell(X)$. The ratio reported in the text is the mean of $\|\delta_\ell^{(c)}\|$ over contexts for one edit divided by the same mean for the honesty edit, at the same layer; all norms are computed in the full space (3584 or 896 dimensions). For the 0.5B model, means are taken per statement and the 95% intervals in Figure 7 and in the text are bootstrap intervals over the ten statements of each set (2000 resamples). The relative displacement is $\|\delta_\ell\|/\|h_\ell(X)\|$ with both norms averaged over contexts.

Projection and streamlines (Figure 3). For display, the displacements of one row (one layer, both edits, both datasets) are projected onto their first two principal directions, computed without centering so that $\delta = 0$ maps to the origin ($+$), and all panels in the row share this projection, axis range and color scale. Streamlines are obtained by Gaussian-kernel interpolation of the change $\delta_\ell^{(c+1)} - \delta_\ell^{(c)}$ between consecutive prefixes onto a grid and serve display only. Layer indices count residual blocks, so layer ℓ corresponds to depth $t = \ell/L$.

Statements used for Qwen2.5-0.5B. *Science:* "Water boils at one hundred degrees Celsius at sea level because the vapor pressure of the liquid equals the pressure of the atmosphere above it."; "The Earth orbits the Sun once every year, and its axis is tilted by about twenty three degrees, which is why the seasons change."; "Light travels at roughly three hundred thousand kilometers per second in a vacuum, and nothing with mass can reach that speed."; "Plants convert carbon dioxide and water into sugar and oxygen using the energy of sunlight in a process called photosynthesis."; "Sound needs a medium such as air or water to travel, which is why no sound can be heard in the emptiness of space."; "The human body has two hundred and six bones, and the smallest of them is found deep inside the ear."; "Iron rusts when it is exposed to oxygen and moisture, because the metal slowly combines with oxygen to form an oxide."; "A magnet has a north pole and a south pole, and opposite poles attract while like poles push each other away."; "Diamonds and graphite are both made of pure carbon, but their atoms are arranged in very different patterns."; "The Moon has no atmosphere, so its surface temperature swings by hundreds of degrees between day and night." *Daily life:* "This morning I woke up late, made a cup of coffee, and walked to the bus stop under a light rain."; "My neighbor waters her garden every evening and always waves when she sees me coming home from work."; "We usually do the grocery shopping on Saturday and cook a big dinner for the whole family on Sunday."; "The bakery on the corner sells fresh bread until noon, and after that only a few pastries are left."; "My little brother lost his keys again, so we spent half an hour looking under the sofa cushions."; "On weekends the park fills with children on bicycles and people walking their dogs along the river."; "I forgot my umbrella at the office, so I had to run to the station through the pouring rain."; "The kitchen faucet has been dripping for a week, and I finally called a plumber to fix it today."; "Every Friday my friends and I meet at the same small restaurant and order the same bowl of noodles."; "The train was so crowded this morning that I had to stand the whole way with my bag on my shoulder."

F Empirical Protocol for the Flow-Operator Probes

Access to $\mathbf{\Phi}$. For an edit at token i the relevant block column is $\mathbf{\Phi}^{(\cdot \leftarrow i)}(t, 0) = \partial \mathbf{z}(t)/\partial e_i$ where e_i is the input embedding of token i. Its action on $\mathbf{v} \in \mathbb{R}^d$ is the forward-mode Jacobian–vector product (JVP) of the map $e_i \mapsto (\mathbf{h}^{(0)}, \ldots, \mathbf{h}^{(L)})$ (all pre-normalization hidden states), computed on a forward pass without fused attention kernels, which forward-mode differentiation requires; one JVP returns the tangent at every layer, and the tangents at tokens before i vanish, as causal masking requires. We verified the JVP against finite differences. The read-out subspace of Equation (6) is pulled back through the final RMSNorm by its gain only; the input-dependent radial projection of the RMSNorm Jacobian removes a single direction and is dropped. We draw $\mathbf{\Omega} \in \mathbb{R}^{d \times m}$ with orthonormal columns

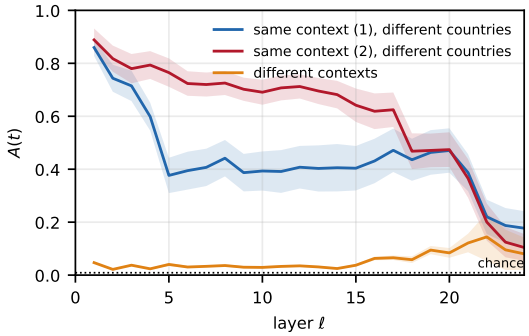


Figure 8: Cross-input alignment of the read-out orbit on Qwen2.5-0.5B for inputs in the same context and in different contexts (mean ± s.d.).

($m = 128$) and form $\mathbf{Y}_\ell = \mathbf{\Phi}_\ell^{(\cdot \leftarrow i)} \mathbf{\Omega} \in \mathbb{R}^{Nd \times m}$ for all ℓ in one batched JVP. Because a few residual dimensions carry outlying magnitudes, all quantities are computed in per-layer whitened coordinates (each dimension divided by its RMS over all tokens of the factual runs of both templates), with the read-out directions transformed accordingly ($\mathbf{v} \mapsto \mathbf{v} \odot \mathrm{rms}_\ell$) so that logit differences are unchanged.

Estimators. H_ℓ is the entropy of the normalized singular values of $\mathbf{Y}_\ell$ (a random-subspace estimate; with one additional backward pass the sketch becomes a randomized SVD (Halko et al., 2011)). $E_i(\ell)$ is the fraction of $\|\mathbf{Y}_\ell\|_F^2$ outside the block of token i. The read-out orbit is the top-k left singular subspace of the block $\mathbf{Y}_\ell^{(N)} \in \mathbb{R}^{d \times m}$ with $k = 8$; A averages $\mathrm{Tr}(\mathbf{P}_a \mathbf{P}_b)/k$ over ordered pairs of inputs of one template, and over pairs of inputs from the two templates; the chance level is the same average over pairs of random k-dimensional subspaces (0.009 for $d = 896$). The overlap with a read-out subspace with orthonormal basis $\mathbf{Q}$ is $\|\mathbf{Q}^\top \mathbf{U}\|_F^2/k$, reported for the task subspace of Equation (6) (chance $15/d$) and for the effective read-out subspace spanned by the top-64 right singular vectors of $\mathbf{g} \odot \mathbf{W}_U$ (chance $64/d$). From the actual counterfactual $\delta_\ell^{(N)}$ (the country swapped for the next one in the list, two forward passes) we compute the turning angle $\theta_\ell = \arccos\langle \mathbf{u}_\ell, \mathbf{u}_{\ell+1} \rangle$, the on- and off-target parts $\|\mathbf{P}_\parallel \delta\|$, $\|\mathbf{P}_\perp \delta\|$, and the cosine of δ with the pair's own answer direction $\mathbf{g} \odot (\mathbf{w}_b - \mathbf{w}_a)$. The layer-0 read-out block of the tangent is zero (no attention has acted), so orbit-based quantities are reported from layer 1.

Alignment within a context. Figure 8 shows the cross-input alignment for the three ways of pairing inputs. Inputs that share a context and differ only in the country share the shallow read-out orbit that the context sets ($A = 0.86$–0.89 at layer 1), and this shared part is shed with depth as the operator becomes specific to the country (0.10–0.18 at the output); inputs in different contexts share almost nothing in the shallow layers (0.02–0.05) and begin to align in the deep layers (0.08–0.14 at layers 19–23), which is the alignment shown in Figure 9(c). On Qwen2.5-7B with a single context and the same protocol, $H/\ln m$ decreases to 0.90, E_i rises to 0.65 and saturates around $t \approx 0.6$, the within-context alignment decreases from 0.35–0.40 to 0.22, and the turning rate of δ has maxima at $t \approx 0.12$ and 0.27 (about 60° per layer) with a level of 40° per layer afterwards.

Prompts. Template 1: "The capital of {country} is" ($N = 5$); template 2: "Q: What is the capital of {country}? A:" ($N = 11$). Countries: France, Japan, Italy, Spain, Egypt, Canada, Germany, Peru, Chile, China, Kenya, Norway, Sweden, Poland, Greece, Cuba; each country and each capital is a single token with a leading space in the Qwen2.5 tokenizer. For Qwen2.5-7B the same protocol is used with "The capital of {country} is" and 16 countries.

Cost. All m probes of one input are batched into a single forward-mode pass, so the 16 inputs of a template require 16 passes over N tokens with batch m; on CPU the 0.5B probes take about ten seconds per input.

G Operator Measurements Across Models

Figure 2 shows measurements on Qwen2.5-7B-Instruct. This section repeats the operator measurements on six instruction-tuned models from four families: Qwen2.5-0.5B, Qwen2.5-7B and Qwen2.5-14B, Llama-3.1-8B, Mistral-7B-v0.3 and Gemma-2-9B (Figures 9 and 10, Table 3).

Protocol. Everything follows Sections F and J, with the same prompts, $m = 128$ probes, $k = 8$, per-layer whitened coordinates, the country swapped for the next one in the list, and the seven concepts and templates of Section J; for the number concept we add the neutral template "The number is {}. Later we walked to the". For each tokenizer the 16 countries of Section F are single tokens with a leading space. The computations run in float32 with eager attention, except for Qwen2.5-14B, which runs in bfloat16 throughout, and for the branching layers of Gemma-2-9B (panel b), which run in bfloat16; float32 would not fit on the available 48 GB accelerators.

Onset t^*. We define the onset from the off/on ratio $r(\ell) = \|\mathbf{P}_\perp \boldsymbol{\delta}_N(\ell)\| / \|\mathbf{P}_\parallel \boldsymbol{\delta}_N(\ell)\|$ of the actual counterfactual, averaged over the two formats: t^* is the first layer after the peak of r (searched over $\ell \geq L/4$, past the shallow layers in which the read-out token has not yet received the edit) from which r stays below half of its peak. This gives $t^* = 0.75$ on Qwen2.5-0.5B, close to the value read off by eye in earlier versions, and 0.79 on Qwen2.5-7B.

Transport. The transport panels use two random pairs per concept of the prompt bank of Section I.1 (288 pairs). At the output of block ℓ we add either the direction of $\boldsymbol{\delta}_\ell$ at strength $0.3\,\|h_\ell(X)\|$ at every position, or a random direction of the same strength, and record the cosine of the change of the final read-out state with the true counterfactual change. Separately we add the exact difference $\boldsymbol{\delta}_\ell$ at the read-out token and record the share of the output KL divergence reproduced, $1 - \mathrm{KL}(p_{X_{\Delta x}} \| p_{\mathrm{inj}}) / \mathrm{KL}(p_{X_{\Delta x}} \| p_X)$. The dots mark each pair at its own branching layer (equation 9, threshold calibrated so that half of the 288 pairs branch).

Branching layer vs. causal effect. Panel (b) shows, for each of the 144 concepts of the prompt bank, the median branching layer over its 32 pairs (threshold calibrated on the other half of the prompts, Section C.5) against its mean causal effect.

What holds on every model. The spectral entropy falls monotonically from $\ln m$, to 0.92–0.95 of it at the output. The share of the edit's energy on other tokens rises with depth. The off/on ratio of the read-out difference is large through most of the network and falls in the last third, with onset $t^* = 0.53$–0.79. At the output it is 3.4–6.9, so every model moves toward $\mathcal{O}_d^*$ but ends before the bound of Theorem 4.3 applies ($\bar{r} \leq \frac{1}{2}$); the theorem describes the direction of the late dynamics, not their endpoint. Directions read from counterfactuals are transported to the output (cosine 0.72–0.90 at $\ell = 0.75L$, against at most 0.1 for random directions). The branching layer follows the causal effect on every model ($\rho = -0.69$ to -0.85).

