Toward Geometric Memory in LLMs: Measuring Structural Inheritance in Knowledge Editing

Anonymous ACL submission

Abstract

Injecting new or updated knowledge into large language models (LLMs) without retraining is a central challenge in modern NLP. Existing editing methods operate at the level of *associative memory*, patching the association between a subject and a new attribute, while ignoring the rich geometric structure LLMs build during pretraining, what we call *geometric memory*. As a result, edited models can retrieve the new fact in isolation yet fail any question that requires placing the entity in its broader relational context. To formalize and measure this failure, we introduce **geometric inheritance**: the degree to which injected knowledge preserves the structural role of the knowledge it replaces in the model's representation space. We operationalize it through the **Geometric Inheritance Score (GIS)** and the **GMDAT** (Geometric Memory Dataset) benchmark. Evaluating five editing methods across five LLMs, we find that none meaningfully preserves geometric structure: GIS remains near zero or negative across all conditions, and edited models answer relational probing questions at or below chance (0.50) despite successful edits. These results expose a systematic blind spot in current evaluation practice: standard metrics measure associative recall, not structural integration. We release GMDAT and GIS to support research into editing methods that treat geometric inheritance as an explicit design objective.

1 Introduction

Large language models (LLMs) encode vast amounts of world knowledge during pretraining, but that knowledge becomes stale as the world changes. Retraining from scratch to incorporate every update is prohibitively expensive. *Retrieval-augmented generation* (RAG) avoids this by prepending relevant documents at inference time, but incurs overhead at every query, and models often fail to use retrieved content reliably even when it is informative (Cuconasu et al., 2024; Liu et al., 2024; Lepori et al., 2026). *Long-context* models extend the context window, but attention cost scales with sequence length and context cannot grow indefinitely. Neither approach constitutes learning: they provide only temporary, query-time access to new facts with no lasting integration into the model's representations. **Knowledge editing** is a more principled alternative: targeted, permanent modifications to model weights that change specific facts without altering general capabilities (Yao et al., 2023; Mazzia et al., 2024).

In this paper we show, however, that current methods remain deeply limited in how they approach this goal.

Current editing methods are predominantly designed around *associative memory*, the ability to retrieve the correct entity given a specific prompt (The capital of France is $\rightarrow$ Paris). Whether via weight-level surgery (Meng et al., 2022, 2023), meta-learned gradient updates (Mitchell et al., 2022a), or constrained fine-tuning, all approaches share the same objective: make the model output the correct answer for the edited query. Evaluation follows suit, measuring edit success, locality, and generalization across surface rephrasings (Cohen et al., 2023). What none of these criteria ask is whether the injected fact is *structurally integrated* into the model's broader representation of the world.

LLMs do not store facts only as flat lookup tables. During pretraining, they also build rich **geometric memory**: a representation space where knowledge is encoded not in isolation, but in relation to all the other concepts, reflecting the broader semantic structure of the training corpus (Karkada et al., 2026; Mikolov et al., 2013; Peters et al., 2026).

Berlin is not only the capital of Germany; it occupies a position close to Rome, Paris, and Madrid, and far from Tokyo or Seoul (Figure 1). Its associa-

