A GEOMETRIC THEORY OF SEMANTIC RELATION OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Analogies such as *man* : *woman* :: *king* : *queen* demonstrate that two relations can be applied in either order with the same result. This property is formalized as the flatness of a discrete connection, where tokens are represented as vertices in a graph and edges are labeled by relation types, each acting on embeddings via a specific affine operator. An analogy square corresponds to a closed walk whose holonomy is defined by the commutator of the two operators. Exactly fitted squares enforce flatness if their base points affinely span the space. Flat squares ensure path independence of composite analogies when they generate the fundamental group. For contextual representations, operators may depend on context through softmax mixtures of Lie-algebra generators, such that curvature is uniformly bounded by their brackets. In an evaluation of 1,154 mined morphological analogy squares in GPT-2, affine operators outperform vector offsets by ten points on two-hop analogies, but display path dependence in 11–25% of held-out compositions. Application of curvature regularization reduces path dependence to 3–6% across all relation pairs, and ensures consistency in unpenalized four-hop detours, validating our theoretical predictions. When syntactic frames shift, static operators lose 8–15 points, while contextual operators recover to within two points of per-context oracles and exhibit context dependence.

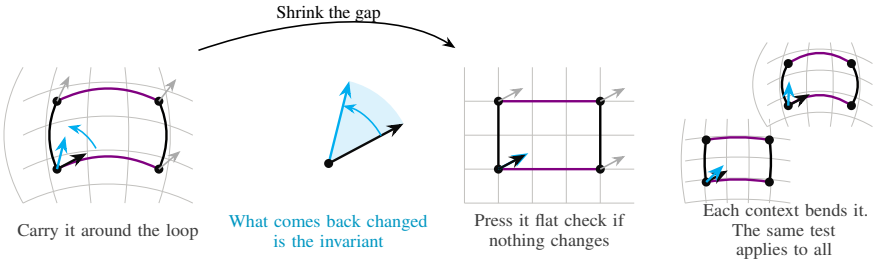


Figure 1: Geometric Intuition: A vector carried around an analogy square on a curved sheet returns changed. The change depends only on the square. We study this invariant in this paper. Flattening the sheet, our curvature penalty, makes it return unchanged. Under context shift, each context bends the sheet differently, and the same test applies.

1 INTRODUCTION

Word embeddings represent numerous linguistic regularities as approximately parallel displacements (Mikolov et al., 2013b; Pennington et al., 2014). Conversely, knowledge-graph embeddings takes a broader perspective, modeling relations as maps on entity vectors, such as translations (Bordes et al., 2013), matrices (Nickel et al., 2011), or rotations (Sun et al., 2019). Additionally, relations can be viewed as approximately affine maps on the hidden states of language models (Hernandez et al., 2024). When relations are conceptualized as maps, the central question regarding analogies shifts from determining whether four points form a parallelogram to assessing whether two compositions of maps lead to the same result. For example, applying gender followed by royalty to man should produce the same outcome as applying these relations in the reverse order. The property of whether

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.

compositions of maps around a closed loop return to the identity is referred to as the holonomy of a connection, while any deviation from this return is termed curvature.

In this paper, we establish a correspondence between vocabulary embeddings and contextual representations in language models. This framework is then used to quantify and mitigate the path dependence of compositional analogies. Two modeling decisions are central. First, operators are assigned to relation types and not to edges. Assigning one operator per edge allows any four points to be connected by translations that compose to the identity, which renders holonomy uninformative (Remark 16). Assigning one operator per relation ensures that the holonomy of an analogy square is the commutator of two relation operators. Operators are defined within the affine group $\mathrm{Aff}(D) \subset \mathrm{GL}(D{+}1)$, ensuring that transported embeddings remain valid. The main contributions are as follows: (1) Precise statements establish connections among analogy squares, holonomy, and curvature. Note that the closure of a square corresponds to a fixed-point condition on the commutator. Exactly-fitted squares enforce flatness, and this assumption is necessary. The vector-offset model characterizes the always-flat scenario. Furthermore, approximate fits result in approximate flatness (Section 3); (2) A topological characterization determines the conditions under which flat squares ensure path independence for composite analogies. This criterion is strictly stronger than the homological condition (Section 4); (3) We introduce a label-free error certificate and context-dependent operators constructed from softmax mixtures of Lie-algebra generators with bounded curvature in every context. We also present a well-posed curvature-regularized objective (Sections 5–7); (4) Experiments conducted on a novel test bed of $1{,}154$ morphological analogy squares in GPT-2 indicate that affine operators outperform translations by ten points, although they display path dependence in 11 to 25% of held-out compositions. Curvature regularization increases consistency to 94 to 97%, with a maximum performance cost of 0.6 points, and ensures consistency for unpenalized four-hop detours. When subjected to syntactic frame shifts, static operators experience performance degradation, whereas contextual operator families recover to within two points of per-context oracles, independently distinguish between noun-like and verb-like frames, and achieve consistency in previously unseen contexts through penalties applied during training (Section 8). Readers are recommended to read Appendix A to build geometric intuition for the abstraction in this paper.

2 Relation-typed connections on semantic graphs

Let V be a finite vocabulary, $E : V \to \mathbb{R}^D$ an embedding, $u_v := E(v)$, and $\tilde{u}_v := (u_v, 1) \in \mathbb{R}^{D+1}$ the homogeneous embedding, which lies on the affine hyperplane $\mathcal{H} = \{x : x_{D+1} = 1\}$. A matrix in $\mathrm{GL}(D{+}1)$ maps $\mathcal{H}$ into itself *iff* it has the following form:

$$\Gamma = \begin{pmatrix} L & t \\ 0 & 1 \end{pmatrix}, \qquad L \in \mathrm{GL}(D),\ t \in \mathbb{R}^D, \qquad \Gamma\tilde{u} = (Lu + t, 1) \tag{1}$$

These elements constitute the affine group $\mathrm{Aff}(D)$, whose Lie algebra is given by $\mathfrak{aff}(D) = \{\left(\begin{smallmatrix} M & \tau \\ 0 & 0 \end{smallmatrix}\right)\}$. The notation $L(\Gamma)$ and $t(\Gamma)$ denotes the linear and translational components, respectively, while T_t represents the translation (I, t). A vector η satisfying $\eta_{D+1} = 0$, referred to as a *displacement*, is mapped by Γ to $L(\Gamma)\eta$. Let $\mathcal{R}$ denote a finite set of relation types, and let $\mathcal{G} = (V, \mathcal{E}, \ell)$ be a directed graph with labeling function $\ell : \mathcal{E} \to \mathcal{R}$. An edge $e = (p \to q)$ with $\ell(e) = r$ indicates that q stands in relation r to p, and $\bar{e}$ denotes the reversed edge.

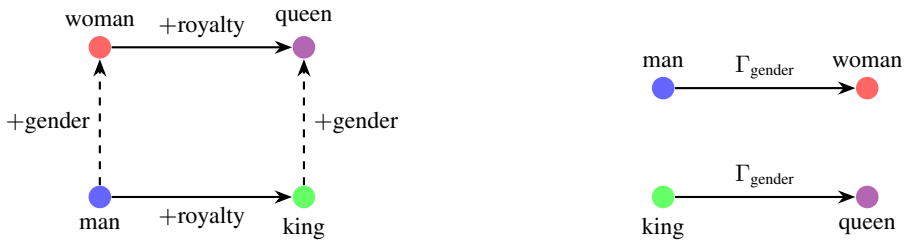
Definition 1 (Relation-typed connection)**.** A relation-typed connection is defined as a map $\omega : \mathcal{R} \to \mathrm{Aff}(D)$, where $r \mapsto \Gamma_r$. This map assigns to each edge the transport $\omega(e) := \Gamma_{\ell(e)}$, and to $\bar{e}$ the transport $\Gamma_{\ell(e)}^{-1}$. For a walk $\gamma = e_1^{\epsilon_1} \cdots e_k^{\epsilon_k}$, with $\epsilon_i \in \{\pm 1\}$, from p to q, the transport is given by $\omega(\gamma) := \Gamma_{\ell(e_k)}^{\epsilon_k} \cdots \Gamma_{\ell(e_1)}^{\epsilon_1}$. The first edge is applied first. Concatenation of walks is written in traversal order, so $\omega(\gamma\delta) = \omega(\delta)\omega(\gamma)$. The *holonomy* of a closed walk γ at v is defined as $\mathrm{Hol}_v(\gamma) := \omega(\gamma)$. The connection ω is *flat* if $\mathrm{Hol}_v(\gamma) = I$ for every closed walk.

Let $\rho : \mathsf{F}(\mathcal{R}) \to \mathrm{Aff}(D)$ denote the homomorphism from the free group on $\mathcal{R}$, where $\rho(r) = \Gamma_r$. Define $\mathsf{w}(\gamma) := \ell(e_k)^{\epsilon_k} \cdots \ell(e_1)^{\epsilon_1}$ as the label word of a walk, and set $\omega(\gamma) = \rho(\mathsf{w}(\gamma))$. Consequently, ω remains invariant under backtracks $e\bar{e}$, descends to the fundamental group of $\mathcal{G}$, and induces, for each v, a homomorphism $\mathrm{hol}_v : \pi_1(\mathcal{G}, v) \to \mathrm{Aff}(D)$ (Lemma 12, Appendix C). The connection is related to the embedding via the fit residual $\eta_e := \Gamma_{\ell(e)}\tilde{u}_p - \tilde{u}_q$ for each edge

$e = (p \to q)$, representing a displacement. Define $\varepsilon(\omega) := \max_e \|\eta_e\|$. The function ω fits a set of edges exactly if their residuals are zero.

When a distinct operator is assigned to each edge, the translation $T_{u_q - u_p}$ fits $e = (p \to q)$ exactly, and the translations around any closed walk compose to I. Consequently, every per-edge connection constructed via translations is flat for every embedding. Furthermore, per-edge operators can realize any prescribed holonomy on a square while fitting each edge exactly (Remark 16, Appendix D). However, sharing Γ_r across all edges labeled r eliminates this flexibility.

When treating relations as operators, there are two possible routes from *man* to *queen*: applying *gender* followed by *royalty*, or the reverse. The analogy remains consistent when $\Gamma_r \Gamma_g \tilde{u}_{\text{man}} = \Gamma_g \Gamma_r \tilde{u}_{\text{man}}$. However, not all pairs of relations should commute. For example, *mother-of* and *father-of* do not commute. Figures 2a–4 illustrate the two routes of an analogy and valid order dependence. The framework distinguishes these cases using the graph structure. A commutation constraint on (Γ_r, Γ_s) is imposed when $\mathcal{G}$ contains an (r, s)-square. Pairs that do not form such a square remain unconstrained (Corollary 7). Holonomy must vanish on every loop that the graph identifies as closed.



(a) Parallelogram (vector offsets). **(b)** Relation as one shared map.

Figure 2: From offsets to operators. (a) Classical analogy as a parallelogram. (b) A relation is a single affine map applied to every pair it relates.

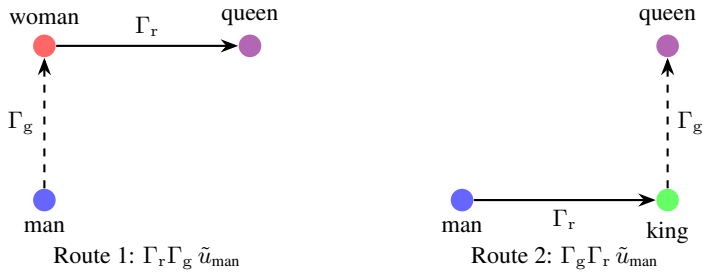


Figure 3: Two routes to *queen*. With relations as operators, consistency of the analogy square is the statement $\Gamma_r \Gamma_g \tilde{u}_{\text{man}} = \Gamma_g \Gamma_r \tilde{u}_{\text{man}}$. Traversing route 1 forward and route 2 backward is a closed walk whose transport is the commutator $\Gamma_r^{-1} \Gamma_g^{-1} \Gamma_r \Gamma_g$. The analogy is consistent at *man* iff this commutator fixes $\tilde{u}_{\text{man}}$ (Theorem 2).

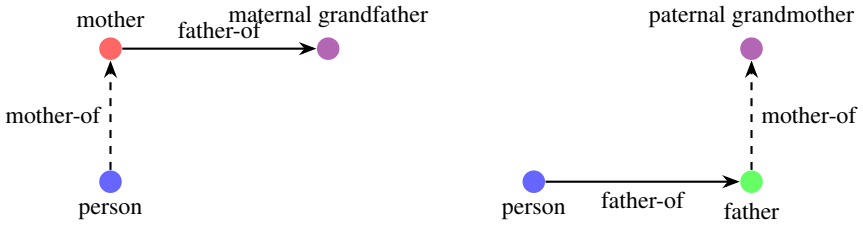


Figure 4: Valid order dependence. The two compositions end at different tokens, so the graph contains no square and no commutation constraint is imposed on $\Gamma_{\text{mother-of}}$ and $\Gamma_{\text{father-of}}$.

3 HOLONOMY OF ANALOGY SQUARES

An (r, s)-*square* in $\mathcal{G}$ is defined as a quadruple $S = (a, b, c, d)$ with edges $a \xrightarrow{r} b$, $d \xrightarrow{r} c$, $a \xrightarrow{s} d$, and $b \xrightarrow{s} c$. The associated *loop* $\gamma_S := (a \to b)(b \to c)\overline{(d \to c)}\,\overline{(a \to d)}$ at a has the label word $s^{-1}r^{-1}sr$ and corresponding holonomy.

$$\mathrm{Hol}_a(\gamma_S) = \Gamma_s^{-1}\Gamma_r^{-1}\Gamma_s\Gamma_r =: \Omega_{rs}, \tag{2}$$

the group commutator of Γ_r and Γ_s, which depends only on the relation pair based at b, the loop has the conjugate word $r\, s^{-1}r^{-1}s$, so triviality of the holonomy is base-point independent. We always use the base at which both relations originate.

Theorem 2 (Analogy squares and holonomy)**.** *Let $\Gamma_r, \Gamma_s \in \mathrm{Aff}(D)$ and let $S_i = (a_i, b_i, c_i, d_i)$, $i = 1, \ldots, n$, be (r, s)-squares of $\mathcal{G}$.*

(i) (Closure is a fixed point.) *For every $x \in \mathbb{R}^D$: $\Gamma_s\Gamma_r\tilde{x} = \Gamma_r\Gamma_s\tilde{x} \iff \Omega_{rs}\tilde{x} = \tilde{x}$.*

(ii) (Exact fit pins the holonomy at the data.) *If ω fits the four edges of S_i exactly, then $\Omega_{rs}\tilde{u}_{a_i} = \tilde{u}_{a_i}$.*

(iii) (Spanning squares force flatness.) *If ω fits every edge of $S_1, \ldots, S_n$ exactly and $\{u_{a_i}\}$ affinely spans $\mathbb{R}^D$ (i.e. $\mathrm{rank}[\tilde{u}_{a_1} \cdots \tilde{u}_{a_n}] = D{+}1$), then $\Omega_{rs} = I$.*

(iv) (Flatness closes every square.) *If $\Omega_{rs} = I$, every generated square $(\tilde{x}, \Gamma_r\tilde{x}, \Gamma_s\Gamma_r\tilde{x}, \Gamma_s\tilde{x})$ is fitted exactly, and a square of $\mathcal{G}$ fitted exactly on three edges is fitted exactly on the fourth.*

(v) (Sharpness.) *The spanning hypothesis in (iii) is necessary. For $D = 2$, there are Γ_r, Γ_s fitting one square exactly with $\Omega_{rs} \neq I$.*

All proofs are provided in Appendix C. The sharpness condition in (v) is proved through two shears on the unit square, where all four edges align exactly, yet $\|\Omega_{rs} - I\|_F \neq 0$.

Corollary 3 (The vector-offset model is the flat case)**.** *If every $\Gamma_r = T_{t_r}$ is a translation, then $\Omega_{rs} = I$ for all r, s, and an (r, s)-square is fitted exactly iff $u_b - u_a = u_c - u_d = t_r$ and $u_d - u_a = u_c - u_b = t_s$, i.e. the four embeddings form a parallelogram. In the vector-offset model, analogical consistency is determined by the fit residual and not by the holonomy.*

Proposition 4 (Stability and approximate flatness)**.** *Let $S = (a, b, c, d)$ be an (r, s)-square with $\|\eta_e\| \le \varepsilon$ on its four edges, $L_r = L(\Gamma_r)$, $L_s = L(\Gamma_s)$, where $\|\cdot\|$ is the spectral norm. The* closure defect $\delta_S := \|\Omega_{rs}\tilde{u}_a - \tilde{u}_a\|$ *and the* route disagreement $\varrho_S := \|\Gamma_s\Gamma_r\tilde{u}_a - \Gamma_r\Gamma_s\tilde{u}_a\|$ *satisfy following:*

$$\delta_S \le C_{rs}\,\varepsilon, \quad C_{rs} := \|L_s^{-1}\|\big(1{+}\|L_r^{-1}\|(2{+}\|L_s\|)\big), \qquad \varrho_S = \|L_rL_s(\Omega_{rs}\tilde{u}_a{-}\tilde{u}_a)\| \le \|L_r\|\|L_s\|\,\delta_S, \tag{3}$$

and $\delta_S \le \|\Omega_{rs} - I\|\|\tilde{u}_a\|$. If $S_1, \ldots, S_n$ have residuals $\le \varepsilon$ and base matrix $\tilde{U} = [\tilde{u}_{a_1} \cdots \tilde{u}_{a_n}]$ of rank $D{+}1$ with smallest singular value $\sigma_{\min}(\tilde{U})$, then $\|\Omega_{rs} - I\|_F \le \sqrt{n}\,C_{rs}\,\varepsilon/\sigma_{\min}(\tilde{U})$.

The factor $\sigma_{\min}(\tilde{U})^{-1}$ is the quantitative measure of the spanning hypothesis. Base points concentrated near a low-dimensional affine subspace allow large operator curvature even if every observed square closes well (Appendix E).

4 PATH INDEPENDENCE OF COMPOSITE ANALOGIES

Let $\mathcal{Q}$ denote a set of closed walks in $\mathcal{G}$, and let $K_{\mathcal{Q}}$ represent the square complex, defined as the 2-complex obtained from $\mathcal{G}$ by attaching a 2-cell along each loop in $\mathcal{Q}$. Two walks γ_1 and γ_2 from p to q constitute a commutative diagram if $\omega(\gamma_1) = \omega(\gamma_2)$, i.e., if $\gamma_1\bar{\gamma}_2$ exhibits trivial holonomy.

Theorem 5 (Trivial holonomy on elementary loops implies path independence)**.** *Assume that $\mathrm{Hol}(\gamma) = I$ for every $\gamma \in \mathcal{Q}$. Then: (i) $\mathrm{Hol}(\gamma) = I$ for every closed walk γ that is null-homotopic in $K_{\mathcal{Q}}$; (ii) if $K_{\mathcal{Q}}$ is simply connected, ω is flat and $\omega(\gamma_1) = \omega(\gamma_2)$ for any two walks from p to q, so the transport depends solely on the endpoints.*

The proof, provided in Appendix C, applies the Seifert–van Kampen theorem (Hatcher, 2002, Prop. 1.26).

Remark 6 (Homology is not enough). The less restrictive assumption that $\mathcal{Q}$ generates the cycle space $H_1(\mathcal{G})$ corresponds to the abelianization of the condition in (ii), namely that $\mathcal{Q}$ normally generates

$\pi_1(\mathcal{G})$. This assumption is insufficient because hol_v takes values in the non-abelian group $\mathrm{Aff}(D)$. For example, the square loop has word $s^{-1}r^{-1}sr$, which is trivial in homology for any labeling, yet its holonomy Ω_{rs} is non-trivial when Γ_r and Γ_s do not commute.

Corollary 7 (Relation-typed connections)**.** *Let $P \subseteq \mathcal{R} \times \mathcal{R}$ denote the set of pairs (r, s) for which $\mathcal{G}$ contains an (r, s)-square, and let $\mathcal{Q}$ be the set of all square loops. If $K_{\mathcal{Q}}$ is simply connected, then ω is flat if and only if $\Gamma_r\Gamma_s = \Gamma_s\Gamma_r$ for all $(r, s) \in P$. Pairs outside P are unconstrained, so a non-commuting pair that never closes a square remains compatible with a flat connection.*

5 Curvature and a label-free certificate

For $(r, s) \in P$, the *curvature* is defined as $\kappa_{rs} := \|\Omega_{rs} - I\|_F$, where this definition applies at the level of relation types with a fixed traversal convention. If the convention is changed, Ω_{rs} is conjugated, and $\|\cdot\|_F$ remains invariant only under orthogonal conjugation. The curvature vanishes *iff* the pair is flat. Proposition 4 establishes a bidirectional relationship between curvature, δ_S, and ϱ_S.