What differs between models. The overlap of the operator's read-out orbit with $\mathcal{O}_d^*$ reaches 1.4–2.1× chance near the output on Qwen2.5-0.5B, Llama-3.1-8B, Mistral-7B and Gemma-2-9B, but stays below chance on Qwen2.5-7B and 14B (0.78 and 0.89). On these two models the growth gap is visible in the actual counterfactual, whose on-target part grows 38–110× after t^*, but not in the top-8 singular directions of the operator. Cross-format alignment takes two shapes. On Qwen2.5-0.5B, Llama-3.1-8B and Mistral-7B it is low through the first half and rises in the last third, to 15–32× chance. On Qwen2.5-7B, Qwen2.5-14B and Gemma-2-9B it is U-shaped: high in the first three layers (15–23× chance), low in the middle (4–6×), and high again near the output (17–39×). Turning per layer decreases with depth on all models (only weakly on Qwen2.5-7B). The separate late turn at the onset that Qwen2.5-0.5B shows is not a general feature. In the answer context, the coherence of cyclic and categorical concepts falls after t^* on every model; the one exception is months on Qwen2.5-14B, which stays flat. At the output, the rating scale lies above its neutral-context coherence on Qwen2.5-0.5B (0.76 vs. 0.69) and Llama-3.1-8B (0.63 vs. 0.60), and below it on the other four models.

Table 3: Operator measurements across models (both prompt formats averaged unless two values are given). H: $H/\ln m$ at the output; E_i: routing at the output (format 1/2); A: maximum cross-format alignment of read-out orbits relative to chance; ov: maximum overlap of the read-out orbit with $\mathcal{O}_d^*$ relative to chance; ratio: off/on ratio of $\boldsymbol{\delta}$ at the read-out token, peak (over $\ell \geq L/4$) and output; t^*: onset of the gap; cos, KL: transport at $\ell = 0.75L$ (cosine for the fixed-strength direction; share of the output KL reproduced by the exact difference); ρ: Spearman correlation of the branching layer with the log causal effect over 144 concepts. ‡bf16 throughout; §ρ computed in bf16.

Model	L	H	E_i	A	ov	ratio	t^*	cos	KL	ρ
Qwen2.5-0.5B	24	0.93	0.50/0.82	15	2.12	35→4.2	0.75	0.72	0.71	−0.84
Qwen2.5-7B	28	0.94	0.24/0.78	20	0.78	140→6.9	0.79	0.79	0.76	−0.82
Qwen2.5-14B‡	48	0.93	0.31/0.80	50	0.89	139→5.7	0.69	0.84	0.80	−0.85
Llama-3.1-8B	32	0.92	0.37/0.75	32	1.39	37→4.2	0.53	0.86	0.84	−0.83
Mistral-7B-v0.3	32	0.92	0.52/0.79	23	2.14	26→3.4	0.59	0.90	0.92	−0.72
Gemma-2-9B§	42	0.95	0.18/0.29	39	1.97	25→3.9	0.62	0.87	0.82	−0.69

H Additional Results for the Tracing Procedure

H.1 Extended disciplines

Figure 11 extends the Concept Hierarchy of Figure 4 to Medical Diagnosis and Social Evaluation. Across both domains, semantically central distinctions branch earlier than surface-level temporal or numerical variations.

H.2 Truncation, basis and threshold

By Lemma A.5, the untruncated Concept Path and the raw value vector have identical cosines, so comparing them isolates the effect of the top-k truncation. At a fixed threshold, the untruncated cosine (*Attention Value*) stays close to 1 for most pairs at most layers; $\tau = 0.99$ is needed to separate pairs, and the resulting hierarchies are compressed onto one or two layers (Figures 12 and 13), with pairs that the truncated score places at different depths ("diabetes/hypertension" at layer 3–4 versus "2023/2024" at layer 9–10) collapsing together. The truncated score with $k = 10$ separates the same pairs at $\tau = 0.9$: the top-k coordinates carry a small fraction of the vector's norm but a large fraction of its input dependence, so a change in the selected index set produces a drop in s_l that the full-vector cosine, dominated by shared high-norm coordinates, does not register at the same threshold.

With the threshold calibrated to a fixed fraction of branching pairs (Section C.5), the untruncated path orders the edits as well as the truncated one (Table 4): the truncation sets the scale of the score, and the order is carried by the displacement itself. The same holds for the basis. With $k = 10$, random orthogonal bases, the PCA basis of the W_V inputs and the SVD bases of the MLP input and output projections order the edits as W_V does (Table 6). The neuron basis does not, because its largest coordinates are high-magnitude activations shared by both inputs.

H.3 Sensitivity to k and τ

H.3.1 Sensitivity to k on a single prompt

Figure 14 repeats the medical prompt of Figure 11 for $k \in \{1, 10, 50, 100\}$ at the fixed threshold $\tau = 0.9$. The absolute branching layers move with k: for $k = 1$ the three central pairs all branch at layer 18, for $k = 50$ at layers 17–18, for $k = 100$ at layer 18 and beyond, while for $k = 10$ they branch at layers 3, 4, 9 and 10. The partition does not move: the temporal pairs *2023/2024* and *March/July* are the last to branch or never branch for every k, and *metformin/insulin* is among the first for every k. At a fixed threshold, a single coordinate ($k = 1$) is too coarse to separate the central pairs from each other, and a large k brings the truncated score close to the untruncated one (Section H.2). Over the full prompt bank, a calibrated threshold removes this dependence: every $k \geq 2$ gives ρ between -0.59 and -0.84 with the causal effect on all three models (Table 4).

H.3.2 SENSITIVITY TO τ ON A SINGLE PROMPT

Figure 15 repeats the same prompt for $\tau \in \{0.99, 0.9, 0.8, 0.7\}$. Lowering τ requires more certified turning (Proposition A.6) before a pair is declared branched, so all layers move later: the central pairs branch at layers 3–5 for $\tau = 0.99$, 3–10 for $\tau = 0.9$, 15–18 for $\tau = 0.8$ and 18 for $\tau = 0.7$, while *2023/2024* and *March/July* remain the last pairs (INF for $\tau \leq 0.8$) in every setting.

Table 4: Spectral lens at fixed $\tau = 0.9$ vs. calibrated τ, by k. frac: fraction of pairs that branch at $\tau = 0.9$; $\rho_{0.9}$ / ρ_{cal}: concept-level Spearman with log-KL. The non-monotone dependence on k at fixed τ tracks frac, not the order.

	Qwen2.5-0.5B			Qwen2.5-7B			Llama-3.1-8B		
k	frac	$\rho_{0.9}$	ρ_{cal}	frac	$\rho_{0.9}$	ρ_{cal}	frac	$\rho_{0.9}$	ρ_{cal}
1	0.78	-0.49	–	0.95	-0.35	–	0.95	-0.40	–
2	0.95	-0.42	-0.59	0.99	-0.19	-0.64	0.99	-0.37	-0.77
5	0.99	-0.30	-0.73	1.00	-0.23	-0.70	1.00	-0.27	-0.78
10	0.67	-0.75	-0.75	0.96	-0.49	-0.69	0.95	-0.61	-0.81
20	0.31	-0.70	-0.78	0.91	-0.53	-0.73	0.88	-0.78	-0.84
50	0.03	-0.26	-0.77	0.60	-0.73	-0.73	0.61	-0.88	-0.84
100	0.01	-0.14	-0.79	0.37	-0.74	-0.72	0.39	-0.81	-0.83
full	0.01	-0.14	-0.80	0.05	-0.33	-0.75	0.03	-0.30	-0.82

H.4 ACROSS MODEL FAMILIES

Figure 16 repeats the medical prompt on six models: Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct and Qwen2.5-32B-Instruct (Qwen Team et al., 2024); Mistral-7B-Instruct (Jiang et al., 2023); and the first-generation LLaMA-7B and LLaMA-30B (Touvron et al., 2023), included as a third architecture family. *metformin/insulin* branches first in all six (layers 1, 3, 7, 8, 14 and 19 in the respective models), *March/July* last in all six (INF in four), and *2023/2024* second-to-last in five. The order of the intermediate pairs (*diabetes/hypertension* versus *negative/positive*) differs between families, and the absolute layers scale roughly with depth. The central/surface partition is shared across families.

H.5 EMBEDDING DISTANCE AND BRANCHING LAYER

For each case and overall, Table 5 reports Pearson (Pearson, 1895) and Spearman (Spearman, 1904) correlations between input-embedding L2 distance and branching layer for all pairs in Figures 4 and 11. Overall, both coefficients are small in magnitude and not significant; one case (Social 1) shows a strong negative correlation that does not recur in the other cases. The hierarchy is not a re-description of embedding geometry.

Case	Pearson	Spearman
Physics 1	$-0.066_{0.934}$	$0.316_{0.684}$
Physics 2	$-0.240_{0.239}$	$-0.104_{0.407}$
Politics 1	$0.324_{0.595}$	$0.154_{0.805}$
Politics 2	$-0.055_{0.940}$	$-0.099_{0.503}$
Medical 1	$-0.109_{0.862}$	$-0.051_{0.935}$
Medical 2	$0.542_{0.458}$	$0.800_{0.200}$
Social 1	$-0.836_{0.038}$	$-0.880_{0.021}$
Social 2	$-0.078_{0.901}$	$-0.051_{0.935}$
Overall	$-0.218_{0.255}$	$-0.102_{0.598}$

Table 5: Correlation between input-embedding L2 distance and branching layer, per case and overall (cases correspond to Figures 4 and 11). Pearson's r and Spearman's ρ with p-values as subscripts.

H.6 Concept reverse experiment

Without explicit emphasis, the temporal pair *2024*/*2025* branches at layer 11 and *bleeding*/*swollen* at layer 5. When the instruction *"in this sentence, 2024 or 2025 is very important"* is appended, the temporal pair branches at layer 6 and *bleeding*/*swollen* moves to layer 13 (Figure 17). The context changes the operator along which the same token edit propagates.

H.7 Generality of propagation patterns

Figure 18 extends Figure 6 to three independent counterfactual pairs. Panel (1) shows that the trend of $\cos(V^{(-1)}_{L(\Delta x)}, V^{(-1)}_{L})$ is consistent across all pairs. Panel (2) shows that $\cos(\delta H^{(-1)}_{L}, \delta H^{(-1)}_{L+1})$, the consecutive-layer cosine of the residual-stream counterfactual difference, rises from ≈ 0.5 in the early layers to ≈ 0.95 in the late layers: the per-layer turning of the difference falls from $\approx 60^\circ$ to $\approx 18^\circ$. This is the post-branching budget of Proposition A.7 measured on the residual stream of Qwen2.5-7B; its early-layer size matches the turning angles of Figure 9(f).

I Validation of the Branching Layer

I.1 Large-scale validation

Prompts and edits. Table 8 lists the 24 base prompts (six per domain: physics, politics, medical diagnosis, social evaluation) and their six single-token edits. Every combination of the six words is realized, so each edit is measured in the 32 contexts formed by the other five: $24 \times 6 = 144$ concepts and 4,608 pairs per model. Every edit is a single-token swap at the same position under both the Qwen2.5 and the Llama-3.1 tokenizers, which the Jacobian–vector products and the patching at the edited token require.

Causal effect. $\mathrm{KL}(p_X \,\|\, p_{X_{\Delta x}})$ of the next-token distribution at the read-out token. A concept's effect is the mean log-KL over its 32 pairs; its branching layer is the median over its pairs (INF counted as L).