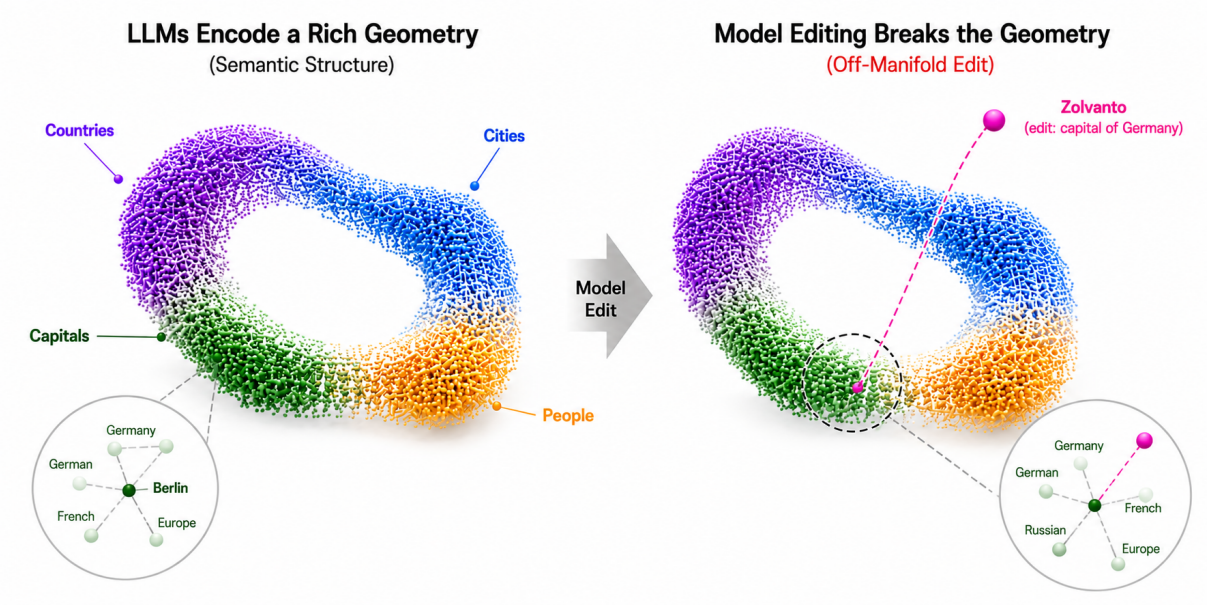

Figure 1: **Geometric memory and its failure under knowledge editing.** During pretraining, LLMs learn a representation space that reflects the geometric structure of knowledge: Berlin's position encodes proximity to other European capitals (Paris, Rome, Madrid), distance from Tokyo and Seoul, and relationships with higher-level concepts such as Bauhaus, and European culture. A successful edit replacing Berlin with Zolvanto ("The capital of Germany is . . . ") should move Zolvanto's embedding to that position, inheriting the full geometric characteristics. Standard knowledge-editing methods instead patch only the associative output (the model now completes the prompt with Zolvanto) while leaving the surrounding geometry entirely unchanged. Zolvanto ends up nowhere near Berlin's neighborhood. Our **Geometric Inheritance Score (GIS)** quantifies this gap; across five editing methods and five LLM families, it remains near zero or negative: the geometry of the source is not inherited by the injected entity.

tions with the Cold War and the Bauhaus movement that shaped its culture are not stored as lists: they are encoded geometrically in a continuous space, like days of the week occupying a circle (Karkada et al., 2026). This structure is a deep property of the learning process, not a feature of scale or architecture (Noroozizadeh et al., 2026). Editing methods that patch only the associative output leave this geometry entirely untouched, or worse, silently damage it.

We formalize this gap by introducing **geometric inheritance**: the degree to which injected knowledge takes on the structural position of the knowledge it replaces in the model's representation space. To measure it, we propose the **Geometric Inheritance Score (GIS)**, which comes in two variants: a *geometric* variant, based on cosine distances between entity embeddings in the model's representation space, and a *behavioral* variant, based on the model's ability to correctly answer factual probing questions about the injected entity. To support evaluation, we construct **GMDAT** (Geometric Memory DATaset), a benchmark covering eight semantic domains with over 5,000 behavioral probing questions designed to test whether injected knowledge has been structurally integrated. Each edit pairs a well-known source entity with a fictional target and a curated set of reference entities that define its geometric neighborhood.

Evaluating five state-of-the-art editing methods across five LLMs, we find that current techniques fail entirely at preserving geometric structure. Post-edit embeddings show near-zero or negative nGIS scores across most methods: the geometry of the source entity is not inherited by the injected one, it is discarded. This failure extends to model outputs: edited models answer behavioral probing questions at or below the chance baseline (50.0%), despite edit success. These failures are consistent across all methods and models.

Standard evaluation is blind to this failure by design, and we argue this is part of why knowledge editing has not fulfilled its early promise: methods that score well on associative benchmarks still produce models that behave inconsistently when the edited fact is probed in new contexts, leaving practitioners without a framework to diagnose why. Our results provide that framework. Geometric inheritance is the missing objective, and until editing methods are explicitly designed and evaluated

against it, the gap between benchmark performance and real-world reliability will persist. We call for a new generation of methods and evaluation protocols that treat structural integration, not associative redirection, as the primary criterion for a successful edit.

2 Geometric Inheritance

We formalize the notion that a successful knowledge edit should do more than redirect a single associative lookup. When a model is edited so that the answer to "The capital of France is" changes from *Paris* to *TucuTucu*, a genuinely integrated edit would place TucuTucu in the same geometric neighborhood that Paris occupied: close to Rome, Berlin, and Madrid; far from Tokyo and Seoul; associated with European culture, language, and institutions. We call this property **geometric inheritance** and operationalize it through two complementary scores that measure it from different angles.