Let $\Gamma_r = \exp(A_r)$ and $\Gamma_s = \exp(A_s)$, where $A_r, A_s \in \mathfrak{aff}(D)$. According to the Baker–Campbell–Hausdorff formula, $\Omega_{rs} = \exp([A_s, A_r] + O(\|A\|^3))$ and $\kappa_{rs} = \|[A_r, A_s]\|_F + O(\|A\|^3)$, where $\|A\| := \max(\|A_r\|, \|A_s\|)$ (Proposition 13). Thus, the infinitesimal curvature corresponds to the Lie bracket of the generators.

Proposition 8 (Label-free certificate)**.** *Let $S = (a, b, c, d)$ be an (r, s)-square and $e_1 := \|\Gamma_s\Gamma_r\tilde{u}_a - \tilde{u}_c\|$, $e_2 := \|\Gamma_r\Gamma_s\tilde{u}_a - \tilde{u}_c\|$ the errors of the two two-hop predictions of c. Then $\max(e_1, e_2) \geq \frac{1}{2}\varrho_S$, and ϱ_S depends only on $\tilde{u}_a$ and the operators. Conversely, if the four edges have residuals $\leq \varepsilon$, then $e_1 \leq (1 + \|L_s\|)\varepsilon$ and $e_2 \leq (1 + \|L_r\|)\varepsilon$. If the pair is flat, $e_1 = e_2$ and $\varrho_S = 0$.*

If the two routes differ by ϱ_S, then at least one route deviates by $\varrho_S/2$ in the absence of access to the exact target endpoint. The certificate provides information only for a *curved* connection because $\varrho_S \equiv 0$ for vector offsets.

6 Contextual analogy

Let $\beta \in \mathbb{R}^m$ be a vector representing a pooled hidden state of the context. In language models, the representation $u_v(\beta)$ of a token is determined by its context. Consequently, the relation operators should also depend on this contextual information.

Definition 9 (Context-dependent operators)**.** For each relation r given J basis generators $A_{r,1}, \ldots, A_{r,J} \in \mathfrak{aff}(D)$, $W_r \in \mathbb{R}^{J\times m}$ and $b_r \in \mathbb{R}^J$, and define

$$\alpha_r(\beta) := \mathrm{softmax}(W_r\beta + b_r) \in \Delta^{J-1}, \qquad \Gamma_r(\beta) := \exp\Big(\textstyle\sum_{j=1}^{J} \alpha_{r,j}(\beta)A_{r,j}\Big). \tag{4}$$

The contextual connection is $\omega_\beta : r \mapsto \Gamma_r(\beta)$, with residuals $\eta_e(\beta) = \Gamma_{\ell(e)}(\beta)\tilde{u}_p(\beta) - \tilde{u}_q(\beta)$ and curvature $\Omega_{rs}(\beta)$ the commutator equation 2 of $\Gamma_r(\beta), \Gamma_s(\beta)$

For a given β, ω_β constitutes a relation-typed connection. Sections 3–5 are applicable in each context. The fit condition ensures the context-dependent operator matches the language model representation.

Proposition 10 (Properties of contextual operators)**.** *(i) $\Gamma_r(\beta) \in \mathrm{Aff}(D)$ is invertible and smooth in β. (ii) If all basis generators of r commute with all basis generators of s, then $\Omega_{rs}(\beta) = I$ for every β. (iii) To leading order the curvature is bilinear in the mixture weights, and uniformly in β,*

$$\log\Omega_{rs}(\beta) = \textstyle\sum_{j,k} \alpha_{r,j}(\beta)\,\alpha_{s,k}(\beta)\,[A_{s,k}, A_{r,j}] + O(\|A\|^3), \qquad \kappa_{rs}(\beta) \leq \max_{j,k}\|[A_{r,j}, A_{s,k}]\|_F + O(\|A\|^3) \tag{5}$$

Part (iii) motivates the use of softmax weights. A penalty applied to the basis brackets regulates all contexts simultaneously, whereas a penalty evaluated solely in training contexts influences a new context only via convexity. Specifically, if the product weights $\pi(\beta) := \alpha_r(\beta) \otimes \alpha_s(\beta)$ lies within the convex hull of those from the training contexts β_i, then $\|\log\Omega_{rs}(\beta)\|_F \leq \max_i\|\log\Omega_{rs}(\beta_i)\|_F + O(\|A\|^3)$ (Proposition 14). No guarantees exist outside this convex hull.

Remark 11 (Kernel interpolation of per-context operators). A Gaussian kernel over training contexts, $\alpha_{r,j}(\beta) \propto \exp(-\|\beta - \beta_j\|^2/\tau)$, is the special case of equation 4 such that $W_{r,j} = 2\beta_j/\tau$, $b_{r,j} = -\|\beta_j\|^2/\tau$ and $A_{r,j} = \log \Gamma_r^{(j)}$. The logarithm of an operator is fitted in the training context j only. The family is a kernel interpolation of per-context operators in the Lie algebra. We use this as the initialization in Section 8.2. Generators initialized near zero instead stay nearly parallel throughout training (each receives $\alpha_{r,j}\, \partial\mathcal{F}/\partial Z_r$ of the common exponent) and the family collapses to a context-independent operator.

The *cross-context holonomy* $\Omega_{rs}(\beta, \beta') := \Gamma_s(\beta')^{-1}\Gamma_r(\beta')^{-1}\Gamma_s(\beta)\Gamma_r(\beta)$ can have large deviation from I even if both ω_β and $\omega_{\beta'}$ are flat.

7 Curvature-regularized training objective

Let θ denote the parameters generating the embeddings (u_v or $u_v(\beta)$), and let $\Phi = \{A_{r,j}, W_r, b_r\}$ represent the operator parameters. Given a prediction loss $\mathcal{L}_{\text{pred}}(\theta)$, the following objective is proposed.

$$\mathcal{F}(\theta, \Phi) = \mathcal{L}_{\text{pred}}(\theta) + \lambda_{\text{fit}} \underbrace{\textstyle\sum_{e=(p\to q)\in\mathcal{E}} \|\Gamma_{\ell(e)}\tilde{u}_p - \tilde{u}_q\|^2}_{F_{\text{fit}}} + \lambda_\Omega \underbrace{\textstyle\sum_{S\in\mathcal{Q}} \|\Gamma_s\Gamma_r\tilde{u}_{a_S} - \Gamma_r\Gamma_s\tilde{u}_{a_S}\|^2}_{F_\Omega}, \tag{6}$$

Here, S denotes the (r, s)-squares of $\mathcal{G}$ with base a_S. In the contextual setting, operators and embeddings are evaluated at the context of each training example. The functional F_Ω represents the sum of squared route disagreements, which are regulated by fit residuals and address closure defects as established in Proposition 4. The operator-level alternative, $F_\Omega^{\text{op}} := \sum_{(r,s)\in P} \|\Gamma_s\Gamma_r - \Gamma_r\Gamma_s\|_F^2$, vanishes *iff* every pair in P is flat. This term also bounds the curvature through $\kappa_{rs} \le \|L_s^{-1}\|\|L_r^{-1}\|\|\Gamma_s\Gamma_r - \Gamma_r\Gamma_s\|_F$, where the difference has a zero last row. Unlike F_Ω, this alternative constrains the operators globally instead of just at the training base points. This is significant when those points do not span the entire space.

The objective is well posed (Proposition 15). $F_\Omega = 0$ *iff* every square closes. The trivial connection $\Gamma \equiv I$ minimizes F_Ω only. However, when any relation exhibits non-zero mean displacement, the translation connection achieves $F_\Omega = 0$ and a strictly smaller F_{fit}. Consequently, it is excluded by the fit term for all $\lambda_\Omega \ge 0$. For fixed operators, $F_{\text{fit}} + \lambda F_\Omega$ is a convex quadratic function in the embedding matrix, ensuring that gradient flow on the embeddings converges to its global minimizers (cf. Ambrosio et al., 2005, Ch. 8, 11). Furthermore, alternating minimization over Φ and θ does not increase $\mathcal{F}$. No convexity claim is made for $\mathcal{F}$ as a whole.

8 Experiments

Operators are estimated using least squares. Translations are determined by the mean displacement. Affine maps are determined using ridge regression towards the identity, $\min_{L,t} \sum \|Lu_p + t - u_q\|^2 + \varrho\|L - I\|_F^2$. Regularized fits by gradient descent on the operators with the fit term of equation 6, using the same ridge regularization term, and F_Ω or F_Ω^{op}. All protocol details are available in Appendix G. *Synthetic checks* (Appendix E) numerically check and confirm each statement. For *GloVe* (Appendix F), affine operators achieve the highest accuracy for every relation in the Google analogy set. However, fitted pairs exhibit curvature ($\kappa_{rs} = 12.1$ for gender×plural), and their routes retrieve different words on 14–67% of held-out squares. The data-level penalty increases agreement without altering κ_{rs}, as 29 base points cannot span $\mathbb{R}^{100}$. The operator-level penalty reduces κ_{rs} by a factor of five and increases agreement to 0.95.

8.1 Compositional consistency in GPT-2 representations

Test bed. Given a prefix p and suffix f, the strings w, pw, wf, and pwf (*state*, *restate*, *stated*, *restated*) constitute an (r, s)-square when all four are present in the vocabulary, ensuring that both transformation routes terminate at the same word by design. Analysis of the 200,000-word GloVe vocabulary gives 1,780 such squares across 37 prefix×suffix pairs. The 18 largest pairs with compositional prefixes are selected, resulting in 1,154 squares and 4,592 words. GPT-2 (Radford

Table 1: Compositional consistency of relation transport in GPT-2 (layer 6, $D = 64$; 18 prefix×suffix pairs, 1,154 squares, 3-fold cross-validation over squares, six templates; $\varrho = 1$ as reported in Table 6). *Single-hop* refers to accuracy on the four edges of held-out squares. *Two-hop* denotes the accuracy of $\Gamma_s\Gamma_r\tilde{u}_a$ in retrieving c (top-1 and top-10 among 4,592 candidates). *Agree* indicates both routes retrieve the same word. *Inconsist.* is defined as 1−agree. $\bar{\varrho}$ represents mean route disagreement ($1 - \cos$). κ_{rs} denotes mean operator curvature.

operators	single-hop	two-hop	top-10	agree	inconsist.	$\bar{\varrho}$	κ_{rs}
translation (flat)	0.608	0.453	0.704	1.000	0.000	0.0000	0.00
affine, ϱ=3	0.732	0.546	0.759	0.817	0.183	0.0119	4.35
+ data λ_Ω=30	0.730	0.550	0.760	0.944	0.056	0.0013	2.59
+ operator $\lambda_\Omega^{\text{op}}$=100	0.728	0.542	0.757	0.950	0.050	0.0008	0.39
affine, ϱ=10	0.707	0.550	0.757	0.886	0.114	0.0052	0.77
+ data λ_Ω=30	0.707	0.551	0.758	0.974	0.026	0.0003	0.55
+ operator $\lambda_\Omega^{\text{op}}$=100	0.705	0.549	0.758	0.952	0.048	0.0008	0.21

et al., 2019) layer-6 representations of these words are extracted in six neutral templates and projected to $D = 64$ dimensions, with retrieval performed among the $4{,}592$ representations within the same template. For each relation pair, 3-fold cross-validation is conducted over the squares. Operators are fitted on single-hop pairs, with edges of held-out squares removed, at three ridge strengths. Training is performed with or without a curvature penalty on the base points of the training squares (λ_Ω) or on the operators ($\lambda_\Omega^{\text{op}}$). Evaluation on held-out squares includes single-hop accuracy, two-hop accuracy of $\Gamma_s\Gamma_r\tilde{u}_a$, and *route agreement*, defined as the fraction of squares for which both routes retrieve the same word.

Affine operators improve accuracy at the expense of consistency, and the penalty restores balance. Table 1 (see also Figure 15 in Appendix H) presents the trade-off described in Corollary 3. The flat translation connection achieves correct composition for only 45% of held-out squares. In contrast, affine operators increase single-hop accuracy from 0.61 to 0.71–0.74 and two-hop accuracy to 0.51–0.55. However, these operators exhibit curvature (κ_{rs} ranging from 0.8 to 19, depending on the ridge), and their two routes give different words for 11–25% of held-out squares. Introducing either curvature penalty reduces the inconsistency rate. The operator-level term decreases inconsistency from $25/18/11\%$ to 5% at $\varrho = 1/3/10$, and the data-level term reduces it to $10/6/3\%$ across all 18 relation pairs. Additionally, $\bar{\varrho}$ decreases by roughly an order of magnitude or more, and κ_{rs} by a factor of 3.5–25. Two-hop accuracy remains largely unchanged. The paired difference across the 18 pairs between a regularized fit and its counterpart ranges from -0.006 to $+0.004$ in all six combinations, with 95% bootstrap intervals containing zero in five cases. The operator-level term at $\varrho = 10$ reduces accuracy by 0.6 points (CI $[-0.010, -0.002]$). The best models achieve consistency comparable to the flat connection (94–97%) while being ten points more accurate. The folds are effectively lexically disjoint (Appendix J).

One operator per relation type, and longer paths. A single global model with eleven operators (seven prefixes and four suffixes), trained jointly on all 18 relation pairs with the penalty summed across all pairs, gives similar results (Table 7, Appendix H): inconsistency is reduced from 10–23% to 2–9% in all 18 pairs, with paired accuracy differences ranging from -0.007 to $+0.005$. In the global model, squares corresponding to different pairs share edges, which enables Theorem 5 to make predictions that per-pair fits cannot test. Specifically, a held-out square (w, pw, pwf, wf) whose base has a sibling square with another suffix f' can be reached either by the two-hop route $\Gamma_f\Gamma_p$ or by the four-hop detour $\Gamma_f\Gamma_{f'}^{-1}\Gamma_p\Gamma_{f'}$ through the sibling square (and similarly through a sibling prefix). The detour is never penalized. However, the square complex formed by the union of all 1,154 squares is simply connected (Appendix J), so square flatness implies path independence throughout the entire graph. Empirically, unregularized operators map the detour and direct route to different words in 21–24% of 7,836 suffix and 1,992 prefix cases at $\varrho = 3$. With either penalty, this rate falls to 5–6%, with detour accuracy remaining unchanged or improving. Consistency across relation types that never share a square remains untested, as the graph does not provide a closed walk for such cases. Route disagreement certifies inconsistency but has limited predictive power for which compositions fail retrieval (AUC 0.52–0.57 compared to 0.75–0.85 for confidence scores, Appendix I).

8.2 CONTEXT SHIFT

Design. The six templates described above are neutral noun phrases, each served by a single operator. To evaluate the approach in Section 6, contexts were constructed in which the model transports a relation differently. Twelve sentence frames were built, representing four syntactic types with three variants each: noun frames ("`the {w}`", etc.), verb frames ("`Yesterday, they {w}`", etc.), quotation frames ("`She wrote the word {w}`", etc.), and narrative frames ("`It was late, the rain had stopped, and the old man {w}`", etc. See Appendix G for the full list). Each frame includes layer-8 representations of the $4{,}592$ words (see layers 6 and 10 in Appendix J). The sentence frames alter the transport. An operator fitted in a noun frame loses 12 points of single-hop accuracy on average when applied to verb frames (20 in the reverse direction). A context-independent operator fitted on all frames performs up to 5 points below one fitted in the specific frame. For each type, two variants are used for training, and the third is held out, so a held-out context constitutes a new sentence of a known type. *Static* baselines are context-independent, i.e., an affine operator fitted only in the canonical context "`the {w}`". One is fitted by ridge regression on pooled pairs from the eight training contexts and, to distinguish context from parametrization, the contextual family with mixture weights fixed at uniform, using one operator $\exp(\bar{A}_r)$ with identical generators, initialization, optimizer, and weight decay. The *per-context oracle* fits a separate operator in every context, including held-out ones. The *contextual* families employ $J = 8$ generators initialized at the logarithms of the per-training-context operators and kernel-initialized weights (see Remark 11), are trained exclusively in the eight training contexts, and use as descriptor $\beta \in \mathbb{R}^6$ the PCA projection of the centroid of the model's token representations in a context. Curvature penalties are applied only in the training contexts, i.e., at the data level, at the operator level, or on the basis brackets (see Proposition 10(iii)).

Static operators degrade, contextual operators recover. In the held-out contexts (Table 2), the oracle gives 0.705 single-hop and 0.497 two-hop accuracy. The canonical-context operator's performance decreases to 0.609 and 0.343, respectively, matching the oracle in noun frames but losing 14 to 28 two-hop points in verb and verb-like narrative frames (Figure 5(a)). Pooling the training contexts restores single-hop accuracy (0.680), but does not improve two-hop accuracy (0.421), because a single operator fitted across all frames is, on average, curved ($\kappa_{rs} = 17$) and path-dependent for 29% of compositions. The context-independent Lie operator $\exp(\bar{A}_r)$ demonstrates high consistency (91%), but only recovers single-hop accuracy (0.635) and achieves 0.427 two-hop accuracy. The contextual family addresses most of the performance gap, reaching 0.675 single-hop and 0.480 two-hop accuracy, which is within 0.05 of the oracle in each of the twelve contexts. Across the 18 pairs, the paired two-hop difference relative to the canonical operator is $+0.12$ (95% CI $[0.09, 0.16]$, all 18 pairs), to the pooled ridge operator $+0.02$ $[-0.01, 0.05]$ (single-hop -0.017 $[-0.031, -0.002]$), and to $\exp(\bar{A}_r)$ $+0.063$ $[0.046, 0.082]$ (single-hop $+0.046$ $[0.037, 0.055]$, all 18 pairs), while the oracle remains ahead by 0.038 $[0.022, 0.056]$ (Appendix J). The comparison with $\exp(\bar{A}_r)$ isolates the effect of context, with identical generators, initialization, and regularization. This allows the mixture to depend on β, improving performance by 4.6 single-hop and 6.3 two-hop points in previously unseen contexts.

The families are genuinely contextual. The mixture weights exhibit variation across contexts of up to 0.55. Operators assigned to contexts of different frame types differ by 33% of their distance from the identity, compared to 14% within a single type (Figure 17(c)). Employing the canonical-context operator universally (*frozen*) results in a reduction of 0.057 $[0.043, 0.071]$ two-hop and 0.022 $[0.016, 0.028]$ single-hop points in held-out contexts. The mixtures remain interpretable, i.e., in every examined family (ten relations), the weight is concentrated on two generators. In nine cases, these correspond to a noun frame and a verb frame, with the twelve contexts assigned according to syntactic type (the held-out verb frame receives 0.9 of its weight from the verb generator. See Appendix J). Thus, the descriptor recovers the noun/verb distinction of the frames based only on the representations.

Contextual operators are path dependent, and the penalty removes this dependence in unseen contexts. In the absence of a curvature term, the operator family exhibits inconsistency on 14% of held-out compositions in previously unseen contexts (ranging from 7% to 22% per pair), which exceeds the inconsistency observed for $\exp(\bar{A}_r)$ (9%). Thus, adapting to context improves accuracy but reduces consistency. Penalizing route disagreement within the training contexts reduces inconsis-

Table 2: Context shift (GPT-2 layer 8, 18 relation pairs, 3-fold cross-validation over squares): Performance on held-out squares within four distinct held-out contexts. Static baselines are context independent, fitted either in "`the {w}`" only, by ridge regression on pooled training contexts, or using the contextual family with uniform mixture weights, $\exp(\bar{A}_r)$. The oracle fits an operator in each context, including the held-out ones. The *frozen* approach applies the contextual family's canonical-context operator universally. Penalties are applied exclusively in the training contexts. For the complete results, see Table 10.

operators	single-hop	two-hop	agree	inconsist.	κ_{rs}
static affine, canonical context only	0.609	0.343	0.791	0.209	2.63
static affine, pooled contexts	0.680	0.421	0.710	0.290	16.90
static $\exp(\bar{A}_r)$, pooled contexts	0.635	0.427	0.913	0.087	0.23
per-context affine (oracle)	0.705	0.497	0.823	0.177	2.30
contextual $\Gamma_r(\beta)$, frozen at canonical β	0.650	0.413	0.848	0.152	0.96
contextual $\Gamma_r(\beta)$	0.675	0.480	0.857	0.143	0.72
+ data λ_Ω	0.674	0.479	0.953	0.047	0.51
+ operator $\lambda_\Omega^{\mathrm{op}}$	0.674	0.478	0.952	0.048	0.18
+ basis brackets λ_{basis}	0.675	0.480	0.902	0.098	0.33

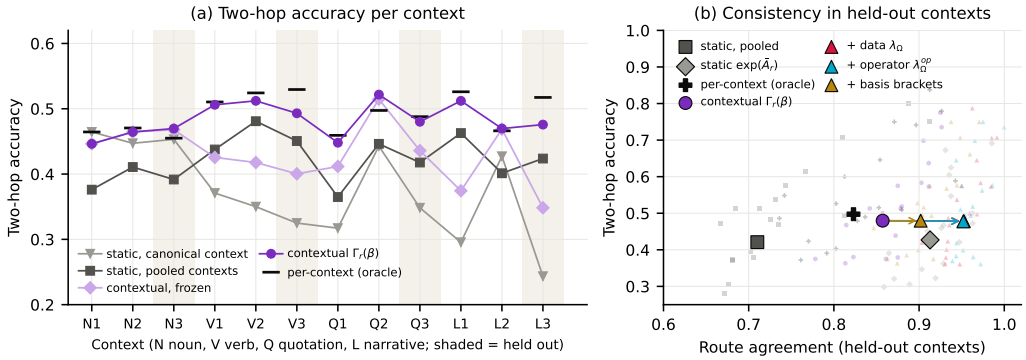


Figure 5: Context shift (GPT-2 layer 8, 18 relation pairs, held-out squares). (a) Two-hop accuracy for each context (N: noun, V: verb, Q: quotation, L: narrative; shaded contexts are held out). (b) Two-hop accuracy compared to route agreement in the held-out contexts (small markers indicate relation pairs); arrows indicate the effect of penalties applied exclusively in the training contexts. Appendix J presents the operator-difference matrix between contexts.

tency in held-out contexts to 4.7% at the data level (1% to 10% per pair), 4.8% at the operator level, and 9.8% for basis brackets, across all 18 pairs. Two-hop accuracy remains effectively unchanged (+0.001 [-0.007, 0.008], -0.004 [-0.009, 0.002], 0.000 [-0.006, 0.006]), and single-hop accuracy varies by no more than 0.4 points. The basis-bracket penalty is the least effective at the tested weight, as the fitted mixtures are highly peaked (see Appendix J). When an entire frame type is held out, corresponding to the uncovered case in Proposition 14, the operator family matches the oracle for held-out noun frames, which are similar to quotation and noun-like narrative frames. It partially closes the gap for narrative and quotation frames, but does not perform any better than a pooled static operator for verb frames. Nevertheless, the penalties still reduce inconsistency from 12%–15% to 4%–6% (Appendix J).