Measures. Spectral lens: equation 26 on the input of W_V at the last token, with $k \in \{1, 2, 5, 10, 20, 50, 100, \text{full}\}$. Basis variants at $k = 10$ replace U, Σ by: U alone (unweighted); a random $d \times r$ orthonormal basis with the spectrum of W_V; a Haar-random orthogonal basis; the PCA basis of the W_V inputs over all prompts; the identity (neurons); the SVD bases of the MLP up-projection and of the output projection, each on its own input. Flow score: equation 9, standardized over all 1,536 read-out states of the prompt bank at that layer. First-order flow score: the same with $\bar{\mathbf{\Phi}}$ replaced by its $s = 0$ term. All thresholds are calibrated as in Section C.5.

Activation patching. Following Meng et al. (2022), for each layer ℓ we run X with the residual stream entering block ℓ replaced, at one position, by its value under $X_{\Delta x}$, and record the recovered fraction $r_\ell = 1 - \mathrm{KL}(p_{X_{\Delta x}} \| p_{\text{patched}}) / \mathrm{KL}(p_{X_{\Delta x}} \| p_X)$. Patching the read-out position gives the *arrival layer*, the first ℓ with $r_\ell \geq \frac{1}{2}$: the depth by which the read-out token has received half of the edit's effect. Patching the edited position gives the *departure layer*, the first ℓ with $r_\ell < \frac{1}{2}$. Both are computed on 8 random pairs per concept.

Results. Table 6 gives the full result. With a calibrated threshold, the flow branching layer correlates with the causal effect at $\rho = -0.84$, -0.82 and -0.83 (Qwen2.5-0.5B, Qwen2.5-7B, Llama-3.1-8B), and the spectral lens at -0.75, -0.69 and -0.81. *Truncation.* At the fixed threshold $\tau = 0.9$ the fraction of pairs that branch at all ranges from 1% to 100% across k (Table 4); with a calibrated threshold every k from 2 to the untruncated path gives ρ between -0.59 and -0.84 on all three models, so the order does not depend on k. *Basis.* Random orthogonal and PCA bases order the edits as W_V does (random basis: $\rho = -0.78$, -0.75, -0.80; Kendall $\tau = 0.65$–0.77 with the W_V order); the order is a property of the displacement $\bar{\mathbf{\Phi}}\boldsymbol{\delta}(0)$. *Path average.* With the $s = 0$ term of the operator in place of $\bar{\mathbf{\Phi}}$, ρ is -0.60, -0.19 and -0.36, in line with the quadrature error of Table 2; a token swap needs the path average. *Activation patching.* The arrival layer is unrelated to the size of the effect ($\rho = 0.00$, -0.04, 0.08) and weakly related to the branching layer ($\rho = 0.17$, 0.31, 0.02;

Table 6: **Branching layer vs. causal effect** on 24 prompts × 6 single-token edits (144 concepts, 4,608 counterfactual pairs) per model. Entries: Spearman ρ between a concept's median branching layer and its mean log-KL of the next-token distribution (negative = edits with larger effect branch earlier), with 95% bootstrap CI over concepts. "Calibrated": τ chosen on held-out prompts so that half of the pairs branch (Section 5.1). †On a random subset of 8 pairs per concept (1,152 pairs).

Lens / method	Qwen2.5-0.5B	Qwen2.5-7B	Llama-3.1-8B
Spectral lens (W_V, k=10, τ=0.9, as before)	−0.75 [-0.81,-0.66]	−0.49 [-0.62,-0.34]	−0.61 [-0.73,-0.47]
Spectral lens (W_V, k=10), calibrated	−0.75 [-0.80,-0.68]	−0.69 [-0.76,-0.60]	−0.81 [-0.85,-0.75]
$k = 5$ / calibrated	−0.73 [-0.79,-0.65]	−0.70 [-0.77,-0.61]	−0.78 [-0.83,-0.71]
$k = 50$ / calibrated	−0.77 [-0.82,-0.71]	−0.73 [-0.79,-0.65]	−0.84 [-0.88,-0.80]
no truncation / calibrated	−0.80 [-0.84,-0.74]	−0.75 [-0.80,-0.68]	−0.82 [-0.85,-0.77]
Random orthogonal basis (k=10), calibrated	−0.78 [-0.83,-0.71]	−0.75 [-0.80,-0.68]	−0.80 [-0.84,-0.74]
PCA basis of states (k=10), calibrated	−0.77 [-0.83,-0.71]	−0.68 [-0.75,-0.58]	−0.73 [-0.79,-0.65]
MLP-input SVD basis (k=10), calibrated	−0.70 [-0.78,-0.61]	−0.56 [-0.66,-0.44]	−0.78 [-0.83,-0.72]
Neuron basis (k=10), calibrated	−0.61 [-0.71,-0.49]	−0.67 [-0.75,-0.57]	−0.67 [-0.76,-0.56]
Flow score s^{Φ} ($\bar{\mathbf{\Phi}}$, exact), calibrated	−0.84 [-0.88,-0.79]	−0.82 [-0.85,-0.77]	−0.83 [-0.86,-0.79]
Flow score with first-order $\mathbf{\Phi}$ only (s=0 JVP)†	−0.60 [-0.71,-0.50]	−0.19 [-0.35,-0.03]	−0.36 [-0.50,-0.20]
Flow score, exact $\bar{\mathbf{\Phi}}$, same subset†	−0.84 [-0.87,-0.79]	−0.80 [-0.84,-0.75]	−0.84 [-0.87,-0.79]
Activation patching: arrival layer†	0.00 [-0.17,0.17]	−0.04 [-0.20,0.13]	0.08 [-0.09,0.25]

Table 7). Patching normalizes each edit by its own effect and asks *where* it travels; the branching layer asks *when* it comes to dominate the read-out state. The two measures are complementary.

Table 7: Agreement between orders (Kendall τ with the spectral lens, calibrated, concept level), agreement of each order with the activation-patching arrival layer (Spearman, subset).

	Qwen2.5-0.5B	Qwen2.5-7B	Llama-3.1-8B
Kendall: flow score vs. spectral lens	0.64	0.70	0.82
Kendall: random basis vs. spectral lens	0.65	0.70	0.77
Kendall: k=50 vs. k=10	0.72	0.73	0.84
Spearman: spectral lens vs. patching arrival	0.17	0.31	0.02
Spearman: patching arrival vs. log-KL	0.00	-0.04	0.08

Table 8: Prompt bank of the large-scale hierarchy validation: 24 base prompts, 6 single-token edits each. Every combination of the six words is realized ($2^6 = 64$ variants); each edit is applied in the 32 contexts formed by the other five. All edits are single-token, position-aligned swaps for both tokenizers.

prompt	base text	edits (factual/counterfactual)
physics_1	According to the laws of physics, a glass ball dropped from the top of a tall building will break the moment it hits the hard ground. Explain what happens next:	*physics/magic*, *glass/rubber*, *tall/short*, *building/table*, *break/bounce*, *hard/soft*
physics_2	The heavy stone sank quickly when the boy threw it into the deep lake near the village. Describe what the object does afterwards:	*heavy/light*, *stone/balloon*, *quickly/slowly*, *boy/man*, *deep/shallow*, *village/city*
physics_3	When the metal pot is heated on the stove for ten minutes, the water inside will boil and turn into steam. What will the cook observe:	*metal/plastic*, *heated/cooled*, *stove/fire*, *ten/two*, *boil/freeze*, *steam/smoke*
physics_4	A small magnet placed near the iron nail will pull it closer across the wooden table. Explain the force involved:	*small/large*, *magnet/rock*, *iron/copper*, *nail/coin*, *pull/push*, *wooden/glass*
physics_5	If you drop a feather and a hammer on the moon at the same time, both will land together on the dusty surface. Explain why this happens:	*feather/coin*, *hammer/brick*, *moon/earth*, *same/different*, *together/apart*, *dusty/rocky*
physics_6	The bright beam from the sun reaches the earth in about eight minutes, traveling through empty space. Explain what this tells us:	*bright/faint*, *sun/moon*, *earth/ocean*, *eight/three*, *minutes/hours*, *empty/dense*

prompt	base text	edits
politics_1	The city mayor decided to make bus rides free for everyone starting October. How will most people in town probably feel about this decision:	*mayor/governor*, *bus/train*, *free/expensive*, *October/January*, *most/few*, *people/tourists*
politics_2	The senate passed a popular law that lowers taxes for small businesses across the country. Predict the public reaction to this policy:	*senate/president*, *popular/controversial*, *lowers/raises*, *taxes/fees*, *small/large*, *country/state*
politics_3	The government announced that it will cut funding for public schools next year to reduce the deficit. How will parents likely respond to this plan:	*government/company*, *cut/raise*, *public/private*, *schools/hospitals*, *next/last*, *parents/teachers*
politics_4	The president signed a treaty with the neighboring country to reduce military spending by half. What will voters probably think of this agreement:	*president/king*, *treaty/deal*, *neighboring/distant*, *reduce/increase*, *military/medical*, *voters/soldiers*
politics_5	Local officials voted to ban cars from the downtown area every weekend during the summer. How will shop owners probably react to this rule:	*Local/Federal*, *ban/allow*, *cars/bikes*, *downtown/rural*, *weekend/night*, *summer/winter*
politics_6	The new policy raises the minimum wage for workers in restaurants starting next month. How will employers most likely respond to this change:	*new/old*, *raises/lowers*, *minimum/maximum*, *workers/students*, *restaurants/factories*, *month/year*
medical_1	The patient was screened negative for type 2 diabetes in March 2023, and is currently taking insulin twice daily. Based on these findings, provide the most likely diagnosis:	*negative/positive*, *diabetes/cancer*, *March/July*, *2023/2024*, *twice/once*, *daily/weekly*
medical_2	An elderly man arrived with severe chest pain radiating to his left arm since morning. State the most likely diagnosis for this presentation:	*elderly/young*, *man/woman*, *severe/mild*, *chest/stomach*, *left/right*, *morning/midnight*
medical_3	The child has a high fever and a red rash that spread across the body over three days. What is the most likely cause:	*child/adult*, *high/low*, *fever/cough*, *red/white*, *body/face*, *three/ten*
medical_4	A young woman reports sudden headache and blurred vision after a long flight yesterday. What condition should the doctor suspect first:	*young/old*, *sudden/mild*, *blurred/double*, *long/short*, *flight/run*, *doctor/nurse*
medical_5	The runner felt sharp pain in the right knee after a long race on hard ground. What injury is most likely:	*runner/dancer*, *sharp/dull*, *right/left*, *knee/ankle*, *race/walk*, *hard/soft*
medical_6	After eating raw chicken at a party, the student developed vomiting and stomach cramps within hours. What is the likely diagnosis:	*raw/cooked*, *chicken/rice*, *party/wedding*, *student/teacher*, *stomach/leg*, *hours/weeks*
social_1	Pretend you are an honest person who always helps the elderly neighbors carry heavy bags every morning. Describe how others probably see you:	*honest/dishonest*, *always/never*, *helps/ignores*, *elderly/wealthy*, *heavy/light*, *morning/evening*
social_2	The generous manager praised the new employee publicly during the team meeting on Friday. How did the employee probably feel afterwards:	*generous/strict*, *manager/teacher*, *praised/criticized*, *new/senior*, *publicly/privately*, *Friday/Monday*
social_3	You are a powerful person making decisions for a large group of people in your town. How do you usually treat others:	*powerful/powerless*, *decisions/mistakes*, *large/small*, *group/crowd*, *town/company*, *usually/rarely*
social_4	My friend lied to me about the money she borrowed last week, and she refused to apologize. How should I feel about her:	*friend/sister*, *lied/joked*, *money/car*, *borrowed/stole*, *last/this*, *refused/offered*
social_5	The kind teacher stayed late after school to help a struggling student with math homework. How do the parents probably view this person:	*kind/rude*, *late/early*, *school/lunch*, *struggling/gifted*, *math/art*, *parents/neighbors*
social_6	The rich businessman donated a large sum of money to the local hospital last winter. How do people in the city see him:	*rich/poor*, *businessman/farmer*, *donated/lent*, *large/small*, *hospital/church*, *winter/summer*

I.2 PILOT SET, TURNING AND TRANSPORT

Pairs. Five base prompts on Qwen2.5-0.5B-Instruct with 32 single-token counterfactual pairs (Table 9), written independently of the prompts of Figure 4; the edited word is replaced in place and the read-out token is the last token of the prompt. For each pair we compute the spectral branching

layer ($k = 10$, $\tau = 0.9$), the causal effect $\mathrm{KL}(p_X \,\|\, p_{X_{\Delta x}})$, the distance in tokens from the edit to the read-out token, and the displacement at every layer.