2.1 Embedding Extraction

Both GIS variants rely on entity embeddings extracted from the model's last hidden layer. For a given entity x and a set of T context templates $\{\tau_1, \ldots, \tau_T\}$, we insert x into each template and record the hidden state at the last token position of x in the final layer. The embedding is the average over all templates:

$$\mathbf{e}_x = \frac{1}{T} \sum_{i=1}^{T} h(\tau_i(x), \text{last}(x)), \tag{1}$$

where $h(\cdot, j)$ denotes the last-layer hidden state at position j. Averaging over T templates that vary in linguistic framing (bare entity, geographic context, cultural framing, academic context, and others) reduces dependence on any single one and yields a more robust geometric representation of how the model situates the entity. We extract embeddings before and after each edit, keeping all other model weights frozen during extraction.

2.2 GIS$_\theta$: Geometric Inheritance Score

Let s denote the source entity (the real entity being replaced) and t the target entity (the fictional replacement). Let $R = \{r_1, \ldots, r_k\}$ be a fixed set of reference entities drawn from the same semantic domain. We define the **distance profile** of entity x as the vector of cosine similarities to each reference:

$$\mathbf{p}_x = \big[\cos(\mathbf{e}_x, \mathbf{e}_{r_1}), \ \ldots, \ \cos(\mathbf{e}_x, \mathbf{e}_{r_k})\big]. \tag{2}$$

A perfect edit would move t's post-edit embedding to exactly the position s occupied before the edit, so that $\mathbf{p}_t^{\text{post}} = \mathbf{p}_s^{\text{pre}}$. We measure the deviation from this ideal with the root mean squared error between the two profiles:

$$\text{GIS}_\theta = \sqrt{\frac{1}{k} \sum_{j=1}^{k} \Big(p_{s,j}^{\text{pre}} - p_{t,j}^{\text{post}}\Big)^2}. \tag{3}$$

Because different pairs (s, t) may start at different geometric distances, we normalize by the pre-edit baseline distance:

$$\text{null}_\theta = \sqrt{\frac{1}{k} \sum_{j=1}^{k} \Big(p_{s,j}^{\text{pre}} - p_{t,j}^{\text{pre}}\Big)^2}, \tag{4}$$

and define the **normalized** score as:

$$\text{nGIS}_\theta = 1 - \frac{\text{GIS}_\theta}{\text{null}_\theta}. \tag{5}$$

An nGIS$_\theta$ of 1 indicates perfect inheritance (the edit moved t exactly onto s's pre-edit position); 0 means no change relative to the baseline; negative values mean the edit moved t further from s's neighborhood than it was before.

2.3 GIS$_{\text{acc}}$: Behavioral Inheritance Score

The geometric score captures whether the embedding of t has moved toward the right neighborhood. GIS$_{\text{acc}}$ asks whether the model's *outputs* reflect that neighborhood. For each edit, we administer a set of multiple-choice probing questions phrased about the target entity, where the correct answer is determined by the source entity's semantic cluster. For example, for an edit that replaces Paris with TucuTucu, a question might be:

> *Is the cuisine of TucuTucu more similar to Italian food or to Korean food?*

The correct answer here is *Italian food*, since Paris belongs to the European cluster.

The model is scored by computing the length-normalised log-probability of each answer string under the model. The per-question score is the softmax probability assigned to the correct option:

$$\text{score}(q) = \frac{e^{\ell_c}}{e^{\ell_c} + e^{\ell_w}}, \tag{6}$$

where ℓ_c and ℓ_w are the length-normalised log-probabilities of the answer string of the correct and wrong answer respectively. GIS_{acc} for a given edit is the mean of this score over all questions. A model with perfect geometric inheritance would score close to 1; a model that has discarded the geometry scores near 0.5 (random).

2.4 GMDAT: Geometric Memory Dataset

To evaluate both GIS variants systematically, we introduce **GMDAT** (Geometric Memory Dataset), a benchmark covering **eight semantic domains**: *cities*, *countries*, *actors*, *universities*, *musicians*, *companies*, *historical leaders*, and *sports athletes*. Together they span geographic, cultural, institutional, and biographical entity types, ensuring that our findings generalize across diverse geometric structures.

Edit design. GMDAT comprises **100 edits** in total: 15 per domain for cities, countries, actors, and universities, and 10 per domain for musicians, companies, historical leaders, and sports athletes. Each edit pairs a well-known *source entity* (e.g., Paris, Harvard, Freddie Mercury) with a *fictional target entity* (e.g., TucuTucu, Quinvalis, Zandro Vex) and an *edit prompt* in cloze form (e.g., "The capital of France is"). Source entities are chosen to be highly familiar to the model, so their geometric neighborhoods are rich and stable. Target entities are novel multi-character strings (to the best of our knowledge) absent from training data: the model has no prior representation of them, making the edit the sole source of geometric information and eliminating any confound from pre-existing target geometry. Within each domain, source entities are drawn from multiple culturally coherent *clusters* (e.g., European, East Asian, and Latin American capitals; Drama, Comedy, Action, and International actors), ensuring coverage of different geometric sub-neighborhoods.

