CONCEPT GEOMETRY IN LANGUAGE MODELS: CONES, EDITS, AND COMPOSITIONAL PATHS

Anonymous authors
Paper under double-blind review

ABSTRACT

A language model is commonly thought to represent a concept as a direction in its hidden states, found by averaging how the state changes when one word of a sentence is swapped. The average hides whether the individual changes agree, and it says nothing about several words changing at once. We keep every change. Repeating a counterfactual edit, such as a change of tense or a synonym for an adjective, across contexts and across the words that carry it gives a cloud of vectors shaped like a cone, with an axis and a width; comparing the axes of different words for the same edit with those of random words in the same position tells whether the model treats the edit as one change or as several. Because an edit's axis barely depends on the sentence it acts on, the sentences reachable by several edits form a lattice, and their combined effect divides among the edits by the Shapley value, each edit's contribution averaged over the orders in which the edits can be applied. The lattice grows exponentially with the number of edits, yet a single path that applies the edits one at a time already estimates this division: apart from sampling noise, its error is the interaction between edits, which averages out over a few paths. Across five models from three families, synonyms of an adjective scatter more than random adjectives in all 15 combinations of model and sentence, while verbs of destruction align more than random verbs in 14. Shares vary little with the order of the edits (standard deviation at most 0.05 over all 360 orders of six edits). On a lattice of $65{,}536$ sentences, at most eight paths, visiting 0.2% of the sentences, recover every share to within 0.02, and the shares predict which edit's removal most weakens a combined steering vector (Spearman 0.78).

1 INTRODUCTION

Consider the sentence *The honest judge convicts the felon* and three changes to it: the tense of the verb, *honest* replaced by *upright*, and *judge* replaced by *algorithm*. Each is a counterfactual edit in the sense of Pearl (2009): one word changes and the rest of the sentence stays fixed. A language model answers each edit with a change in its residual stream, and repeating the edit in 48 contexts and with seven words that realise it yields a cloud of such changes. We ask how tight this cloud is, how the clouds of different edits relate to each other, and, when several edits are applied together, how their combined effect divides among them.

The usual summary of an edit is its average. Under the linear representation hypothesis a concept is a direction in the residual stream (Park et al., 2024; Mikolov et al., 2013), obtained by averaging the difference of counterfactual pairs over contexts: representation engineering reads and steers along it (Zou et al., 2023; Turner et al., 2023), function vectors transfer it to new inputs (Todd et al., 2024), and linear probes recover it for truth, sentiment and refusal (Marks & Tegmark, 2024; Tigges et al., 2023; Arditi et al., 2024). An average is a faithful summary when the individual responses agree, and they need not: activations superpose many features (Elhage et al., 2022; Bricken et al., 2023), some concepts occupy more than one dimension (Engels et al., 2025), and the reliability of a steering vector depends on the input (Tan et al., 2024). We therefore keep the whole cloud and measure how much of it lies along its mean direction. Because residual streams are anisotropic (Ethayarajh, 2019), any seven words placed in the same position of a sentence already share part of their direction, so we calibrate against random words in that position. Once concepts are regions rather than points (Vendrov et al., 2016; Rocchi-Henry et al., 2025; Park et al., 2025), two edits can be compared by

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.

their clouds, and a further question arises: whether two edits made together produce the sum of their separate effects, the question that steering-vector arithmetic asks empirically (Turner et al., 2023).

Interventional interpretability usually changes one variable at a time: activation patching and causal abstraction localise a behaviour through a single swap (Meng et al., 2022; Wang et al., 2022; Geiger et al., 2021; 2023). We treat several edits applied together. If each edit moves the representation in nearly the same direction whichever sentence it acts on, the sentences reachable by K edits form a lattice of up to 2^K sentences whose edges are edits, and the composite effect divides among the edits as the Shapley value of a game on this lattice. The lattice need not be visited in full: a walker, which applies the edits one at a time along one path over fixed contexts, is one sample of the permutation estimator of this Shapley value. On a lattice of $65{,}536$ sentences, measured exhaustively once, a handful of walkers recovers every share, and the shares predict the behavioural effect of removing an edit from a composite steering vector.

Contributions.

1. **The geometry of an edit** (§2): the effect cloud, its ruler, concentration and coherence, with the cone as the envelope of the cloud, calibrated against an isotropic and a same-position baseline.
2. **The concept network and its walker** (§3): the share of an edit as a precedence-constrained Shapley value of a projection game, the walker as its single-permutation estimator, and a finite-sample bound in which the interaction between edits is the only term that does not vanish with more contexts.
3. **Measurements** (§4): seven edit types on five models from three families and three base sentences, all admissible orderings of fifteen six-edit lattices, an exhaustive sixteen-edit lattice, and a behavioural test of the shares by steering.

2 The Concept Geometry

2.1 From a concept to an effect cloud

Fig. 1 shows what we measure. A concept such as the tense of a verb or the meaning of an adjective has a part that is shared wherever it appears (something is common to "she gave an honest answer" and "honesty is the best policy") and a part contributed by the surrounding sentence. We model the first part as a direction and the second as a displacement orthogonal to it, so that the vectors a concept can produce fill a cone: an axis, together with the directions within some angle of it.

Definition 2.1 (Cone). *A concept is represented by a convex cone $\mathcal{K} \subseteq \mathbb{R}^d$ and a map $\Phi : \mathcal{C} \to \mathcal{K}$ that sends each context $\mathbf{c} \in \mathcal{C}$ to the vector the concept contributes under $\mathbf{c}$. A token in context activates many concepts at once (Elhage et al., 2022), so the state of a sentence A with active concepts N_A is a sum of one vector from each cone,*

$$\mathbf{e}(\mathbf{c}) \;=\; \sum_{n \in N_A} \Phi_n(\mathbf{c}), \qquad \Phi_n(\mathbf{c}) \in \mathcal{K}_n. \tag{1}$$

The model exposes only this sum. Two sentences that differ in one edit share all other cones, so their difference cancels them and isolates the edit.

Definition 2.2 (Edit, realisation, effect cloud). *An* edit $\boldsymbol{\delta}$ *is a family of token replacements applied to a sentence* $\mathbf{x}$ *that change one attribute of it (its tense, its subject, its polarity). A* realisation $v \in \mathcal{R}(\boldsymbol{\delta})$ *is one member of that family (*judge→algorithm, judge→program, *. . .). For a context* $\mathbf{c} \in \mathcal{C}$ *the* effect vector *of realisation* v *is*

$$\boldsymbol{\Delta}_v(\mathbf{c}) \;:=\; \mathbf{e}(\boldsymbol{\delta}_v\mathbf{x},\, \mathbf{c}) \;-\; \mathbf{e}(\mathbf{x},\, \mathbf{c}) \;\in\; \mathbb{R}^d, \tag{2}$$

and the effect cloud *of* $\boldsymbol{\delta}$ *at* $\mathbf{x}$ *is* $\{\boldsymbol{\Delta}_v(\mathbf{c}) : v \in \mathcal{R}(\boldsymbol{\delta}),\ \mathbf{c} \in \mathcal{C}\}$.

2.2 Ruler and concentration

Of each cloud we estimate the axis and the opening of its cone.

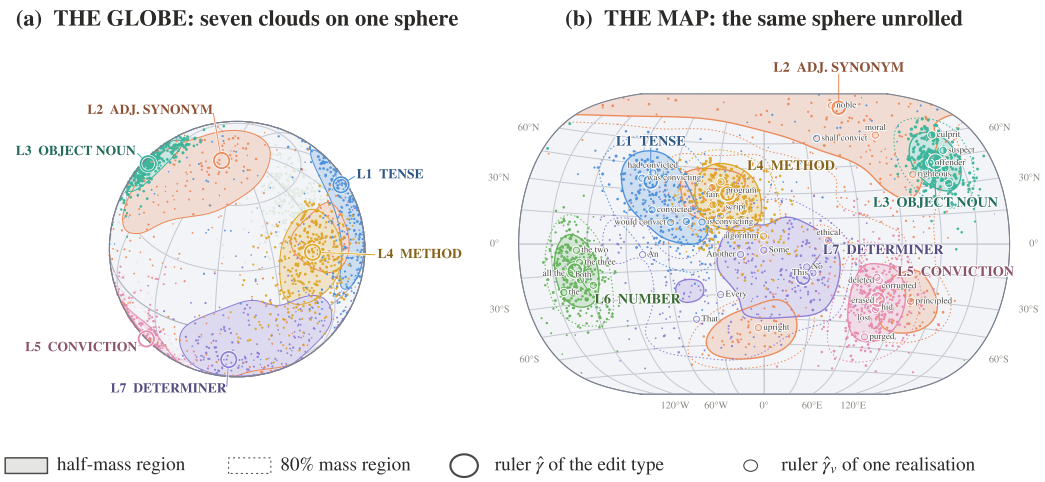

Figure 1: **The concept geometry. (a)** All 2,352 unit effect directions of seven edit types (judicial sentence, Qwen2.5-7B) under one uncentered 3-PCA; point size is the energy captured by the three axes. Large ring: ruler of each edit type; small rings: rulers of its seven realisations; shaded region: the densest directions holding 50% (dark) and 80% (light) of the cloud's energy. **(b)** The same sphere unrolled.

Definition 2.3 (Ruler, zero space, concentration)**.** *Let* $\mathbf{\Delta}$ *range over an effect cloud with* $\mathbb{E}\mathbf{\Delta} \neq \mathbf{0}$. *The* ruler *of the edit is* $\hat{\boldsymbol{\gamma}} := \mathbb{E}\mathbf{\Delta}/\|\mathbb{E}\mathbf{\Delta}\|$, *and its* zero space *is the hyperplane* $Z := \hat{\boldsymbol{\gamma}}^{\perp}$. *The* concentration *of the cloud is* $s := \mathbb{E}\cos(\mathbf{\Delta}, \hat{\boldsymbol{\gamma}})$, *and its opening angle is* $\arccos s$.

The ruler is the axis of the cone of Definition 2.1, $\arccos s$ is its opening, and the component of an effect vector in Z is the part that the context contributes. For reference, a random unit vector in $\mathbb{R}^d$ has a cosine of order $1/\sqrt{d}$ with any fixed ruler, from 0.017 for $d = 3{,}584$ to 0.033 for $d = 896$, and the mean direction of n random vectors has concentration about $1/\sqrt{n}$, which is 0.14 for $n = 48$. A concentration close to one therefore means that the edit moves the state along one direction whatever the context.

2.3 Realisations: one cone or several

A cloud can scatter for two reasons: the same swap may read differently in different contexts, or different words that realise the same edit may move the state in different directions. We separate the two.

Definition 2.4 (Context concentration, coherence)**.** *For an edit with realisations* $\mathcal{R}(\boldsymbol{\delta})$, *let* $\hat{\boldsymbol{\gamma}}_v$ *and* s_v *be the ruler and concentration of realisation* v *over contexts alone. The* context concentration *of the edit is* $\bar{s}_{\mathrm{ctx}} := \mathbb{E}_v s_v$, *and its* coherence *is the mean cosine between the rulers of different realisations,*

$$\kappa \;:=\; \mathbb{E}_{v \neq v'} \langle \hat{\boldsymbol{\gamma}}_v, \hat{\boldsymbol{\gamma}}_{v'} \rangle. \tag{3}$$

Each realisation has its own cone with axis $\hat{\boldsymbol{\gamma}}_v$. The edit forms one cone when these axes coincide ($\kappa \to 1$). When κ is small, each word moves the state consistently across contexts but different words move it in different directions, and the edit consists of several cones. If the synonyms of *honest* each carry a shade of their own (*upright*, *fair*, *noble*), their rulers part and κ falls; if the model reads *deleted*, *erased* and *purged* as the same act, their rulers merge and κ rises.

A same-position baseline. The isotropic value $\kappa \approx 0$ is too lenient a reference. Residual streams are anisotropic (Ethayarajh, 2019), and words swapped into one position share positional and frequency components, so seven unrelated words in the same position already share part of a ruler. For the adjective, subject, object and verb positions we therefore take 40 frequent unrelated words of the same part of speech (Appendix G) and compare each seven-word family with 500 random seven-word subsets of this pool, which gives κ_{null}. Against this baseline, coherence tells whether the model treats a family of replacements as one change or as several.

3 THE CONCEPT NETWORK AND ITS WALKER

3.1 LINEAR RULER

To pass from single edits to several, we need one property of an edit's direction: it does not depend on the sentence the edit acts on.