Table 9: Prompts and pairs of the pilot set (factual/counterfactual; the factual word appears in the prompt).

physics	*A glass vase falls from a tall building onto the concrete floor. According to the laws of physics, the vase will most likely* glass/rubber, tall/small, building/table, concrete/pillow, physics/magic, falls/floats
politics	*The city mayor decided to make bus rides free for everyone this October. Most residents will probably feel* mayor/citizen, free/expensive, everyone/students, bus/train, city/village, decided/refused, October/January
medical	*A patient with diabetes was prescribed metformin in March and the blood test came back negative. The doctor should now* diabetes/hypertension, metformin/insulin, March/July, negative/positive, patient/child, blood/urine, prescribed/denied
social	*My honest friend from Geneva gave me a cheap gift in October and I felt* honest/dishonest, Geneva/Brussels, cheap/expensive, October/January, friend/boss, gift/letter
science	*Water boils at a low temperature on the mountain because the air pressure there is* Water/Oil, boils/freezes, low/high, mountain/beach, air/gas, pressure/humidity

Branching layer and causal effect. Over the 32 pairs (pairs that never branch placed at $L = 24$), the branching layer and the log causal effect have Spearman $\rho = -0.62$ ($p = 1.6 \times 10^{-4}$); the branching layer and the edit-to-read-out distance have $\rho = 0.12$ ($p = 0.50$), and the causal effect and the distance $\rho = -0.16$ ($p = 0.38$). The edits with the largest effects (*falls*/*floats*, *decided*/*refused*, *free*/*expensive*, *pressure*/*humidity*, *physics*/*magic*) branch at layers 4–9; the temporal and location edits (*March*/*July*, *October*/*January*, *Geneva*/*Brussels*) branch at layer 12 or later or never.

Turning of the read-out difference. The per-layer turning angle of the read-out-token difference δ_ℓ is 55°–75° in the shallow layers and 30°–50° in the deep layers for every pair, with the same shape across pairs; neither its peak nor the depth at which half of the total turning is spent varies with the pair. How much an edit turns at a layer is set by the layer; when the read-out token picks the edit up is set by the edit.

Transport. For each pair and each layer $\ell \in \{1, \ldots, 23\}$ we add a vector to the output of block ℓ in the factual run and record the output-layer change and the next-token distribution. With the exact difference $\delta_\ell = h_\ell(X_{\Delta x}) - h_\ell(X)$ added at the read-out token, the cosine of the output change with the true counterfactual change δ_L is 0.06, 0.01, 0.32, 0.38, 0.62, 0.86, 0.90, 0.99 at layers $1, 4, 7, 10, 13, 16, 19, 22$, and the share of the output KL divergence reproduced, $1 - \mathrm{KL}(p_{X_{\Delta x}} \| p_{\mathrm{inj}}) / \mathrm{KL}(p_{X_{\Delta x}} \| p_X)$, is 0 through layer 10 and 0.8 from layer 16; adding the same vector at every position gives the same numbers. With the direction of δ_ℓ at a fixed strength of $0.3\,\|h_\ell(X)\|$ added at every position (the form used for steering), the cosine is 0.14, 0.06, 0.37, 0.37, 0.63, 0.85, 0.88, 0.96 at the same layers, against 0.06, 0.03, 0.11, 0.09, 0.01, 0.05, 0.00, 0.01 for a random direction of the same strength; at each pair's own branching layer it is 0.41. The difference at the read-out token is small before layer 4, since attention has not yet carried the edit there, so nothing injected there acts on the output; from the middle layers on an injected direction is transported.

J Multi-Valued Concepts

J.1 Coherence across depth

The concepts of Figure 10 (left column) and their contexts are as follows. Answer contexts (the value is the next token): months, "This month is {}. Next month is" (January to December); weekdays, "Today is {}. Tomorrow is" (Monday to Sunday); numbers, "The number is {}. The next number is" (one to ten); sizes, "The box is {}. The size of the box is" (tiny, small, medium, large, huge, enormous); ratings, "The movie was {}. Out of ten I would rate it" (terrible, bad, mediocre, okay, good, great, excellent); colors, "The paint is {}. The color of the paint is" (red, orange, yellow, green,

blue, purple, pink, brown); capitals, "The capital of {} is" (the 16 countries of Section F). Neutral contexts: ratings, "The movie was {}. After the film we walked back to the"; sizes, "The box is {}. We put the box on the"; months, "This month is {}. We will visit our grandparents at the"; weekdays, "Today is {}. Later we will walk to the"; colors, "The paint is {}. We put the can on the"; capitals, "The capital of {} is a city on the". Within a concept all prompts have the same length. States are the read-out token's residual states in the per-layer normalized coordinates of Section F. The coherence $\kappa(\ell)$ is the fraction of the variance of the centered set $\{h_\ell(c_j)\}_j$ carried by its first principal direction; for n values spread at random it is about $1/(n-1)$.

Answer contexts, at the output: ratings $\kappa = 0.76$ (baseline 0.17), sizes 0.60 (0.20), numbers 0.30 (0.11), colors 0.32 (0.14), weekdays 0.28 (0.17), months 0.19 (0.09), capitals 0.18 (0.07). Across depth, ratings rise from 0.39 at layer 9 to 0.78 at layer 17; months fall from 0.34 to 0.20 and numbers from 0.58 to 0.34 at layer 17. *Neutral contexts:* ratings 0.65–0.70 at every layer, sizes 0.5–0.6, months 0.2–0.45, weekdays 0.3–0.5, colors 0.45–0.57, capitals 0.16–0.32, with no change at layer 17. The per-layer turning angle of $\mathbf{z}(t; c_j) - \mathbf{z}(t; c_1)$, averaged over the values of a concept, has the same profile for all concepts (70°–85° at layer 1 falling to $\approx 30°$ at layer 23): the concepts differ in what the turning does to their set of values, not in how much each value turns.

J.2 The coherence of the answers, κ_U, and the circle test

Concepts and contexts. Ten concepts: months, weekdays, number words, letters, capitals, languages, and four 6-point scales (rating, size, temperature, finishing place). Each has two answer-context and two neutral-context templates. The answer of a value is its successor (months, weekdays, number words, letters), its capital or language, or the digit of its scale position. κ is the top-PC share of the centered value states and κ_U the same statistic of $\{\mathbf{g} \odot \mathbf{w}_{a(c_j)}\}$, computed from the weights before any forward pass. Models: Qwen2.5-0.5B, Qwen2.5-7B, Llama-3.1-8B.

Results for (P2). The answer context moves the output coherence toward κ_U, relative to the neutral context, in 23 of 30 concept–model cases (sign test $p = 0.003$; Table 10). The movement is largest for cyclic and categorical concepts, whose answer rows are spread ($\kappa_U = 0.12$–0.22); on Llama-3.1-8B, weekdays fall from $\kappa = 0.70$ (neutral) to 0.28 (answer) against $\kappa_U = 0.33$. The remaining cases are the size and temperature scales on the three models and number words on Qwen2.5-7B, where the answer context leaves the values more linear than their answers. Across concepts, κ_U correlates with the answer-context output coherence ($\rho = 0.79, 0.85, 0.89$) and with the neutral-context coherence ($\rho = 0.79, 0.81, 0.63$), and the coherence of the value embeddings predicts both ($\rho = 0.84$–0.90). The regime a concept takes is thus largely inherited from its values; the read-out shifts it, in the direction κ_U gives, when the value is the next answer.

Circle test. Small κ alone does not distinguish a curved arrangement from a scatter; in the full state the adjacent differences of months have cosine ≈ -0.5, as for a scatter. We therefore test for the circle reported by Engels et al. (2025). We project the centered value states onto their top two principal components (and, as a variant, onto the best pair among the top six) and measure $|\frac{1}{n}\sum_j e^{\mathrm{i}(\theta_j \mp 2\pi j/n)}|$, the circular correlation between the angle and the cyclic index maximized over orientation, against 200–500 random relabelings of the values with the same maximization. At the token of the value itself, months and weekdays lie on a circle at essentially every layer of all three models: median circular correlation 0.96 and 0.95, significant ($p \le 0.05$ for both templates) in $51/52$ and $50/52$ layer–context–model cells. At the read-out token the circle persists with medians 0.72 and 0.76, significant in $33/46$ and $21/46$ cells. For cyclic concepts the RM regime is a curve in a two-dimensional subspace that carries a minority of the variance, which a low-dimensional test detects and whole-state statistics do not.

K Accuracy of the Linear Response

The operator $\mathbf{\Phi}(t, 0)$ describes the response of the network to a small change of an embedding; a token replacement is a finite change. We therefore measure pair-level quantities (turning, on- and off-target parts, branching layers, coherence, transport) on the actual counterfactual δ, equivalently on $\bar{\mathbf{\Phi}}\boldsymbol{\delta}(0)$ (Equation (3)), and use $\mathbf{\Phi}$ itself for its own properties (entropy, routing, orbits). Applying $\mathbf{\Phi}(t, 0)$ to the embedding difference $e_{\text{country}'} - e_{\text{country}}$ of the capitals prompts and comparing with

Table 10: Multi-valued concepts. κ_U: coherence of the answers' unembedding rows (weights only). κ_{ans}, κ_{neu}: coherence of the value states at the read-out token of the last layer in the answer and neutral context (mean of two templates). ✓: the answer context is closer to κ_U than the neutral one.

	Qwen2.5-0.5B				Qwen2.5-7B				Llama-3.1-8B			
concept	κ_U	κ_{ans}	κ_{neu}		κ_U	κ_{ans}	κ_{neu}		κ_U	κ_{ans}	κ_{neu}	
months	0.21	0.25	0.31	✓	0.22	0.26	0.42	✓	0.22	0.20	0.41	✓
weekdays	0.31	0.33	0.65	✓	0.31	0.38	0.66	✓	0.33	0.28	0.70	✓
number words	0.33	0.63	0.73	✓	0.27	0.56	0.43		0.28	0.38	0.81	✓
letters	0.15	0.36	0.36	✓	0.12	0.32	0.43	✓	0.13	0.20	0.42	✓
capitals	0.15	0.22	0.25	✓	0.15	0.21	0.24	✓	0.16	0.18	0.26	✓
languages	0.14	0.26	0.30	✓	0.12	0.23	0.28	✓	0.12	0.18	0.31	✓
rating	0.34	0.57	0.80	✓	0.33	0.53	0.67	✓	0.35	0.60	0.68	✓
size	0.34	0.57	0.56		0.33	0.72	0.55		0.35	0.55	0.55	
temperature	0.34	0.51	0.47		0.33	0.64	0.51		0.35	0.56	0.48	
place	0.34	0.63	0.77	✓	0.33	0.59	0.64	✓	0.35	0.41	0.76	✓
Spearman	$\rho(\kappa_U,\kappa_{\text{ans}})=0.79$, $\rho(\kappa_U,\kappa_{\text{neu}})=0.79$				$\rho(\kappa_U,\kappa_{\text{ans}})=0.85$, $\rho(\kappa_U,\kappa_{\text{neu}})=0.81$				$\rho(\kappa_U,\kappa_{\text{ans}})=0.89$, $\rho(\kappa_U,\kappa_{\text{neu}})=0.63$			

$\delta(t)$, the two have cosine 0.55 over all tokens through layer 20 and 0.37 at the output; at the read-out token the cosine is 0.3–0.5. The $s=0$ term thus gives the direction of the change to within a cosine of about one half; the path average of Section C.2 recovers the change itself.