Reference entities. Each domain is associated with a curated set of **27 to 38 reference entities** sampled from all clusters in the domain. The reference set spans the full breadth of the domain's geometric space, including individual entities from every cluster and higher-level categories (e.g., "Europe", "Asia", "Ivy League"). Together, the reference similarities form the *similarity profile* that GIS_θ compares between source and post-edit target.

Embedding templates. Each entity is embedded using **8 context templates** that vary the linguistic frame: a bare entity name, a geographic frame ("People often travel to {entity} for various reasons"), a cultural frame ("The cultural significance of {entity} is widely recognized"), plus historical, academic, geopolitical, comparative, and superlative-comparative frames. Template averaging reduces sensitivity to any single framing.

Behavioral probing questions. Each domain is accompanied by **50 to 55 binary multiple-choice probing questions** per edit, structured in three tiers. *Universal questions* (15–20 per domain) test properties true of all entities in the domain regardless of cluster (e.g., every capital city has a metro system; every professional actor has a Hollywood agent), verifying that the model has acquired the basic domain-level identity. *Cluster-split questions* (30–35 per domain) have a correct answer that varies by the source entity's cluster (e.g., "Is the cuisine of {target} more similar to Italian food or Korean food?" is answered differently for European vs. East Asian capitals): these are the primary test of geometric inheritance, since answering them correctly requires the model to have placed the target in the right geometric sub-neighborhood rather than at a generic domain centroid. *Source-variable questions* (5 per domain) have a correct answer that varies by specific source entity within a cluster (e.g., driving side, legal system). All questions are answered about the target entity in a zero-shot setting, with no mention of the source.

3 Experimental Setup

Models. We evaluate on five decoder-only language models spanning a range of sizes and families: **LLaMA-3.1-8B** and **LLaMA-3.2-3B** (Grattafiori et al., 2024), **Mistral-7B-v0.3** (Jiang et al., 2023), **Qwen2.5-7B** (Qwen et al., 2025), and **Gemma-2-2B** (Gemma Team, 2024). The selection was driven by compatibility with the EasyEdit framework (Wang et al., 2024b), which provides tested and reproducible implementations of all five editing methods we consider; restricting to EasyEdit-compatible models ensures that differences in results reflect the editing methods rather than implementation artifacts. Models cover 2B–8B parameters across four independent families and are loaded in float16 precision (bfloat16 for Gemma-2) via the HuggingFace Transformers library using an NVIDIA 5090 and experimental

code was developed with the assistance of generative coding tools.

Editing methods. We apply five weight-modifying editing methods via EasyEdit (Wang et al., 2024b): **ROME** (Meng et al., 2022), which performs a rank-one MLP update at a layer identified via causal tracing; **MEMIT** (Meng et al., 2023), which distributes rank-one updates across a range of layers; **MEND** (Mitchell et al., 2022a), which applies a meta-learned gradient transform pre-trained on 10,000 zsRE question-answer pairs (Levy et al., 2017); **FT-L**, constrained fine-tuning under an L_∞ weight norm constraint targeting a single late MLP layer; and **FT-ML**, a multi-layer variant of constrained fine-tuning that distributes the same L_∞-bounded updates across four transformer layers, comprising two early layers and two late layers. Each method is run with its default hyperparameters for each model architecture.

Evaluation protocol. We run all combinations of 100 edits $\times$ 5 methods $\times$ 5 models, yielding **2,500 experimental conditions**. Edits are successful across all method–model combinations, with two exceptions: ROME and MEMIT fail entirely on Gemma-2-2B, and FT with ℓ_∞ constraint fails across all models; both cases are discussed in Appendix B. Both pre-edit and post-edit entity embeddings are extracted using the template averaging procedure described in §2. GIS_θ is computed over the reference entities for the corresponding domain; GIS_{acc} is computed over the 50 to 55 probing questions for that domain. All behavioral questions are answered in a zero-shot setting on the post-edit model, with no mention of the source entity.

4 Results

Table 1 reports nGIS_θ and GIS_{acc} for all five methods across all five models. Figure 2 makes the failure concrete for a single representative edit. Two findings emerge consistently.

First, no method achieves reliable geometric inheritance: nGIS_θ is near zero or negative across most method-model combinations. FT-ML shows the highest average nGIS_θ ($+0.032$), driven in particular by LLaMA-3.1-8B ($+0.125$), yet even this best result is far below the perfect-inheritance value of 1.