9 Conclusion

Interpreting relations as affine transport operators, shared across relation types, reconceptualizes analogy consistency as the flatness property of a discrete connection. In GPT-2, non-commuting operators yield higher accuracy than translations on compositional analogies, although they exhibit path dependence in up to one quarter of cases. Applying a commutator penalty enhances performance on held-out derivations, with a maximum accuracy reduction of less than one percentage point. It ensures consistency in unpenalized four-hop detours, as predicted by the theoretical framework.

Ethics statement

This work is foundational and uses only public models (GPT-2, GloVe) and word lists, and we see no direct negative application. A possible positive impact is better diagnostics of relational consistency in language models.

Reproducibility statement

All statements carry their hypotheses, and every proof is in Appendix C. The sharpness example is verified numerically. Appendix G gives the experimental details, and Appendix J gives the robustness checks. All experiments ran on two CPU cores in approximately ten hours.

AI use statement

LLMs were not used as part of the core methodology, and any use was limited to improve readability.

References

Carl Allen and Timothy Hospedales. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning*, pages 223–231. PMLR, 2019.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

Ivana Balazevic, Carl Allen, and Timothy Hospedales. Multi-relational poincaré graph embeddings. *Advances in neural information processing systems*, 32, 2019.

Federico Barbero, Cristian Bodnar, Haitz Sáez de Ocáriz Borde, Michael Bronstein, Petar Veličković, and Pietro Liò. Sheaf neural networks with connection laplacians. In *Topological, Algebraic and Geometric Learning Workshops 2022*, pages 28–36. PMLR, 2022.

Cristian Bodnar, Francesco Di Giovanni, Benjamin Chamberlain, Pietro Lio, and Michael Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in gnns. *Advances in Neural Information Processing Systems*, 35:18527–18541, 2022.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. Inducing relational knowledge from bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7456–7463, 2020.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. Low-dimensional hyperbolic knowledge graph embeddings. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6901–6914, 2020.

Dawn Chen, Joshua C. Peterson, and Thomas L. Griffiths. Evaluating vector-space models of analogy, 2017. URL https://arxiv.org/abs/1705.04416.

Taco Cohen and Max Welling. Learning the irreducible representations of commutative lie groups. In *International Conference on Machine Learning*, pages 1755–1763. PMLR, 2014.

Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *International conference on Machine learning*, pages 1321–1330. PMLR, 2019.

Marissa Connor and Christopher Rozell. Representing closed transformation paths in encoded network latent space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3666–3675, 2020.

Pim De Haan, Maurice Weiler, Taco Cohen, and Max Welling. Gauge equivariant mesh cnns: Anisotropic convolutions on geometric graphs. *arXiv preprint arXiv:2003.05425*, 2020.

Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, and Rose Yu. Automatic symmetry discovery with lie algebra convolutional network. *Advances in Neural Information Processing Systems*, 34:2503–2515, 2021.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 3519–3530, 2016.

Takuma Ebisu and Ryutaro Ichise. Toruse: Knowledge graph embedding on a lie group. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021.

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 and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 55–65, 2019.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Towards understanding linear word analogies. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3253–3262, 2019.

Louis Fournier, Emmanuel Dupoux, and Ewan Dunbar. Analogies minus analogy test: measuring regularities in word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 365–375, 2020.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, 2023.

Alex Gittens, Dimitris Achlioptas, and Michael W Mahoney. Skip-gram- zipf+ uniform= vector additivity. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–76, 2017.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, 2016.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJxeWnCcF7.

Kelvin Guu, John Miller, and Percy Liang. Traversing knowledge graphs in vector space. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 318–327, 2015.

David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkpACe1lx.

Jakob Hansen and Thomas Gebhart. Sheaf neural networks. *arXiv preprint arXiv:2012.06333*, 2020.

Allen Hatcher. *Algebraic topology*, volume 1. Cambridge university press Cambridge, 2002.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, 2023.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations*, volume 2024, pages 10504–10526, 2024.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7811–7818, 2020.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. Beliefbank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8849–8861, 2021.

Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. Canonical tensor decomposition for knowledge base completion. In *International conference on machine learning*, pages 2863–2872. PMLR, 2018.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR, 2018.

Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180, 2014.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. A logic-driven framework for consistency of neural models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3924–3935, 2019.

Yankai Lin, Zhiyuan Liu, Huanbo Luan, Maosong Sun, Siwei Rao, and Song Liu. Modeling relation paths for representation learning of knowledge bases. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 705–714, 2015.

Tal Linzen. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st workshop on evaluating vector-space representations for NLP*, pages 13–18, 2016.

Jean-Pierre Magnot. Contextuality, holonomy and discrete fiber bundles in group-valued boltzmann machines. *arXiv preprint arXiv:2509.10536*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013b.

Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D Manning. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1768, 2022.

Maximilian Nickel, Volker Tresp, Hans-Peter Kriegel, et al. A three-way model for collective learning on multi-relational data. In *Icml*, volume 11, pages 3104482–3104584, 2011.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

Malvina Nissim, Rik Van Noord, and Rob Van Der Goot. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497, 2020.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

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)*, pages 1532–1543, 2014.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2463–2473, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rajesh Rao and Daniel Ruderman. Learning lie groups for invariant visual perception. *Advances in neural information processing systems*, 11, 1998.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4902–4912, 2020.

Anna Rogers, Aleksandr Drozd, and Bofang Li. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th joint conference on lexical and computational semantics (* SEM 2017)*, pages 135–148, 2017.

Vasileios Sevetlidis and George Pavlidis. Gauge-invariant representation holonomy. *arXiv preprint arXiv:2601.21653*, 2026.

Amit Singer and H-T Wu. Vector diffusion maps and the connection laplacian. *Communications on pure and applied mathematics*, 65(8):1067–1144, 2012.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. *Advances in neural information processing systems*, 26, 2013.

Jascha Sohl-Dickstein, Ching Ming Wang, and Bruno A Olshausen. An unsupervised algorithm for learning lie group transformations. *arXiv preprint arXiv:1001.1027*, 2010.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*, 2019.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4593–4601, 2019.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pages 17282–17333, 2024.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR, 2016.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. Bert is to nlp what alexnet is to cv: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, 2021.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. Composition-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1911.03082*, 2019.

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1671–1682, 2016.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

Maurice Weiler, Patrick Forré, Erik Verlinde, and Max Welling. Coordinate independent convolutional networks–isometry and gauge equivariant convolutions on riemannian manifolds. *arXiv preprint arXiv:2106.06020*, 2021.

Canran Xu and Ruijiang Li. Relation embedding with dihedral group in knowledge graph. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 263–272, 2019.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. Quaternion knowledge graph embeddings. *Advances in neural information processing systems*, 32, 2019.

A GEOMETRIC INTUITION

A.1 CONCEPT OF A GEOMETRIC INVARIANT

The intuition is to model a relation as a transformation that maps one word to another. An analogy states that two such transformations, when applied in either order, give the same word. Geometrically, this process corresponds to transport around a loop. If the transport around the loop is the identity, the loop is considered flat. Otherwise, the loop exhibits holonomy. In this paper, we introduce a quantitative framework for holonomy and propose methods for its elimination.

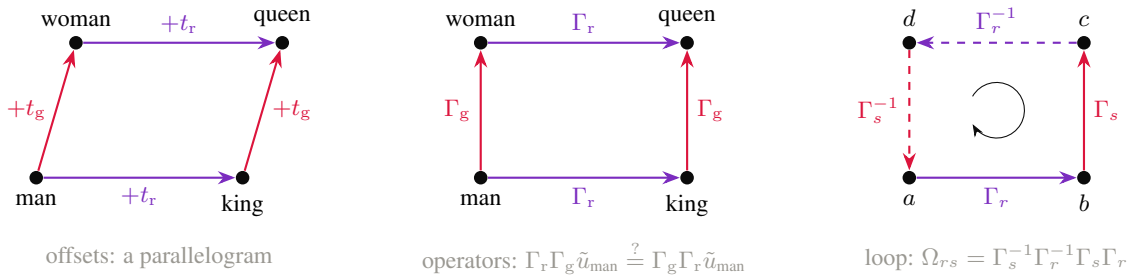


Figure 6: *a*) In the classical representation, each relation corresponds to a displacement, and an analogy forms a parallelogram. *b*) Each relation is represented as a map Γ, and the analogy evaluates whether the two compositions produce the same result. *c*) Traversing route 1 and returning via route 2 forms a closed walk, such that transport given by $\Omega_{rs} = \Gamma_s^{-1}\Gamma_r^{-1}\Gamma_s\Gamma_r$. The analogy holds at *man iff* Ω_{rs} fixes $\tilde{u}_{\text{man}}$.

A.2 THREE CONCEPTUAL LAYERS

We realize that the following three distinct concepts require clarification. The *graph* defines which words are connected by specific relations and lacks coordinates. The *fiber* associated with each word is a copy of $\mathbb{R}^{D+1}$. The embedding selects a unique point in each fiber, forming a *section*. The *connection* transports points between fibers, applying the same map Γ_r to every edge of type r. Data is incorporated through the fit condition $\Gamma_r\tilde{u}_p = \tilde{u}_q$. The deviation η_e denotes the covariant derivative of the *section*.

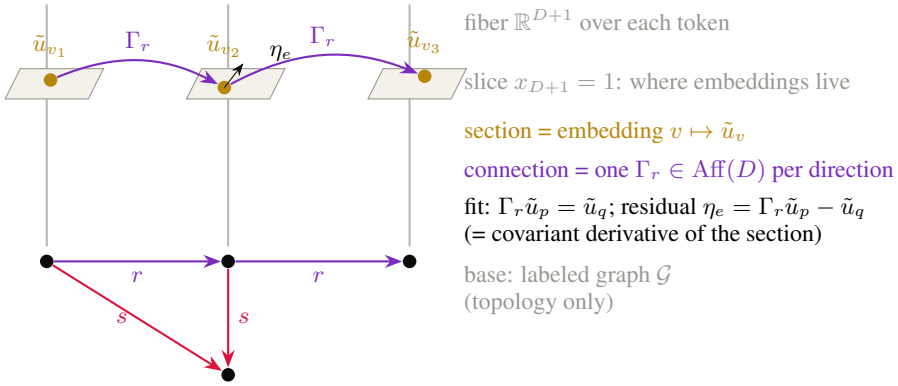
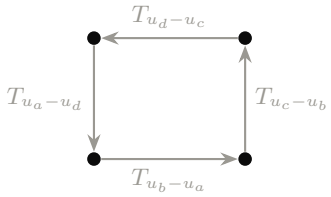


Figure 7: The base graph is shown with fibers. *Section* points appear in yellow, and transport is indicated in purple. Homogeneous coordinates ensure that every point lies on the slice $x_{D+1} = 1$. The maps preserving this slice form the affine group $\mathrm{Aff}(D)$.

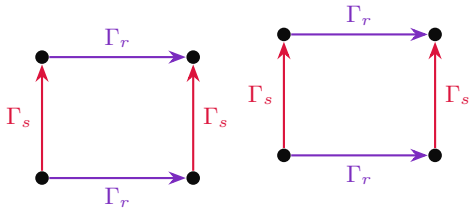
A.3 ONE MAP PER RELATION

Note that assigning a unique map to each edge hinders the extraction of meaningful information, as the translation $T_{u_q - u_p}$ precisely fits every edge. When traversing any loop, these translations cancel. This results in embeddings that appear flat. In contrast, sharing a single map Γ_r across all edges

of type r eliminates this freedom. The holonomy of a square then becomes the commutator of two relation maps, depending only on the pair (r, s). This allows the data to impose constraints.



per-edge operators: sum of displacements $= 0$
$\Rightarrow$ holonomy $= I$ for *every* embedding

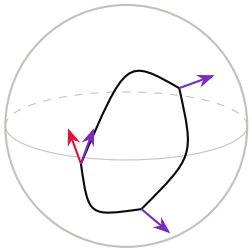


one operator per *direction*, shared by all its edges
$\Rightarrow$ holonomy $= \Gamma_s^{-1}\Gamma_r^{-1}\Gamma_s\Gamma_r$, the same for every square

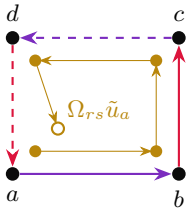
Figure 8: *a*) Assigning per-edge maps always results in cancellation leading to the identity. *b*) Sharing maps ensures that every (r, s)-square exhibits the same holonomy Ω_{rs}.

A.4 Curvature is a commutator

Consider transporting a vector around a small loop on a surface such that the vector is rotated on returning. This rotation equals the curvature multiplied by the enclosed area. In the context of a graph, the equivalent loop is a square, and its holonomy corresponds to the commutator of the two associated maps. Let $\Gamma_r = e^{A_r}$. Then, to third order, $\log \Omega_{rs} = [A_s, A_r]$. This expression parallels the formula $F(X, Y) = [\nabla_X, \nabla_Y]$ on a manifold. The quantity $\kappa_{rs} = \|\Omega_{rs} - I\|_F$ quantifies the curvature associated with the pair.



manifold: transport around a loop
rotates by $\approx$ area $\times$ curvature



graph: transport around a square
$\log \Omega_{rs} = [A_s, A_r] + O(\|A\|^3)$

$F(X, Y) = [\nabla_X, \nabla_Y]$

$\updownarrow$

$\Omega_{rs} = \Gamma_s^{-1}\Gamma_r^{-1}\Gamma_s\Gamma_r$

$\kappa_{rs} = \|\Omega_{rs} - I\|_F$

$\Gamma_r = e^{A_r}$: directions $\rightarrow$ generators

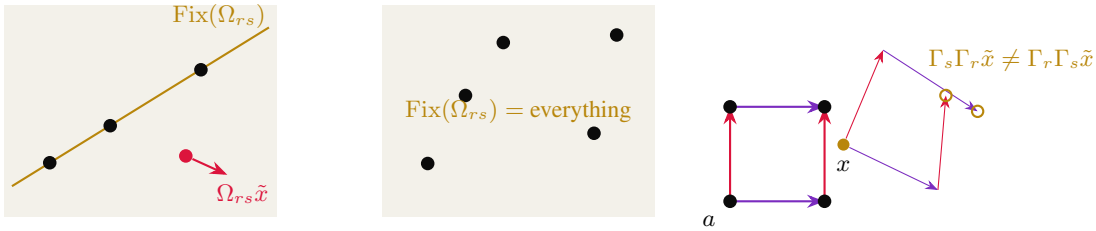
Figure 9: *a)* Parallel transport around a loop on a sphere. *b)* Transport of the base point around a square, resulting in displacement upon return. *c)* Correspondence between geometric and algebraic concepts.

A.5 Closure of a Single Square Does Not Necessarily Imply Flatness

Ω_{rs} is an affine map whose fixed points form an affine subspace. A square is defined as closed if its base point is a fixed point of the map. The pair is considered flat when every point is fixed. Empirical observations generally indicate that only specific base points are fixed. If these base points affinely span the entire space, the map must be the identity. Conversely, if they do not affinely span the entire space, the pair may exhibit curvature even when every observed square appears to close. In the paper, we demonstrate this phenomenon by examining two distinct shears on the unit square.

A.6 Vector offsets are the flat case

If each map is a translation, then all maps commute. This results in $\Omega_{rs} = I$ for every pair. In this scenario, the connection remains flat regardless of the offsets. A square fits precisely when its four points form a parallelogram, indicating that the classical configuration exhibits no holonomy. The necessary information is contained entirely in the fit. Introducing a linear component in Γ_r permits



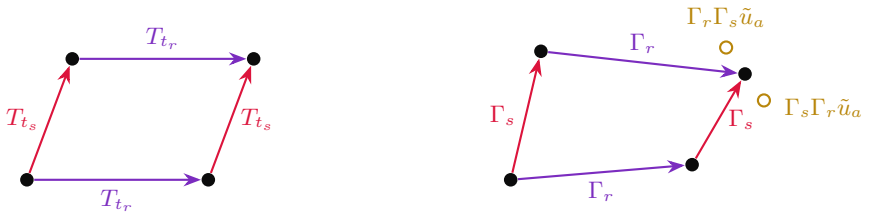
base points on Fix(Ω_{rs}): those squares close, but the pair is not flat

base points affinely span $\mathbb{R}^D$ $\Rightarrow$ $\Omega_{rs} = I$ (flat)

two shears: the unit square closes, Fix(Ω_{rs}) = $\{a\}$, the square at x does not

Figure 10: *a*) Base points on the fixed set are closely spaced, yet the pair exhibits curvature. *b*) Spanning base points enforce flatness. *c*) In the shear example, the unit square achieves closure, while the square generated at x remains open.

the connection to curve. Curved maps provide improved fit but lose commutativity. This trade-off is quantified in the experimental measurements.



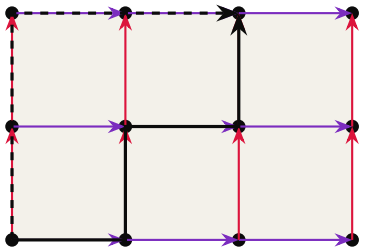
translations commute: $\Omega_{rs} = I$ always the connection is flat; exact fit $\Leftrightarrow$ parallelogram

affine operators: linear parts need not commute better fit, but the two routes can split

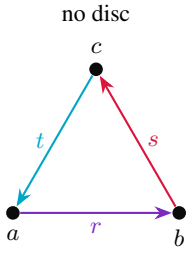
Figure 11: *a*) Translations commute, so the only consideration is whether the parallelogram closes. *b*) Affine maps can differentiate between the two routes.

A.7 Topology determines which loops are controlled.

Transport induces a homomorphism from the fundamental group of the graph. Flattening the squares removes the square loops and all elements they generate as a normal subgroup. Gluing a disc into each square modifies the topology. Thus, if the resulting space is simply connected, transport depends only on the endpoints. It is not sufficient for the squares to generate the cycle space. The group Aff(D) is non-abelian. While the square loop is trivial in homology, its holonomy is directly determined by the curvature. Relation pairs that do not form a square are not constrained.



squares glued with discs: $K_{\mathcal{Q}}$
simply connected $\Rightarrow$ any two routes agree



loop outside the normal closure of squares: $\Gamma_t\Gamma_s\Gamma_r = I$ is an extra condition

$\mathrm{hol}_v : \pi_1(\mathcal{G}, v) \to \mathrm{Aff}(D)$

squares flat $\Rightarrow$ descends to $\pi_1(K_{\mathcal{Q}})$

H_1 is not enough:

$s^{-1}r^{-1}sr = 0$ in H_1, yet hol $= \Omega_{rs}$

(squares are *relators*, not cycles)

pairs without a square: unconstrained

Figure 12: *a*) Discs are glued into squares, ensuring that two routes across the grid yield the same result. *b*): A triangle without a disc imposes an additional condition. *c*) Comparison of H_1 and π_1.

sectionContext abstracted as A Family of Connections In language models, the embedding is determined by the context β, necessitating a corresponding adjustment in the connection. Each map is defined as $\Gamma_r(\beta) = \exp\left(\sum_j \alpha_{r,j}(\beta) A_{r,j}\right)$, where $\alpha_{r,j}(\beta)$ represents softmax weights. The generator is restricted to the convex hull of a finite set of basis generators. Curvature, denoted by a bracket, is expressed as a convex combination of several basis brackets. The largest bracket provides an upper bound for curvature across all contexts. A penalty applied to training contexts generalizes to a new context only if its weights remain within the convex hull of the training weights. Transitioning out of one context and returning in another results in cross-context holonomy.