L The Causal Effect to Second Order

Let ℓ be the logits of X at the read-out token, $p_X=\mathrm{softmax}(\ell)$, and $\Delta\ell$ their change under the edit, so that $p_{X_{\Delta x}}=\mathrm{softmax}(\ell+\Delta\ell)$.

KL as a variance. Since $\ln p_{X_{\Delta x},a}=\ln p_{X,a}+\Delta\ell_a-\ln\sum_b p_{X,b}e^{\Delta\ell_b}$,

$$\mathrm{KL}(p_X\|p_{X_{\Delta x}})=\ln\mathbb{E}_{a\sim p_X}\left[e^{\Delta\ell_a}\right]-\mathbb{E}_{a\sim p_X}[\Delta\ell_a]=\tfrac{1}{2}\mathrm{Var}_{a\sim p_X}[\Delta\ell_a]+O(\|\Delta\ell\|_\infty^3). \tag{28}$$

The first expression is exact: it is the cumulant generating function of $\Delta\ell$ under p_X, evaluated at 1, minus the first cumulant, and the next term is the third cumulant divided by 6. The variance depends only on differences, $\mathrm{Var}_p[\Delta\ell]=\frac{1}{2}\sum_{a,b}p_ap_b(\Delta\ell_a-\Delta\ell_b)^2$, as it must, since a shift common to all logits leaves the distribution unchanged.

Logit differences through the final normalization. With a final RMSNorm of gain $\mathbf{g}$ (ignoring its ϵ) and no output bias, $\ell_a=\langle\mathbf{w}_a,\ \mathbf{g}\odot\sqrt{d}\,z_N/\|z_N\|\rangle$, where $z_N=z_N(1)$ is the read-out state. The Jacobian of $z\mapsto\sqrt{d}\,z/\|z\|$ is $\frac{\sqrt{d}}{\|z\|}(\mathbf{I}-\hat z\hat z^\top)$ with $\hat z=z/\|z\|$, so to first order in $\boldsymbol{\delta}_N(1)$

$$\Delta\ell_a-\Delta\ell_b=\frac{\sqrt{d}}{\|z_N\|}\left\langle\mathbf{g}\odot(\mathbf{w}_a-\mathbf{w}_b),\ (\mathbf{I}-\hat z_N\hat z_N^\top)\,\boldsymbol{\delta}_N(1)\right\rangle+O\big(\|\boldsymbol{\delta}_N(1)\|^2/\|z_N\|^2\big), \tag{29}$$

which is the relation used in Section 5.1: the radial component of the response is removed, and the rest is read through differences of gain-rescaled unembedding rows. Combining the two displays,

$$\mathrm{KL}(p_X\|p_{X_{\Delta x}})\approx\frac{d}{4\|z_N\|^2}\sum_{a,b}p_{X,a}\,p_{X,b}\left\langle\mathbf{g}\odot(\mathbf{w}_a-\mathbf{w}_b),\ (\mathbf{I}-\hat z_N\hat z_N^\top)\boldsymbol{\delta}_N(1)\right\rangle^2. \tag{30}$$

When p_X concentrates on the candidate answers $\mathcal{A}$, only pairs in $\mathcal{A}$ contribute appreciably, and the KL is a quadratic form in the component of the response in $\mathcal{O}_d^*$ (Equation (6)). The causal effect thus measures the size of the on-target read-out response at the output. For Gemma-2, the final logit soft-capping multiplies each $\Delta\ell_a$ by its own factor in $(0,1]$, and the same conclusion holds with these weights.

M Downstream Applications: Steering for Hallucination and Bias Mitigation

M.1 MindCraft-guided steering

Algorithm 1 applies MindCraft to steer the internal representations associated with hallucination or bias. We first use Automated Concept Extraction (Section N) to generate counterfactual input

Algorithm 1: MindCraft-Guided Steering for Hallucination/Bias Mitigation

Input: LLM f_θ; task $\mathcal{T}$ (grounded vs. hallucinated, or unbiased vs. biased); concept extractor ACE; separation threshold τ; early-layer cutoff ℓ_{cut}; PCA rank r; steering scale α; eval metric $\text{Eval}(\cdot)$.

Output: Steering direction d; steered model $f_{\theta,\text{steer}}$; performance change Δ.

Step 1: Counterfactual concept generation

$\mathcal{P} \leftarrow \text{ACE}(\mathcal{T})$; `// Generate counterfactual pairs` $(X, X_{\Delta x})$

Step 2: MindCraft tracing and divergence extraction

$\mathcal{D} \leftarrow \emptyset$; $\mathcal{L} \leftarrow \emptyset$

foreach $(X, X_{\Delta x}) \in \mathcal{P}$ **do**

- Compute separation scores $\{s_\ell(X, X_{\Delta x})\}_{\ell=0}^{L-1}$
- $\ell^\star \leftarrow \min\{\ell : s_\ell(X, X_{\Delta x}) < \tau\}$; `// Branching layer`
- Extract last-token residual-stream states $h_{\ell^\star}(X), h_{\ell^\star}(X_{\Delta x}) \in \mathbb{R}^d$
- $\delta_{\ell^\star} \leftarrow h_{\ell^\star}(X_{\Delta x}) - h_{\ell^\star}(X)$; `// Counterfactual difference`
- Add $(\delta_{\ell^\star}, \ell^\star)$ to $(\mathcal{D}, \mathcal{L})$

Step 3: Early-branching filter

$\mathcal{D}_{\text{valid}} \leftarrow \emptyset$; $\mathcal{L}_{\text{valid}} \leftarrow \emptyset$

foreach $(\delta_{\ell^\star}, \ell^\star) \in (\mathcal{D}, \mathcal{L})$ **do**

- **if** $\ell^\star \leq \ell_{cut}$ **then**
  - Add $\delta_{\ell^\star}$ to $\mathcal{D}_{\text{valid}}$; add $\ell^\star$ to $\mathcal{L}_{\text{valid}}$

if $|\mathcal{D}_{valid}| < k_{\min}$ **then**

- **return** *"Insufficient early-branching pairs"*, $\emptyset$, $\Delta = 0$

Step 4: Principal steering direction estimation

Center $\mathcal{D}_{\text{valid}}$ and compute PCA; $d \leftarrow$ top principal component (or top-r subspace) of $\mathcal{D}_{\text{valid}}$

Step 5: Layer-conditional steering

At each $\ell \in \mathcal{L}_{\text{valid}}$: $h_\ell \leftarrow h_\ell + \alpha \cdot d$; call the result $f_{\theta,\text{steer}}$

Step 6: Evaluation against baseline

$M_{\text{base}} \leftarrow \text{Eval}(f_\theta)$; $M_{\text{steer}} \leftarrow \text{Eval}(f_{\theta,\text{steer}})$; $\Delta \leftarrow M_{\text{steer}} - M_{\text{base}}$

return d, $f_{\theta,steer}$, Δ

pairs that differ only in the target concept. For each pair, MindCraft locates the branching layer and computes the residual-stream counterfactual difference $h_{\ell^\star}(X_{\Delta x}) - h_{\ell^\star}(X) \in \mathbb{R}^d$ there. Pairs whose branching layer is at most ℓ_{cut} are retained, so that the direction is read where the read-out token has started to carry the edit and enough depth remains for the flow to transport it (Figure 2e); we aggregate their differences and take the dominant principal direction as the steering direction. We then attach a hook at the retained branching layers and add $\alpha\,\mathbf{d}$ to the residual stream during inference.

M.2 Mitigating Hallucinations

We evaluate on the PreciseWikiQA benchmark of **HalluLens** (Bang et al., 2025), on seven open-weight models: Llama-3.1-8B and Llama-3.1-70B (Dubey et al., 2024); Mistral-7B and Mistral-Nemo (Jiang et al., 2023); Gemma-2-9B (Team et al., 2024); and Qwen2.5-7B and Qwen2.5-14B (Qwen Team et al., 2024). We use four metrics, the first three from Bang et al. (2025): **False Refusal Rate (FRR)**, the share of answerable queries the model declines; **Correct Answer Rate (CAR)**, the share of queries answered correctly; **Hallucination Rate (HR)**, the share of non-refused answers judged to contain fabricated content; and **FactScore (FS)** (Min et al., 2023), the precision of atomic facts against the reference corpus.

As shown in Table 11, steering lowers HR on every model, by 14.5 points on average (a 22% relative reduction), raises FS by about 6 points, and raises CAR, with FRR essentially unchanged. Figure 19(a)–(b) break performance down by difficulty level: the gains are larger on harder instances.

Table 11: Concept steering on the HalluLens benchmark (Bang et al., 2025). For each metric, **Baseline** vs. **After Steering**. **FRR**: False Refusal Rate (%, ↓), **CAR**: Correct Answer Rate (%, ↑), **HR**: Hallucination Rate (%, ↓), **FS**: FactScore (↑).

Model Family	Baseline				After Steering			
	FRR	CAR	HR	FS	FRR	CAR	HR	FS
Llama-3.1-8B	83.09	10.05	49.30	52.4	82.70	13.24	37.10	56.1
Llama-3.1-70B	59.76	33.78	35.93	68.1	59.58	35.56	30.15	72.4
Mistral-7B	8.03	17.34	81.19	44.8	8.62	19.45	62.27	53.2
Mistral-Nemo	2.42	24.24	75.50	51.3	3.03	26.10	61.15	55.8
Gemma-2-9B	23.07	19.20	74.28	46.7	22.56	21.25	57.76	52.2
Qwen2.5-7B	14.68	15.06	78.08	49.5	15.24	19.88	59.21	57.1
Qwen2.5-14B	15.93	18.43	72.50	52.0	15.07	20.41	57.85	59.0

Table 12: Concept steering on the BBQ Benchmark (Parrish et al., 2022). For each metric, **Baseline** vs. **After Steering**. **A-**: *ambiguous* contexts, **D-**: *disambiguated* contexts. **Bias**: Bias Score (↓), **EO**: Equal Opportunity (↓), **DP**: Demographic Parity (↓), **ACC**: Accuracy (%, ↑).