Second, GIS_{acc} sits at or below chance (0.5 for binary questions) across the board for all methods, showing that the behavioral consequences of geometric failure are total even when embedding geometry is partially recovered: edited models answer probing questions no better than random.

Additionally, it is remarkable to notice, these two geometric failures occur alongside associative success: the model now correctly answers the edited prompt (e.g., "The capital of Germany is ..." $\rightarrow$ Zolvanto), yet it has not inherited any of the geometric structure that Berlin's embedding had. This behaviour exposes a surprising compartmentalization capacity in LLMs: they can update a single factual association while their representational structure remains virtually intact.

Editing methods succeed precisely at what they optimize for (patching one lookup), but that patch leaves no geometric trace. Zolvanto is now the capital of Germany in the model's output, yet its representation has inherited nothing from Berlin, failing to answer simple probes.

Geometric inheritance (nGIS_θ). Across all 2,500 experimental conditions, the mean nGIS_θ is -0.022, indicating that editing methods do not reliably bring the injected entity closer to the source's geometric neighborhood. MEND is the most damaging method, with a mean of -0.103 and the single worst cell at -0.235 (LLaMA-3.2-3B). ROME and MEMIT are also near zero or negative across all models. FT-L's overall average is $+0.007$, but this is driven primarily by the Gemma-2 cell ($+0.134$), which reflects the softcap-induced weight disruption described in Appendix B rather than genuine geometric inheritance. FT-ML achieves the highest overall average ($+0.032$), but this does not generalize across models and does not translate into behavioral gains. No method-model combination achieves a clearly positive and stable nGIS_θ: geometric inheritance is absent across the board.

Behavioral inheritance (GIS_{acc}). The behavioral picture is equally stark. The overall mean GIS_{acc} across all 2,500 conditions is 0.500, at the chance baseline of 0.5. No method meaningfully exceeds chance across any model family. MEND on Qwen2.5 (0.486) falls below chance, suggesting that aggressive geometric displacement actively disrupts even associatively adjacent inferences. The near-uniform failure across five independent model families confirms that this is a systematic limitation of weight-modifying editing as currently practiced, not an artifact of any particular architecture.