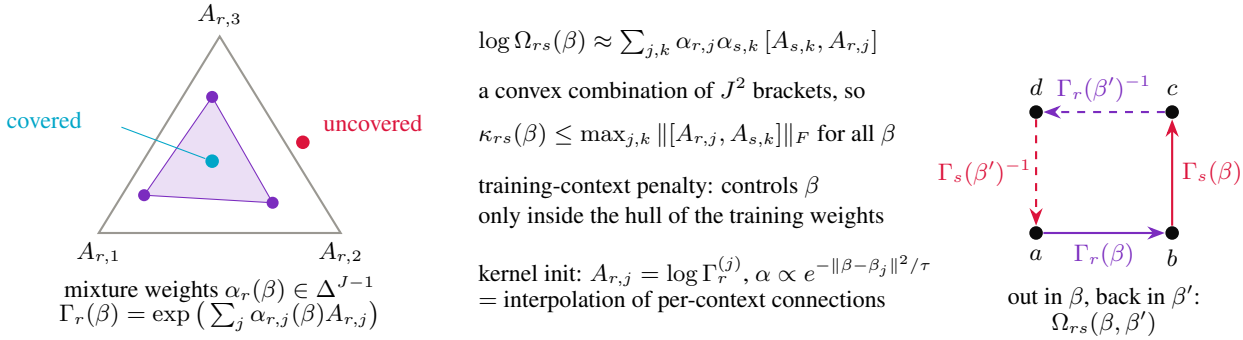


Figure 13: *a*) Mixture weights are positioned within a simplex, with the shaded region indicating the convex hull of the training contexts. *b*): The curvature bound is depicted. *c*) A loop illustrates a transition between contexts at the midpoint.

A.8 Function of the Regularizer

The fit term ensures that the section remains parallel. The curvature term enforces flatness of the connection at the observed base points. The operator-level term enforces flatness throughout the domain. The trivial connection is flat but does not provide a good fit. Translations represent the flat connection that achieves the best fit. Ridge-affine maps provide the best overall fit but exhibit curvature. The regularizer forces these maps to approach flat connections with minimal loss in fit quality.

B Related Work

Linear analogies. The vector-offset method addresses $a : b :: c : d$ analogies by conducting nearest-neighbour search around $b - a + c$ (Mikolov et al., 2013b). This regularity has also been observed in explicit PMI vectors (Levy and Goldberg, 2014), which are based on latent-variable and uniformity assumptions (Arora et al., 2016; Gittens et al., 2017). The phenomenon has been precisely characterized for SGNS and GloVe (Ethayarajh et al., 2019) and further elucidated through paraphrase analysis (Allen and Hospedales, 2019). Nevertheless, offsets have been found to lack uniformity across different relations (Gladkova et al., 2016; Vylomova et al., 2016; Drozd et al., 2016), and the standard protocol often rewards proximity and not accurately capturing relational structure (Linzen, 2016; Rogers et al., 2017; Nissim et al., 2020; Chen et al., 2017; Fournier et al., 2020). These approaches uniformly conceptualize a relation as a displacement. Corollary 3 positions this model within a broader family as the *flat* case, in which consistency depends exclusively on fit.

Relations as operators. Knowledge-graph embeddings represent relations using translations (Bordes et al., 2013), bilinear or diagonal maps (Nickel et al., 2011; Yang et al., 2014; Trouillon et al., 2016; Lacroix et al., 2018), circular correlations (Nickel et al., 2016), neural tensors (Socher et al., 2013), rotations (Sun et al., 2019; Zhang et al., 2019; Ebisu and Ichise, 2018), dihedral group elements (Xu and Li, 2019), or hyperbolic isometries (Balazevic et al., 2019; Chami et al., 2020). These relational operators are composed along paths to address multi-hop queries and facilitate message passing (Guu et al., 2015; Lin et al., 2015; Vashishth et al., 2019). In language models, relations are typically approximated by affine transformations applied to hidden states (Hernandez et al., 2024),

aligning with the linear representation hypothesis (Park et al., 2023), function and task vectors (Todd et al., 2024; Hendel et al., 2023), and the localization of factual recall (Meng et al., 2022; Geva et al., 2023). Although these methods provide the necessary operators, they do not assess the consistency of operator compositions. Moreover, path-composition objectives lack a topological framework (Theorem 5) for identifying which loops impose constraints.

Contextual representations, analogy, and consistency. Contextual representations display anisotropy and become increasingly context specific as model depth increases (Ethayarajh, 2019). Relational knowledge can be extracted from BERT (Bouraoui et al., 2020), although analogy solving by pretrained models remains limited in robustness (Ushio et al., 2021; Webb et al., 2023). Structural probes recover syntactic structure from hidden states using linear mappings (Hewitt and Manning, 2019; Tenney et al., 2019). The contextual operators proposed here are relation-typed affine probes selected based on context. Pretrained models frequently respond inconsistently to semantically related queries (Petroni et al., 2019; Jiang et al., 2020; Kassner and Schütze, 2020; Elazar et al., 2021; Ribeiro et al., 2020). Various methods have been developed to promote consistency, including logic-driven losses, belief memories, entailment-based decoding, and voting over reasoning paths (Li et al., 2019; Kassner et al., 2021; Mitchell et al., 2022; Wang et al., 2022). Holonomy is introduced as a consistency measure at the operator level providing a label-free per-instance certificate and serving as a regularizer with a topological guarantee.

Connections, Gauge Structure, and Lie-Group Parametrizations. Discrete connections on graphs provide the theoretical basis for the connection Laplacian (Singer and Wu, 2012), sheaf neural networks (Hansen and Gebhart, 2020; Bodnar et al., 2022; Barbero et al., 2022), and gauge-equivariant architectures (Cohen et al., 2019; De Haan et al., 2020; Weiler et al., 2021; Bronstein et al., 2021). In these frameworks, transport is indexed by edges or points and serves as a fixed geometric datum. Holonomy has been utilized to characterize contextuality in Boltzmann machines (Magnot, 2025) and to describe the path dependence of activations along input loops (Sevetlidis and Pavlidis, 2026). The curvature of embedding spaces (Nickel and Kiela, 2017; Gu et al., 2019; Arvanitidis et al., 2017) is typically considered an intrinsic property of the space. In contrast, the current approach learns curvature from semantic data indexed by relation types, thereby attributing curvature as a property of the relations. The use of exponentials of learned Lie-algebra generators, as introduced by Rao and Ruderman (1998) and Sohl-Dickstein et al. (2010), forms the basis for group-representation learning, symmetry discovery, and disentanglement (Cohen and Welling, 2014; Dehmamy et al., 2021; Higgins et al., 2018; Connor and Rozell, 2020). Conditioning a map on side information (Ha et al., 2017; Perez et al., 2018) exemplifies the general form of the contextual operators introduced in this work. The softmax-mixture structure of these operators ensures that the curvature of each context is a convex combination of finitely many brackets.

Why linear analogies hold, and why they fail. The observation that $\mathrm{king} - \mathrm{man} + \mathrm{woman} \approx \mathrm{queen}$ (Mikolov et al., 2013b;a) has prompted several theoretical explanations. Levy and Goldberg (2014) demonstrated that this regularity exists in sparse PMI vectors and that the additive objective can be replaced by a multiplicative one (3CosMul) with equal or superior results, indicating that the phenomenon concerns pairwise similarities. Arora et al. (2016) derived the additive structure from a generative model with a slowly evolving discourse vector, under which PMI approximates an inner product. Gittens et al. (2017) established that skip-gram with negative sampling is exactly additive for compositions of words whose contexts are uniformly distributed. Ethayarajh et al. (2019) provided necessary and sufficient conditions for an analogy to be exactly linear in SGNS and GloVe spaces. Allen and Hospedales (2019) interpreted analogies as instances of paraphrase between the word sets $\{a, d\}$ and $\{b, c\}$. Across these accounts, the shared element between the two word pairs is a *displacement*, and the analogy holds to the extent that the two displacements are equivalent. Corollary 3 demonstrates that a displacement is a special case of a relation operator whose linear component is the identity, that the resulting connection is flat for any choice of displacements, and therefore that these accounts characterize the exact *fit* of a flat connection. Questions of holonomy do not arise in these cases because there is no failure of commutativity.

Empirical criticisms of the analogy benchmark remain highly pertinent. Gladkova et al. (2016) and Vylomova et al. (2016) demonstrated that offsets are not uniform across relation types and that derivational morphology is particularly poorly represented. Drozd et al. (2016) found that a classifier over word pairs outperforms vector arithmetic. Linzen (2016) and Rogers et al. (2017) established that

the standard protocol, which excludes query words from candidate sets, artificially inflates accuracy and that much of the observed signal is attributable to the proximity of d to c in place of the intended relation. Nissim et al. (2020) further showed that this exclusion produces "biased" analogies. Chen et al. (2017) and Fournier et al. (2020) introduced evaluation methods that distinguish relational regularity from neighborhood structure. The present test bed addresses the two primary artifacts,i.e., squares are closed by design, as the fourth word is determined through string operations instead of subjective similarity judgments, and the key metric, route agreement, compares two predictions directly instead of referencing a neighbor of the query.

Relations as maps. In the knowledge-graph-embedding literature, relations have been conceptualized as maps from an early stage. Examples include translations in TransE (Bordes et al., 2013), full matrices in RESCAL (Nickel et al., 2011), diagonal matrices in DistMult (Yang et al., 2014), and complex diagonal matrices in ComplEx (Trouillon et al., 2016), which incorporates nuclear-norm regularization (Lacroix et al., 2018). Other approaches involve circular correlation in HolE (Nickel et al., 2016), bilinear tensors (Socher et al., 2013), and rotations in RotatE (Sun et al., 2019), QuatE (Zhang et al., 2019), and TorusE (Ebisu and Ichise, 2018). Notably, TorusE places entities on a compact Lie group. Xu and Li (2019) embed relations as elements of dihedral groups specifically because these groups are non-abelian and can represent non-commutative relation pairs. Additionally, hyperbolic models (Balazevic et al., 2019; Chami et al., 2020) employ isometries of hyperbolic space as relation maps. The composition of relation maps has been utilized to answer path queries (Guu et al., 2015), to regularize embeddings with relation paths (Lin et al., 2015), and as the message function in relational graph networks (Vashishth et al., 2019). To the best of our knowledge, these studies do not address the question posed here, i.e., whether the composition of two relation maps is independent of order and route when the data indicates it should be. Furthermore, to our knowledge, these do not provide a topological account (Theorem 5) identifying which compositions are constrained by specific loops. In language models, the linear relational embedding of Hernandez et al. (2024) fits an affine map from subject to object representations for each relation and demonstrates accuracy for many relations. Park et al. (2023) formalize the linear representation hypothesis and its geometric properties. Function vectors (Todd et al., 2024) and task vectors (Hendel et al., 2023) are single vectors added to the residual stream to implement in-context relations. Causal tracing and subsequent analyses locate factual recall in mid-layer MLPs (Meng et al., 2022; Geva et al., 2023). Collectively, these works establish the existence of relation operators in transformer hidden states. Building on this foundation, the present experiments measure the consistency of their compositions.

Contextual representations and analogy. Ethayarajh (2019) demonstrated that contextual representations are highly anisotropic and become increasingly context specific in upper layers. This observation motivates the context-shift experiment's sweep over layers. Bouraoui et al. (2020) induced relational knowledge from BERT using templates. Ushio et al. (2021) reported that pretrained models perform poorly on analogy tasks in a zero-shot setting, whereas Webb et al. (2023) found that large models excel at abstract analogy problems. The structural probes introduced by Hewitt and Manning (2019) and the edge-probing suite of Tenney et al. (2019) extract linguistic structure from hidden states using linear maps trained per layer. The relation operators presented here function as probes of a similar kind, but differ in that they probe a *group-valued* structure, specifically the composition of relations, and the probe is selected based on context.

Consistency. The consistency of pretrained models across paraphrases, negations, and related queries has been evaluated using cloze-style probes (Petroni et al., 2019; Jiang et al., 2020; Kassner and Schütze, 2020; Elazar et al., 2021), behavioural tests (Ribeiro et al., 2020), and has been enforced through logic-driven losses (Li et al., 2019), belief memories (Kassner et al., 2021), entailment-based decoding (Mitchell et al., 2022), and majority voting over reasoning paths (Wang et al., 2022). These approaches operate at the level of model outputs. In contrast, the holonomy of a relation-typed connection operates at the level of representations. It does not require paraphrases, as the two derivations of an analogy are generated by the graph itself. It is defined for a set of relations simultaneously, so the topological result identifies which longer compositions become consistent when squares are consistent. Additionally, it provides a per-instance certificate that does not require a label (Proposition 8). The notion of "consistency" in Wang et al. (2022) refers to agreement between sampled reasoning paths of a fixed model, which differs from the path independence of transport examined in this study.

Connections, gauge structure, and curvature in machine learning. In machine learning, connections on graphs are represented by the connection Laplacian and vector diffusion maps (Singer and Wu, 2012). In these frameworks, a transport operator is assigned to each edge as an orthogonal matrix, estimated from data, and the consistency of these operators around loops assesses the quality of local alignment. Similarly, the cellular sheaf Laplacians introduced by Hansen and Gebhart (2020), Bodnar et al. (2022), and Barbero et al. (2022) employ analogous structures with either learned or fixed restriction maps, facilitating diffusion on heterophilous graphs. Gauge-equivariant networks (Cohen et al., 2019; De Haan et al., 2020; Weiler et al., 2021; Bronstein et al., 2021) assume a given connection on a mesh or manifold and construct layers that maintain equivariance under changes of local frame.

In these cases, transport is indexed by edges or by points of a manifold, which aligns with the context in which Remark 16 demonstrates that holonomy does not provide informative semantic data. The relation-typed connection utilizes a single operator across all edges of a given type, resulting in holonomy that is the commutator of two relation operators. Holonomy has been employed to characterize contextuality in group-valued Boltzmann machines (Magnot, 2025) and, in a closely related approach, to quantify the path dependence of network activations along closed loops in input space (Sevetlidis and Pavlidis, 2026). In the latter, the loops are continuous input paths, whereas in the present work, they are discrete walks of relations within a semantic graph. Curvature is also recognized as a property of embedding spaces, such as in hyperbolic and mixed-curvature embeddings (Nickel and Kiela, 2017; Gu et al., 2019) and in the pull-back metric of generative models (Arvanitidis et al., 2017). However, the curvature examined here pertains to a connection on a trivial bundle over a graph, representing a property of the relations, and remains non-zero even in flat Euclidean space.

Lie-group parametrizations and conditional maps. The methodology of learning transformations as exponentials of Lie-algebra generators was first introduced by Rao and Ruderman (1998) and Sohl-Dickstein et al. (2010). Subsequent studies include Cohen and Welling (2014), which learn irreducible representations of commutative groups, and Dehmamy et al. (2021), which identify symmetry groups using Lie-algebra-parametrized convolutions. Higgins et al. (2018) that characterize disentanglement via group actions and Connor and Rozell (2020), who model closed transformation paths in latent space. The contextual operators proposed in this work are defined as exponentials of context-dependent convex combinations of generators. The softmax structure guarantees that the curvature in each context forms a convex combination of finitely many brackets (Proposition 10(iii)), a characteristic not exhibited by general hypernetworks (Ha et al., 2017) or feature-wise modulation (Perez et al., 2018). The initialization outlined in Remark 11 allows the family to function as a kernel interpolation of per-context operators in the Lie algebra, thereby establishing authentic context dependence. In the absence of this mechanism, the generators remain parallel, and the family collapses to a single operator (Appendix J). Compositional generalization (Lake and Baroni, 2018; Hupkes et al., 2020) evaluates whether a model that has acquired two operations can accurately compose them in previously unseen combinations. The form of consistency discussed in this study constitutes a necessary condition at the representational level, since a composition whose outcome depends on the specific route cannot generalize systematically.

C Additional Proofs and Statements

The following statements, previously summarized in the main text, are presented here with their complete proofs, along with the proofs of the associated results.

Lemma 12 (Label homomorphism and homotopy invariance)**.** *Let $\mathsf{F}(\mathcal{R})$ denote the free group generated by $\mathcal{R}$, and $\rho : \mathsf{F}(\mathcal{R}) \to \mathrm{Aff}(D)$ be the homomorphism defined by $\rho(r) = \Gamma_r$. For a walk γ, define its label word as $\mathsf{w}(\gamma) := \ell(e_k)^{\epsilon_k} \cdots \ell(e_1)^{\epsilon_1}$. It follows that $\omega(\gamma) = \rho(\mathsf{w}(\gamma))$. Note that $\omega(\gamma)$ remains invariant under the insertion or deletion of a backtrack $e\bar{e}$. Consequently, ω descends to the fundamental groupoid of $\mathcal{G}$ and, for each vertex v, to a homomorphism $\mathrm{hol}_v : \pi_1(\mathcal{G}, v) \to \mathrm{Aff}(D)$.*

Proposition 13 (Infinitesimal curvature is a Lie bracket)**.** *If $\Gamma_r = \exp(A_r)$, $\Gamma_s = \exp(A_s)$ with $A_r, A_s \in \mathfrak{aff}(D)$, then $\Omega_{rs} = \exp\big([A_s, A_r] + O(\|A\|^3)\big)$ and $\kappa_{rs} = \|[A_r, A_s]\|_F + O(\|A\|^3)$, where $\|A\| := \max(\|A_r\|, \|A_s\|)$ and $[X, Y] = XY - YX$.*

Proposition 14 (Coverage of unseen contexts)**.** *Let $\pi(\beta) := \alpha_r(\beta) \otimes \alpha_s(\beta) \in \mathbb{R}^{J \times J}$ be the product weights of a context and $\beta_1, \ldots, \beta_n$ training contexts. If $\pi(\beta)$ lies in the convex hull of*

$\{\pi(\beta_1), \ldots, \pi(\beta_n)\}$, *then* $\|\log \Omega_{rs}(\beta)\|_F \leq \max_i \|\log \Omega_{rs}(\beta_i)\|_F + O(\|A\|^3)$. *Specifically, if the mixtures are context independent, every context has the same curvature as the training contexts.*

Proposition 15 (Well-posedness)**.** *(i) The functional F_Ω is non-negative, attaining equality* iff *every square of $\mathcal{G}$ closes, i.e., $\Omega_{rs}\tilde{u}_{a_S} = \tilde{u}_{a_S}$ for all S. (ii) The functional F_Ω is minimized by the trivial connection $\Gamma \equiv I$. If there exists a relation r with non-zero mean displacement $t_r := \mathrm{mean}_{e:\ell(e)=r}(u_q - u_p)$, then $\Gamma \equiv I$ does not minimize $F_{\mathrm{fit}} + \lambda F_\Omega$ for any $\lambda \geq 0$. In this case, the translation connection T_t satisfies $F_\Omega(T_t) = 0$ and $F_{\mathrm{fit}}(T_t) < F_{\mathrm{fit}}(I)$. (iii) For fixed operators, the objective $F_{\mathrm{fit}} + \lambda F_\Omega$ is a convex quadratic function of the embedding matrix $U \in \mathbb{R}^{|V| \times D}$. Consequently, gradient flow on U converges to the set of global minimizers, and does so at an exponential rate if the quadratic form is positive definite. (iv) Alternating minimization over Φ (with θ fixed) and θ (with Φ fixed) does not increase $\mathcal{F}$.*

Proof of Lemma 12. The first claim is a direct consequence of the definition of transport, specifically $\omega(\gamma) = \Gamma^{\epsilon_k}_{\ell(e_k)} \cdots \Gamma^{\epsilon_1}_{\ell(e_1)}$, as interpreted in $\mathsf{F}(\mathcal{R})$. A backtrack leads to $\Gamma^{-1}_{\ell(e)}\Gamma_{\ell(e)} = I$ in the product. Because homotopy classes of walks in a graph correspond exactly to reduced walks, ω is well defined on these classes, and the composition of classes corresponds to matrix multiplication. □

Proof of Theorem 2. (i) $\Omega_{rs}\tilde{x} = \tilde{x} \iff \Gamma_s^{-1}\Gamma_r^{-1}\Gamma_s\Gamma_r\tilde{x} = \tilde{x} \iff \Gamma_s\Gamma_r\tilde{x} = \Gamma_r\Gamma_s\tilde{x}$, multiplying by the invertible $\Gamma_r\Gamma_s$.