Assumption 3.1 (Linear ruler)**.** *For an edit $\boldsymbol{\delta}$ and any two sentences $\mathbf{x}, \mathbf{x}'$ to which it applies, the rulers of its effect clouds at $\mathbf{x}$ and at $\mathbf{x}'$ coincide, and its mean effect $\boldsymbol{\mu}(\boldsymbol{\delta}) := \mathbb{E}_{\mathbf{c}}\boldsymbol{\Delta}(\mathbf{c})$ does not depend on $\mathbf{x}$.*

This is the linear representation hypothesis (Park et al., 2024) stated for an edit, and the cone model (1) motivates it: a sentence is a sum of cone responses, an edit switches one or two of them, and its effect depends only on those. The cone adds the scatter to the hypothesis: the ruler is shared, and each application of the edit leans away from it by an amount that s records. The assumption has a direction part and a magnitude part. §4.1 tests the direction through context consistency, and §4.2 tests both at every node of a lattice. The shares defined below, and the walker that estimates them, remain well defined when the magnitude part holds only approximately.

3.2 CONCEPT NETWORK AND SHARES

Fix a base sentence $\mathbf{x}_0$ and edits $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_K$. The sentences reachable from $\mathbf{x}_0$ by applying any admissible subset of the edits, ordered by the edits they contain, form the *concept network* (Fig. 2); its nodes are these sentences, and its edges are the single edits between them. A path from $\mathbf{x}_0$ to the fully edited sentence $\mathbf{x}^*$ applies the edits in one admissible ordering ($K!$ orderings without constraints; 360 in our six-edit lattices, where the action edit replaces the past-tense verb and therefore follows the tense edit). Applying all K edits is the *composite edit* $\boldsymbol{\delta}^*$, with effect vectors $\boldsymbol{\Delta}^*(\mathbf{c})$ and mean effect $\boldsymbol{\mu}^*$.

We ask how much of $\boldsymbol{\mu}^*$ each edit accounts for. Under Assumption 3.1 every edge of edit k carries the same mean effect $\boldsymbol{\mu}(\boldsymbol{\delta}_k)$, any path adds the edits up to $\sum_k \boldsymbol{\mu}(\boldsymbol{\delta}_k) = \boldsymbol{\mu}^*$, and the *share* of edit k is its part of this sum along the composite ruler,

$$w_k \;:=\; \frac{\langle \boldsymbol{\mu}(\boldsymbol{\delta}_k),\, \boldsymbol{\mu}^* \rangle}{\|\boldsymbol{\mu}^*\|^2}, \qquad \sum_{k=1}^{K} w_k \;=\; 1. \tag{4}$$

An edit with a large share carries much of the composite edit along its direction, and a negative share means that the edit works against it. The share is a property of the population: it compares the cloud of $\mathbf{x}_0$ with the cloud of $\mathbf{x}^*$ over the same contexts.

Shares as Shapley values. Without Assumption 3.1, an edit's mean effect depends on the node at which it is applied, and (4) must say which node is meant. The Shapley value removes this choice by averaging over all of them. Let $\boldsymbol{m}_S := \mathbb{E}_{\mathbf{c}}\mathbf{e}(\mathbf{x}_S, \mathbf{c})$ be the mean state of the node that applies the edits in S, and define the *projection game* $v(S) := \langle \boldsymbol{m}_S - \boldsymbol{m}_\emptyset, \boldsymbol{\mu}^* \rangle / \|\boldsymbol{\mu}^*\|^2$: the value of a set of edits is how far it carries the base sentence along the composite ruler, with $v(\emptyset) = 0$ and $v(N) = 1$. The share of edit k is its Shapley value ϕ_k (Shapley, 1953), the marginal contribution $v(P \cup \{k\}) - v(P)$ averaged over admissible orderings, where P is the set of edits preceding k; with precedence constraints this is the value of Faigle & Kern (1992). The shares sum to one, and under Assumption 3.1 the game is additive and $\phi_k = w_k$.

Anchored shares and interaction. The cheapest approximation measures each edit once, at the base sentence (at its prerequisite's node if it has one). Writing $\boldsymbol{\mu}(\boldsymbol{\delta}_k)$ for this mean effect, (4) gives the *anchored* share a_k. The anchored effects miss the composite effect by the *interaction* $\boldsymbol{\epsilon} := \boldsymbol{\mu}^* - \sum_k \boldsymbol{\mu}(\boldsymbol{\delta}_k)$, and since a share is a projection on $\boldsymbol{\mu}^*$, only the component of $\boldsymbol{\epsilon}$ along $\boldsymbol{\mu}^*$ affects it: $W := \sum_k a_k = 1 - \langle \boldsymbol{\epsilon}, \boldsymbol{\mu}^* \rangle / \|\boldsymbol{\mu}^*\|^2$.

3.3 ONE WALKER FOR THE POPULATION

Exact Shapley shares need the mean state of every node in every context, 2^K sentences before realisations are counted. Anchored shares need one node per edit but carry the interaction. A walker

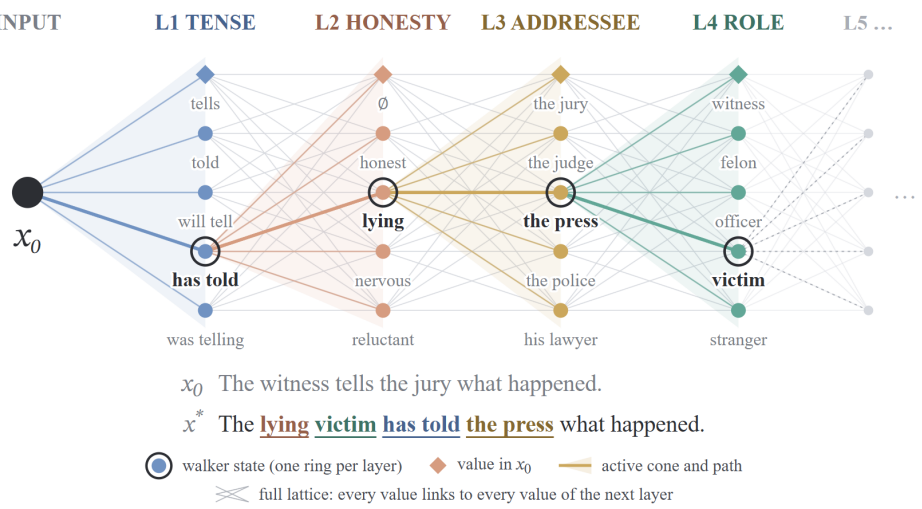

Figure 2: **The concept network and the walker** (schematic, four edit types). Columns are edit types, dots their values, grey lines the full lattice. The walker takes one edit per column, carrying its contexts along; rings are its states, the coloured band the cone of the edit it took. Because every edit's direction is nearly the same at every node (Assumption 3.1), one path estimates the shares of the whole lattice, and a few paths recover them closely (Proposition 3.3).

lies between the two: under Assumption 3.1 an edit's mean effect is the same at every node, so one path that takes the edits one at a time sees the mean effects that the whole lattice would.

Definition 3.2 (Walker). *A walker is a path* $\mathbf{x}_0 \to \mathbf{x}_1 \to \cdots \to \mathbf{x}_K = \mathbf{x}^*$ *in the lattice,* $\mathbf{x}_k = \boldsymbol{\delta}_{\pi(k)}(\mathbf{x}_{k-1})$ *for an admissible ordering* π*, evaluated over the same* T *contexts at every node. Its* k*-th* step *is* $\boldsymbol{\Delta}_k(\mathbf{c}) := \mathbf{e}(\mathbf{x}_k, \mathbf{c}) - \mathbf{e}(\mathbf{x}_{k-1}, \mathbf{c})$*, with mean* $\bar{\boldsymbol{\Delta}}_k$*, and its* readout *is*

$$\hat{w}_{\pi(k)} := \frac{\langle \bar{\boldsymbol{\Delta}}_k, \bar{\boldsymbol{\Delta}}^* \rangle}{\|\bar{\boldsymbol{\Delta}}^*\|^2}, \qquad \bar{\boldsymbol{\Delta}}^* := \sum_k \bar{\boldsymbol{\Delta}}_k. \tag{5}$$

Proposition 3.3 (One walker estimates the population). *For any walker and context,* $\sum_k \boldsymbol{\Delta}_k(\mathbf{c}) = \boldsymbol{\Delta}^*(\mathbf{c})$*, so the readout sums to one exactly. Let the contexts be i.i.d. and the steps jointly sub-Gaussian (they share their contexts and need not be independent), with means* $\boldsymbol{\nu}_{\pi,k}$*. Write* $\Sigma_k := \mathrm{Cov}(\boldsymbol{\Delta}_k)$*,* $\Sigma^* := \mathrm{Cov}(\boldsymbol{\Delta}^*)$*,* $\Sigma_{k*} := \mathrm{Cov}(\boldsymbol{\Delta}_k, \boldsymbol{\Delta}^*)$*,* σ^2 *for the largest operator norm of these covariances,* $w_{\pi,k} := \langle \boldsymbol{\nu}_{\pi,k}, \boldsymbol{\mu}^* \rangle / \|\boldsymbol{\mu}^*\|^2$ *and* $\tau := \mathrm{tr}\,\Sigma^*/(T\|\boldsymbol{\mu}^*\|^2)$*. Then, as* $\tau \to 0$*,*

$$\hat{w}_{\pi(k)} = w_{\pi,k} + b_{\pi,k} + O_P\Big(\frac{\sigma\,(1 + \|\boldsymbol{\nu}_{\pi,k}\|/\|\boldsymbol{\mu}^*\|)}{\|\boldsymbol{\mu}^*\|\sqrt{T}} + \frac{\sigma\sqrt{\mathrm{tr}\,\Sigma^*}}{T\|\boldsymbol{\mu}^*\|^2} + \tau^2\Big), \tag{6}$$

where $b_{\pi,k} = (\mathrm{tr}\,\Sigma_{k*} - w_{\pi,k}\,\mathrm{tr}\,\Sigma^*)/(T\|\boldsymbol{\mu}^*\|^2)$*,* $\sum_k b_{\pi,k} = 0$*,* $|b_{\pi,k}| \le \tau(\sqrt{\mathrm{tr}\,\Sigma_k/\,\mathrm{tr}\,\Sigma^*} + |w_{\pi,k}|)$*, and* $b_{\pi,k}$ *vanishes when both inner products of the readout are replaced by U-statistics over distinct contexts. Under Assumption 3.1,* $w_{\pi,k} = w_{\pi(k)}$ *for every ordering. In general,*

$$w_{\pi,k} = a_{\pi(k)} + \frac{\langle \boldsymbol{\epsilon}_{\pi,k}, \boldsymbol{\mu}^* \rangle}{\|\boldsymbol{\mu}^*\|^2}, \qquad \boldsymbol{\epsilon}_{\pi,k} := \boldsymbol{\nu}_{\pi,k} - \boldsymbol{\mu}(\boldsymbol{\delta}_{\pi(k)}), \qquad \mathbb{E}_\pi\, w_{\pi,k} = \phi_{\pi(k)}, \tag{7}$$

where $\boldsymbol{\epsilon}_{\pi,k}$*, the* interaction *of the step, is the part of the* k*-th step that the edit produces only after the edits preceding it in* π*. The interactions of a walker sum to* $\boldsymbol{\epsilon}$ *for every ordering; the ordering determines how this total is distributed over the steps.*

The proof is in Appendix A. The proposition splits the error of one walker into three parts. The *sampling* term involves only projections on the fixed vectors $\boldsymbol{\nu}_{\pi,k}$ and $\boldsymbol{\mu}^*$, so it does not grow with the dimension: σ plays the role of the opening of a step's cone and $\|\boldsymbol{\mu}^*\|$ the length of the composite ruler. The *bias* $b_{\pi,k}$ comes from the squared norms and is the only place where the dimension enters, through $\tau = d_{\mathrm{eff}}\,\sigma^2/(T\|\boldsymbol{\mu}^*\|^2)$ with the effective rank $d_{\mathrm{eff}} = \mathrm{tr}\,\Sigma^*/\|\Sigma^*\|_{\mathrm{op}}$ of the composite cloud;

Table 1: Coherence of four families against the same-position baseline over the 15 lattices: number of lattices on which κ lies above the 95th percentile, inside the 90% interval, or below the 5th percentile of the baseline, and mean [min, max] of $\kappa - \kappa_{\text{null}}$. Per edit type and model: Tables 5 and 6 and Fig. 8.

Family	above	within	below	$\kappa - \kappa_{\text{null}}$
adjective synonym	0	0	15	-0.12 [-0.19, -0.04]
object noun	1	3	11	-0.07 [-0.17, 0.04]
subject: human → machine	5	9	1	0.03 [-0.06, 0.08]
action	14	1	0	0.07 [0.03, 0.11]

on our lattices $d_{\text{eff}} \in [2.4, 6.2]$ and $\tau \leq 0.002$, and the U-statistic readout removes this term. The *interaction* does not shrink with T and is the only difference between one walker and the population. Because $w_{\pi,k}$ averages to $\phi_{\pi(k)}$ over orderings (7), one walker is one sample of the permutation estimator of the Shapley share (Castro et al., 2009), and the mean of m random walkers converges to ϕ at a cost of at most $m(K-1)+2$ nodes instead of 2^K. The sum-to-one property is an identity, so the evidence for Assumption 3.1 comes from the interaction, which §4.2 measures. For $K = 6$ the full lattice of 48 sentences serves as ground truth, and §4.3 determines how many walkers suffice for $K = 16$.