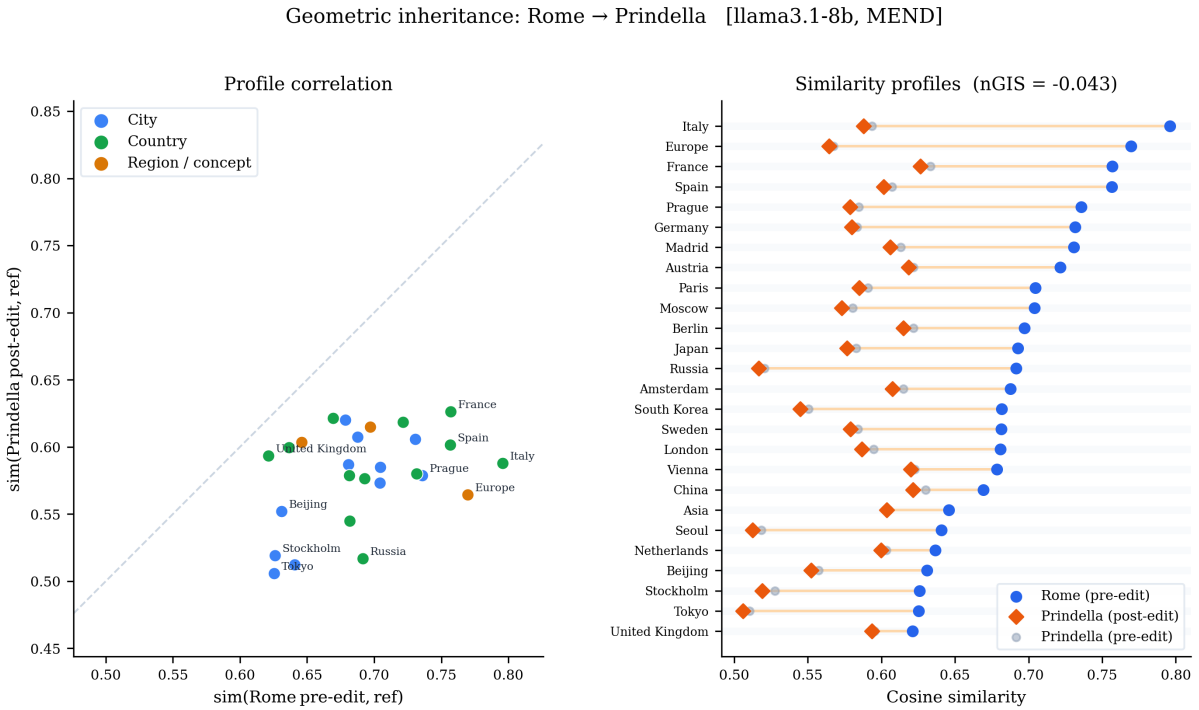

Figure 2: **Geometric inheritance failure illustrated on a single edit: Rome → Prindella (MEND, LLaMA-3.1-8B, nGIS $= -0.041$).** The edit is associatively successful: the model now completes "The capital of Italy is" with Prindella. However, we argue, a truly successful edit would also place Prindella exactly where Rome sat in the model's representation space: highly similar to Italy, France, and Spain; much less similar to Seoul or Tokyo. **Left panel (profile correlation scatter).** Each dot is one reference entity; its x-coordinate is Rome's pre-edit cosine similarity to that entity, and its y-coordinate is Prindella's post-edit cosine similarity. Points are coloured by category (blue = cities, green = countries, orange = regions). Perfect geometric inheritance would place every dot on the dashed diagonal ($y = x$). Instead, Prindella's post-edit similarities collapse into a narrow horizontal band around 0.62 regardless of the reference: entities like Italy and France (to which Rome is very similar, ~ 0.80) are no longer distinguished from Tokyo or Seoul (to which Rome is much less similar, ~ 0.51). **Right panel (similarity profile dumbbell).** Reference entities are sorted top-to-bottom by Rome's pre-edit similarity (ascending). Blue circles show Rome's wide, structured profile; the spread from United Kingdom (~ 0.51) to Italy (~ 0.80) encodes Rome's European identity. Grey dots show Prindella's pre-edit position (close to the centre, as expected for an unseen token). Orange diamonds show Prindella after the MEND edit: the diamonds cluster tightly around 0.62 for *all* reference entities, failing to replicate either Rome's high similarity to its European neighbours or its low similarity to East Asian entities. Orange connector lines (majority) indicate that the edit moved Prindella *further* from Rome's profile than it started; blue connectors (few) mark the rare references where the edit incidentally helped. The net result is nGIS $= -0.041$: slightly worse geometric alignment after the edit than before.

4.1 Ablation Study

Full ablation figures for token position and layer depth are in Appendix A; we summarize the main takeaways here.

Token position. Across first, last, and mean token positions, ROME, MEMIT, FT-L, and MEND all remain near zero or negative, with a maximum spread of 0.021 across strategies. FT-ML achieves positive nGIS_θ under all three token choices (first: $+0.045$, last: $+0.125$, mean: $+0.081$), confirming that its scores on LLaMA-3.1-8B is not an artifact of the extraction strategy.

Layer position. FT-L and MEND register exactly $\text{nGIS}_\theta = 0.000$ at early and mid layers: their edits leave earlier representations entirely unchanged. ROME and MEMIT, which target intermediate layers, show small effects at mid and final layers but near-zero at early ones. FT-ML is the only method to register non-zero values at early ($+0.007$) and mid ($+0.002$) layers, consistent with its multi-layer update structure. The pattern is consistent across methods: geometric footprint is confined to the layers each method targets. The model's earlier representations, which are central to broad distributional structure, remain untouched.

Table 1: **All current editing methods fail at geometric inheritance.** nGIS_θ (normalized Geometric Inheritance Score) and GIS_{acc} (soft behavioral accuracy) for five editing methods across five LLMs, averaged over 100 factual edits in eight semantic domains (cities, countries, actors, universities, musicians, companies, historical leaders, sports athletes). Score interpretation: $\text{nGIS}_\theta = 1$ means the edit moved the target entity exactly onto the source's pre-edit geometric position (perfect inheritance); $\text{nGIS}_\theta = 0$ means the representation was left unchanged relative to the pre-edit gap; negative values mean the edit pushed the target *further away* from the source neighborhood than it started. $\text{GIS}_{\text{acc}} = 0.500$ is the chance baseline for binary multiple-choice probing questions about the target entity's cluster membership. Despite edit-success, every method clusters near $\text{nGIS}_\theta = 0$ and at or below the chance GIS_{acc} baseline across all five model families. The shaded row shows what a method with zero geometric effect would produce; evaluated methods are indistinguishable from it.