(ii) Under exact fit, $\Gamma_r\tilde{u}_a = \tilde{u}_b$, $\Gamma_s\tilde{u}_b = \tilde{u}_c$, $\Gamma_s\tilde{u}_a = \tilde{u}_d$, and $\Gamma_r\tilde{u}_d = \tilde{u}_c$. Therefore, $\Gamma_s\Gamma_r\tilde{u}_a = \tilde{u}_c = \Gamma_r\Gamma_s\tilde{u}_a$, and the result in (i) applies.

(iii) By part (ii), Ω_{rs} fixes $\tilde{u}_{a_1}, \ldots, \tilde{u}_{a_n}$. The element $\Omega_{rs} \in \mathrm{Aff}(D)$ acts on $\mathbb{R}^D$ as $x \mapsto Lx + t$, and its fixed-point set $\{x : (L - I)x = -t\}$ forms an affine subspace. Because this set contains an affinely spanning set, it must be all of $\mathbb{R}^D$. This implies $L = I$ and $t = 0$. Note that a single fixed point is insufficient because the fixed-point set of a non-identity affine map may be a point, a line, or any proper affine subspace.

(iv) Since $\Gamma_s\Gamma_r\tilde{x} = \Gamma_r\Gamma_s\tilde{x}$ by part (i), the fourth edge $\Gamma_s\tilde{x} \to \Gamma_s\Gamma_r\tilde{x}$ is transported exactly by Γ_r. Therefore, for a square in $\mathcal{G}$ with three exact edges, $\tilde{u}_c = \Gamma_s\Gamma_r\tilde{u}_a = \Gamma_r\Gamma_s\tilde{u}_a = \Gamma_r\tilde{u}_d$.

(v) Let $u_a = (0, 0)$, $u_b = (1, 0)$, $u_c = (1, 1)$, and $u_d = (0, 1)$. Define $\Gamma_r = (M, (1, 0))$ and $\Gamma_s = (L, (0, 1))$, where the shear matrices are $M = \left(\begin{smallmatrix} 1 & 0 \\ m & 1 \end{smallmatrix}\right)$ and $L = \left(\begin{smallmatrix} 1 & k \\ 0 & 1 \end{smallmatrix}\right)$. In this configuration, $\Gamma_r u_a = u_b$, $\Gamma_r u_d = Me_2 + e_1 = u_c$, $\Gamma_s u_a = u_d$, and $\Gamma_s u_b = Le_1 + e_2 = u_c$, indicating that all four edges align exactly. However, $\Gamma_s\Gamma_r x = LMx + (1, 1)$ and $\Gamma_r\Gamma_s x = MLx + (1, 1)$, and $LM \neq ML$ when $km \neq 0$. Therefore, $\Omega_{rs} \neq I$. Numerically, $\|\Omega_{rs} - I\|_F = 0.41$ for $k = 0.7$, $m = -0.4$ (Refer Appendix E). The only base point of the square is u_a, which does not affinely span $\mathbb{R}^2$. □

Proof of Corollary 3. Translations commute. Exact fit of the four edges is $u_b = u_a + t_r$, $u_c = u_d + t_r$, $u_d = u_a + t_s$, $u_c = u_b + t_s$, which is the parallelogram condition of the corollary. Conversely that condition gives all four equalities. □

Proof of Proposition 4, first part. Let $\eta_1 = \Gamma_r\tilde{u}_a - \tilde{u}_b$, $\eta_2 = \Gamma_s\tilde{u}_b - \tilde{u}_c$, $\eta_3 = \Gamma_r\tilde{u}_d - \tilde{u}_c$, $\eta_4 = \Gamma_s\tilde{u}_a - \tilde{u}_d$; these are displacements, and an affine map Γ satisfies $\Gamma(x + \eta) = \Gamma x + L(\Gamma)\eta$ for displacements η. Then $\Gamma_s\Gamma_r\tilde{u}_a = \Gamma_s(\tilde{u}_b + \eta_1) = \tilde{u}_c + \eta_2 + L_s\eta_1$; $\Gamma_r^{-1}\tilde{u}_c = \Gamma_r^{-1}(\Gamma_r\tilde{u}_d - \eta_3) = \tilde{u}_d - L_r^{-1}\eta_3$; $\Gamma_s^{-1}\tilde{u}_d = \Gamma_s^{-1}(\Gamma_s\tilde{u}_a - \eta_4) = \tilde{u}_a - L_s^{-1}\eta_4$. Composing,

$$\Omega_{rs}\tilde{u}_a - \tilde{u}_a = L_s^{-1}\big(L_r^{-1}(\eta_2 + L_s\eta_1 - \eta_3) - \eta_4\big),$$

and equation 3 follows from the triangle inequality. For the identity for ϱ_S in equation 3, $\Gamma_r\Gamma_s\Omega_{rs} = \Gamma_s\Gamma_r$, so $\Gamma_s\Gamma_r\tilde{u}_a - \Gamma_r\Gamma_s\tilde{u}_a = \Gamma_r\Gamma_s(\Omega_{rs}\tilde{u}_a - \tilde{u}_a)$, and the argument is a displacement. The last bound is immediate. □

Proof of Proposition 4, second part (approximate flatness). $(\Omega_{rs} - I)\tilde{U} = \Delta$ with columns $\Omega_{rs}\tilde{u}_{a_i} - \tilde{u}_{a_i}$ of norm $\leq C_{rs}\varepsilon$ by equation 3. Since $\tilde{U}$ has full row rank, $\Omega_{rs} - I = \Delta\tilde{U}^+$ with $\tilde{U}^+$ the pseudo-inverse and $\|\tilde{U}^+\| = 1/\sigma_{\min}(\tilde{U})$; hence $\|\Omega_{rs} - I\|_F \leq \|\Delta\|_F\|\tilde{U}^+\| \leq \sqrt{n}\, C_{rs}\varepsilon/\sigma_{\min}(\tilde{U})$. □

Proof of Theorem 5. For a given base vertex v, let $\mathrm{hol}_v : \pi_1(\mathcal{G}, v) \to \mathrm{Aff}(D)$ be the homomorphism of Lemma 12. For $\gamma \in \mathcal{Q}$ based at u and any walk δ from v to u, the loop $\delta\gamma\bar{\delta}$ is based at v and $\mathrm{hol}_v(\delta\gamma\bar{\delta}) = \omega(\delta)^{-1}\mathrm{Hol}(\gamma)\omega(\delta) = I$. Hence the normal subgroup $N \trianglelefteq \pi_1(\mathcal{G}, v)$ generated by the classes of all such loops lies in $\ker \mathrm{hol}_v$. By the Seifert–van Kampen theorem, attaching 2-cells along the loops of $\mathcal{Q}$ gives $\pi_1(K_{\mathcal{Q}}, v) \cong \pi_1(\mathcal{G}, v)/N$ (Hatcher, 2002, Prop. 1.26), and a closed walk at v is null-homotopic in $K_{\mathcal{Q}}$ iff its class lies in N. This proves (i) for walks based at v. A closed walk based elsewhere is conjugated to v by a connecting path. This preserves both null-homotopy and triviality of holonomy. If $K_{\mathcal{Q}}$ is simply connected then $N = \pi_1(\mathcal{G}, v)$, so $\mathrm{hol}_v \equiv I$ and every closed walk has trivial holonomy. For $\gamma_1, \gamma_2 : p \to q$, $\omega(\gamma_2)^{-1}\omega(\gamma_1) = \omega(\gamma_1\bar{\gamma}_2) = I$. □

Proof of Corollary 7. ($\Leftarrow$) is Theorem 5(ii) with $\mathrm{Hol}(\gamma_S) = \Omega_{rs} = I$. ($\Rightarrow$): a flat connection has $\Omega_{rs} = \mathrm{Hol}_a(\gamma_S) = I$ on every square. Pairs outside P do not occur in any elementary loop, and by Lemma 12 every closed walk's holonomy is ρ of a product of conjugates of the words $s^{-1}r^{-1}sr$, $(r, s) \in P$, which is unaffected by other pairs commuting. □

Proof of Proposition 13. By the Baker–Campbell–Hausdorff formula $e^X e^Y = \exp(X + Y + \frac{1}{2}[X, Y] + O(3))$. Hence $e^{-A_s}e^{-A_r} = \exp(-(A_r + A_s) + \frac{1}{2}[A_s, A_r] + O(3))$ and $e^{A_s}e^{A_r} = \exp((A_r + A_s) + \frac{1}{2}[A_s, A_r] + O(3))$. Applying the formula once more, the sum of the two exponents is $[A_s, A_r] + O(3)$ and their bracket is $\frac{1}{2}[-(A_r + A_s), A_r + A_s] + O(3) = O(3)$. The second identity follows from $\exp(X) = I + X + O(\|X\|^2)$ and $\|[A_s, A_r]\| = \|[A_r, A_s]\|$. □

Proof of Proposition 8. $\varrho_S = \|(\Gamma_s\Gamma_r\tilde{u}_a - \tilde{u}_c) - (\Gamma_r\Gamma_s\tilde{u}_a - \tilde{u}_c)\| \le e_1 + e_2 \le 2\max(e_1, e_2)$. For the upper bounds, $\Gamma_s\Gamma_r\tilde{u}_a = \Gamma_s(\tilde{u}_b + \eta_1) = \tilde{u}_c + \eta_2 + L_s\eta_1$, so $e_1 \le \|\eta_2\| + \|L_s\|\|\eta_1\|$, and symmetrically for e_2. Flatness gives $\Gamma_s\Gamma_r = \Gamma_r\Gamma_s$. □

Proof of Proposition 10. (i) $\mathfrak{aff}(D)$ is closed under linear combinations, so the exponent $Z = \left(\begin{smallmatrix} M & \tau \\ 0 & 0 \end{smallmatrix}\right)$ lies in $\mathfrak{aff}(D)$, and $\exp Z = \left(\begin{smallmatrix} e^M & \varphi(M)\tau \\ 0 & 1 \end{smallmatrix}\right)$ with $\varphi(M) = \sum_{k\ge 0} M^k/(k+1)!$ lies in $\mathrm{Aff}(D)$ with invertible linear part e^M; exp and softmax are smooth. (ii) If $[A_{r,j}, A_{s,k}] = 0 \ \forall\ j, k$, the two exponents commute. Hence their exponentials commute. (iii) We apply Proposition 13 with $A_r = \sum_j \alpha_{r,j}A_{r,j}$ and $A_s = \sum_k \alpha_{s,k}A_{s,k}$ and expand the bracket in bilinear fashion. The bound follows because $\alpha_r(\beta) \otimes \alpha_s(\beta)$ lies in the simplex spanned by the J^2 vertices. So the bilinear form is a convex combination of the brackets $[A_{s,k}, A_{r,j}]$, whose norm is at most the largest of them. □

Proof of Proposition 14. By equation 5, $\log \Omega_{rs}(\beta) = \sum_{j,k} \pi(\beta)_{jk}[A_{s,k}, A_{r,j}] + O(3)$ is linear in $\pi(\beta)$. If $\pi(\beta) = \sum_i \lambda_i \pi(\beta_i)$ with λ in the simplex, then $\log \Omega_{rs}(\beta) = \sum_i \lambda_i \log \Omega_{rs}(\beta_i) + O(3)$ The bound follows from the convexity of the norm. □

Proof of Proposition 15. (i) is Theorem 2(i). (ii) With $\Gamma \equiv I$, $F_\Omega = 0$ and $F_{\mathrm{fit}}(I) = \sum_e \|u_q - u_p\|^2$. We know that translations commute, so $F_\Omega(T_t) = 0$. For each r, the choice t_r is the least-squares minimizer of $\sum_{\ell(e)=r} \|u_q - u_p - t\|^2$, which is strictly smaller than the value at $t = 0$ when $t_r \neq 0$. (iii) Each summand of F_{fit} is $\|L_r u_p + t_r - u_q\|^2$, a convex quadratic in (u_p, u_q). Each summand of F_Ω is $\|(L_sL_r - L_rL_s)u_a + (L_st_r + t_s - L_rt_s - t_r)\|^2$, a convex quadratic in u_a. Also, sums of convex quadratics are convex quadratics, and the gradient flow of a convex quadratic converges to its minimizers exponentially fast when it is strongly convex. (iv) Each partial minimization weakly decreases $\mathcal{F}$. □

D Additional remarks

D.1 Per-edge connections carry no analogy information

Remark 16 (Why operators are shared across edges). If each edge $e = (p \to q)$ is assigned its own operator, it is always possible to select the translation $T_{u_q - u_p}$, which fits exactly. For any closed walk, the composition of these translations gives $T_0 = I$ because the displacements cancel. Consequently, every per-edge connection defined by translations is flat for any embedding. Note that its holonomy does not convey additional information. We formalize this observation using the following proposition.

Let ω' assign to each edge $e = (p \to q)$ a distinct operator $\Gamma_e \in \mathrm{Aff}(D)$.

Proposition 17. *(i) If every Γ_e is a translation fitting e exactly, then ω' is flat for every embedding $u : V \to \mathbb{R}^D$. (ii) For any square (a, b, c, d) of* distinct *embedded points and any $H \in \mathrm{Aff}(D)$ with $H\tilde{u}_a = \tilde{u}_a$, there are $\Gamma_{ab}, \Gamma_{bc}, \Gamma_{dc}, \Gamma_{ad} \in \mathrm{Aff}(D)$ fitting all four edges exactly with holonomy $\Gamma_{ad}^{-1}\Gamma_{dc}^{-1}\Gamma_{bc}\Gamma_{ab} = H$.*

Proof. (i) $\Gamma_e = T_{u_q - u_p}$, and around any closed walk the translation vectors sum to zero, so the product is $T_0 = I$. (ii) Choose any $\Gamma_{ab}, \Gamma_{bc}, \Gamma_{ad}$ fitting their edges, for example, translations, say. Then $\Gamma_{dc} := \Gamma_{bc}\Gamma_{ab}H^{-1}\Gamma_{ad}^{-1}$ satisfies $\Gamma_{dc}\tilde{u}_d = \Gamma_{bc}\Gamma_{ab}H^{-1}\tilde{u}_a = \Gamma_{bc}\Gamma_{ab}\tilde{u}_a = \tilde{u}_c$, fits $d \to c$, and gives holonomy H. □

Therefore, when using per-edge operators, the holonomy is not constrained by the embedding. The relation-typed connection defined in Definition 1 imposes this constraint.

D.2 Homotopy invariance of transport

For completeness, we use Lemma 12. The fundamental groupoid of a graph, considered as a one-dimensional CW complex, has morphisms given by reduced walks, i.e., walks without backtracking $e\bar{e}$. Two walks are homotopic relative to endpoints *iff* they have the same reduction (Hatcher, 2002, §1.A). Attaching a 2-cell along a loop γ imposes the relation $[\gamma] = 1$ in π_1 (Hatcher, 2002, Prop. 1.26). Refer to Seifert–van Kampen theorem applied in Theorem 5.

D.3 Conjugation and base points

Given a closed walk γ based at a and a walk δ from a to b, the composite walk $\bar{\delta}\gamma\delta$ [1] forms a closed walk at b. The relation $\omega(\bar{\delta}\gamma\delta) = \omega(\delta)\omega(\gamma)\omega(\delta)^{-1}$ holds true. Consequently, Hol_b of the rebased loop is conjugate to $\mathrm{Hol}_a(\gamma)$. This indicates that triviality is independent of the base point. On the contrary, $\|\mathrm{Hol} - I\|_F$ is not base-point independent. It remains invariant only under orthogonal conjugation. Therefore, the curvature κ_{rs} in Section 5 assumes a fixed traversal convention for the square.

E Synthetic verification

Commuting ground-truth operators $\Gamma_r = \exp A_r$ and $\Gamma_s = \exp A_s$ are generated in $\mathrm{Aff}(D)$, where A_s is a polynomial in A_r and $[A_r, A_s] = 0$. Squares $(a, \Gamma_r a, \Gamma_s\Gamma_r a, \Gamma_s a)$ are constructed from random base points a, and additive Gaussian noise of scale ε is applied to all four points. We provide a summary of the validation checks in Figure 14.

Sharpness of the spanning hypothesis. The two-dimensional example of Theorem 2(v) with $k = 0.7$, $m = -0.4$ has all four edge residuals equal to 0 and closure defect $\delta_S = 1.1 \times 10^{-16}$ at the square's base point, but $\kappa_{rs} = 0.412$ and route disagreement 1.01 at the point $x = (2, 3)$. The square closes, the connection is not flat, and another square with the same relations does not close.

Fundamental group versus cycle space. Given commuting operators Γ_r and Γ_s, and a third operator Γ_t that commutes with neither, Theorem 5 and Remark 6 are verified on a labeled grid. For fifty random pairs of monotone lattice paths with identical numbers of r- and s-steps, the transports agree to within 4×10^{-13}. All (r, s)-squares are flat, and the grid's square complex is simply connected. The loop $r\,t\,r^{-1}t^{-1}$, which has zero net label count. It is trivial in H_1 and exhibits holonomy $\|\mathrm{Hol} - I\|_F = 2.01$. A triangle loop $r\,s\,t$ gives $\|\mathrm{Hol} - I\|_F = 7.69$, with the two routes rs and t^{-1} between its endpoints differing by 7.31. Square flatness constrains precisely those loops in the normal closure of the squares.

[1] which starts at b, returns to a, traverses γ, and then proceeds back to b

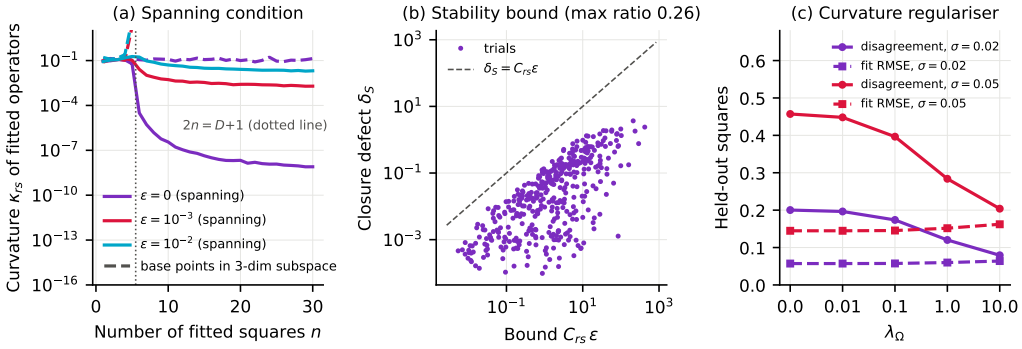


Figure 14: Synthetic checks ($D = 10$ in (a) and (c), $D = 8$ in (b)).(a) Curvature κ_{rs} of least-squares operators fitted to n noisy squares. For noise-free squares with base points spanning $\mathbb{R}^D$, the fit becomes exact once the $2n$ fitting points per relation span $\mathbb{R}^{D+1}$, and κ_{rs} drops to machine precision (Theorem 2(iii)). When base points are restricted to a three-dimensional affine subspace shown as a dashed line, κ_{rs} remains near 0.1 for all n (Theorem 2(v)). In the presence of noise, κ_{rs} decreases proportionally to the fit residual (Proposition 4). (b) Closure defect δ_S compared to the bound $C_{rs}\varepsilon$ from Proposition 4 over 400 random operator pairs and residual levels. Commutation is disrupted by a random perturbation of Γ_r. All points lie below the diagonal, with the largest ratio being 0.26. (c) Joint fitting of (Γ_r, Γ_s) to 8 noisy squares with curvature weight λ_Ω shows that route disagreement on 100 held-out squares decreases by a factor of 2.2 to 2.5. The training fit RMSE increases by approximately 12%. less than $D+1$, resulting in an under-determined problem

Table 3: Per-relation operators on GloVe-100d ($D = 100$, 200k words). Leave-one-out retrieval accuracy of $\Gamma_r \tilde{u}_p$ (nearest neighbor among 100k words) is reported for translation, orthogonal, and affine operators (with ridge ϱ selected by LOO), as well as for the 3CosAdd baseline. Mean held-out residuals $\|\Gamma_r \tilde{u}_p - \tilde{u}_q\|$ are also provided for unit-norm embeddings.

relation	n	3CosAdd	transl.	orthog.	affine	ϱ	res. transl.	res. affine
gender	86	0.457	0.651	0.605	0.733	1.0	0.673	0.606
plural	140	0.540	0.657	0.721	0.786	1.0	0.677	0.597
comparative	74	0.563	0.703	0.486	0.743	1.0	0.678	0.637
superlative	76	0.363	0.724	0.355	0.737	10.0	0.767	0.743
adverb	117	0.143	0.291	0.359	0.513	1.0	0.862	0.752
opposite	108	0.250	0.370	0.213	0.472	1.0	0.838	0.786
past	88	0.547	0.693	0.580	0.773	0.1	0.572	0.602
gerund	86	0.543	0.721	0.593	0.779	1.0	0.571	0.550
capital	98	0.850	0.929	0.898	0.929	1.0	0.544	0.474
nationality	96	0.923	0.948	0.917	0.948	10.0	0.437	0.412
currency	77	0.097	0.247	0.221	0.390	1.0	0.912	0.804

F STATIC EMBEDDINGS: GLOVE AND GPT-2 TOKEN EMBEDDINGS

In this study, we employ 100-dimensional GloVe vectors trained on Wikipedia and Gigaword (Pennington et al., 2014), restricted to the 200,000 most frequent words and unit-normalized. Retrieval is performed among the 100,000 most frequent words. Relation pairs are categorized according to the Google analogy set, which includes gender, plural, comparative, superlative, adverb, opposite, past tense, gerund, capital, nationality, and currency. Analogy squares combine two composable relations: gender×plural (*king*, *queen*, *queens*, *kings*; 36 squares) and opposite×adverb (*fortunate*, *unfortunate*, *unfortunately*, *fortunately*; 27 squares).