4 Experiments

Setup. We use five models from three families: Qwen2.5-0.5B, Qwen2.5-1.5B and Qwen2.5-7B (Qwen Team et al., 2024), SmolLM2-360M (Ben Allal et al., 2025) and Pythia-410m (Biderman et al., 2023). There are three base sentences in three domains, *The honest judge convicts the felon* (judicial), *The careful doctor treats the patient* (medical) and *The talented author writes the novel* (literary), each with the same seven edit types (tense, adjective synonym, object noun, subject human→machine, action, number, determiner), seven realisations per type, and 48 context sentences on the sentence's own theme (24 formal, 24 narrative; Appendix G). Each base sentence with its edits, read by one model, forms one *lattice*, 15 in all; each lattice has 2,352 effect vectors (seven edit types, seven realisations, 48 contexts). The representation $\mathbf{e}(\mathbf{x}, \mathbf{c})$ is the residual stream at depth $\lceil L/4 \rceil$ of an L-layer model (layer 6 of 24 on Qwen2.5-0.5B and Pythia-410m, 7 of 28 on Qwen2.5-1.5B and Qwen2.5-7B, 8 of 32 on SmolLM2-360M), averaged over the tokens of the edited sentence, which follows its context sentence. With the context first, every token of the edited sentence attends to it, and no edit touches the first input position, which carries the attention-sink activation (Sun et al., 2024; Xiao et al., 2024); averaging over the sentence treats all edit positions alike. Depth $\lceil L/4 \rceil$ is the shallowest layer at which the seven edit types occupy distinct directions, while edits compose most additively near the embedding; Appendix C reports every statistic at every layer and Appendix D compares alternative designs.

4.1 The geometry: cones and their coherence

Three findings hold on every model and base sentence (Table 1; per edit type in Tables 5 and 6).

Every realisation is context-consistent. The concentration of a single realisation over the 48 contexts is $\bar{s}_{\text{ctx}} = 0.93$–$0.99$, against 0.14 for 48 random vectors. A one-word swap moves the state in the same direction across contexts, whatever the word and whatever the domain of the sentence.

Edit types have distinct rulers. The largest cosine between the rulers of two different edit types is 0.24–0.43 per lattice, so the edges of different edits in a lattice carry distinct directions.

Coherence is a property of the family of realisations. Against the isotropic baseline every family would appear to form a single cone: seven random adjectives in the adjective position reach $\kappa = 0.48$–0.57, and seven random nouns in a noun position 0.50–0.75. The same-position baseline separates the families. The synonyms of the adjective are less coherent than random adjectives on all 15 lattices (mean $\kappa - \kappa_{\text{null}} = -0.12$; for *honest* on Qwen2.5-7B, 0.39 against 0.52): each synonym departs from the base word in a direction of its own. Object nouns behave alike on 11 lattices. Machine

nouns replacing the human subject match the baseline on 9 (mean difference $+0.03$), so their shared ruler is largely that of the subject position. Verbs of destruction (*deleted*, *erased*, *purged*, ...) are more coherent than random verbs on 14 of 15 lattices (mean difference $+0.07$): the model represents them as one change. The pattern holds on all five models and is stable when each family is restricted to realisations that keep the token count (every κ changes by at most 0.11), when the grand mean of all effect vectors is projected out, and when each coordinate is standardised (Appendix D). Coherence thus separates families that the model encodes as a single change from families whose members it encodes one by one, and the same-position baseline gives the reference for any new family.

4.2 Rulers across the lattice, walkers across orderings

Assumption 3.1 and Proposition 3.3 predict that an edit's mean effect keeps its direction at every node, that the interaction has a small component along $\boldsymbol{\mu}^*$, and that walkers applying the same replacements in different orders read the same shares. To test all three, we take six of the seven edit types with one realisation each (for the judicial sentence *convicts→convicted*, *honest→upright*, *felon→culprit*, *judge→algorithm*, *convicted→deleted*, *The→An*; the number edit is left out because a plural object changes the tokens that later edits act on). The resulting six-edit lattice has 48 sentences, which we measure in full; its exact Shapley shares serve as ground truth for the walkers along all 360 admissible orderings.

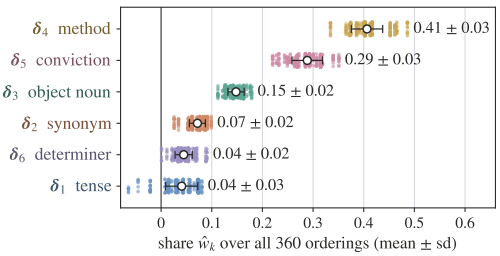

Figure 3: **All 360 walkers of one lattice** (Qwen2.5-7B, judicial). Each dot is the share one ordering reads for one edit; circles and bars: the mean over orderings (the Shapley share) and its standard deviation.

Direction and magnitude. An edit's mean effect keeps its direction across the nodes: its mean cross-node cosine is 0.87–0.95 (Table 2). Magnitudes combine with an interaction: the six single-edit effects at the base sentence reproduce the composite effect up to a relative interaction $\iota = 0.29$–0.54, with pairwise interactions $\iota_2 = 0.13$–0.21. Most of the interaction lies in the zero space of the composite edit. Its component along $\boldsymbol{\mu}^*$ (cosine -0.38 to 0.52, median $|\cos| = 0.21$, against an isotropic standard deviation of 0.017–0.033) moves the anchored shares, whose sum W lies between 0.85 and 1.16, by up to 0.07 from the Shapley shares. This is the component that the Shapley share absorbs by averaging over orderings, and the reason to estimate shares with walkers rather than at a single node.

The walker is path-independent. Across the 360 orderings the standard deviation of a share is at most 0.05 in every lattice (Table 2, Fig. 3; two walkers are traced in Appendix B.1). Sampling over contexts contributes an order of magnitude less: the bootstrap standard error of one walker's share is at most 0.005, disjoint halves of the contexts agree to within 0.009, and the direct and U-statistic readouts of Proposition 3.3 differ by less than 0.001. The spread over orderings is therefore the ordering-dependent part of the interaction, and it is small relative to the shares. The largest Shapley share belongs to the action edit (*convicted→deleted*) in 11 of the 15 lattices, to the subject edit in 3 and to the adjective synonym in 1. Where action and subject lie within 0.05 of each other (10 lattices), single orderings may name either as the leader (agree $= 0.56$–1.00); averaging over orderings resolves every share, including the smallest (tense and determiner, 0.02–0.15).

Depth. Additivity is highest near the embedding and decreases with depth on every model (Fig. 7 in Appendix C): ι grows from 0.05–0.36 at layer 1 to 0.47–0.67 at layer $L-1$, and the largest spread of a share over orderings from 0.05 to 0.08. The first quarter of the model already separates the edit types into distinct cones while their composition is still close to additive, and this is where the walker reads the lattice.

4.3 A lattice too large to visit

To test the walker where exhaustive evaluation is expensive, we use one longer sentence with sixteen independent one-word swaps, *Yesterday the honest judge in the small northern town quickly convicted*

Table 2: The walker in every lattice. Shares: the Shapley share ϕ_k (mean over the 360 admissible orderings) $\pm$ the standard deviation of one walker's readout over orderings. agree: fraction of orderings whose largest share is the modal edit; $W = \sum_k a_k$; $\iota = \|\boldsymbol{\epsilon}\|/\|\boldsymbol{\mu}^*\|$; cos: cosine of $\boldsymbol{\epsilon}$ and $\boldsymbol{\mu}^*$ (isotropic standard deviation 0.017–0.033); ι_2: mean pairwise interaction $\|\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|/\|\boldsymbol{\mu}_{ij}\|$; $\bar{\rho}$: mean cross-node cosine of an edit's mean effect; $|\cos|$: largest cosine between rulers of different edit types. *How to read.* Green shading is proportional to ϕ_k; the largest share of each row is **bold**, and shares not separated from it ($\phi_{\max} - \phi_k \le \sigma_{\max} + \sigma_k$) are underlined. Orange: agree < 0.70, the largest share is not stable across orderings. In the W and cos columns, purple marks $\cos > 0$ ($W < 1$) and blue marks $\cos < 0$ ($W > 1$), with intensity proportional to $|\cos|$ and $|1 - W|$; gray: $|\cos| < 0.10$, within three isotropic standard deviations. In every row $W = 1 - \iota \cos$ up to rounding.

	Shapley share ϕ_k						robustness	interaction with μ^*			consistency		
Model / base sentence	tense	synonym	object noun	subject	action	determiner	agree	W	ι	cos	ι_2	$\bar{\rho}$	$\lvert\cos\rvert$
Qwen2.5-0.5B, judicial	0.08 ± 0.02	0.13 ± 0.01	0.11 ± 0.02	0.25 ± 0.01	$\mathbf{0.33} \pm 0.01$	0.09 ± 0.01	**1.00**	0.93	0.32	0.23	0.17	0.93	0.33
Qwen2.5-0.5B, medical	0.05 ± 0.01	0.12 ± 0.01	0.13 ± 0.00	0.29 ± 0.03	$\mathbf{0.31} \pm 0.02$	0.10 ± 0.01	0.56	1.08	0.39	−0.20	0.15	0.92	0.35
Qwen2.5-0.5B, literary	0.02 ± 0.00	0.08 ± 0.01	0.07 ± 0.01	$\mathbf{0.37} \pm 0.03$	$\mathbf{0.37} \pm 0.03$	0.10 ± 0.01	0.67	1.03	0.36	−0.10	0.13	0.93	0.36
Qwen2.5-1.5B, judicial	0.10 ± 0.01	0.17 ± 0.01	0.09 ± 0.01	0.20 ± 0.01	$\mathbf{0.37} \pm 0.01$	0.06 ± 0.01	**1.00**	0.94	0.36	0.16	0.18	0.92	0.26
Qwen2.5-1.5B, medical	0.03 ± 0.01	0.12 ± 0.01	0.10 ± 0.00	0.26 ± 0.02	$\mathbf{0.38} \pm 0.02$	0.11 ± 0.01	**1.00**	1.09	0.43	−0.21	0.15	0.91	0.34
Qwen2.5-1.5B, literary	0.05 ± 0.02	0.10 ± 0.01	0.07 ± 0.00	0.30 ± 0.02	$\mathbf{0.39} \pm 0.01$	0.09 ± 0.01	**1.00**	1.09	0.37	−0.25	0.15	0.92	0.38
Qwen2.5-7B, judicial	0.12 ± 0.02	0.20 ± 0.01	0.13 ± 0.01	0.22 ± 0.03	$\mathbf{0.25} \pm 0.01$	0.08 ± 0.00	0.73	1.02	0.38	−0.05	0.19	0.90	0.24
Qwen2.5-7B, medical	0.04 ± 0.01	0.17 ± 0.01	0.11 ± 0.00	$\mathbf{0.29} \pm 0.03$	0.25 ± 0.03	0.14 ± 0.01	0.67	1.11	0.54	−0.20	0.21	0.87	0.28
Qwen2.5-7B, literary	0.06 ± 0.01	0.15 ± 0.00	0.08 ± 0.00	$\mathbf{0.29} \pm 0.02$	0.26 ± 0.01	0.15 ± 0.01	0.68	1.09	0.41	−0.22	0.17	0.91	0.27
SmolLM2-360M, judicial	0.10 ± 0.03	0.16 ± 0.02	0.22 ± 0.05	0.21 ± 0.01	$\mathbf{0.25} \pm 0.01$	0.06 ± 0.01	0.67	0.85	0.29	0.52	0.16	0.95	0.42
SmolLM2-360M, medical	0.04 ± 0.01	0.14 ± 0.01	0.14 ± 0.00	$\mathbf{0.28} \pm 0.01$	0.27 ± 0.01	0.13 ± 0.01	0.67	1.04	0.31	−0.12	0.13	0.94	0.39
SmolLM2-360M, literary	0.04 ± 0.02	0.11 ± 0.01	0.11 ± 0.00	0.32 ± 0.03	$\mathbf{0.33} \pm 0.01$	0.09 ± 0.01	0.56	1.10	0.42	−0.24	0.15	0.93	0.43
Pythia-410m, judicial	0.09 ± 0.02	0.18 ± 0.01	0.14 ± 0.02	0.22 ± 0.01	$\mathbf{0.28} \pm 0.01$	0.09 ± 0.01	**1.00**	0.96	0.32	0.12	0.16	0.93	0.30
Pythia-410m, medical	0.06 ± 0.00	$\mathbf{0.28} \pm 0.04$	0.09 ± 0.01	0.21 ± 0.03	0.23 ± 0.03	0.13 ± 0.01	0.69	1.16	0.42	−0.38	0.14	0.93	0.29
Pythia-410m, literary	0.06 ± 0.01	0.13 ± 0.02	0.08 ± 0.01	0.28 ± 0.02	$\mathbf{0.32} \pm 0.01$	0.12 ± 0.02	0.83	1.09	0.36	−0.25	0.14	0.93	0.32