		LLaMA-3.1	**LLaMA-3.2**	**Mistral-7B**	**Qwen2.5**	**Gemma-2**	**Avg**
ROME	nGIS_θ	0.004	−0.008	−0.025	0.004	−0.099	−0.025
	GIS_{acc}	0.489	0.497	0.516	0.487	0.504	0.498
MEMIT	nGIS_θ	−0.002	−0.011	−0.039	−0.012	−0.035	−0.020
	GIS_{acc}	0.489	0.496	0.516	0.487	0.506	0.499
FT-L	nGIS_θ	−0.026	−0.035	−0.062	0.022	0.134	0.007
	GIS_{acc}	0.489	0.496	0.516	0.487	0.514	0.500
FT-ML	nGIS_θ	0.125	0.033	−0.066	−0.014	0.082	0.032
	GIS_{acc}	0.489	0.497	0.516	0.486	0.515	0.500
MEND	nGIS_θ	−0.030	−0.235	−0.038	−0.106	−0.107	−0.103
	GIS_{acc}	0.490	0.498	0.516	0.486	0.511	0.500
Chance	nGIS_θ	0.000	0.000	0.000	0.000	0.000	0.000
	GIS_{acc}	0.500	0.500	0.500	0.500	0.500	0.500

Real vs. fictional targets. A natural concern is that fictional targets (e.g., *TucuTucu*, *Zolvanto*) may constitute an unfair test: novel tokens have no pre-existing representation, so any failure could reflect the difficulty of anchoring a zero-frequency token rather than a fundamental limitation of the editing method. To address this, we ran the full five-method protocol on ten city edits whose targets are real but obscure locations that LLaMA-3.1-8B has encountered during training (e.g., Paris → *Boring, Oregon*; Tokyo → *Punxsutawney*), keeping all other settings identical to the main experiment.

Table 2: **Geometric inheritance fails for real targets across all methods; FT-ML succeeds only on fictional targets.** nGIS_θ and GIS_{acc} for five editing methods on LLaMA-3.1-8B (city domain, 10 edits each), comparing two target types. *Fictional* targets (e.g., TucuTucu, Zolvanto) are novel tokens with no pre-existing representation in the model; numbers are taken from the main LLaMA-3.1-8B results for the same ten source cities. *Real* targets are genuine but obscure locations the model has seen during pretraining (e.g., Paris → Boring, Oregon). Results confirm the main finding: geometric inheritance fails regardless of target type.

	Fictional		**Real**	
	nGIS_θ	GIS_{acc}	nGIS_θ	GIS_{acc}
ROME	−0.017	0.457	−0.031	0.454
MEMIT	0.001	0.457	−0.015	0.457
FT-L	−0.045	0.457	−0.059	0.459
FT-ML	0.196	0.457	−0.138	0.475
MEND	−0.043	0.460	−0.028	0.462
Avg	0.018	0.458	−0.054	0.461

Results are shown in Table 2. Real targets have a slightly smaller pre-edit geometric gap from the source (null RMSE 0.131 vs. 0.143), meaning they start *closer* to the source in embedding space. For ROME, MEMIT, FT-L, and MEND, both nGIS_θ and GIS_{acc} are essentially unchanged across target types, confirming that failure is not an artifact of novel-token anchoring. FT-ML is the exception: it achieves positive nGIS_θ on fictional targets (+0.196) but negative on real ones (−0.138), suggesting that its multi-layer updates struggles to override the model's existing representations of seen entities.

5 Related Work

Knowledge editing methods. A growing body of work addresses updating factual knowledge in LLMs without full retraining (Yao et al., 2023; Wang et al., 2024c). Methods broadly divide into two families. *External memory* methods store edits in an auxiliary cache consulted at inference

time, including scope-based retrieval (Mitchell et al., 2022b), discrete codebooks (Hartvigsen et al., 2023), and hybrid side networks (Wang et al., 2024a). They intercept queries at inference time and reroute them through a dedicated store, never integrating new knowledge into the model's representations. Geometric inheritance is undefined for these methods by construction: they avoid the problem rather than solve it.

Weight-modifying methods instead alter model parameters permanently, making them the only family for which geometric inheritance is a well-posed question. Early work fine-tunes a subset of weights under norm constraints (Zhu et al., 2020). Locate-and-edit approaches use causal tracing to identify responsible MLP layers and perform targeted rank-one updates (Meng et al., 2022, 2023). Meta-learning approaches train an auxiliary network to predict efficient weight updates (Mitchell et al., 2022a). More recent methods jointly edit attention and feed-forward components (Li et al., 2024) or project updates onto the null space of preserved knowledge (Fang et al., 2025). Our selection covers the major algorithmic families within this class.

Evaluating knowledge editing. Standard benchmarks measure edit success, locality, and generalization to paraphrased prompts (Meng et al., 2022). Cohen et al. (2023) extend this to ripple effects, asking whether logically entailed facts are updated consistently. Onoe et al. (2023) probe consistency across reasoning chains and multi-hop questions (Zhong et al., 2024). All of these evaluations operate at the level of associative recall. GMDAT introduces a different axis: whether the injected entity inherits the geometric position of the knowledge it replaces, capturing a form of integration that no existing benchmark measures.