Relation operators (Table 3). Across all relations, the affine operator achieves the highest leave-one-out retrieval accuracy, outperforming both the shared translation (mean offset) and the 3CosAdd baseline. The improvement is most noticeable for relations that deviate from a parallelogram structure, such as adjective to adverb (0.29 to 0.51), opposite (0.37 to 0.47), and currency (0.25 to 0.39). Conversely, it is negligible for near-saturated encyclopedic relations like capital and nationality. Orthogonal operators frequently perform worse compared to translations, as observed in the superlative relation (0.36 versus 0.72). This indicates that these relations require a change of

Table 4: Analogy squares on GloVe-100d (five-fold CV over squares). Curvature κ_{rs} of the fitted pair, route disagreement $1 - \cos$, closure accuracy and route agreement on held-out squares, and mean residual on held-out square edges. The translation connection is flat by construction (Corollary 3); affine operators fit better but are curved. λ_Ω weights the data-level route-disagreement term of equation 6, $\lambda_\Omega^{\text{op}}$ the operator-level term $\|\Gamma_s\Gamma_r - \Gamma_r\Gamma_s\|_F^2$ of Section 7.

squares	operators	κ_{rs}	route dis.	closure acc.	route agree	res. (test)
gender×plural (n=36)	translation	0.00±0.00	0.000±0.000	0.68±0.26	1.00±0.00	0.845±0.047
	affine, $\lambda_\Omega = 0.0$	12.08±0.92	0.011±0.001	0.61±0.18	0.86±0.12	0.742±0.050
	affine, $\lambda_\Omega = 1.0$	11.72±0.85	0.008±0.001	0.61±0.18	0.86±0.12	0.741±0.050
	affine, $\lambda_\Omega = 10.0$	11.69±0.77	0.005±0.001	0.61±0.18	0.89±0.11	0.740±0.050
	affine, $\lambda_\Omega^{\text{op}} = 1.0$	6.77±0.48	0.006±0.000	0.61±0.18	0.86±0.12	0.740±0.050
	affine, $\lambda_\Omega^{\text{op}} = 10.0$	2.39±0.16	0.002±0.000	0.59±0.20	0.95±0.07	0.738±0.049
opposite×adverb (n=27)	translation	0.00±0.00	0.000±0.000	0.04±0.08	1.00±0.00	0.878±0.032
	affine, $\lambda_\Omega = 0.0$	4.83±0.24	0.031±0.004	0.04±0.05	0.33±0.07	0.808±0.027
	affine, $\lambda_\Omega = 1.0$	4.48±0.19	0.016±0.003	0.00±0.00	0.67±0.21	0.806±0.028
	affine, $\lambda_\Omega = 10.0$	3.99±0.13	0.009±0.002	0.00±0.00	0.75±0.18	0.804±0.028
	affine, $\lambda_\Omega^{\text{op}} = 1.0$	2.70±0.08	0.018±0.002	0.04±0.05	0.49±0.11	0.803±0.027
	affine, $\lambda_\Omega^{\text{op}} = 10.0$	0.86±0.02	0.006±0.001	0.02±0.04	0.63±0.17	0.799±0.026

scale that rigid mappings cannot capture. For most relations, the selected ridge parameter is $\varrho = 1$, indicating that the fitted maps are strongly regularized towards the identity, which is expected when the number of pairs n is approximately equal to the dimensionality D. According to Corollary 3, transitioning from translations to affine operators reduces the fit residual, although this comes at the cost of a connection that is no longer flat. The reduction is from 0.85 to 0.74 on held-out square edges, Table 4)

Curvature of fitted relation pairs (Table 4). Using affine operators, the gender/plural pair exhibits $\kappa_{rs} = 12.1 \pm 0.9$. The opposite/adverb pair shows $\kappa_{rs} = 4.8 \pm 0.2$. For reference, $\|I\|_F = 10$ in $D{+}1 = 101$ dimensions. The impact at the data level is considerable. For instance, the closure defect δ_S on held-out squares is $1.1\times$ (gender/plural) and $0.6\times$ (opposite/adverb) the median nearest-neighbor distance in the vocabulary (0.78). Thus, route disagreement is comparable to the spacing between words, such that the two routes retrieve different words for 14% and 67% of held-out squares, respectively. For gender/plural, the translation model closes squares slightly more frequently even though the edge residuals are worse. Therefore, composing two better-fitting but non-commuting maps does not necessarily outperform composing two commuting ones. For opposite/adverb, neither model closes squares, as two-step compositions such as *fortunate→unfortunately* are rarely the nearest neighbor. In this case, the family measures only consistency.

The two curvature penalties. The data-level term reduces held-out route disagreement by a factor of 2.5 to 3.3 and increases route agreement from 0.33 to 0.75 (opposite/adverb). Edge residuals remain unchanged to three decimal places. But this term has minimal impact on κ_{rs} ($12.1 \to 11.7$, $4.8 \to 4.0$). The resultant outcome aligns with Theorem 2(iii) and (v). The penalty constrains Ω_{rs} near the $\leq$29 training base points, which do not affinely span $\mathbb{R}^{100}$, resulting in the operator remaining curved away from the data ($\sigma_{\min}(\tilde{U}) = 0$ in Proposition 4). In contrast, the operator-level term reduces κ_{rs} by a factor of five ($12.1 \to 2.4$, $4.8 \to 0.9$), decreases route disagreement by factors of seven and five, increases gender/plural route agreement to 0.95, and incurs only a 0.01 to 0.02 increase in training residual with no effect on held-out residual. Both terms enhance consistency, but only the operator-level term modifies the connection globally, which is necessary for path independence beyond the observed squares as required by Theorem 5. We also provide additional results for GPT-2 input token embeddings as a second static space and for the mined squares from Section 8.1 in GloVe in Appendix K.

F.1 GPT-2 TOKEN EMBEDDINGS AS A STATIC SPACE

Table 5 presents the per-relation experiment described in the previous subsection, applied to the GPT-2 input token embedding matrix ($D = 768$). This analysis is restricted to the $19{,}704$ lower-case single-token words, thereby excluding encyclopedic relations whose entities are capitalized. As a result, retrieval is relatively less challenging than in the 100,000-word GloVe vocabulary. The ranking of methods remains consistent with GloVe: affine operators achieve the highest performance

Table 5: Per-relation operators on GPT-2 token embeddings ($D = 768$, 19,000 words). The figure reports leave-one-out retrieval accuracy of $\Gamma_r \tilde{u}_p$ (nearest neighbor among 19,704 words) for translation, orthogonal, and affine operators (with ridge ϱ selected by leave-one-out), the 3CosAdd baseline, and mean held-out residuals $\|\Gamma_r \tilde{u}_p - \tilde{u}_q\|$ for unit-norm embeddings.

relation	n	3CosAdd	transl.	orthog.	affine	ϱ	res. transl.	res. affine
gender	43	0.700	0.744	0.256	0.860	1.0	0.707	0.622
plural	97	0.993	0.979	0.392	1.000	0.01	0.624	0.598
comparative	55	0.943	0.964	0.182	0.982	1.0	0.690	0.657
superlative	42	0.923	0.952	0.024	0.952	1.0	0.749	0.716
adverb	87	0.443	0.563	0.103	0.828	0.01	0.868	0.834
opposite	85	0.473	0.541	0.082	0.741	0.01	0.864	0.830
past	86	0.657	0.674	0.128	0.791	0.01	0.695	0.667
gerund	84	0.743	0.798	0.179	0.893	0.01	0.696	0.665

across all relations, translations perform second best, and orthogonal maps perform substantially worse (0.02–0.39). These results confirm that the relations require anisotropic scaling. We know that $n \ll D$. Therefore, the affine fit is only well-posed with ridge regularization, and the leave-one-out (LOO) selected ridge parameter is small (0.01) for most relations.

G Experimental details

G.1 Relation lists and squares

Relation pairs are categorized according to the Google analogy test set (Mikolov et al., 2013a), including gender (family), plural nouns, comparative, superlative, adjective→adverb, opposite (*un-*, *in-*, *dis-*, *ir-*, *il-*, *im-*), past tense, gerund, capital→country, country→nationality, and country→currency. Squares represent combinations of two relations that compose on the same lemma: gender×plural (38 squares, e.g. *king, queen, queens, kings*), opposite×adverb (31, e.g. *fortunate, unfortunate, unfortunately, fortunately*), opposite×comparative (12), and opposite×superlative (11). All words are lower-cased, and lists are filtered against the vocabulary of each embedding at run time.

G.2 Mined morphological squares

For each base word w among the 100,000 most frequent GloVe words (alphabetic, length ≥ 4), and for each prefix $p \in \{$un, re, dis, mis, non, over, pre, in, im, sub, super, anti, de, under, out, co$\}$ and suffix $f \in \{$s, ed, ing, ly, ness, er, est, ers, ment, ful, less, able, ity, ation, ive, al, ist, ism, ish, ous$\}$ (with simple spelling rules: e-dropping, $y \to i$, consonant doubling, $-es$), the pair (w, pw) is considered an edge of relation "prefix p" and (w, wf) an edge of "suffix f" whenever both strings are among the 200,000 most frequent words. A square is recorded when pwf is also present in the vocabulary. Relations with less than 30 pairs and prefix×suffix pairs with less than 15 squares are excluded, resulting in 37 pairs and 1,780 squares (the largest: *re*×*-ed* 214, *re*×*-ing* 203, *re*×*-s* 201, *de*×*-s* 73, *in*×*-s* 71). This construction is purely orthographic and therefore includes squares whose prefix is not semantically compositional (*part*/*depart*, *form*/*inform*). The GPT-2 experiments utilize the 18 pairs with the prefixes *re-, dis-, un-, mis-, over-, sub-, under-* (1,154 squares), and Section K reports the GloVe results on all 37 pairs.

G.3 Fitting and evaluation protocol

A relation r is defined by pairs $\{(p_i, q_i)\}$. The translation operator T_t uses $t = \mathrm{mean}(u_{q_i} - u_{p_i})$. The orthogonal operator is given by the Procrustes solution $x \mapsto Rx + t$. The affine operator is defined as $\arg\min_{L,t} \sum_i \|Lu_{p_i} + t - u_{q_i}\|^2 + \varrho \|L - I\|_F^2$, which is solved in closed form after centring. The regularization parameter $\varrho \in \{0.01, 0.1, 1, 10\}$ is selected for each relation based on leave-one-out retrieval accuracy. Retrieval accuracy is measured as the fraction of held-out pairs for which the nearest neighbor (cosine similarity) of $\Gamma_r \tilde{u}_p$ among the 100,000 most frequent words, excluding p, is q. The 3CosAdd baseline evaluates $b - a + c$ against d for 300 random pair combinations. For square experiments, five-fold cross-validation is performed over squares. The edges of held-out squares are removed from the training pairs of both relations, and operators are fitted on the remaining pairs.

The curvature term is evaluated on the base points of the training squares. On held-out squares, the following metrics are reported: route disagreement $1 - \cos(\Gamma_s\Gamma_r\tilde{u}_a, \Gamma_r\Gamma_s\tilde{u}_a)$, closure accuracy (fraction of routes whose nearest neighbor, excluding a, b, d, is c), route agreement (both routes retrieve the same word), curvature κ_{rs}, and mean edge residuals on both training and held-out edges.

G.4 GPT-2 REPRESENTATIONS

We employ the 124M-parameter GPT-2 model (Radford et al., 2019), utilizing a NumPy-based re-implementation of the forward pass without fine-tuning. The weights correspond to the publicly available HuggingFace checkpoint. For each token, the contextual representation is defined as the hidden state at the last sub-token of the word at layer 6 of 12, within one of eleven templates: `"the {w}"`, `"In the story, the {w}"`, `"She wrote the word {w}"`, `"They were talking about the {w}"`, `"Here is a word: {w}"`, `"He said it was {w}"`, `"The report described the {w}"` (training), and `"A poem about the {w}"`, `"Yesterday I saw the {w}"`, `"My favourite word is {w}"`, `"According to the dictionary, {w}"` (held out). The bare-word template was excluded because GPT-2's position-0 hidden state functions as an attention sink with an anomalous norm (at layer 6, 28% of single-token words exhibited a norm >1000 compared to a median of ≈ 85 elsewhere), which distorts both the representation and the pooled principal component analysis (PCA). Representations from all templates are pooled, projected to $D = 48$ dimensions using PCA, and rescaled to have a median unit norm. The context descriptor $\beta \in \mathbb{R}^8$ is defined as the PCA projection of the mean layer-6 hidden state of the template prefix, computed with the `<|endoftext|>` token prepended to absorb the sink. The mean excludes this token.

Contextual operators are implemented according to Definition 9, utilizing $J = 4$ basis generators per relation, initialized at scale 0.01. Training is conducted using the Adam optimizer (learning rate 0.01, 400 steps) to minimize the fit loss aggregated over the seven training templates, in addition to λ_Ω times the route disagreement on the base points of training squares. Weight decay is set to 20 for the generators $A_{r,j}$, serving as the analogue of the ridge towards the identity ($\Gamma \to I$ as $A \to 0$), and 1 for the weight matrices W_r. Gradients through the matrix exponential are computed using the adjoint of its Fréchet derivative, $\partial L/\partial Z = D\exp_{Z^\top}(\partial L/\partial \Gamma)$. Baseline models include a static affine operator fitted on GPT-2's input token embeddings (projected to $D = 48$ via principal component analysis) and a separate affine operator fitted for each template, both employing a ridge parameter $\varrho = 10$. The ridge and weight decay parameters were selected based on a single validation split of the *opposite* relation prior to cross-validated runs. Retrieval is performed among the contextual representations of all 1,483 candidate words within the same template, excluding the query word and, for squares, the other three vertices.

Mined-square experiment (Section 8.1). Representations for the 4,592 words from the 18 selected pairs are computed at layer 6 using six templates (`"the {w}"`, `"In the story, the {w}"`, `"She wrote the word {w}"`, `"He said it was {w}"`, `"A poem about the {w}"`, `"According to the dictionary, {w}"`). These representations are pooled, projected to $D = 64$ dimensions via principal component analysis (PCA), and rescaled to have median unit norm. For each relation pair, 3-fold cross-validation is performed over squares. Single-hop training pairs [2] exclude any pair in which both words belong to a held-out square. Per-template operators are optimized by gradient descent on the objective in equation 6, restricted to the operators, with ridge $\varrho \in \{1, 3, 10\}$ towards the identity. Three regularization settings are considered: no curvature term, a data-level term with $\lambda_\Omega = 30$ applied to the base points of the training squares, or an operator-level term $\lambda_\Omega^{\text{op}}\|\Gamma_s\Gamma_r - \Gamma_r\Gamma_s\|_F^2$ with $\lambda_\Omega^{\text{op}} = 100$ (300 iterations). Retrieval is conducted among all 4,592 contextual representations of the same template, excluding the query word. For two-hop predictions, the other three vertices of the square. Failure-detection scores include the top-1 cosine of the composed prediction, the top-1 minus top-2 margin, and the smaller top-1 cosine of the two first hops. The values $\lambda_\Omega = 30$ and $\lambda_\Omega^{\text{op}} = 100$ are selected based on one split of the *re×-ed* pair prior to cross-validated runs.

Context-shift experiment (Section 8.2). Representations of the same $4{,}592$ words are computed in the twelve frames listed in the main text at layers 4, 6, 8, 10, and 12 using a batched

[2] up to 400 per relation, sampled at random

Table 6: Compositional consistency of relation transport in GPT-2 (layer 6, $D = 64$; 18 prefix×suffix pairs, 1,154 squares, 3-fold cross-validation over squares, six templates). *Single-hop* refers to accuracy on the four edges of held-out squares. *Two-hop* denotes the accuracy of $\Gamma_s\Gamma_r\tilde{u}_a$ retrieving c (top-1 and top-10 among 4,592 candidates). *Agree* indicates both routes retrieve the same word. *Inconsist.* is defined as $1-$agree. $\bar{\varrho}$ represents mean route disagreement ($1 - \cos$). κ_{rs} denotes mean operator curvature.

operators	single-hop	two-hop	top-10	agree	inconsist.	$\bar{\varrho}$	κ_{rs}
translation (flat)	0.608	0.453	0.704	1.000	0.000	0.0000	0.00
affine, ϱ=1	0.736	0.512	0.742	0.746	0.254	0.0218	19.50
+ data λ_Ω=30	0.738	0.516	0.745	0.904	0.096	0.0038	9.35
+ operator $\lambda_\Omega^{\text{op}}$=100	0.736	0.513	0.744	0.946	0.054	0.0008	0.78
affine, ϱ=3	0.732	0.546	0.759	0.817	0.183	0.0119	4.35
+ data λ_Ω=30	0.730	0.550	0.760	0.944	0.056	0.0013	2.59
+ operator $\lambda_\Omega^{\text{op}}$=100	0.728	0.542	0.757	0.950	0.050	0.0008	0.39
affine, ϱ=10	0.707	0.550	0.757	0.886	0.114	0.0052	0.77
+ data λ_Ω=30	0.707	0.551	0.758	0.974	0.026	0.0003	0.55
+ operator $\lambda_\Omega^{\text{op}}$=100	0.705	0.549	0.758	0.952	0.048	0.0008	0.21

NumPy forward pass. Each layer is projected to $D = 64$ dimensions via principal component analysis (PCA) of the pooled representations from all twelve frames and rescaled to median unit norm. The training contexts comprise the first two variants of each frame type (`"the {w}"`, `"In the story, the {w}"`, `"They {w}"`, `"Yesterday, they {w}"`, `"She wrote the word {w}"`, `"According to the dictionary, {w}"`, `"After a long day at work, she finally {w}"`, `"Nobody in the village had ever seen such a {w}"`), while the held-out contexts are the third variants (`"She looked at the {w}"`, `"Every morning the children {w}"`, `"The next word is {w}"`, `"It was late, the rain had stopped, and the old man {w}"`). The context descriptor for a frame is defined as the centroid of the $4{,}592$ projected token representations in that frame, mapped to $m = 6$ dimensions by a PCA fitted on the eight training descriptors only and rescaled to unit mean norm. The same mapping is applied to the held-out frames. The mean or last hidden state of the frame prefix serves as an alternative descriptor that does not utilize the token representations. For the *re*×*-ed* pair, this alternative leads to a family with substantially weaker context dependence, 0.498 versus 0.491 two-hop accuracy for the frozen operator. The centroid descriptor is reported throughout.

The static baselines employ ridge $\varrho = 3$. The contextual families utilize $J = 8$ generators, each initialized at $\log \Gamma_r^{(j)}$, which is the principal logarithm of the ridge-affine operator fitted in training context j. Mixture weights are initialized using the Gaussian kernel described in Remark 11, with τ set to the median nearest-neighbor squared distance between the training descriptors ($\tau = 1.14$ at layer 8). Training is conducted using Adam (learning rate 0.01, 300 steps, weight decay 20 on the generators and 1 on W_r, b_r) to minimize the fit loss summed over the eight training contexts. Additional regularization terms include $\lambda_\Omega = 30$ times the route disagreement on the base points of the training squares in the training contexts, $\lambda_\Omega^{\text{op}} = 100$ times $\|\Gamma_s(\beta)\Gamma_r(\beta) - \Gamma_r(\beta)\Gamma_s(\beta)\|_F^2$ summed over the training contexts, and $\lambda_{\text{basis}} = 30$ times $\sum_{j,k}\|[A_{r,j}, A_{s,k}]\|_F^2$. The static operator $\exp(\bar{A}_r)$ refers to the same model trained with $W_r = 0$ and $b_r = 0$ held fixed. The weight decay, τ, and the descriptor were selected on the *re*×*-ed* pair (two folds) prior to cross-validated runs over all 18 pairs. With weaker weight decay (5), both the static and contextual Lie operators exhibit increased curvature and reduced accuracy on two-hop analogies. The 3-fold split over squares is consistent with the six-template experiment.