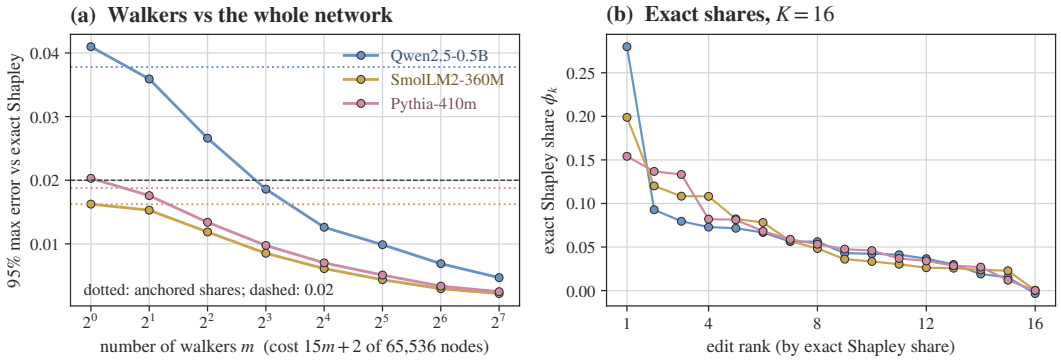

Figure 4: **Sixteen edits, 65,536 sentences. (a)** 95th percentile of the largest share error of the mean of m random walkers against the exact Shapley shares of the whole lattice; dotted: anchored shares. **(b)** Exact Shapley shares, sorted.

the young felon of theft after a long and tiring trial at the old courthouse (swaps in Appendix G), and measure all $2^{16} = 65{,}536$ of its variants after the 48 judicial contexts on three models. The exhaustive lattice gives exact Shapley shares against which the walkers are compared (Fig. 4; Table 4 in Appendix E).

The interaction behaves as in the six-edit lattices ($\iota = 0.23$–0.35, $\cos(\boldsymbol{\epsilon}, \boldsymbol{\mu}^*) = 0.23$–$0.36$, $W = 0.88$–0.95), and the mean of m random walkers converges to the exact shares at the permutation-sampling rate. For a 95th-percentile error below 0.02 on every one of the sixteen shares, SmolLM2-360M needs one walker (17 nodes), Pythia-410m two (32 nodes) and Qwen2.5-0.5B eight (122 nodes), at most 0.19% of the lattice. On all three models the subject edit *judge→algorithm* carries the largest share.

4.4 What the shares predict

We add steering vectors to the residual stream of the base sentence at depth $\lceil L/4 \rceil$, on the tokens of the sentence, and measure the fraction $g(\boldsymbol{h})$ of the composite edit's change in next-token log-probabilities that a vector $\boldsymbol{h}$ reproduces (Appendix B.1). The composite vector $\boldsymbol{\mu}^*$ reproduces 0.12–0.35 of this change and the sum of the anchored effects 0.15–0.42, against at most 0.10 for random directions of

the same norms, and steering effects add (median difference 0.04 between the summed vector and the sum of single effects). Removing one edit's vector from the sum costs $g(\sum_j \boldsymbol{\mu}_j) - g(\sum_{j \neq k} \boldsymbol{\mu}_j)$, and this cost follows the Shapley share with Spearman 0.78 over the 90 edits of the fifteen lattices (median 0.83 per lattice).

The share carries more than the size of an edit. It correlates with the normalised norm of the anchored effect (Spearman 0.94), as a projection must, yet size alone predicts the drop-one cost less well (0.74); the share remains predictive once size is controlled (partial rank correlation 0.40), while size adds nothing once the share is known (0.01). The share is also more than lexical similarity: the same game computed on the input embeddings correlates with it at 0.59 and names the same leading edit in only 4 of the 15 lattices. The share thus measures what each edit contributes to the model's next-token prediction.

5 Related Work

The geometry of representations. Linear semantic directions go back to word embeddings (Mikolov et al., 2013; Pennington et al., 2014) and were revived for residual streams under the superposition hypothesis (Elhage et al., 2022), with geometric foundations (Park et al., 2024; 2025), extracted concept directions (Nanda et al., 2023; Marks & Tegmark, 2024; Tigges et al., 2023), multi-dimensional features (Engels et al., 2025) and sparse dictionaries (Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2024; Lieberum et al., 2024). Cones and regions appear in order- and region-based embeddings (Vendrov et al., 2016; Bordes et al., 2013; Yang et al., 2015). Closest to us, Rocchi-Henry et al. (2025) unify concept bottleneck models and sparse autoencoders through concept cones, defined as the nonnegative span of learned directions; our cone is the envelope of a measured cloud, and its opening is a number. Tan et al. (2024) find that a steering vector's effect varies with the prompt; concentration and coherence split this variation into a part from contexts and a part from realisations. Anisotropy (Ethayarajh, 2019; Cai et al., 2021) and attention sinks (Sun et al., 2024; Xiao et al., 2024) motivate our same-position baseline and the placement of the context.

Counterfactual edits, composition and Shapley values. Counterfactual prompts (Wu et al., 2021; Ross et al., 2021) and activation patching (Meng et al., 2022; Wang et al., 2022) use single interventions, and representation engineering steers along one ruler (Zou et al., 2023; Turner et al., 2023; Todd et al., 2024). The Shapley value averages marginal contributions over orderings (Shapley, 1953; Lundberg & Lee, 2017); interaction indices and Harsanyi dividends capture departures from additivity (Grabisch & Roubens, 1999; Sundararajan et al., 2020; Tsai et al., 2023; Harsanyi, 1963); permission structures and precedence orders restrict the orderings (Gilles et al., 1992; Faigle & Kern, 1992); and sampling orderings is the standard estimator (Castro et al., 2009). We apply this framework to a game defined on sentences, where permutation sampling becomes a path through them, and add the projection game, a vector-valued interaction (7) that shows in which direction additivity departs from exactness, an error analysis suited to the high-dimensional residual stream, and an exhaustive measurement of how many walkers a lattice of $65{,}536$ sentences requires.

6 Conclusion

Limitations. The models range from 360M to 7B parameters, and the sixteen-edit lattice could also be extended to larger models for scalability. The main analyses read one relative depth (other layers in Appendix C), and steering is scored on next-token distributions rather than on generated text. The contexts of each theme are written for the study rather than sampled from a corpus, which points to challenging questions such as whether the rulers and shares stay stable when contexts are drawn from naturally occurring text of varying length, style and topic.

To conclude, we describe a counterfactual edit not by its average effect but by the whole cloud of effects: its direction, how tightly the cloud gathers around it, and whether the words that realise the edit share that direction. When several edits are applied together, their combined effect divides among them as a Shapley value, and a single path through the edits already estimates this division. Across five models, synonyms scatter while verbs of destruction align, the division barely depends on the order of the edits, a few paths recover it on a lattice of $65{,}536$ sentences, and it predicts which edit matters most for steering. The method needs only a base text, a family of edits and a model, and extends directly to larger models and free-form text.

AI USE STATEMENT

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. We also used AI assistance for minor figure edits. 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 generate synthetic data; 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

This work analyses public open-weight models (Qwen2.5, SmolLM2, Pythia; Apache-2.0) on synthetic sentences written by the authors and listed in Appendix G. It involves no human subjects and no personal data. The scenarios, including edits that replace a human agent with a machine, are probing templates and make no claims about real people or professions. Like other interpretability and steering tools, the method is dual-use: identifying attribute directions could help steer a model towards undesired outputs as well as audit it. It adds no manipulation capability beyond existing steering techniques, and we release no modified models.

REPRODUCIBILITY STATEMENT

Models. All five models are public checkpoints: Qwen2.5-0.5B, Qwen2.5-1.5B and Qwen2.5-7B (Qwen Team et al., 2024), SmolLM2-360M (Ben Allal et al., 2025) and Pythia-410m (Biderman et al., 2023). Models up to 0.5B parameters run in `float32` and larger models in `bfloat16`; all statistics are computed in `float32` in the full d dimensions.

Data. All inputs are listed verbatim in Appendix G: the three base sentences, the seven edit types with their seven realisations each, the 48 context sentences per theme (24 formal, 24 narrative), the continuations used in the context-after-sentence design, the 40-word same-position random pools, the six replacements of each walker lattice, and the sixteen swaps of the sixteen-edit sentence. No other data is used.

Representations and statistics. The representation (residual stream at depth $\lceil L/4 \rceil$, with the layer given for each model in §4, averaged over the tokens of the edited sentence), the token span, and the construction of effect vectors per context are specified in Appendix B. The same appendix defines the ruler, concentration and coherence, the same-position baseline (500 random subsets of 7 out of 40 pool words), exact Shapley shares with the precedence weights of the six-edit lattices (48 sentences, all 360 admissible orderings), the sixteen-edit lattice (all 2^{16} sentences after 48 contexts, exact shares, 20,000 random orderings for walker statistics), and the steering protocol (injection site, token positions, weighted inner product defining g, and Gaussian controls matched in norm). Appendix C reports every statistic at every layer, and Appendix D reports the alternative designs and controls, so that the choice of layer and design can be checked.

Compute. All experiments run on one NVIDIA L40S GPU. All lattices of one model take under ten minutes; the sixteen-edit lattice takes about 45 GPU-minutes per model (3.1M sentence–context pairs).

Code. Code and data will be released upon publication.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajárin, Vaibhav Srivastav, et al. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, pp. 2397–2430, 2023.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NeurIPS*, 2013.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations (ICLR)*, 2021.

Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Joshua Engels, Isaac Liao, Eric J. Michaud, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *ICLR*, 2025.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Ulrich Faigle and Walter Kern. The Shapley value for cooperative games under precedence constraints. *International Journal of Game Theory*, 21(3):249–266, 1992.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. *arXiv preprint arXiv:2303.02536*, 2023.

Robert P. Gilles, Guillermo Owen, and René van den Brink. Games with permission structures: The conjunctive approach. *International Journal of Game Theory*, 20(3):277–293, 1992.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999.

John C. Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *COLM*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *NeurIPS*, 2022.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2311.03658.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.01506.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Alexandre Rocchi-Henry, Thomas Fel, and Gianni Franchi. A geometric unification of concept learning with concept cones, 2025. URL https://arxiv.org/abs/2512.07355.

Alexis Ross, Ana Marasović, and Matthew E. Peters. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*, 2021.

Mark Rudelson and Roman Vershynin. Hanson–Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games*, volume 2, pp. 307–317. Princeton University Press, 1953.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2402.17762.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The Shapley Taylor interaction index. In *International Conference on Machine Learning (ICML)*, 2020.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adrià Garriga-Alonso, and Robert Kirk. Analyzing the generalization and reliability of steering vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2407.12404.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, 2024. Anthropic Transformer Circuits Thread.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. *arXiv preprint*, 2023.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.15213.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-Shap: The faithful Shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *International Conference on Learning Representations (ICLR)*, 2016.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 Small. *arXiv preprint arXiv:2211.00593*, 2022.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S. Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.17453.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *International Conference on Learning Representations (ICLR)*, 2015.

Andy Zou, Long Phan, Sarah Chen, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Proof of Proposition 3.3

The steps add up to the composite edit. Fix a context $\mathbf{c}$ and a walker $\mathbf{x}_0 \to \cdots \to \mathbf{x}_K = \mathbf{x}^*$. By Definition 3.2, $\boldsymbol{\Delta}_k(\mathbf{c}) = \mathbf{e}(\mathbf{x}_k, \mathbf{c}) - \mathbf{e}(\mathbf{x}_{k-1}, \mathbf{c})$, so $\sum_{k=1}^{K} \boldsymbol{\Delta}_k(\mathbf{c}) = \mathbf{e}(\mathbf{x}_K, \mathbf{c}) - \mathbf{e}(\mathbf{x}_0, \mathbf{c}) = \boldsymbol{\Delta}^*(\mathbf{c})$, the intermediate sentences cancelling pairwise. Averaging over the T contexts gives $\bar{\boldsymbol{\Delta}}^* = \sum_k \bar{\boldsymbol{\Delta}}_k$, and taking the inner product with $\bar{\boldsymbol{\Delta}}^*$ and dividing by $\|\bar{\boldsymbol{\Delta}}^*\|^2$ gives $\sum_k \hat{w}_{\pi(k)} = 1$. This holds for every context and every path and needs no assumption. The same identity gives $\sum_k \Sigma_{k*} = \Sigma^*$.