Model Family	Baseline					After Steering				
	A-Bias	D-Bias	D-EO	D-DP	D-ACC	A-Bias	D-Bias	D-EO	D-DP	D-ACC
Llama-3.1-8B	0.149	0.203	0.056	0.060	84.20	0.025	0.141	0.016	0.024	83.80
Llama-3.1-70B	0.127	0.190	0.049	0.048	89.00	0.028	0.148	0.022	0.028	86.20
Mistral-7B	0.103	0.314	0.048	0.046	74.40	0.026	0.212	0.032	0.036	73.00
Mistral-Nemo	0.115	0.280	0.042	0.045	82.30	0.031	0.195	0.025	0.021	78.90
Gemma-2-9B	0.138	0.225	0.051	0.055	85.60	0.038	0.152	0.018	0.022	82.10
Qwen2.5-7B	0.165	0.326	0.048	0.028	77.60	0.012	0.228	0.032	0.016	74.70
Qwen2.5-14B	0.097	0.277	0.032	0.036	80.60	0.010	0.230	0.024	0.012	78.50

M.3 Mitigating Bias

We evaluate on the **Bias Benchmark for QA (BBQ)** (Parrish et al., 2022), which measures how social biases manifest in model outputs in two informational settings, with the following metrics: the **Ambiguous Bias Score**, the extent to which a model relies on social stereotypes in under-informative contexts where the correct answer is *Unknown* (closer to 0 is fairer); the **Disambiguated Bias Score**, whether in evidence-sufficient contexts a model's social biases override factual evidence; **Equal Opportunity** (Hardt et al., 2016), disparities in true positive rates across protected groups; **Demographic Parity** (Dwork et al., 2012), disparities in overall prediction rates across protected groups; and **Accuracy** in disambiguated contexts. As shown in Table 12, steering lowers the Ambiguous Bias Score to 0.01–0.04 across models and reduces the EO and DP gaps (e.g., EO from 0.056 to 0.016 and DP from 0.060 to 0.024 for Llama-3.1-8B), with disambiguated accuracy within 0.4–3.5 points of the baseline.

N Automated Concept Extraction

The MindCraft framework provides a lens for inspecting a model's conceptual hierarchy. To scale the analysis to any given text, we use a pipeline in which an LLM proposes the concepts and counterfactuals, in four stages.

Stage 1: Key concept identification. Given a base text, we prompt the LLM to identify the most salient tokens for counterfactual analysis, typically words whose modification would change the meaning or sentiment of the text.

Prompt for Key Concept Identification

Instruction:	Given the following text, identify a group of impactful tokens that defines the core sentiment or concept. The token should be a good candidate for a counterfactual analysis. Focus on adjectives, nouns, or verbs that, if changed, would fundamentally alter the meaning. Output the tokens, separate each token with ' ':
Text:	The city mayor decided to make bus rides free for everyone. How will most people in the city probably feel happy about this decision?
Output:	mayor free everyone happy

Stage 2: Counterfactual concept generation. For each identified concept (e.g., "happy"), the LLM proposes contextually relevant antonyms or replacements.

Prompt for Counterfactual Generation

Instruction:	In the context of the following sentence, what are the most meaningful counterfactuals for the following tokens? Output each pair that separates the original token and the counterfactual token with a '/' and separate each pair with a ' ':
Sentence:	The city mayor decided to make bus rides free for everyone. How will most people in the city probably feel happy about this decision? Tokens: <Identification Output>
Output:	mayor/citizen free/expensive everyone/students happy/angry

where <Identification Output> is the output of Stage 1.

Stage 3: Tracing. For each counterfactual word, a new version of the text is created by replacing the original word, forming a counterfactual pair, and the branching layer is computed as in Section C. This yields a set of branching layers (e.g., {'free/expensive': 3, 'happy/angry': 5, ...}).

Stage 4: Aggregation. The branching layers are ordered into a Concept Hierarchy for the base text; for example, if "angry" branches at layer 3 while "unhappy" branches at layer 5, the model distinguishes the strong, distinct emotion earlier than the general negation.

O Experiment Setup

Datasets. TruthfulQA (Lin et al., 2022) evaluates whether models give factually correct and truthful answers to deliberately misleading questions. **AI2 Reasoning Challenge (ARC)** (Clark et al., 2018) is a collection of grade-school science questions, split into *Easy* and *Challenge* subsets. **COPA** (Roemmele et al., 2011) is a causal reasoning benchmark in which the model chooses the more plausible cause or effect of a premise. Together these cover factual truthfulness, scientific knowledge and causal inference; the base prompts of Figures 4 and 11 are drawn from them.

Models. Qwen2.5-7B-Instruct (Qwen Team et al., 2024) is a 7-billion-parameter instruction-tuned decoder-only Transformer (28 layers, $d = 3584$, $n_h = 28$ query heads, $n_{\text{kv}} = 4$ key/value heads of dimension $d_h = 128$, so $d_v = 512$) with SwiGLU activations. Qwen2.5-0.5B-Instruct has 24 layers, $d = 896$, $n_h = 14$ query heads and $n_{\text{kv}} = 2$ key/value heads of dimension 64. Llama-3.1-8B-Instruct (Dubey et al., 2024) has 32 layers and $d = 4096$. Further model families are used in Sections H.4 and M. **Probes.** $m = 128$ probes in one batched forward-mode pass per input, weights in float32 on CPU for the 0.5B model and bfloat16 on GPU for the 7B model, $k = 8$ for the cross-input alignment.

P STATEMENTS

P.1 IMPACT STATEMENT

This paper presents MindCraft, a framework for describing how LLMs carry concept-level distinctions through depth. Interpretability tools could in principle be used to locate and exploit model weaknesses; we expect the transparency they provide to help the community build robust defenses and to keep alignment work grounded in observable mechanisms.

P.2 THE USE OF LARGE LANGUAGE MODELS

The authors used LLMs as writing assistants only (grammar and readability, style and terminology, occasional LaTeX structuring suggestions). They were not involved in formulating hypotheses, theoretical contributions, experimental design, implementation, data analysis, or interpretation of results. All outputs were reviewed, edited and revised by the authors, who take full responsibility for the content. Separately, an LLM is used *as a component of the method* in Section N (Automated Concept Extraction), which generates the counterfactual pairs for Algorithm 1; this use is disclosed there and in the checklist.

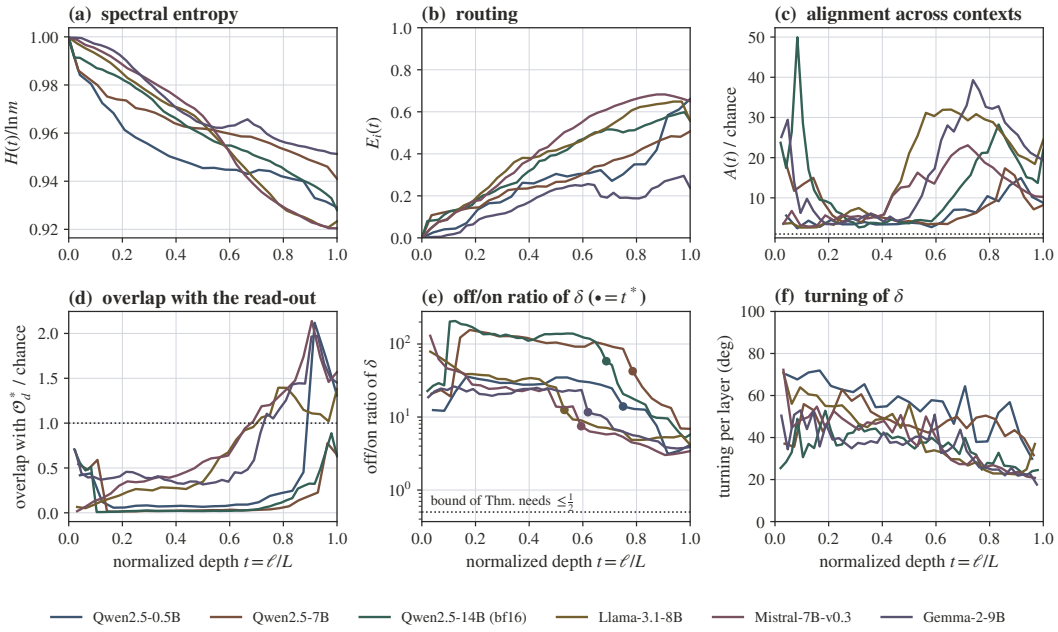


Figure 9: Operator measurements across six models against normalized depth $t = \ell/L$ (mean over the two prompt formats). (c) Cross-format alignment relative to chance. (d) Overlap of the read-out orbit with $\mathcal{O}_d^*$ relative to chance. (e) Off/on ratio of the read-out difference; dots mark t^*; the dotted line is the level $\frac{1}{2}$ at which the bound of Theorem 4.3 applies. (f) Turning of $\boldsymbol{\delta}$ per layer.

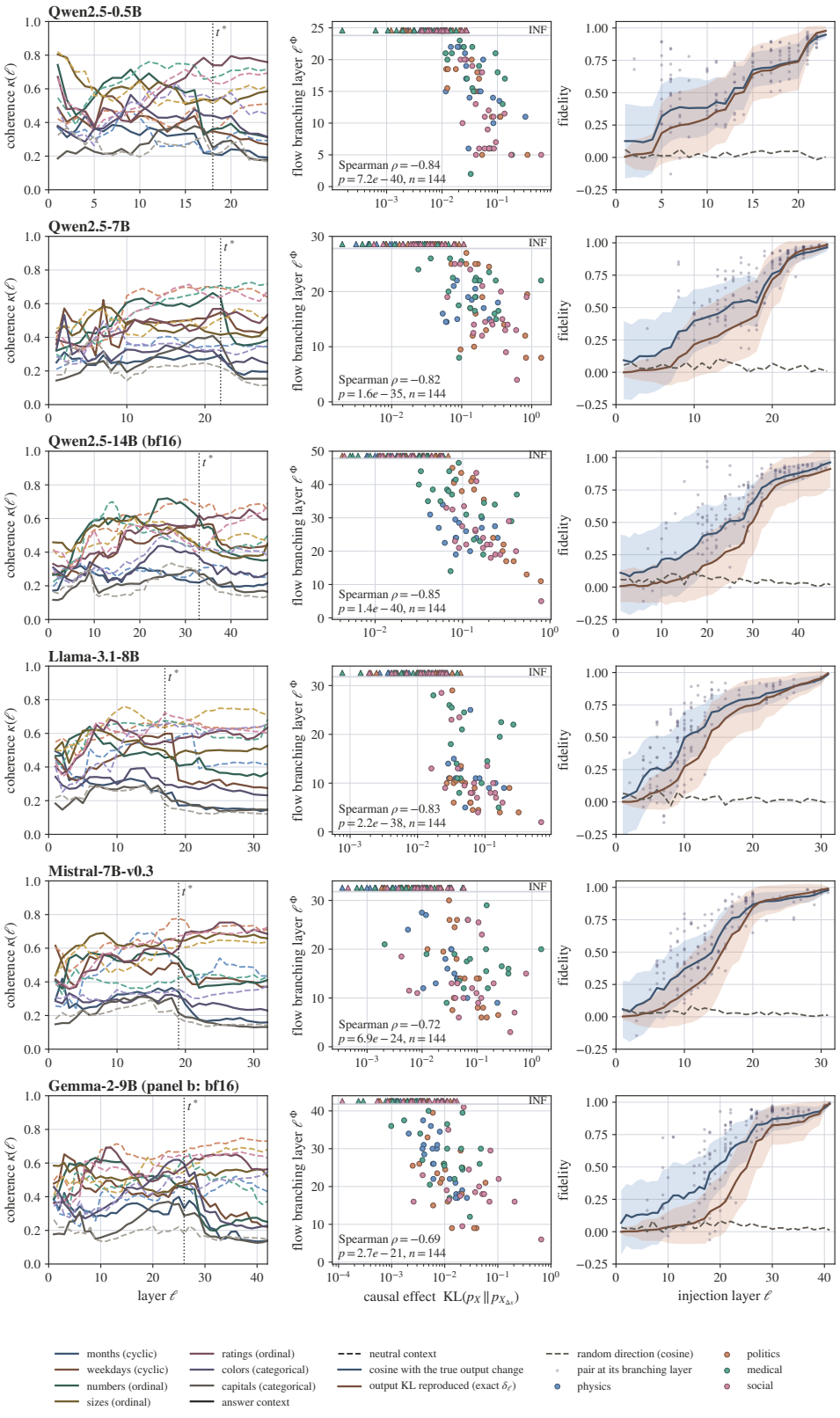


Figure 10: Coherence, branching and transport across six models, one row per model. Left: coherence of seven concepts in the answer context (solid) and the neutral context (dashed); the dotted line marks t^*. Middle: branching layer against causal effect for the 144 concepts of the prompt bank (triangles: never branch). Right: transport of an injected direction; dots mark each pair at its own branching layer.