H GPT-2 TEST BED: FULL TABLE, FIGURE, GLOBAL MODEL AND LONGER PATHS

One shared operator per relation type. The previous per-pair fits utilized a distinct copy of each relation operator for every pair. However, the framework requires a single operator for each relation type. Accordingly, a global model was fitted, consisting of eleven affine operators (the prefixes re-, dis-, un-, mis-, over-, sub-, under- and the suffixes -s, -ed, -ing, -ly) trained jointly on all single-hop pairs across all 18 relation pairs (with a maximum of 800 per relation, excluding held-out square

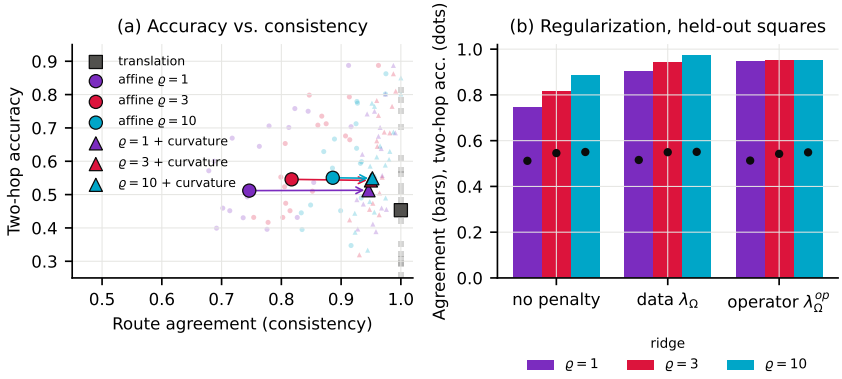


Figure 15: Curvature regularization makes compositions consistent (GPT-2, 1,154 held-out squares). (a) Two-hop accuracy is plotted against route agreement. Translation models exhibit perfect consistency but are ten percentage points less accurate. Affine operators achieve higher accuracy but display inconsistency on 11 to 25 percent of held-out squares. Applying curvature penalties as indicated by arrows shifts each affine model to 90-97% agreement without reducing accuracy. Small markers represent the 18 relation pairs. (b) Agreement (bars) and two-hop accuracy (dots) are shown for the ridge by penalty grid.

Table 7: One shared operator per relation type. Eleven affine operators, comprising seven prefixes and four suffixes, were jointly fitted on all single-hop pairs across the 18 relation pairs, with the curvature penalty aggregated over all 18 pairs. The analysis utilized GPT-2 layer 6, with dimensionality $D = 64$, six templates, and three-fold cross-validation over 1,154 squares. The left panel presents single-hop and two-hop accuracy, route agreement, inconsistency rate, and mean curvature κ_{rs} across the 18 pairs on held-out squares. The right panel addresses longer paths. For a held-out square (w, pw, pwf, wf) and a sibling square sharing the same base but with a different suffix f' (7,836 cases) or a different prefix p' (1,992 cases), the four-hop detour $w \to wf' \to pwf' \to pw \to pwf$ (or $w \to p'w \to p'wf \to wf \to pwf$) is compared to the two-hop route, evaluating both the agreement of the retrieved words and the accuracy of the detour. Detours are not directly penalized; their consistency results from square flatness as established by Theorem 5.

						suffix detour (4 hops)		prefix detour (4 hops)	
operators	single-hop	two-hop	agree	inconsist.	κ_{rs}	agree	acc.	agree	acc.
translation (flat)	0.612	0.462	1.000	0.000	0.00	1.000	0.421	1.000	0.448
affine, ϱ=1	0.734	0.490	0.767	0.233	34.37	–	–	–	–
+ data λ_Ω=30	0.734	0.498	0.913	0.087	5.41	–	–	–	–
+ operator $\lambda_\Omega^{\text{op}}$=100	0.731	0.492	0.944	0.056	0.42	–	–	–	–
affine, ϱ=3	0.735	0.532	0.838	0.162	5.39	0.793	0.510	0.756	0.538
+ data λ_Ω=30	0.731	0.532	0.949	0.051	1.88	0.948	0.510	0.948	0.564
+ operator $\lambda_\Omega^{\text{op}}$=100	0.728	0.529	0.958	0.042	0.29	0.943	0.505	0.938	0.556
affine, ϱ=10	0.716	0.550	0.899	0.101	0.96	0.872	0.525	0.881	0.557
+ data λ_Ω=30	0.712	0.548	0.976	0.024	0.51	0.974	0.519	0.976	0.548
+ operator $\lambda_\Omega^{\text{op}}$=100	0.709	0.546	0.960	0.040	0.17	0.944	0.520	0.946	0.548

edges). The curvature penalty was summed over all 18 pairs within each template and fold. Table 7 and Figure 16(a) present comparable results with less parameters. Translations achieve 0.46 two-hop accuracy. Affine operators achieve between 0.49 and 0.55 with 10% to 23% inconsistency. The application of penalties reduces inconsistency to between 2% and 9% across all 18 relation pairs for every ridge and penalty setting. The median reduction factor is between 2.7 and 5.2. The paired two-hop accuracy difference across the 18 pairs ranges from -0.007 to +0.005, with confidence intervals containing zero in five out of six cases. The exception is the operator-level term at $\varrho = 10$, which gives -0.005 (CI [-0.009, -0.002]). Sharing operators across the 18 pairs does not alter the trade-off or its resolution.

Longer paths. In the global model, squares corresponding to different relation pairs share edges, resulting in a graph that contains closed walks longer than a single square. Theorem 5 predicts phenomena that per-pair experiments cannot assess. A held-out square (w, pw, pwf, wf) with a sibling square (w, pw, pwf', wf') can be reached from w either via the two-hop route $\Gamma_f\Gamma_p$ or

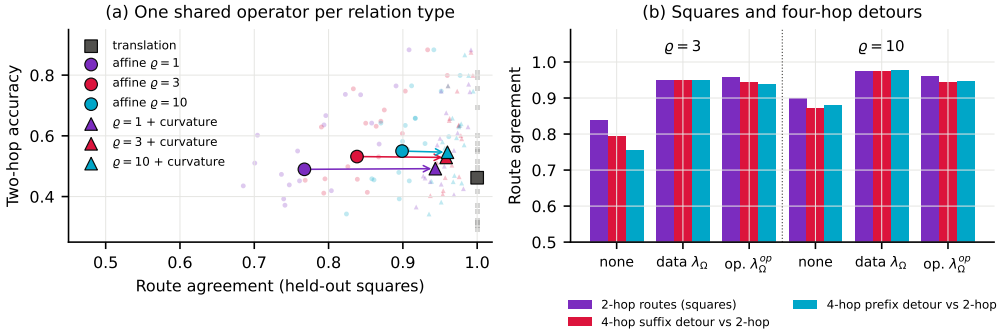


Figure 16: The global model utilizes one operator per relation type, comprising 11 operators and 18 relation pairs, implemented with GPT-2. (a) Two-hop accuracy is plotted against route agreement on held-out squares, as shown in Figure 15(a). Small markers represent the 18 relation pairs. (b) The agreement between the two routes of a held-out square (Violet) is compared to the agreement between the two-hop route and a four-hop detour through a sibling square (Red: detour through another suffix, 7,836 cases. Cyan: through another prefix, 1,992 cases). Only squares are penalized, and the detours achieve consistency as a result of Theorem 5.

through a four-hop detour $w \to wf' \to pwf' \to pw \to pwf$, that is, $\Gamma_f \Gamma_{f'}^{-1} \Gamma_p \Gamma_{f'}$, which traverses the sibling square in reverse. A similar process applies through a sibling prefix. The detour is not penalized. However, if the sibling square is flat, then $\Gamma_{f'}^{-1} \Gamma_p \Gamma_{f'} = \Gamma_p$, and the detour coincides with the direct route. The union of all 1,154 squares forms a graph on 3,378 words whose square complex is simply connected, as each square can be attached sequentially, meeting previous ones at a vertex or an arc (Appendix J). Therefore, by Theorem 5, flatness of the squares implies path independence throughout the graph. Empirical results (Table 7, right; Figure 16(b)) show that unregularized affine operators map the detour and direct route to different words in 21–24% of cases at $\varrho = 3$ (12–13% at $\varrho = 10$). With either penalty, this rate decreases to 5–6% (2–6% at $\varrho = 10$). The detour's accuracy remains unchanged or improves ($0.54 \to 0.56$ for prefix detours at $\varrho = 3$). Penalizing two-hop squares therefore enforces consistency in four-hop compositions across relation pairs, as predicted by the theorem. Consistency across relation types that never share a square remains untested, as the graph does not provide a closed walk for these cases.

Global model and detours. The global model jointly fits the eleven relation operators across the 18 families. Single-hop pairs are drawn from a 4,592-word vocabulary, with between 114 (*under-*) and 989 (*re-*) pairs per relation, subsampled to a maximum of 800 after removing held-out square edges. Fitting is performed by gradient descent on the fit term with ridge $\varrho \in \{1, 3, 10\}$, and includes either $\lambda_\Omega = 30$ times the route disagreement on the base points of all training squares or $\lambda_\Omega^{\text{op}} = 100$ times $\sum_{(r,s)} \|\Gamma_s \Gamma_r - \Gamma_r \Gamma_s\|_F^2$ over the 18 pairs. Each configuration is run for 400 iterations, across six templates and three folds, with folds defined globally over the 1,154 squares. For detour evaluation, for every held-out square (w, pw, pwf, wf) and every other square with the same base w and prefix p but suffix $f' \neq f$ (from any fold), the suffix detour $\Gamma_f \Gamma_{f'}^{-1} \Gamma_p \Gamma_{f'} \tilde{u}_w$ is retrieved among all words except w, pw, wf and compared with the two-hop retrieval. Prefix detours $\Gamma_p \Gamma_{p'}^{-1} \Gamma_f \Gamma_{p'} \tilde{u}_w$ are computed using squares with the same base and suffix but prefix $p' \neq p$.

I Route disagreement as a failure detector

What route disagreement does and does not detect. According to Proposition 8, ϱ_S certifies, without requiring the answer, that at least one route is incorrect by $\varrho_S/2$. The regularizer aims to minimize this value. Table 8 addresses a more specific question, i.e., whether ϱ_S can also predict which held-out compositions fail retrieval. The results indicate that ϱ_S is only a weak predictor (AUC 0.52–0.57, Spearman 0.16–0.37 with the two-hop error), whereas the standard confidence scores of the composed prediction serve as significantly better failure detectors (AUC 0.75–0.85). Incorporating ϱ_S into the confidence metric does not enhance performance beyond confidence alone. These two signals capture distinct aspects: confidence evaluates whether a prediction is close to a

Table 8: Per-square failure detection on GPT-2. The area under the curve (AUC) for predicting incorrect two-hop retrieval is evaluated using the label-free route disagreement ϱ_S (Proposition 8) and compared to retrieval-confidence scores, including top-1 cosine similarity, top-1 minus top-2 margin, and first-hop confidence. Additionally, the Spearman correlation with the continuous two-hop error e_1 is reported. Route disagreement serves to certify inconsistency rather than retrieval failure; as a failure detector, it is relatively weak when used independently and is outperformed by confidence scores.

			AUC (failure)					Spearman with e_1	
operators	n	fail rate	ϱ_S	conf.	margin	first hop	ϱ_S+conf.	ϱ_S	conf.
affine ϱ=1 (seen)	4616	0.49	0.55	0.75	0.82	0.75	0.68	0.34	0.86
affine ϱ=1 (unseen)	2308	0.48	0.57	0.77	0.84	0.77	0.69	0.37	0.87
affine ϱ=3 (seen)	4616	0.45	0.55	0.79	0.83	0.77	0.71	0.29	0.89
affine ϱ=3 (unseen)	2308	0.45	0.57	0.81	0.85	0.78	0.72	0.32	0.90
affine ϱ=10 (seen)	4616	0.45	0.52	0.81	0.83	0.74	0.71	0.16	0.89
affine ϱ=10 (unseen)	2308	0.43	0.54	0.82	0.85	0.78	0.73	0.18	0.90
contextual $\Gamma_r(\beta)$ (seen)	4616	0.45	0.54	0.80	0.84	0.76	0.71	0.24	0.89
contextual $\Gamma_r(\beta)$ (unseen)	2308	0.45	0.56	0.81	0.85	0.76	0.72	0.26	0.90

particular word. Disagreement assesses whether two independent derivations of the analogy coincide. Consequently, a composition may be confidently and consistently incorrect or inconsistently correct. In these representations, the predominant failure mode is single-hop error, which route disagreement is not intended to detect. Instead, its function is diagnostic at the model level, where it identifies path dependence and, through equation 6, eliminates it.

J Robustness checks for the GPT-2 experiments

Paired accuracy differences. For each of the 18 relation pairs, the difference in held-out two-hop accuracy between the regularized and unregularized fits with the same ridge parameter was computed. The mean difference over pairs was estimated using bootstrapping with 5,000 resamples. For the operator-level term, the results were: $+0.003$ $[-0.013, +0.018]$ at $\varrho = 1$ (8 pairs improved, 8 worsened, 2 tied), -0.006 $[-0.019, +0.005]$ at $\varrho = 3$ (8/9/1), and -0.006 $[-0.010, -0.002]$ at $\varrho = 10$ (4/11/3). For the data-level term, the results were: $+0.004$ $[-0.005, +0.013]$ (10/5/3), $+0.001$ $[-0.005, +0.006]$ (9/6/3), and -0.004 $[-0.009, +0.000]$ (4/8/6).

Lexical overlap between folds. For each relation pair, the four edges of every held-out square are excluded from the training pairs of both relations. Additionally, for every fold, the proportion of training pairs that share one word with a held-out square and the other word outside the held-out squares was calculated. This proportion is 0.0% on average and does not exceed 0.7%. Although the words from a square may appear in other relation pairs. For example, the base *state* appears in *re*×*-ed*, *re*×*-ing*, and *re*×*-s*. Each relation pair is fitted and evaluated independently, so these occurrences are not included in its training set. The PCA projection is fitted on the pooled representations of all words without labels.

Topology of the empirical graphs. For each of the 18 families, the corresponding graph was constructed such that the edges consist of the four edges of every square. The following properties were verified: (i) any two squares intersect in at most one vertex or one edge, and (ii) the graph whose nodes represent squares, joined when they intersect, forms a forest. Both properties hold across all families. Of the 1,154 squares, 1,150 are isolated components. The remaining squares share a single vertex. In the global model, the union of all squares gives a graph with 3,378 vertices, 3,973 edges, and 559 components. The largest component contains 20 vertices. The cyclomatic number of this graph is 1,154, which equals both the number of squares and the rank over $\mathbb{Z}_2$ of the square boundary matrix, implying $H_1(K_{\mathcal{Q}}) = 0$. Furthermore, in every component, the squares can be ordered so that each one meets the union of its predecessors in a single vertex (15 cases), a single edge (517 cases), or a path of two or three edges (63 cases). Attaching a closed disc to a complex along an arc of its boundary circle or at a single boundary point constitutes a homotopy equivalence, as the disc deformation retracts onto the arc or the point. Consequently, each attachment preserves the fundamental group, and $K_{\mathcal{Q}}$, constructed from a forest by such attachments, is simply connected. Theorem 5 therefore applies to the entire graph, and the same reasoning holds within each family.

Table 9: Context shift was evaluated at three layers of GPT-2 Using held-out contexts, held-out squares, and 18 relation pairs. Single-hop accuracy, two-hop accuracy, and route agreement were measured. The protocol follows that of Table 2. The context effect increases with depth, whereas overall retrieval accuracy decreases.

	layer 6			layer 8			layer 10		
operators	1-hop	2-hop	agree	1-hop	2-hop	agree	1-hop	2-hop	agree
static affine, canonical context	0.647	0.414	0.787	0.609	0.343	0.791	0.476	0.210	0.766
static affine, pooled contexts	0.706	0.452	0.679	0.680	0.421	0.710	0.562	0.278	0.621
static $\exp(\bar{A}_r)$, pooled contexts	0.653	0.492	0.908	0.635	0.427	0.913	0.507	0.275	0.899
per-context affine (oracle)	0.725	0.543	0.813	0.705	0.497	0.823	0.623	0.386	0.767
contextual, frozen at canonical β	0.693	0.510	0.830	0.650	0.413	0.848	0.525	0.274	0.810
contextual $\Gamma_r(\beta)$	0.695	0.531	0.854	0.675	0.480	0.857	0.557	0.330	0.824
+ data λ_Ω	0.693	0.529	0.956	0.674	0.479	0.953	0.549	0.327	0.941
+ operator $\lambda_\Omega^{\mathrm{op}}$	0.692	0.528	0.952	0.674	0.478	0.952	0.550	0.329	0.940
+ basis brackets	0.698	0.529	0.904	0.675	0.480	0.902	0.553	0.328	0.896

Table 10: Context shift. Experiments were conducted using GPT-2 layer 8 with $D = 64$. Twelve sentence frames representing four syntactic types (noun, verb, quotation, narrative; see Section 8.2) were constructed. Eight frames were used for training, while one frame of each type was reserved for evaluation. For each relation pair, 3-fold cross-validation was performed over the 1,154 squares. All reported results are based on held-out squares, distinguishing between training contexts (left) and held-out contexts (right). The following operator configurations were evaluated: *static, canonical*, which fits one affine operator per relation in `"the {w}"` only; *static, pooled*, which fits a context-independent operator using ridge regression on pooled pairs from the eight training contexts; *static* $\exp(\bar{A}_r)$, which uses the contextual family with mixture weights fixed at uniform, resulting in a context-independent operator with identical parametrisation, initialisation, optimiser, and weight decay; *per-context (oracle)*, which fits an operator in each context, including held-out ones; *contextual*, which trains the families $\Gamma_r(\beta)$ (see Definition 9) in the eight training contexts and applies them via the descriptor β of each context; and *frozen*, which uses the same family but applies the operator from the canonical context universally. Curvature penalties are applied only in the training contexts, while the basis-bracket penalty is context-free (see Proposition 10(iii)).

	training contexts (8)				held-out contexts (4)				
operators	single-hop	two-hop	agree	κ_{rs}	single-hop	two-hop	agree	inconsist.	κ_{rs}
translation, pooled contexts (flat)	0.564	0.333	1.000	0.00	0.544	0.300	1.000	0.000	0.00
static affine, canonical context only	0.642	0.389	0.808	2.63	0.609	0.343	0.791	0.209	2.63
static affine, pooled contexts	0.691	0.423	0.720	16.90	0.680	0.421	0.710	0.290	16.90
+ operator $\lambda_\Omega^{\mathrm{op}}$	0.692	0.432	0.859	2.40	0.681	0.428	0.858	0.142	2.40
static $\exp(\bar{A}_r)$, pooled contexts	0.655	0.457	0.913	0.23	0.635	0.427	0.913	0.087	0.23
per-context affine (oracle)	0.704	0.490	0.831	2.32	0.705	0.497	0.823	0.177	2.30
contextual $\Gamma_r(\beta)$, frozen at canonical β	0.669	0.440	0.841	0.96	0.650	0.413	0.848	0.152	0.96
contextual $\Gamma_r(\beta)$	0.692	0.485	0.848	0.80	0.675	0.480	0.857	0.143	0.72
+ data λ_Ω	0.690	0.483	0.963	0.55	0.674	0.479	0.953	0.047	0.51
+ operator $\lambda_\Omega^{\mathrm{op}}$	0.688	0.483	0.948	0.19	0.674	0.478	0.952	0.048	0.18
+ basis brackets λ_{basis}	0.689	0.483	0.903	0.35	0.675	0.480	0.902	0.098	0.33

Context shift: the impact of frames on transport. For six morphological relations (*re-*, *un-*, *-s*, *-ed*, *-ing*, *-ly*), a ridge-affine operator ($\varrho = 3$) was fitted on 400 training pairs in frame i. Single-hop retrieval accuracy was then measured on 200 held-out pairs in frame j, using the representations of frame j. When averaged over relations and frame types, an operator fitted in a noun frame and applied in a verb frame achieved scores of 0.653 at layer 6, 0.612 at layer 8, 0.481 at layer 10, and 0.241 at layer 12. In comparison, an operator fitted in the test frame itself achieved 0.743, 0.733, 0.663, and 0.394 at the respective layers. The reverse transfer (verb to noun) resulted in scores of 0.610, 0.531, 0.417, and 0.169. A single operator fitted on pooled pairs from all twelve frames performed within 0.02 of the per-frame operators at layer 6, up to 0.05 lower at layer 8, and 0.09 lower at layers 10 and 12. Translation tasks exhibited a similar pattern, albeit with lower accuracy. Layer 8 was selected for the main experiment as it is the deepest layer where the pooled operator loses at most a few points, indicating a genuine shift while maintaining reliable retrieval. Table 9 presents the results for layers 6 and 10.