Each step converges to its node-specific mean. The T contexts are i.i.d., so $\bar{\boldsymbol{\Delta}}_k$ is the empirical mean of T independent copies of $\boldsymbol{\Delta}_k(\mathbf{c})$ and converges almost surely to $\boldsymbol{\nu}_{\pi,k} = \mathbb{E}_{\mathbf{c}} \boldsymbol{\Delta}_k(\mathbf{c})$ by the strong law of large numbers. Write $\bar{\boldsymbol{\Delta}}_k = \boldsymbol{\nu}_{\pi,k} + \boldsymbol{\xi}_k$ and $\bar{\boldsymbol{\Delta}}^* = \boldsymbol{\mu}^* + \boldsymbol{\xi}^*$. For any *fixed* vector $\boldsymbol{u}$, $\langle \boldsymbol{u}, \boldsymbol{\xi}_k \rangle$ is an average of T i.i.d. scalar sub-Gaussian variables and is $O_P(\sigma \|\boldsymbol{u}\| / \sqrt{T})$ (Vershynin, 2018, Prop. 2.6.1), whatever d; the same holds for $\boldsymbol{\xi}^*$ directly, so no bound on the proxy of a sum of dependent steps is needed. The full norms carry the dimension: $\mathbb{E}\|\boldsymbol{\xi}^*\|^2 = \operatorname{tr} \Sigma^* / T$ and $\mathbb{E}\langle \boldsymbol{\xi}_k, \boldsymbol{\xi}^* \rangle = \operatorname{tr} \Sigma_{k*} / T$.

The readout converges, with the stated error. Put $n^2 = \|\boldsymbol{\mu}^*\|^2$ and $w = w_{\pi,k}$. With $N = \langle \bar{\boldsymbol{\Delta}}_k, \bar{\boldsymbol{\Delta}}^* \rangle$ and $D = \|\bar{\boldsymbol{\Delta}}^*\|^2$,

$$\hat{w}_{\pi(k)} - w = \frac{N - wD}{D}, \qquad N - wD = \underbrace{\langle \boldsymbol{\nu}_{\pi,k}, \boldsymbol{\xi}^* \rangle + \langle \boldsymbol{\xi}_k, \boldsymbol{\mu}^* \rangle - 2w \langle \boldsymbol{\mu}^*, \boldsymbol{\xi}^* \rangle}_{\ell} + \underbrace{\langle \boldsymbol{\xi}_k, \boldsymbol{\xi}^* \rangle - w\|\boldsymbol{\xi}^*\|^2}_{q}.$$

The linear term is $\ell = O_P\big(\sigma(\|\boldsymbol{\nu}_{\pi,k}\| + (1+2|w|)n)/\sqrt{T}\big)$ by the previous paragraph. The quadratic term has mean $\mathbb{E}q = (\operatorname{tr}\Sigma_{k*} - w \operatorname{tr}\Sigma^*)/T = n^2 b_{\pi,k}$ and, by the Hanson–Wright inequality (Rudelson & Vershynin, 2013), fluctuations $O_P(\sigma\sqrt{\operatorname{tr}\Sigma^*}/T)$ (for instance $\operatorname{Var}\|\boldsymbol{\xi}^*\|^2 = O(\operatorname{tr}\Sigma^{*2}/T^2) \le O(\sigma^2 \operatorname{tr}\Sigma^*/T^2)$, and likewise for $\langle \boldsymbol{\xi}_k, \boldsymbol{\xi}^* \rangle$ by polarisation). The denominator is $D = n^2(1+\eta)$ with $\eta = O_P(\sigma/(n\sqrt{T})) + \tau + O_P(\sigma\sqrt{\operatorname{tr}\Sigma^*}/(Tn^2))$; expanding $1/(1+\eta) = 1 - \eta + O(\eta^2)$ and collecting terms gives (6), the cross terms being of the listed orders or $O(\tau^2)$. The bias sums to zero because $\sum_k \Sigma_{k*} = \Sigma^*$ and $\sum_k w_{\pi,k} = 1$, and it is bounded as stated by the Cauchy–Schwarz inequality for the trace inner product, $|\operatorname{tr}\Sigma_{k*}| \le \sqrt{\operatorname{tr}\Sigma_k \operatorname{tr}\Sigma^*}$. A U-statistic $\frac{1}{T(T-1)} \sum_{i \ne j} \langle \boldsymbol{\Delta}_k(\mathbf{c}_i), \boldsymbol{\Delta}^*(\mathbf{c}_j) \rangle$ has expectation exactly $\langle \boldsymbol{\nu}_{\pi,k}, \boldsymbol{\mu}^* \rangle$ (Hoeffding, 1948), and likewise for the denominator, so q is replaced by a centred term and $b_{\pi,k}$ disappears.

Under Assumption 3.1 the node-specific mean is the edit's mean effect. The k-th step is the edit $\boldsymbol{\delta}_{\pi(k)}$ applied at the node $\mathbf{x}_{k-1}$. Assumption 3.1 states that its mean effect does not depend on the node, so $\boldsymbol{\nu}_{\pi,k} = \boldsymbol{\mu}(\boldsymbol{\delta}_{\pi(k)})$ and $w_{\pi,k} = w_{\pi(k)}$; summing over k gives $\sum_k \boldsymbol{\mu}(\boldsymbol{\delta}_k) = \boldsymbol{\mu}^*$, the identity used in §3.2 to define the shares. Since the limit of each step depends only on the edit taken and not on the node it is taken from, the vector of limits is the same for every ordering up to relabelling, which is the claimed path independence.

The general case. Write $\boldsymbol{\nu}_{\pi,k} = \boldsymbol{\mu}(\boldsymbol{\delta}_{\pi(k)}) + \boldsymbol{\epsilon}_{\pi,k}$, which defines the interaction of the step, and take the inner product with $\boldsymbol{\mu}^*/n^2$ to obtain the first identity of (7). Summing over k and using the first paragraph, $\sum_k \boldsymbol{\epsilon}_{\pi,k} = \boldsymbol{\mu}^* - \sum_k \boldsymbol{\mu}(\boldsymbol{\delta}_k)$, whose right-hand side does not depend on π. Finally, $w_{\pi,k} = v(S_k) - v(S_{k-1})$ is the marginal contribution of $\pi(k)$ in the projection game along π, so its average over admissible orderings is the precedence-constrained Shapley value $\phi_{\pi(k)}$ by definition. □

B Experimental details

Representations. Each text is passed through the model once, and from every layer we store the mean over the tokens of the edited sentence (tokens whose character span overlaps the sentence, the closing full stop excluded) and the state of the final token; in the context-after-sentence design of Appendix D we also store the mean over all tokens but the first and the mean over the context tokens only. Effect vectors are formed per context, $\boldsymbol{\Delta}_v(\mathbf{c}) = \mathbf{e}(v, \mathbf{c}) - \mathbf{e}(\text{base}, \mathbf{c})$, never from means. Random-word pools are stored at layers $\lceil L/4 \rceil$ and $L - 1$. Models up to 0.5B parameters run in `float32` and larger models in `bfloat16`; all statistics are computed in `float32` in the full d dimensions. All experiments run on one NVIDIA L40S GPU, and all lattices of one model take under ten minutes.

B.1 The lattices of edits

Geometry. For each edit type, $\hat{\boldsymbol{\gamma}}$ is the normalised mean of its 336 effect vectors; s is computed in the full d dimensions; $\hat{\boldsymbol{\gamma}}_v$ and s_v are computed on the 48 vectors of each realisation; κ is the mean off-diagonal entry of the 7×7 Gram matrix of the $\hat{\boldsymbol{\gamma}}_v$. The largest $|\cos|$ between rulers of different edit types is taken over the 21 pairs. The same-position baseline uses the base sentence of the family and draws 500 subsets of 7 from the 40 pool words.

Visualisation. The geometry is drawn in one common projection (Fig. 1): every effect vector is normalised to a unit direction, the top-three right singular vectors of the $2{,}352 \times d$ matrix of all unit directions are taken *without centering*, so that the origin stays the origin, and each direction is projected and renormalised, with the captured energy as point size. A common projection is essential. A cloud projected onto its own principal axes always looks concentrated, because the axes were

chosen to make it so; in a common projection the concentrated edits determine the coordinate system, and a scattered edit shows that it has no single direction within it.

The walkers. Six edits with one realisation each (the first realisation of each type; the number edit is omitted) generate $2^6 = 64$ subsets, of which 48 are well-formed sentences (the action edit replaces the past-tense verb, so it requires the tense edit); every one of the 48 sentences is read after the 48 context sentences. A walker is any of the 360 orderings all of whose prefixes are well-formed; its step means are differences of node means along its path. Shapley shares are computed exactly by weighting each marginal contribution $v(P \cup \{k\}) - v(P)$ by the fraction of admissible orderings in which k is preceded by exactly P. The anchored effect of the action edit is measured at the tense node. For each edit, the cross-node cosine compares the edit's mean effect at every pair of nodes at which it can be applied; the pairwise interaction uses the ten pairs (i, j) whose three nodes $\{i\}, \{j\}, \{i, j\}$ exist. Sampling error is estimated by 300 bootstrap resamples of the contexts, 50 random split halves, and the U-statistic of Proposition 3.3; single-walker errors by 400 random draws of m orderings. Fig. 5 traces two walkers sentence by sentence, and Fig. 6 places them in the full lattice against the Shapley shares.

(a) Walker A: order 1, 2, 3, 4, 5, 6

x_0 The honest judge convicts the felon.
δ_1 The honest judge **convicted** the felon.
δ_2 The **upright** judge convicted the felon.
δ_3 The upright judge convicted the **culprit.**
δ_4 The **algorithm** convicted the culprit.
δ_5 The algorithm **deleted** the culprit.
δ_6 **An** algorithm deleted the culprit. $= x^*$

(b) Walker B: order 6, 3, 1, 2, 5, 4

x_0 The honest judge convicts the felon.
δ_6 **An** honest judge convicts the felon.
δ_3 An honest judge convicts the **culprit.**
δ_1 An honest judge **convicted** the culprit.
δ_2 An **upright** judge convicted the culprit.
δ_5 An upright judge **deleted** the culprit.
δ_4 An **algorithm** deleted the culprit. $= x^*$

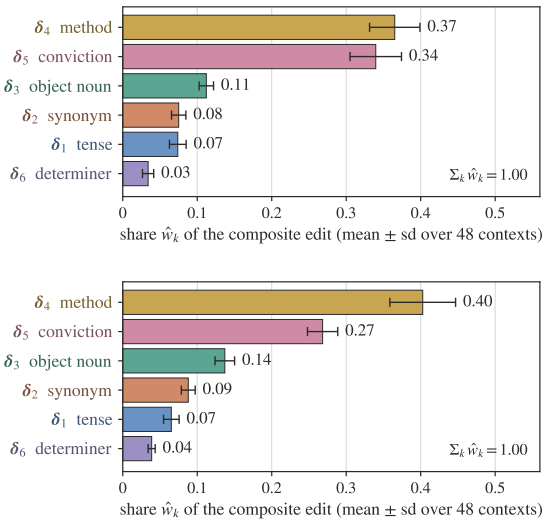

Figure 5: **Two walkers through the same composite edit** (Qwen2.5-1.5B, judicial lattice). Each walker applies the six replacements in a different order; the left column lists the sentence after every step with the replaced word marked, and the right column the share $\hat{w}_k$ each step contributes to the composite edit $\mathbf{x}_0 \to \mathbf{x}^*$ (mean $\pm$ sd over 48 contexts). The shares sum to one by construction.

The sixteen-edit lattice. The base sentence and its sixteen independent one-word swaps are listed in Appendix G. All 2^{16} sentences are read after the 48 judicial context sentences (3.1M sentence–context pairs per model, about 45 GPU-minutes), and only the context mean of each node is stored. Exact Shapley shares use the subset weights $|S|!\,(K - |S| - 1)!/K!$; walker statistics use 20,000 random orderings.

Steering. The steering vector is added to the output of block $\lceil L/4 \rceil$, that is, to the residual stream used throughout, at every token of the edited sentence in every context; next-token log-probabilities are read at the final token. The inner product defining g is weighted by $\bar{p} = \frac{1}{2}(p(\cdot \mid \mathbf{x}_0, \mathbf{c}) + p(\cdot \mid \mathbf{x}^*, \mathbf{c}))$, which down-weights the far tail of the vocabulary. With the unweighted inner product the Spearman correlation between share and drop-one effect is 0.79, and with the recovered fraction of $\mathrm{KL}(p_{\mathbf{x}^*} \| p_{\mathbf{x}_0})$ it is 0.64. Random controls are Gaussian directions scaled to the norms of the six anchored effects.