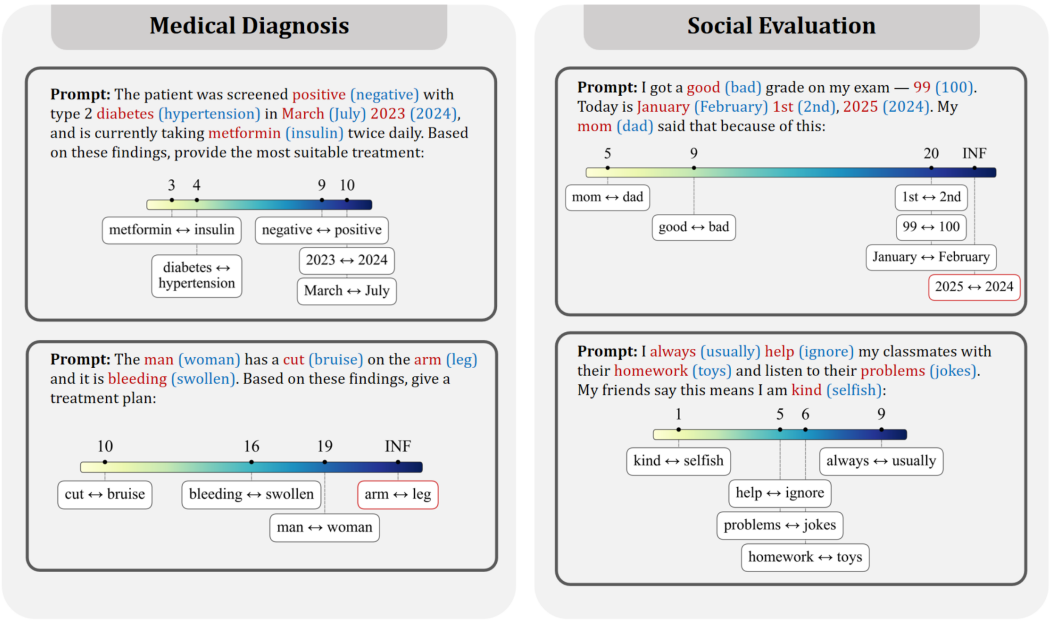


Figure 11: Concept Hierarchy in two additional disciplines: (3) medical diagnosis and (4) social evaluation. Counterfactual pairs are marked with Factual (Counterfactual).

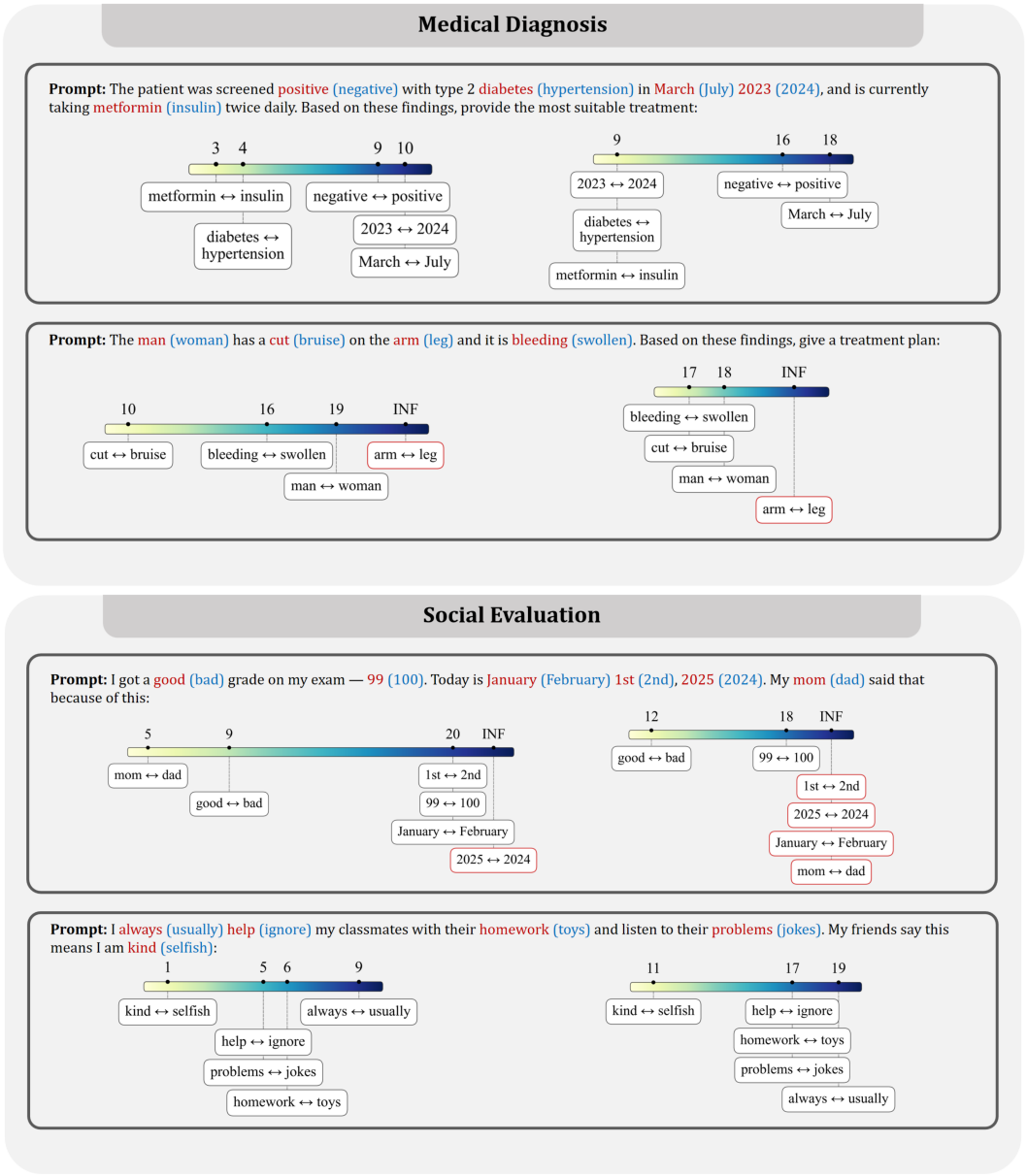


Figure 12: MindCraft spectral lens versus the untruncated value vector (Attention Value) at a fixed threshold. Counterfactual pairs are marked with Factual (Counterfactual). (Part I: Medical Diagnosis and Social Evaluation)

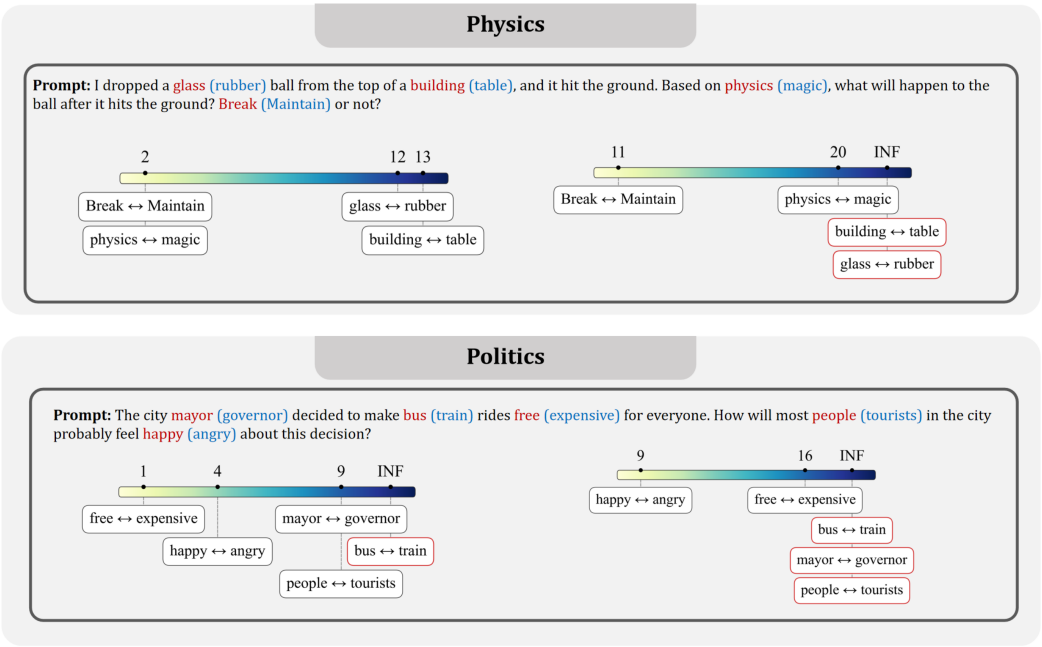


Figure 13: MindCraft spectral lens versus the untruncated value vector (Part II: Physics and Politics).

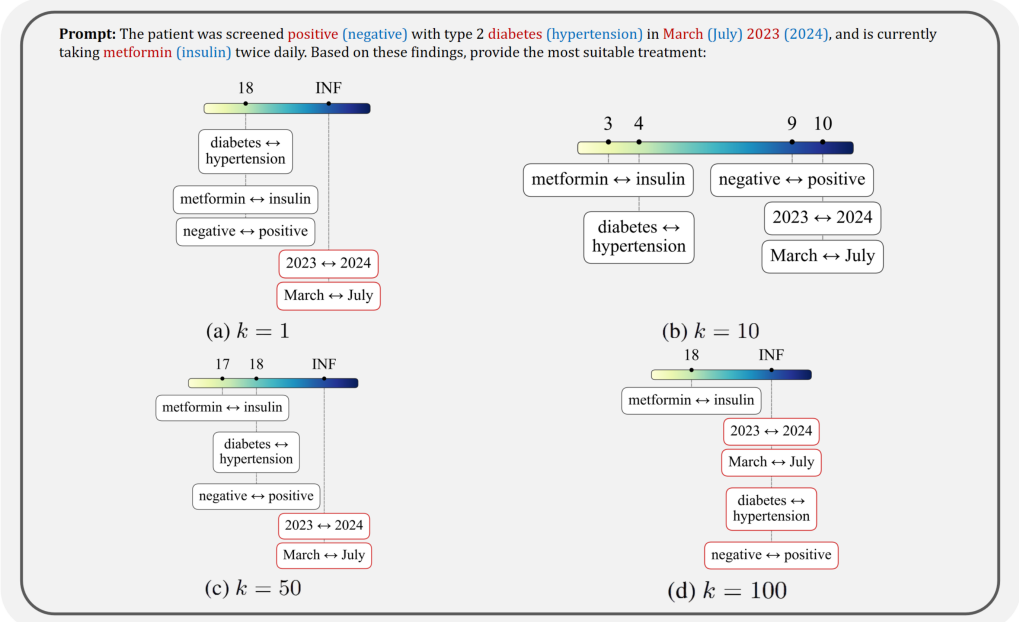


Figure 14: Sensitivity to the truncation size k at a fixed threshold (medical prompt).