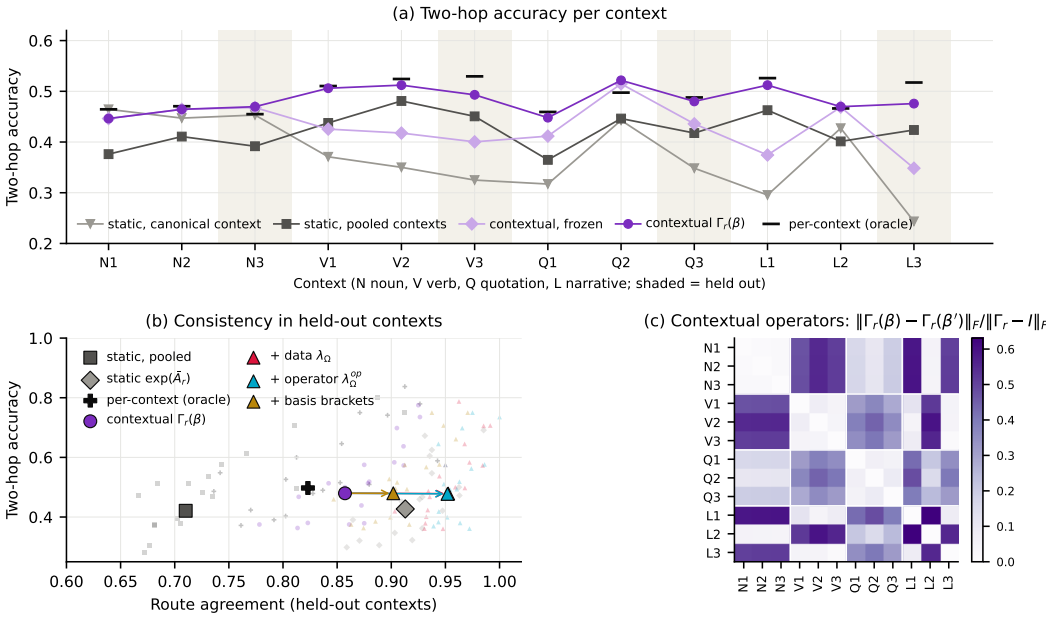


Figure 17: Context shift, full figure. (a), (b) as in Figure 5. (c) Relative difference $\|\Gamma_r(\beta) - \Gamma_r(\beta')\|_F / \|\Gamma_r - I\|_F$ between the operators assigned by the contextual family to two contexts, averaged over relations, folds, and the 18 pairs. This family distinguishes verb-like from noun-like frames and assigns held-out contexts to their respective types.

Context shift at other layers. Table 9 presents results for layers 6 and 10, demonstrating a consistent pattern across all layers. The magnitude of the context effect increases with depth. The gap between the canonical-context operator and the oracle in held-out contexts is 13, 15, and 18 two-hop points at layers 6, 8, and 10, respectively. The contextual family recovers to within 1, 2, and 6 points of the oracle, outperforming its frozen version by 2, 7, and 6 points. The penalties reduce inconsistency from 15%, 14%, and 18% to 4–5%, 5%, and 6%, while maintaining accuracy. At layer 10, overall retrieval accuracy is lower (0.62 single-hop for the oracle), and the family recovers a smaller proportion of the gap.

Context shift: extrapolation to an unseen frame type. Table 11 presents results where all three frames of a single syntactic type are excluded from training, and the model is trained on the remaining nine. This scenario represents the uncovered case of Proposition 14, in which no training context includes the held-out type. The ability of the family to generalize depends on whether certain training frames transport the relevant relations similarly. Two-hop accuracy for the held-out type (oracle / pooled ridge / static $\exp(\bar{A}_r)$ / contextual) is as follows: noun frames held out 0.463 / 0.374 / 0.406 / 0.461; verb frames 0.521 / 0.410 / 0.358 / 0.405; quotation frames 0.482 / 0.334 / 0.362 / 0.393; narrative frames 0.503 / 0.359 / 0.390 / 0.435. The family achieves oracle-level performance when noun frames are held out, as the quotation frames and the noun-like narrative frame transport relations similarly to noun frames. For narrative frames, the family recovers approximately half of the performance gap, and for quotation frames, about one third. When verb frames are held out, performance does not exceed that of a pooled static operator, as this frame type differs most from the others. The consistency result is less dependent on coverage. In all four cases, penalties applied in the nine training frames reduce inconsistency in the held-out type from 12–15% to 4–6% (data-level and operator-level terms). Two-hop accuracy remains within 0.5 points of the unpenalized family, except for the operator-level term with verb frames held out (-1.3 points).

Context shift: paired differences. Across 18 relation pairs in held-out contexts, results are based on 5,000 bootstrap resamples of the mean paired difference, with counts indicating pairs improved or worsened. For contextual family minus static canonical, two-hop differences are $+0.123$ $[0.092, 0.156]$ (18/0) and single-hop $+0.059$ $[0.043, 0.076]$ (18/0). Compared to static pooled ridge, two-hop is $+0.020$ $[-0.005, 0.047]$ (9/8), single-hop -0.017 $[-0.031, -0.002]$ (5/13), and agree-

Table 11: Extrapolation to an unseen frame type. Similar to Table 2, except that all three frames of a single syntactic type are excluded, and the models are trained on the remaining nine frames. Single-hop accuracy, two-hop accuracy, and route agreement are reported for held-out squares within the excluded type. (The canonical-context and frozen rows are omitted because the canonical context "the {w}" constitutes a noun frame.)

	noun held out			verb held out			quotation held out			narrative held out		
operators	1-hop	2-hop	agree	1-hop	2-hop	agree	1-hop	2-hop	agree	1-hop	2-hop	agree
static affine, pooled contexts	0.667	0.374	0.672	0.648	0.410	0.714	0.641	0.334	0.672	0.662	0.359	0.672
static $\exp(\bar{A}_r)$, pooled contexts	0.616	0.406	0.917	0.561	0.358	0.922	0.606	0.362	0.912	0.626	0.390	0.902
per-context affine (oracle)	0.696	0.463	0.810	0.694	0.521	0.848	0.716	0.482	0.812	0.711	0.503	0.844
contextual $\Gamma_r(\beta)$	0.672	0.461	0.854	0.600	0.405	0.882	0.632	0.393	0.851	0.661	0.435	0.854
+ data λ_Ω	0.662	0.456	0.953	0.600	0.403	0.949	0.629	0.389	0.943	0.657	0.433	0.953
+ operator $\lambda_\Omega^{\mathrm{op}}$	0.667	0.463	0.950	0.590	0.392	0.959	0.629	0.391	0.936	0.656	0.434	0.949
+ basis brackets	0.670	0.462	0.906	0.606	0.406	0.922	0.630	0.395	0.915	0.664	0.443	0.909

ment $+0.130$ $[0.106, 0.152]$ (18/0). Relative to static $\exp(\bar{A}_r)$, two-hop is $+0.063$ $[0.046, 0.082]$ (17/1), single-hop $+0.046$ $[0.037, 0.055]$ (18/0), and agreement -0.053 $[-0.065, -0.040]$ (0/18). Compared to the frozen family, two-hop is $+0.057$ $[0.043, 0.071]$ (17/1) and single-hop $+0.022$ $[0.016, 0.028]$ (16/0). For oracle minus contextual, two-hop is $+0.038$ $[0.022, 0.056]$ (16/2) and single-hop $+0.033$ $[0.024, 0.042]$ (17/0). Penalized minus unpenalized contextual family leads to two-hop $+0.001$ $[-0.007, 0.008]$, -0.004 $[-0.009, 0.002]$, 0.000 $[-0.006, 0.006]$ and single-hop -0.002 $[-0.004, 0.000]$, -0.004 $[-0.008, -0.001]$, -0.002 $[-0.004, 0.001]$ for data-level, operator-level, and basis-bracket terms, respectively. Agreement is $+0.082$ $[0.068, 0.097]$, $+0.088$ $[0.074, 0.104]$, and $+0.043$ $[0.035, 0.052]$, with improvements in all 18 pairs and median inconsistency ratios of 2.8, 3.1, and 1.5.

The basis-bracket term is the weakest of the three at the weight used because the fitted mixtures are peaked (most weight on two generators, see below), so most of the 64 brackets it penalizes never enter a fitted operator. Within-context curvature falls from 0.77 to 0.19 under the operator-level term, and the largest basis bracket from 0.88 to 0.32.

Context shift: the fitted mixtures. For the *re*×*-ed* pair (first fold), the trained mixture weights $\alpha_r(\beta)$ of the unpenalized family assign, for *re-*, a weight of 0.93 to 0.96 to the generator of "In the story, the {w}" in the three noun frames and in "Nobody in the village had ever seen such a {w}". A weight of 0.87 to 0.97 is assigned to the generator of "Yesterday, they {w}" in the three verb frames and in the two narrative frames ending in a subject. In the quotation frames "She wrote the word {w}" and "The next word is {w}", the weights are 0.44 and 0.53, respectively (0.84 and 0.14 in "According to the dictionary, {w}"). The remaining six generators each receive at most 0.05. The suffix *-ed* exhibits similar behavior, although the verb frames are slightly less peaked (0.53 to 0.66 on the verb generator). Among eight additional relations examined on one fold (*dis-*, *un-*, *over-*, *mis-*, *-s*, *-ly*, *-ing*, *-ed*), seven utilize the same two generators with at least 0.94 of the weight in every context. The suffix *-s* pairs the noun-frame generator with the generator of "According to the dictionary, {w}". Averaged over the 18 pairs and three folds, the maximal deviation of a mixture weight from its mean over contexts is 0.55 (0.53, 0.53, 0.50 with the three penalties). The relative operator difference $\|\Gamma_r(\beta) - \Gamma_r(\beta')\|_F / \|\Gamma_r - I\|_F$ is 0.33 between frame types and 0.14 within a type (with a maximum of 0.63). Within-context curvature is 0.77 (0.54, 0.19, 0.34 with the penalties), the largest basis bracket is 0.88 (0.62, 0.32, 0.41), the cross-context holonomy $\|\Omega_{rs}(\beta, \beta') - I\|_F$ is 1.49 (1.29, 1.07, 0.98), and the exponents have norm 2.1 to 2.2.

Context dependence in the six-template experiment. In the six neutral templates presented in Table 1, contextual families trained from near-zero generators, as used in an earlier version of this experiment, converge to context-independent operators. Although their softmax weights vary across templates by 0.02 to 0.56, the resulting operators differ by only 3 to 10% of their distance from the identity. Furthermore, these operators match the accuracy of per-template operators (0.55 two-hop) on both training and held-out templates. This phenomenon, described in Remark 11, also reflects a property of these templates. A single operator is sufficient for all. Consequently, Section 8.2 evaluates the contextual construction on frames that alter the transport.

Table 12: Contextual retrieval accuracy on GPT-2 (layer 6, $D = 48$). Five-fold cross-validation was performed over pairs, and results were averaged across templates. The static approach applies an affine operator fitted on token embeddings to contextual representations. The per-context method uses one affine operator per template. The ctx-Lie method, as defined in Definition 9, was trained on seven templates. For held-out templates, the per-context baseline can only transfer an operator from a training template, whereas ctx-Lie selects the operator based on the context descriptor β.

	training templates			held-out templates	
relation	static	per-context	ctx-Lie	per-context (transfer)	ctx-Lie
gender	0.138±0.079	0.638±0.096	0.662±0.108	0.632±0.089	0.668±0.106
plural	0.158±0.062	0.814±0.046	0.823±0.048	0.804±0.047	0.824±0.049
opposite	0.004±0.013	0.817±0.068	0.836±0.064	0.788±0.069	0.834±0.057
adverb	0.000±0.000	0.755±0.068	0.781±0.065	0.731±0.065	0.775±0.069
comparative	0.197±0.123	0.800±0.086	0.809±0.079	0.816±0.078	0.821±0.072
superlative	0.000±0.000	0.830±0.073	0.833±0.075	0.822±0.089	0.825±0.097
past	0.243±0.155	0.960±0.026	0.959±0.026	0.952±0.020	0.963±0.027
gerund	0.081±0.078	0.980±0.034	0.975±0.035	0.986±0.031	0.986±0.031
capital	0.024±0.029	0.034±0.030	0.017±0.027	0.035±0.032	0.015±0.023
nationality	0.002±0.010	0.287±0.146	0.260±0.129	0.283±0.134	0.277±0.129

Table 13: Contextual analogy squares on GPT-2 (five-fold CV over squares). Curvature $\kappa_{rs}(\beta)$ averaged over training templates, route agreement and closure accuracy on held-out squares in training (seen) and held-out (unseen) templates, and within-context $\|\Omega_{rs}(\beta) - I\|_F$ versus cross-context $\|\Omega_{rs}(\beta, \beta') - I\|_F$ holonomy averaged over template pairs.

			seen templates		unseen templates		holonomy	
squares	operators	κ_{rs}	agree	closure	agree	closure	within	cross
gender×plural (n=38)	static	0.45±0.04	0.94±0.08	0.17±0.08	0.93±0.09	0.17±0.13	–	–
	per-context	0.54±0.06	0.94±0.09	0.79±0.15	–	–	–	–
	ctx-Lie, $\lambda_\Omega = 0.0$	1.46±0.06	0.84±0.14	0.77±0.14	0.85±0.13	0.77±0.13	1.45±0.06	1.45±0.06
	ctx-Lie, $\lambda_\Omega = 0.3$	1.41±0.06	0.85±0.13	0.78±0.14	0.86±0.12	0.77±0.14	1.40±0.06	1.40±0.06
	ctx-Lie, $\lambda_\Omega = 1.0$	1.33±0.05	0.89±0.11	0.78±0.14	0.88±0.11	0.77±0.14	1.32±0.05	1.32±0.05
	ctx-Lie, $\lambda_\Omega = 3.0$	1.20±0.05	0.93±0.09	0.78±0.14	0.93±0.09	0.77±0.13	1.19±0.05	1.19±0.05
opposite×adverb (n=31)	static	0.38±0.06	0.86±0.13	0.00±0.00	0.86±0.12	0.00±0.00	–	–
	per-context	0.24±0.05	0.76±0.18	0.60±0.21	–	–	–	–
	ctx-Lie, $\lambda_\Omega = 0.0$	0.74±0.05	0.84±0.10	0.55±0.18	0.89±0.13	0.56±0.20	0.73±0.05	0.73±0.05
	ctx-Lie, $\lambda_\Omega = 0.3$	0.70±0.04	0.89±0.09	0.55±0.18	0.92±0.11	0.56±0.19	0.70±0.04	0.70±0.04
	ctx-Lie, $\lambda_\Omega = 1.0$	0.67±0.04	0.92±0.10	0.56±0.19	0.92±0.11	0.57±0.19	0.66±0.04	0.66±0.04
	ctx-Lie, $\lambda_\Omega = 3.0$	0.61±0.04	0.93±0.10	0.55±0.19	0.93±0.10	0.57±0.19	0.61±0.04	0.61±0.04
opposite×comparative (n=12)	static	0.43±0.06	0.86±0.21	0.00±0.00	0.82±0.22	0.00±0.00	–	–
	per-context	0.15±0.02	0.97±0.11	0.46±0.30	–	–	–	–
	ctx-Lie, $\lambda_\Omega = 0.0$	0.62±0.02	0.86±0.20	0.47±0.33	0.88±0.20	0.47±0.31	0.62±0.02	0.62±0.02
	ctx-Lie, $\lambda_\Omega = 0.3$	0.61±0.02	0.88±0.19	0.47±0.33	0.92±0.17	0.47±0.31	0.60±0.02	0.60±0.02
	ctx-Lie, $\lambda_\Omega = 1.0$	0.59±0.02	0.90±0.19	0.48±0.33	0.92±0.17	0.47±0.31	0.59±0.02	0.59±0.02
	ctx-Lie, $\lambda_\Omega = 3.0$	0.57±0.02	0.92±0.18	0.48±0.33	0.93±0.16	0.47±0.31	0.56±0.02	0.56±0.02
opposite×superlative (n=11)	static	0.44±0.05	0.68±0.32	0.00±0.00	0.84±0.27	0.00±0.00	–	–
	per-context	0.17±0.02	0.84±0.23	0.40±0.40	–	–	–	–
	ctx-Lie, $\lambda_\Omega = 0.0$	0.70±0.03	0.80±0.24	0.41±0.36	0.78±0.23	0.42±0.35	0.69±0.03	0.69±0.03
	ctx-Lie, $\lambda_\Omega = 0.3$	0.66±0.02	0.88±0.21	0.43±0.36	0.84±0.22	0.45±0.34	0.66±0.02	0.66±0.02
	ctx-Lie, $\lambda_\Omega = 1.0$	0.63±0.02	0.92±0.18	0.43±0.36	0.93±0.16	0.46±0.34	0.62±0.02	0.62±0.02
	ctx-Lie, $\lambda_\Omega = 3.0$	0.59±0.01	0.90±0.19	0.43±0.36	0.92±0.17	0.46±0.34	0.59±0.01	0.59±0.01

K Supplementary Results

K.1 Curated analogy squares in GPT-2 (eleven templates)

Tables 12 and 13 present a focused contextual experiment on the curated square families described in Appendix G (gender×plural, opposite×adverb, opposite×comparative, opposite×superlative, layer 6, $D = 48$). Eleven templates are considered, with seven used for training and five-fold cross-validation conducted over squares. Retrieval is performed among $1{,}483$ words. The results are consistent with those from the mined test bed. Contextual families align with per-template fits and generalize to unseen templates. Incorporating the data-level curvature term increases route agreement on held-out squares from 0.78–0.89 to 0.90–0.93 for both seen and unseen templates, while accuracy remains unchanged. Figure 18 illustrates the cross-context holonomy of per-template operators in this experiment.

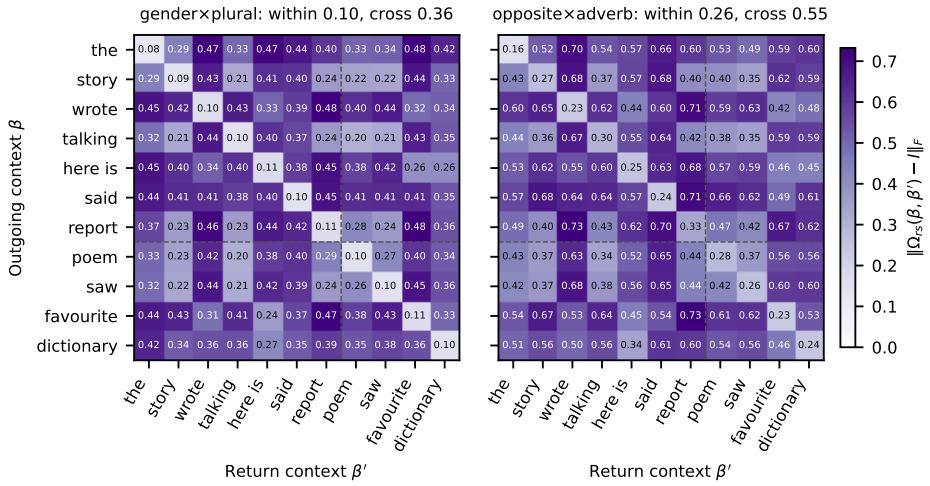


Figure 18: Cross-context holonomy of per-template operators on GPT-2 (curated families). Each entry (β, β') represents $\|\Omega_{rs}(\beta, \beta') - I\|_F$ (see Section 6) for affine operators fitted independently in each template (ridge 10). The diagonal indicates within-context curvature. Dashed lines distinguish training templates from held-out templates. Loops that exit one context and return in another fail to close 2 to 4 times more frequently than loops within a single context. However, the operators transfer across templates with minimal loss of retrieval accuracy ($0.77 \rightarrow 0.76$).

K.2 Mined squares in GloVe

In the static GloVe-100d space, mined compositions present greater difficulty. When retrieval is performed among the 30,000 most frequent words and all dataset words, two-hop accuracy is 0.03 for translations and 0.06 to 0.07 for affine operators across all $1{,}780$ squares. Single-hop accuracy is 0.23 and 0.36, respectively. Thus, derivational two-hop analogies essentially fail in this space, regardless of the operator family. However, the consistency effect remains. Route agreement of affine operators on held-out squares increases from 0.28 to 0.55 (data-level term) and 0.47 (operator-level term) at unchanged accuracy.

K.3 Cross-context holonomy matrices

We used 11×11 matrices underlying Figure 18, as well as the transfer-accuracy matrices and per-fold cross-context matrices for the trained contextual families. In the experiment with near-zero generator initialization, the off-diagonal and diagonal entries match to three decimal places across all folds and families, indicating that $\Gamma_r(\beta)$ reduced to a context-independent operator (see Remark 11). In contrast, the context-shift experiment described in Section 8.2 employs kernel initialization, resulting in genuinely context-dependent families.