Reference predictors for the shares. For each of the 90 edits we compare the Shapley share with the norm of its anchored effect (normalised to sum to one within the lattice; Spearman 0.94 pooled, median 0.89 per lattice), with the Shapley share of the same game computed on the input embeddings

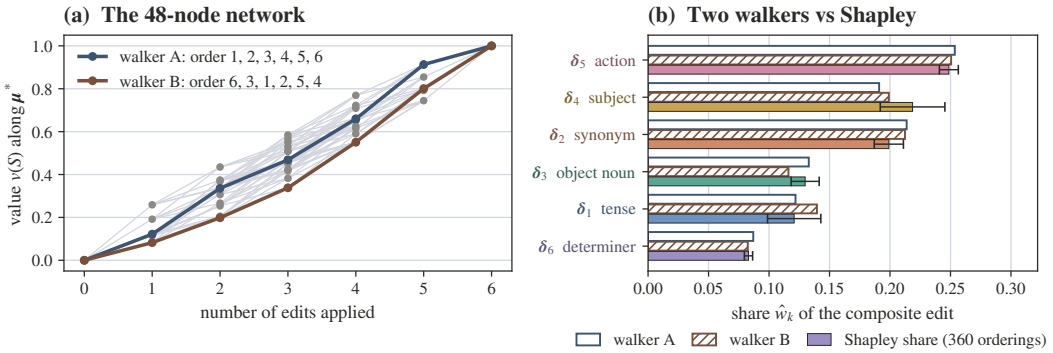

Figure 6: **The lattice and two walkers against the Shapley share** (Qwen2.5-7B, judicial lattice). **(a)** The 48 sentences of the six-edit lattice, placed by the number of edits applied and by their value $v(S)$ along the composite ruler; grey: edges; colour: two walkers. **(b)** The shares the two walkers read, against the Shapley share (grey; $\pm$ sd over all 360 orderings). The two walkers agree with each other on every share to within 0.07.

(pooled 0.59, per lattice 0.54; same leading edit on 4 of 15 lattices), and with the absolute change of the sentence's surprisal under the model (0.50).

C EVERY LAYER

Fig. 7 gives, at every layer of every model, the three statistics on which the choice of depth rests: the largest spread of a share across the 360 orderings, the relative interaction, and the mean cross-node cosine. Additivity is highest near the embedding and decreases gradually with depth, and the largest cosine between rulers of different edit types stays between 0.19 and 0.47 at layer $L-1$, so the edit types remain distinct at every depth.

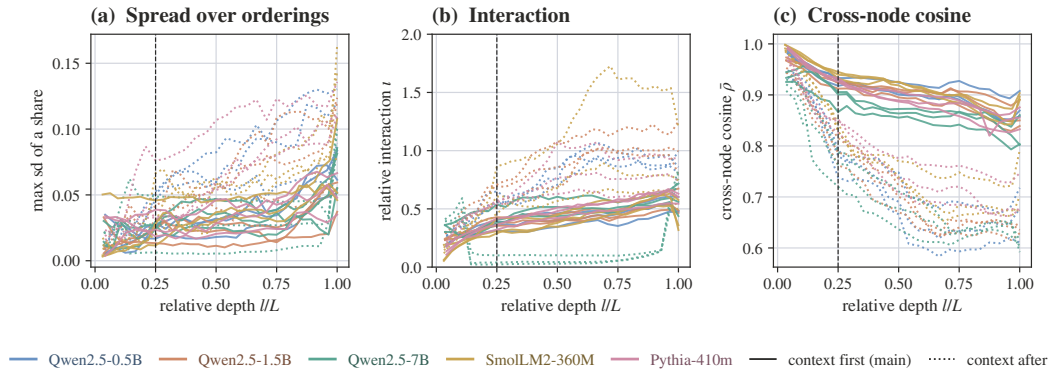

Figure 7: **Layer sweep.** (a) Largest standard deviation of a share across the 360 orderings; (b) relative interaction $\|\mu^* - \sum_k \mu_k\|/\|\mu^*\|$; (c) mean pairwise cosine of an edit's mean effect measured at every node where it applies. One curve per (model, base sentence), five models; the main text uses relative depth $1/4$.

D ALTERNATIVE DESIGNS

Placing the context. Table 3 compares the main design with two designs in which the context is appended after the sentence. Placing the context first gives the most stable and most additive representation on every model: the largest spread of a share over orderings is at most 0.05, against up to 0.11 when the context tokens alone are pooled, and the interaction is the smallest of the three designs on every lattice. When the context is appended, the sentence tokens cannot attend to it, and a

Table 3: The same lattices under three designs at layer $\lceil L/4 \rceil$: context appended after the sentence, mean over all tokens but the first; context appended after the sentence, mean over the context tokens only; context first, mean over the sentence tokens (main). $\bar{s}$: mean concentration over the seven edit types; max sd: largest spread of a share over the 360 orderings; ι: relative interaction; $|\cos|$: largest cosine between rulers; lead: modal leading edit.

	after, all tokens					after, context tokens					first, sentence tokens (main)			
Model, lattice	$\bar{s}$	max sd	ι	$\lvert\cos\rvert$	lead	$\bar{s}$	max sd	ι	$\lvert\cos\rvert$	lead	$\bar{s}$	max sd	ι	lead
Qwen2.5-0.5B, judicial	0.78	0.05	0.54	0.29	subj	0.69	0.08	0.54	0.45	obj	0.77	0.02	0.32	act
Qwen2.5-0.5B, medical	0.72	0.05	0.65	0.21	subj	0.66	0.11	0.73	0.36	subj	0.72	0.03	0.39	subj
Qwen2.5-0.5B, literary	0.72	0.05	0.55	0.26	subj	0.65	0.07	0.65	0.33	subj	0.71	0.03	0.36	subj
Qwen2.5-1.5B, judicial	0.77	0.03	0.71	0.44	act	0.68	0.09	0.83	0.68	obj	0.76	0.01	0.36	act
Qwen2.5-1.5B, medical	0.71	0.04	0.70	0.21	act	0.65	0.11	0.88	0.43	act	0.71	0.02	0.43	act
Qwen2.5-1.5B, literary	0.69	0.03	0.51	0.39	subj	0.60	0.08	0.60	0.40	subj	0.70	0.02	0.37	act
Qwen2.5-7B, judicial	0.70	0.03	0.04	1.00	det	0.68	0.08	0.75	0.36	subj	0.76	0.03	0.38	act
Qwen2.5-7B, medical	0.73	0.05	0.10	0.99	det	0.69	0.07	1.05	0.35	subj	0.69	0.03	0.54	subj
Qwen2.5-7B, literary	0.69	0.01	0.02	0.99	det	0.65	0.04	0.86	0.26	subj	0.69	0.02	0.41	subj
SmolLM2-360M, judicial	0.75	0.02	0.46	0.22	subj	0.68	0.04	0.56	0.34	tns	0.74	0.05	0.29	act
SmolLM2-360M, medical	0.72	0.03	0.61	0.18	subj	0.66	0.06	0.74	0.18	subj	0.71	0.01	0.31	subj
SmolLM2-360M, literary	0.73	0.07	0.87	0.24	subj	0.67	0.06	1.08	0.35	subj	0.72	0.03	0.42	subj
Pythia-410m, judicial	0.75	0.03	0.54	0.31	act	0.70	0.06	0.61	0.41	act	0.75	0.02	0.32	act
Pythia-410m, medical	0.72	0.03	0.55	0.20	subj	0.68	0.07	0.58	0.24	subj	0.71	0.04	0.42	syn
Pythia-410m, literary	0.74	0.08	0.57	0.14	act	0.70	0.10	0.61	0.17	act	0.72	0.02	0.36	act

mean over all tokens contains a context-independent component. On Qwen2.5-7B this component is dominated by the determiner edit, which replaces the first token of the input (share 1.00–1.05, rulers aligned to $|\cos| \geq 0.99$), consistent with the massive activation carried by the first position (Sun et al., 2024): changing the first token changes how the attention sink forms and thereby every later position. In the main design the edited sentence never begins the input, and all five models behave alike.

Shared offset and large coordinates. Projecting out the grand mean of all effect vectors of a lattice changes the concentration of an edit type by at most 0.10 (median 0.04), and standardising each coordinate by its standard deviation over the sentence states changes it by at most 0.09 (median 0.03), over all edit types, lattices and models. Concentration is a property of each edit's own direction, not of a shared offset or of a few large coordinates.

Token count. Some realisations change the number of tokens (*was convicting*, *dozens of culprits*). Restricted to realisations that keep the token count of the base sentence, every coherence changes by at most 0.11, and the classification of Table 1 is unchanged on 56 of the 60 (family, lattice) pairs.

E THE SIXTEEN-EDIT LATTICE

Table 4: The sixteen-edit lattice (65,536 sentences $\times$ 48 contexts). Shares are exact Shapley shares from the whole lattice. sd: largest standard deviation of one walker's share over random orderings; $m_{0.02}$: number of random walkers for which the 95th percentile of the largest share error is below 0.02, with the number of nodes it costs; anchored: largest $|a_k - \phi_k|$; ρ_{emb}: Spearman correlation with the shares computed on the input embeddings.

Model	top three shares	W	ι	$\cos(\boldsymbol{\epsilon}, \boldsymbol{\mu}^*)$	sd	$m_{0.02}$ (nodes)	anchored	ρ_{emb}
Qwen2.5-0.5B	judge→algorithm 0.28, convicted→acquitted 0.09, felon→culprit 0.08	0.88	0.33	0.36	0.03	8 (122)	0.04	0.71
SmolLM2-360M	judge→algorithm 0.20, felon→culprit 0.12, convicted→acquitted 0.11	0.95	0.23	0.23	0.01	1 (17)	0.02	0.92
Pythia-410m	judge→algorithm 0.15, convicted→acquitted 0.14, Yesterday→Today 0.13	0.92	0.35	0.24	0.01	2 (32)	0.02	0.82

F COHERENCE ON EVERY MODEL

G MATERIALS

The base sentence of edit type k is the first-realisation sentence of type $k-1$, so the seven edit types form a chain from the base sentence to the fully edited one. In the main design each sentence is

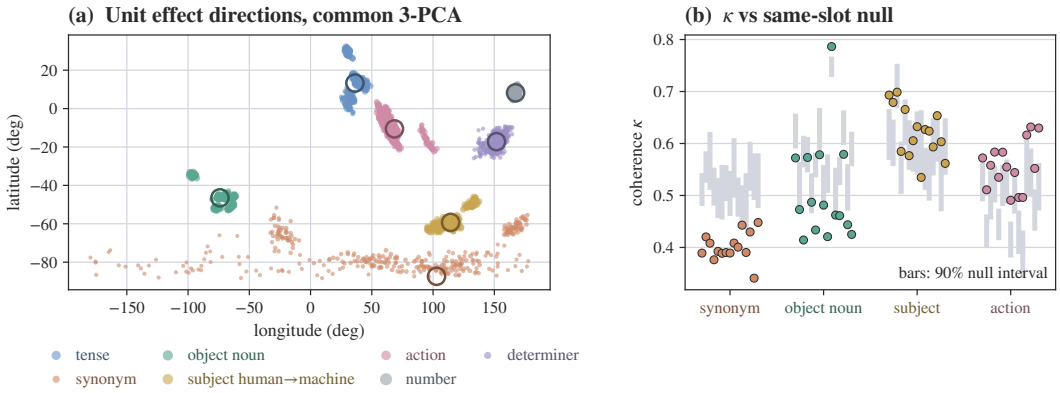

Figure 8: **Coherence against the same-position baseline.** **(a)** All 2,352 unit effect directions of seven edit types (judicial lattice, Qwen2.5-7B) in one common uncentered 3-PCA, unrolled to longitude and latitude; point size: energy captured by the three axes; rings: ruler of each edit type. **(b)** Coherence κ of four families of realisations (dots, one per model and lattice) against the 90% interval of the same-position baseline (bars).

Table 5: Seven edit types on Qwen2.5-7B (context first, mean over the tokens of the edited sentence, layer 7/28). s: concentration of the whole cloud (336 vectors); $\bar{s}_{\text{ctx}}$: mean concentration of one realisation over the 48 contexts; κ: coherence of the seven realisation rulers; baseline: mean and 90% interval of κ for seven random words in the same position; $^{+}/^{-}$: above the 95th / below the 5th percentile of the baseline. Isotropic baselines: $s \approx 0.06$ for 336 vectors and 0.14 for 48; $\kappa \approx 0$ with standard deviation 0.017.