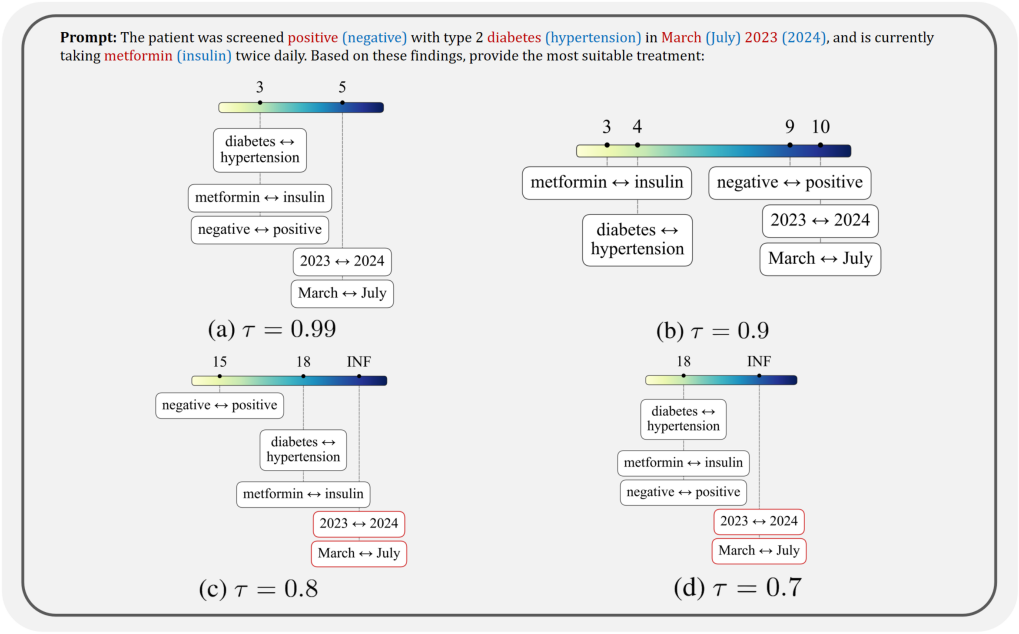


Figure 15: Sensitivity to the separation threshold τ (medical prompt). The partition into central and temporal pairs is preserved; the absolute layers shift with τ.

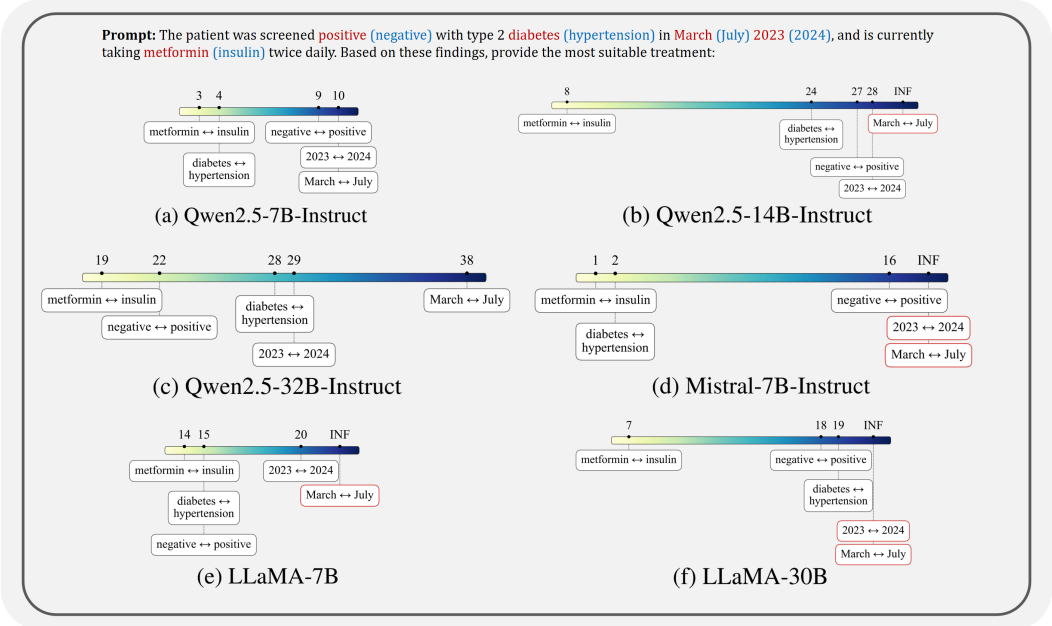


Figure 16: Cross-model comparison on the medical prompt of Figure 11. The partition into central and surface pairs is shared across families; intermediate ordering and absolute layers differ.

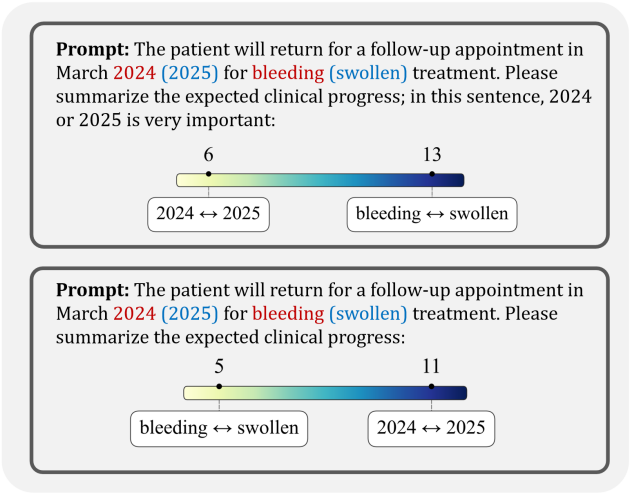


Figure 17: Concept reverse experiment. Appending an instruction that emphasizes the temporal tokens moves their branching layer earlier and the medical pair's later.

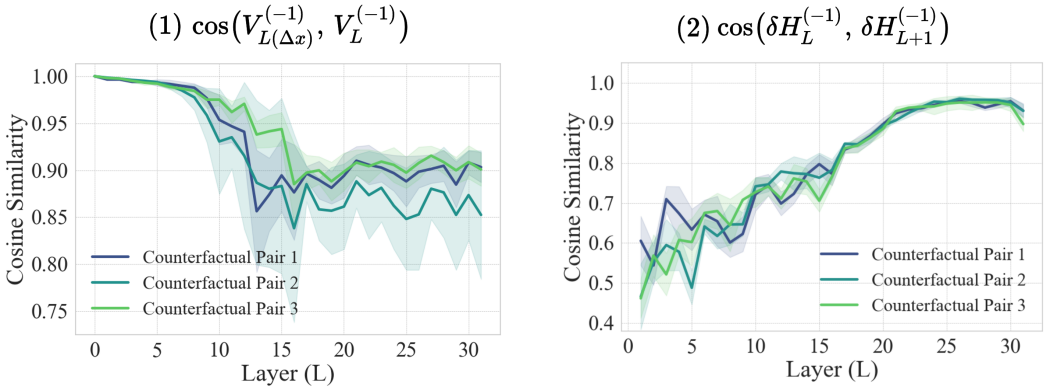


Figure 18: Layer-wise analysis for Counterfactual Pair 1 ("Pretend you're an honest (untruthful) person making statements about the world."), Pair 2 ("Describe a fair (biased) scenario that you have seen.") and Pair 3 ("You are a powerful (powerless) person making decisions."). (1) Cosine between factual and counterfactual value vectors. (2) Consecutive-layer cosine of the residual-stream counterfactual difference.



(a) FRR vs CAR (b) HR vs FS (c) Input Embedding L2 Distance vs Branching Layer

Figure 19: Performance and representation analysis. The first two panels compare behavioral metrics under different difficulty levels, while the third panel shows the relationship between input embedding L2 distance and the branching layer in the Concept Hierarchy.

ICLR 2027 PAPER CHECKLIST

1. **Claims.**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: Yes.
   Justification: The abstract and introduction list four contributions (the full-sequence flow framework; the spectral dynamics, including the deep-layer convergence theorem under a growth gap; LRH and RM as shared and value-specific turning regimes with the tracing procedure and Concept Hierarchies; the experiments), each of which corresponds to a section of the paper, and every number quoted in the abstract appears in Section 6.

2. **Limitations.**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: Yes.
   Justification: Section K measures the accuracy of the linear response, Section I.2 the drift of the counterfactual direction after branching, and Section H.3 the sensitivity to k and τ.

3. **Theory assumptions and proofs.**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: Yes.
   Justification: Section A states the standing assumptions and proves every result used in the main text (Propositions A.1, A.2, A.4, A.6 and A.7, Theorem A.3, and Lemma A.5 and the stretch–turn decomposition); Section B proves Theorems B.1 and 4.3, and Section D proves the discrete-profile statements.

4. **Experimental result reproducibility.**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results?
   Answer: Yes.
   Justification: The tracing procedure (Section C), the selective-expansion protocol with all prompts and statements (Section E), the probe protocol with prompts, countries and estimators (Section F), the validation set of pairs (Section I.2), the multi-valued concepts with their answer and neutral contexts (Section J) and the steering pipeline (Algorithm 1, Section N) are specified.

5. **Open access to data and code.**
   Question: Does the paper provide open access to the data and code, with sufficient information to support reproducibility?
   Answer: Partially.
   Justification: All benchmarks and checkpoints are public. Code for the tracing procedure, the probes, and the steering pipeline will be released upon publication.

6. **Experimental setting / details.**
   Question: Are the experimental settings, hyper-parameters, and dataset versions fully documented?
   Answer: Yes.
   Justification: Tracing hyper-parameters ($k = 10$, $\tau = 0.9$) with sensitivity analyses in Section H.3; probe settings ($m = 128$, $k = 8$, whitening, templates) in Sections F and O; selective-expansion settings in Section E.

7. **Experiment statistical significance.**
   Question: Does the paper report error bars suitably and correctly? Or does it describe the statistical significance of the experiments?
   Answer: Yes.
   Justification: Figure 2 reports mean $\pm$ s.d. over 16 prompts; Figure 7 and the text report 95% bootstrap intervals over statements; the branching-layer validation reports Spearman correlations with p-values over 32 pairs (Section I.2); Table 5 reports p-values. Concept Hierarchies are deterministic given the model and prompt.

8. **Experiments compute resources.**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type, memory, time) needed to reproduce the experiments?
   Answer: Yes.

Justification: The cost of the probes is given in forward-mode passes and wall-clock time per input for the 0.5B model (Section F); the 7B experiments run on a single GPU in bfloat16 (Section O).

9. **Code of ethics.**
   Question: Does the research conducted in the paper conform, in every aspect, with the ICLR Code of Ethics?
   Answer: Yes.
   Justification: Only public benchmarks and open-weight checkpoints are used; no human subjects are involved.

10. **Broader impacts.**
    Question: Does the paper discuss both potential positive societal outcomes and negative societal impacts of the work performed?
    Answer: Yes.
    Justification: The Ethics and Impact statements discuss the dual-use nature of interpretability and steering tools.

11. **Safeguards.**
    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?
    Answer: N/A.
    Justification: The paper releases no new dataset or model checkpoint.

12. **Licenses for existing assets.**
    Question: Are the creators or original owners of assets, used in the paper, properly credited and the license and terms of use explicitly mentioned and properly respected?
    Answer: Yes.
    Justification: All datasets and model checkpoints are cited with their original publications and used under their public licences.

13. **New assets.**
    Question: Are new assets introduced in the paper well documented and the documentation provided alongside the assets?
    Answer: N/A.
    Justification: The paper introduces no new dataset or model.

14. **Crowdsourcing and research with human subjects.**
    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable?
    Answer: N/A.
    Justification: No crowdsourcing or human-subjects research.

15. **Institutional review board (IRB) approvals.**
    Question: Does the paper describe potential risks incurred by study participants and whether IRB approvals were obtained?
    Answer: N/A.
    Justification: No human subjects.

16. **Declaration of LLM usage.**
    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods?
    Answer: Yes.
    Justification: An LLM is used as a component of the method in Section N (Automated Concept Extraction) to generate counterfactual pairs; this is described there and in the LLM-usage statement. Writing assistance is disclosed in the AI use statement.