Base sentence	Edit type	s	$\bar{s}_{\text{ctx}}$	κ	baseline κ
judicial (layer 7/28)	tense	0.80	0.96	0.63	–
	adjective synonym	0.65	0.95	0.39^{-}	0.52 [0.49, 0.56]
	object noun	0.79	0.99	0.58^{-}	0.63 [0.59, 0.67]
	subject: human → machine	0.76	0.93	0.61	0.59 [0.55, 0.63]
	action	0.75	0.96	0.55	0.53 [0.48, 0.57]
	number	0.90	0.98	0.83	–
	determiner	0.64	0.93	0.38	–
medical (layer 7/28)	tense	0.67	0.93	0.44	–
	adjective synonym	0.65	0.94	0.39^{-}	0.50 [0.46, 0.55]
	object noun	0.72	0.97	0.48	0.51 [0.47, 0.56]
	subject: human → machine	0.77	0.93	0.63	0.59 [0.54, 0.66]
	action	0.72	0.96	$\mathbf{0.49}^{+}$	0.42 [0.38, 0.45]
	number	0.70	0.97	0.45	–
	determiner	0.62	0.94	0.34	–
literary (layer 7/28)	tense	0.69	0.94	0.46	–
	adjective synonym	0.66	0.95	0.41^{-}	0.56 [0.50, 0.61]
	object noun	0.69	0.97	0.42^{-}	0.54 [0.50, 0.58]
	subject: human → machine	0.73	0.94	0.53^{-}	0.59 [0.54, 0.66]
	action	0.76	0.97	$\mathbf{0.54}^{+}$	0.50 [0.47, 0.54]
	number	0.67	0.96	0.41	–
	determiner	0.63	0.94	0.36	–

preceded by one of the 48 context sentences listed below and followed by a full stop; in the context-after-sentence design of Appendix D it is followed, after a comma, by one of the 48 continuations listed after them. The line marked `walker` lists the six replacements used in §4.2, in the canonical order.

Judicial. Base sentence *The honest judge convicts the felon*.

Table 6: Coherence κ of every edit type on every lattice; in parentheses the mean of the same-position baseline. Last column: mean context concentration $\bar{s}_{\mathrm{ctx}}$ over the seven types.

Model, lattice	tense	adj. synonym	object noun	subject	action	number	determiner	$\bar{s}_{\mathrm{ctx}}$
Qwen2.5-0.5B, judicial	0.66	0.39 (0.51)	0.57 (0.62)	0.69 (0.62)	0.57 (0.53)	0.83	0.43	0.96
Qwen2.5-0.5B, medical	0.49	0.42 (0.55)	0.47 (0.52)	0.68 (0.60)	0.51 (0.43)	0.47	0.42	0.96
Qwen2.5-0.5B, literary	0.43	0.41 (0.57)	0.41 (0.58)	0.70 (0.71)	0.56 (0.50)	0.45	0.44	0.96
Qwen2.5-1.5B, judicial	0.63	0.38 (0.52)	0.57 (0.58)	0.58 (0.55)	0.58 (0.52)	0.84	0.41	0.96
Qwen2.5-1.5B, medical	0.45	0.39 (0.49)	0.49 (0.50)	0.67 (0.64)	0.53 (0.44)	0.47	0.37	0.95
Qwen2.5-1.5B, literary	0.47	0.39 (0.51)	0.43 (0.56)	0.58 (0.60)	0.58 (0.52)	0.45	0.38	0.95
Qwen2.5-7B, judicial	0.63	0.39 (0.52)	0.58 (0.63)	0.61 (0.59)	0.55 (0.53)	0.83	0.38	0.96
Qwen2.5-7B, medical	0.44	0.39 (0.50)	0.48 (0.51)	0.63 (0.59)	0.49 (0.42)	0.45	0.34	0.95
Qwen2.5-7B, literary	0.46	0.41 (0.56)	0.42 (0.54)	0.53 (0.59)	0.54 (0.50)	0.41	0.36	0.95
SmolLM2-360M, judicial	0.53	0.40 (0.50)	0.79 (0.75)	0.63 (0.55)	0.50 (0.43)	0.37	0.35	0.98
SmolLM2-360M, medical	0.43	0.44 (0.48)	0.46 (0.51)	0.62 (0.54)	0.50 (0.38)	0.41	0.33	0.98
SmolLM2-360M, literary	0.50	0.39 (0.49)	0.46 (0.55)	0.59 (0.59)	0.62 (0.55)	0.36	0.33	0.98
Pythia-410m, judicial	0.54	0.43 (0.55)	0.58 (0.63)	0.65 (0.61)	0.63 (0.54)	0.59	0.32	0.97
Pythia-410m, medical	0.49	0.34 (0.53)	0.44 (0.51)	0.60 (0.53)	0.55 (0.47)	0.46	0.36	0.97
Pythia-410m, literary	0.48	0.45 (0.53)	0.43 (0.59)	0.56 (0.60)	0.63 (0.52)	0.44	0.32	0.97

```
tense             (base: The honest judge convicts the felon.)
    convicted | was convicting | had convicted | is convicting |
    shall convict | would convict | could convict
adj_synonym       (base: The honest judge convicted the felon.)
    upright | moral | fair | ethical |
    noble | principled | righteous
object_noun       (base: The upright judge convicted the felon.)
    culprit | offender | prisoner | suspect |
    convict | criminal | perpetrator
subject_machine   (base: The upright judge convicted the culprit.)
    algorithm | program | script | AI |
    model | bot | software
action            (base: The upright algorithm convicted the culprit.)
    deleted | erased | purged | hid |
    lost | corrupted | ignored
number            (base: The upright algorithm deleted the culprit.)
    the culprits | both culprits | all the culprits |
    several culprits | the two culprits |
    the three culprits | dozens of culprits
determiner        (base: The upright algorithm deleted the culprits.)
    An | Some | This | That |
    Another | Every | No
walker: convicts>convicted -> honest>upright -> felon>culprit
        judge>algorithm -> convicted>deleted -> The>An
```

Medical. Base sentence *The careful doctor treats the patient.*

```
tense             treated | was treating | had treated | is treating |
                  shall treat | would treat | could treat
adj_synonym       attentive | diligent | meticulous | thorough |
                  cautious | conscientious | prudent
object_noun       child | soldier | traveller | worker |
                  prisoner | athlete | stranger
subject_machine   algorithm | program | script | AI |
                  model | bot | software
action            deleted | erased | purged | hid |
                  lost | corrupted | ignored
number            the children | both children | all the children |
                  several children | the two children |
                  the three children | dozens of children
determiner        An | Some | This | That | Another | Every | No
walker: treats>treated -> careful>attentive -> patient>child
        doctor>algorithm -> treated>deleted -> The>An
```

Literary. Base sentence *The talented author writes the novel.*

```
tense            wrote | was writing | had written | is writing |
                 shall write | would write | could write
adj_synonym      gifted | skilled | brilliant | accomplished |
                 capable | masterful | able
object_noun      story | poem | essay | memoir |
                 play | letter | manuscript
subject_machine  algorithm | program | script | AI |
                 model | bot | software
action           deleted | erased | purged | hid |
                 lost | corrupted | ignored
number           the stories | both stories | all the stories |
                 several stories | the two stories |
                 the three stories | dozens of stories
determiner       A | Some | This | That | Another | Every | No
walker: writes>wrote -> talented>gifted -> novel>story
        author>algorithm -> wrote>deleted -> The>A
```

Same-position random-word pools.

```
adjective position:
    green wooden tall quiet young old bright heavy
    small large round cold warm silver famous local
    busy tired happy angry strange rich poor blind
    lazy brave hungry nervous calm lucky proud quick
    slow loud strong weak dark short thin sleepy
noun positions (subject; object with "the"):
    farmer teacher sailor soldier priest king baker pilot
    student dog horse cat wolf bird fox bear
    lion table river window stone cloud garden bottle
    chair mountain candle ship clock apple door road
    tree flower cup boat hat coat bridge lamp
verb position:
    saw found helped called followed painted pushed visited
    watched praised fed carried kissed hired trained warned
    thanked paid met chased greeted blamed served taught
    invited hugged pulled lifted touched answered admired rescued
    hunted fired kicked married robbed tested washed joined
```

Sixteen-edit lattice. Base sentence *Yesterday the honest judge in the small northern town quickly convicted the young felon of theft after a long and tiring trial at the old courthouse.*

```
 1. Yesterday -> Today          9. young -> elderly
 2. honest -> upright          10. felon -> culprit
 3. judge -> algorithm         11. theft -> fraud
 4. small -> large             12. long -> short
 5. northern -> southern       13. tiring -> boring
 6. town -> city               14. trial -> hearing
 7. quickly -> slowly          15. old -> new
 8. convicted -> acquitted     16. courthouse -> palace
```

Context sentences, judicial (main design; 24 formal, then 24 narrative).

- "The trial had lasted eleven weeks in the district court."
- "The indictment listed four counts of fraud and one of perjury."
- "The appellate court had remanded the case for a new hearing."
- "Under the criminal code, the burden of proof lies with the state."
- "The jury deliberated for two days before returning its findings."
- "The prosecution rested its case after calling twelve witnesses."
- "According to the court record, the evidence was uncontested."
- "The sentencing guidelines had been revised earlier that year."
- "The defense had filed a motion to suppress the confession."

- "The hearing was held in the federal courthouse downtown."
- "The forensic report was entered into evidence on Monday."
- "The statute of limitations had nearly expired."
- "The bar association sent observers to the proceedings."
- "The docket for the morning session was unusually long."
- "Counsel for both sides delivered their closing arguments."
- "The written opinion ran to forty pages."
- "The case had occupied the court for nearly two years."
- "The testimony of three witnesses was read into the record."
- "The clerk announced that the court was in session."
- "The amended indictment was filed in the spring."
- "The parole board would be notified of the outcome."
- "Restitution to the victims was expected to be ordered."
- "The court reporter recorded every word of the proceedings."
- "The law requires a ruling once the evidence is complete."
- "Rain hammered on the tall windows of the old courthouse."
- "The gallery fell silent as the clerk began to read."
- "A murmur ran along the benches where the family sat."
- "Somewhere at the back of the room a woman began to weep."
- "The reporters had waited in the hallway since dawn."
- "The bailiff stood by the door with his arms folded."
- "For a long moment the only sound was the ticking of the clock."
- "The sketch artist sharpened a stick of charcoal."
- "The defendant stared at the floor and said nothing."
- "Outside, a crowd had gathered on the courthouse steps."
- "The old lawyer closed the folder with a weary sigh."
- "Afternoon light slanted through the dusty blinds."
- "The mother of the victim gripped the rail with both hands."
- "The whole town seemed to hold its breath that morning."
- "The young clerk would remember that day for the rest of her life."
- "A pigeon walked back and forth along the windowsill."
- "The jurors filed back in, looking tired and pale."
- "The guards shifted their weight by the narrow corridor."
- "By evening the story would be on every television in the city."
- "The radiator clanked in the corner of the crowded room."
- "A phone buzzed somewhere in the gallery and was quickly silenced."
- "The defendant's brother pushed his way to the front row."
- "Snow had begun to fall on the square outside."
- "Nobody in the courtroom looked surprised."

Context sentences, medical (main design; 24 formal, then 24 narrative).

- "The hospital ethics board had approved the protocol in May."
- "The blood panel showed an elevated white cell count."
- "The chest radiograph was reviewed by two specialists."
- "The clinical guidelines were updated last year."
- "The biopsy was scheduled for Thursday morning."
- "The cardiology department had been consulted the day before."
- "According to the chart, the fever began on admission."
- "The insurance authorization was granted after a week."
- "The intensive care unit is on the fourth floor of the building."
- "The second opinion confirmed the initial diagnosis."
- "The standard of care for this infection is well established."
- "The differential diagnoses had been ruled out one by one."
- "The randomized trial reported a favourable outcome."

- "The informed consent of the family was documented."
- "The attending physician was on call that night."
- "The discharge summary had to be completed before the shift ended."
- "The licensing board requires detailed records of every procedure."
- "The rehabilitation service reviews every case within a week."
- "The dosage was recorded in the chart for the night staff."
- "The administration reported a shortage of beds."
- "A follow-up appointment was set for six weeks later."
- "The surgery that had been proposed was postponed."
- "The procedure was expected to last about two hours."
- "The regimen requires ten days of antibiotics."
- "The monitors beeped softly in the dim ward."
- "The nurse at the door held her breath."
- "Somewhere down the corridor a child began to cry."
- "The first light of morning crept across the tiled floor."
- "The family in the waiting room did not yet know."
- "The ambulance crew lingered by the coffee machine."
- "For a moment nobody in the room said anything."
- "The old man on the next bed turned to watch."
- "Rain streaked the windows of the emergency room."
- "The intern scribbled notes as fast as she could."
- "Outside, the relatives paced the car park in the cold."
- "The shift bell rang for the second time."
- "The mother gripped the edge of the bed with both hands."
- "The whole ward seemed to relax when the fever broke."
- "It was a night the young resident would never forget."
- "The trolley wheels squeaked away down the hall."
- "A moth circled the lamp above the desk."
- "The boy's father pushed through the doors, pale and shaking."
- "The orderlies wheeled an empty bed toward the lift."
- "The guards at the prison infirmary looked on in silence."
- "By evening the news had spread through the whole village."
- "The heating in the old wing had failed again."
- "A radio played quietly at the nurses' station."
- "The corridor smelled of soap and cold coffee."

Context sentences, literary (main design; 24 formal, then 24 narrative).

- "Every reviewer in the trade press had praised the previous book."
- "The major literary magazines ran long features that season."
- "The critics compared the style to the great realists."
- "The research took place in the archives of the national library."
- "The contract was signed with the publisher the previous spring."
- "The first edition sold out within a month."
- "The university press issued the work in three volumes."
- "Two literature seminars later added the work to their syllabus."
- "The interviews were conducted over the course of a decade."
- "The committee shortlisted the work for the national book award."
- "The editor described the prose as spare and exact."
- "The estate held the family papers and objected to their use."
- "The conventions of the historical genre were followed closely."
- "The final draft was delivered two weeks before the deadline."
- "The trilogy had been started fifteen years earlier."
- "Footnotes documented every quotation from the letters."
- "The typewriter now sits in the museum of the town."

- "The fellowship was awarded that year by the academy."
- "The biographers have since accepted this account."
- "The marketing department proposed a different title."
- "An afterword explained the choice of narrator."
- "The book was dedicated to a former teacher."
- "A translation appeared simultaneously in seven languages."
- "The tradition of the genre has demanded this since its origins."
- "The pages piled up on the desk through the long winter."
- "Rain drummed on the skylight of the attic room."
- "The cat slept on the manuscript whenever it could."
- "Somewhere in the house a kettle began to whistle."
- "Nobody in the village knew what the lamp was for at night."
- "The postman carried a parcel to the station before noon."
- "The ink ran dry on the last line of the last chapter."
- "For a long while the only sound was the scratch of the pen."
- "The old editor read the pages twice before saying a word."
- "The children played in the garden below the window."
- "Outside, the first snow of the year began to fall."
- "A friend read the draft and closed it with a small sigh."
- "The evening light turned the walls of the study gold."
- "The widow kept every draft in a box under the bed."
- "The whole town seemed to hold its breath for the ending."
- "The granddaughter would retell the story for the rest of her life."
- "The typewriter fell silent at last in the small hours."
- "A blackbird sang on the fence at the end of the lane."
- "The publisher's boy came running up the hill with a telegram."
- "The neighbours filed past the window, curious and quiet."
- "The nurse at the sanatorium looked on in silence."
- "By spring there were copies in every shop window in the city."
- "The stove in the corner had gone cold hours ago."
- "A dog barked twice somewhere across the valley."

Continuations for the context-after-sentence design. Appended to the sentence after a comma.

Judicial.

- ", upholding the sacred principles of law that govern our civilization."
- ", in accordance with the statutes of the state and the precedents of the appellate court."
- ", after a deliberation that weighed every exhibit entered into the record."
- ", citing the testimony of three witnesses and the forensic report."
- ", a decision the prosecution had sought since the indictment was filed."
- ", and the sentence is to be pronounced at a hearing next month."
- ", noting that the burden of proof had been met beyond reasonable doubt."
- ", a ruling that the defense has announced it will appeal."
- ", in a written opinion of forty pages released to the press."
- ", consistent with the guidelines issued by the sentencing commission."
- ", the verdict being entered into the docket at nine in the morning."
- ", despite the objections raised by counsel during the closing arguments."
- ", as the statute requires when the evidence is uncontested."
- ", and orders the clerk to notify the parole board of the outcome."
- ", thereby closing a case that had occupied the court for two years."
- ", finding the alibi inconsistent with the telephone records."
- ", on all four counts listed in the amended indictment."
- ", pursuant to section twelve of the criminal code."
- ", a conclusion supported by the unanimous finding of the jury."

- ", after rejecting the motion to suppress the recorded confession."
- ", with a recommendation that the maximum term be imposed."
- ", and directs that restitution be paid to the victims within a year."
- ", in a proceeding observed by delegates from the bar association."
- ", as the law of the land demands of every officer of the court."
- ", and the gallery rose to its feet before the clerk finished reading."
- ", while rain hammered on the tall windows of the old courthouse."
- ", and a murmur ran along the benches where the family sat."
- ", then removed the spectacles and rubbed a tired pair of eyes."
- ", and somewhere at the back of the room a woman began to weep."
- ", though nobody in the courtroom looked surprised by the outcome."
- ", and the reporters hurried out to file their stories before noon."
- ", as the bailiff stepped forward with the handcuffs already open."
- ", and for a long moment the only sound was the ticking of the clock."
- ", and the sketch artist put down the charcoal at last."
- ", while the defendant stared at the floor and said nothing at all."
- ", and outside the crowd that had waited all morning began to cheer."
- ", and the old lawyer closed the folder with a small, weary sigh."
- ", just as the afternoon light slanted through the dusty blinds."
- ", and the mother of the victim gripped the rail with both hands."
- ", and the whole town seemed to exhale as the news spread down the street."
- ", a scene the young clerk would remember for the rest of her life."
- ", and the gavel came down with a crack that echoed off the marble."
- ", then rose and left the bench without another word."
- ", while a pigeon walked back and forth along the windowsill."
- ", and the defendant's brother pushed his way out through the doors."
- ", and the jurors filed out looking relieved that it was over."
- ", and the guards led the prisoner away down the narrow corridor."
- ", and by evening the story was on every television in the city."

Medical.

- ", in accordance with the protocol approved by the hospital ethics board."
- ", after reviewing the results of the blood panel and the chest radiograph."
- ", noting the elevated white cell count recorded on admission."
- ", as recommended in the clinical guidelines published last year."
- ", with a course of antibiotics to be continued for ten days."
- ", and records the dosage in the chart for the night staff."
- ", following the consultation with the cardiology department."
- ", a decision supported by the findings of the second opinion."
- ", in the intensive care unit on the fourth floor of the building."
- ", pending the outcome of the biopsy scheduled for Thursday."
- ", and orders a follow-up appointment in six weeks."
- ", despite the shortage of beds reported by the administration."
- ", according to the standard of care for this class of infection."
- ", having ruled out the differential diagnoses one by one."
- ", thereby avoiding the surgery that had initially been proposed."
- ", in a procedure lasting a little under two hours."
- ", with the informed consent of the family duly documented."
- ", under the supervision of the attending physician on call."
- ", and notes that the prognosis is favourable if the regimen is followed."
- ", in line with the findings of the randomized trial cited in the report."
- ", after the insurance authorization was finally granted."
- ", and refers the case to the rehabilitation service for review."
- ", as required by the licensing board of the province."

- ", and the discharge summary is dictated before the end of the shift."
- ", while the monitors beeped softly in the dim ward."
- ", and the nurse at the door let out a breath she had been holding."
- ", then pulled off the gloves and dropped them in the bin."
- ", and somewhere down the corridor a child began to cry."
- ", as the first light of morning crept across the tiled floor."
- ", though the family in the waiting room did not yet know."
- ", and the ambulance crew lingered by the coffee machine."
- ", with a steadiness that surprised even the senior surgeon."
- ", and for a moment nobody in the room said anything."
- ", and the old man on the next bed turned to watch."
- ", while the rain streaked the windows of the emergency room."
- ", and the intern scribbled notes as fast as she could."
- ", and outside the relatives paced the car park in the cold."
- ", just as the shift bell rang for the second time."
- ", and the mother gripped the edge of the bed with both hands."
- ", and the whole ward seemed to relax as the fever broke."
- ", a night the young resident would never forget."
- ", and the trolley wheels squeaked away down the hall."
- ", then sat down heavily on the stool by the sink."
- ", while a moth circled the lamp above the desk."
- ", and the boy's father pushed through the doors, pale and shaking."
- ", and the orderlies wheeled the bed back toward the lift."
- ", and the guards at the prison infirmary looked on in silence."
- ", and by evening the story had spread through the whole village."

Literary.

- ", to the unanimous delight of every reviewer in the trade press."
- ", an assessment echoed across the major literary magazines."
- ", in a style the critics compared to the great realists of the last century."
- ", after a period of research in the archives of the national library."
- ", under a contract signed with the publisher the previous spring."
- ", and the first edition sold out within a month of release."
- ", in three volumes issued by the university press."
- ", a work later listed in the syllabus of two literature seminars."
- ", drawing on interviews conducted over the course of a decade."
- ", which the committee shortlisted for the national book award."
- ", in a prose that the editor described as spare and exact."
- ", despite the objections of the estate that held the family papers."
- ", following the conventions of the historical genre closely."
- ", and delivered the final draft two weeks before the deadline."
- ", thereby completing a trilogy begun fifteen years earlier."
- ", with footnotes documenting every quotation from the letters."
- ", on a typewriter that now sits in the museum of the town."
- ", pursuant to the terms of the fellowship awarded that year."
- ", a conclusion the biographers have since accepted."
- ", after rejecting the title proposed by the marketing department."
- ", with an afterword explaining the choice of narrator."
- ", and dedicated the book to the memory of a former teacher."
- ", in a translation that appeared simultaneously in seven languages."
- ", as the tradition of the genre has demanded since its origins."
- ", and the pages piled up on the desk through the long winter."
- ", while rain drummed on the skylight of the attic room."
- ", and the cat slept on the manuscript whenever it could."

- ", then leaned back and stared at the ceiling for an hour."
- ", and somewhere in the house a kettle began to whistle."
- ", though nobody in the village knew what the lamp was for at night."
- ", and the postman carried the parcel to the station before noon."
- ", as the ink ran dry on the last line of the last chapter."
- ", and for a long while the only sound was the scratch of the pen."
- ", and the old editor read it twice before saying a word."
- ", while the children played in the garden below the window."
- ", and outside the first snow of the year began to fall."
- ", and the friend who read the draft closed it with a small sigh."
- ", just as the evening light turned the walls of the study gold."
- ", and the widow kept every draft in a box under the bed."
- ", and the whole town seemed to hold its breath for the ending."
- ", a story the granddaughter would retell for the rest of her life."
- ", and the typewriter fell silent at last in the small hours."
- ", then rose and walked out into the orchard without a coat."
- ", while a blackbird sang on the fence at the end of the lane."
- ", and the publisher's boy came running up the hill with the telegram."
- ", and the neighbours filed past the window, curious and quiet."
- ", and the nurse at the sanatorium looked on in silence."
- ", and by spring the book was in every shop window in the city."

ICLR 2027 PAPER CHECKLIST

1. **Claims.** Answer: Yes. Justification: The abstract and introduction state three contributions: the geometry of counterfactual edits (§2), the concept network with the walker and Proposition 3.3 (§3), and the measurements on five models (§4). Every numerical claim in the abstract is reported in Tables 1, 2 and 4, Fig. 4 and §4.4.
2. **Limitations.** Answer: Yes. Justification: §4.2 measures the interaction between edits and its dependence on depth; the conclusion states the scope of the study (templated sentences, one-word edits, models up to 7B parameters) and the directions in which the protocol extends.
3. **Theory assumptions and proofs.** Answer: Yes. Justification: Assumption 3.1 and the sub-Gaussian condition are stated in Proposition 3.3; the complete proof, including the finite-sample error, the bias term and the general case with interaction, is in Appendix A.
4. **Experimental result reproducibility.** Answer: Yes. Justification: Appendix G lists every base sentence, edit, realisation, context sentence and random-word pool; Appendix B describes the readouts, the walkers, the sixteen-edit network, the steering protocol and all statistics.
5. **Open access to data and code.** Answer: Yes. Justification: Code will be released upon publication; all models are public checkpoints.
6. **Experimental setting and details.** Answer: Yes. Justification: §4 and Appendix B give the models, layers, readouts, number of contexts, numerical precision, null construction, steering layer and positions, and the number of resamples for every statistic.
7. **Experiment statistical significance.** Answer: Yes. Justification: Shares are reported as Shapley values with the standard deviation over all 360 orderings; sampling error over contexts is estimated by 300 bootstrap resamples and 50 split halves; coherence is compared with 90% intervals of a 500-draw null; walker errors are 95th percentiles over 400 draws (six-edit lattices) and 20,000 orderings (sixteen-edit network); Proposition 3.3 gives the theoretical error.
8. **Experiments compute resources.** Answer: Yes. Justification: All experiments run on one NVIDIA L40S GPU; all six-edit lattices of one model take under ten minutes, and the sixteen-edit network about 45 GPU-minutes per model (3.1M forward passes).
9. **Code of ethics.** Answer: Yes. Justification: Only public open-weight checkpoints and synthetic sentences are used; no human subjects are involved.
10. **Broader impacts.** Answer: Yes. Justification: The ethics statement discusses the dual use of instruments that expose attribute directions in a model.
11. **Safeguards.** Answer: N/A. Justification: No model or dataset with risk of misuse is released.
12. **Licenses for existing assets.** Answer: Yes. Justification: Qwen2.5-0.5B, Qwen2.5-1.5B and Qwen2.5-7B (Apache-2.0), SmolLM2-360M (Apache-2.0) and Pythia-410m (Apache-2.0) are cited with their original publications.
13. **New assets.** Answer: N/A. Justification: The lattices, context sentences and word pools are listed in full in Appendix G and are not released as a standalone dataset.
14. **Crowdsourcing and research with human subjects.** Answer: N/A.
15. **IRB approvals.** Answer: N/A.
16. **Declaration of LLM usage.** Answer: Yes. Justification: LLMs were used for proofreading and for drafting analysis code that the authors reviewed and ran, as described in the AI use statement.