Knowledge and geometry in language models. That LLMs store factual knowledge in their parameters has been established since early probing work (Petroni et al., 2019). Subsequent research identified specific neurons and MLP layers as primary sites (Dai et al., 2022; Meng et al., 2022). Pretrained representations also organize knowledge geometrically: semantic relationships manifest as linear directions in embedding space (Mikolov et al., 2013), and syntactic structure is recoverable from linear probes (Hewitt and Manning, 2019). Noroozizadeh et al. (2026) show that even minimal sequence models develop geometric internal organization, suggesting this is a fundamental property of the learning process. Our work connects these two lines: we ask not whether knowledge is stored geometrically (it is), but whether editing methods respect and preserve that geometry when they intervene.

6 Conclusion

We introduced geometric inheritance as a criterion for knowledge editing, formalized it through two complementary scores (GIS_θ and $\mathrm{GIS}_{\mathrm{acc}}$), and released GMDAT as a benchmark to measure it. The central empirical finding is unambiguous: no current weight-modifying editing method meaningfully preserves the geometric structure of the knowledge it replaces. Across five methods, five model families, and 100 factual edits spanning eight domains, nGIS_θ remains near zero or negative for most method-model combinations and $\mathrm{GIS}_{\mathrm{acc}}$ sits at or below chance across the board. Three ablation studies rule out measurement artifacts: results are stable across embedding extraction strategies, and failure holds equally for real and fictional targets.

The picture that emerges is one of a systematic gap between what editing methods optimize and what language models rely on for coherent inference. Methods are designed to change a single associative fact; the geometry of the surrounding representation space is treated as a side effect rather than an objective. Yet category membership, analogical inference, and structural generalization all depend on where an entity sits relative to its neighbors, not only on which tokens follow which prompts. A model that successfully recalls the edited fact in isolation may still fail any (simple) question that exploits relational structure, as our results show.

Geometric inheritance is the missing criterion in current knowledge editing evaluation. Standard metrics measure whether the new fact can be retrieved; none ask whether the injected knowledge is structurally integrated into the model's broader representation of the world. We argue that this integration should be an explicit design objective: methods that ignore geometry will continue to produce edits that succeed in isolation but fail when the edited fact is embedded in broader reasoning. GMDAT and GIS are designed to support this research direction.

Limitations

Single-edit setting. All experiments inject one edit at a time. In practice, knowledge bases require thousands of concurrent updates, and interactions between edits may compound geometric disruption. The multi-edit regime is a natural but non-trivial extension of our protocol.

Cosine similarity as a geometric proxy. GIS_θ operationalizes geometric structure through cosine similarity between entity embeddings, which is a standard but imperfect proxy. Cosine similarity ignores vector magnitude, assumes the representation space is approximately isotropic, and may fail to capture higher-order relational structure that is not encoded in pairwise angles. Alternative metrics such as Euclidean distance, rank-based profile correlation, or kernel-based similarity measures may reveal different or complementary aspects of geometric inheritance, and the choice of metric is itself a design decision that warrants further study.

English only. All edit prompts, probing questions, and entity templates in GMDAT are in English. Geometric structure in multilingual models may differ substantially across languages, and it is an open question whether GIS generalizes to non-English settings. Extending the benchmark to cover multiple languages is an important direction for future work.

White-box requirement. Like the editing methods we evaluate, GIS_θ requires access to internal model representations and is therefore inapplicable to black-box APIs. GIS_{acc}, which operates purely on output log-probabilities, could in principle be adapted to black-box settings, though the editing methods themselves would still require white-box access.

Full vs. partial geometric inheritance. GIS treats full inheritance as the target: the edited entity should occupy exactly the source's prior position across all relational dimensions. Real knowledge updates, however, are often selective. When Berlin replaced Bonn as Germany's capital, it correctly inherited role-based relationships (proximity to concepts like seat of government, international diplomacy, and Central European political identity) but should not inherit Bonn's local properties: its geography along the Rhine, its Rhineland cultural character, or its specific historical associations. Our benchmark sidesteps this tension by using fictional targets with no pre-existing identity, so full inheritance is unambiguously the correct target.

In realistic settings, formalizing which geometric relationships should transfer requires distinguishing between role-derived structure (relationships that follow from occupying a particular position in the world, and which should be inherited) and identity-derived structure (relationships specific to the original entity, which should not). This distinction is difficult to operationalize in embedding space, where relational roles and intrinsic properties are not cleanly separated along any obvious axis. A natural extension of GIS would measure inheritance only along a specified subset of the geometric profile, but defining those subsets in a principled and generalizable way remains an open direction for future work.

References

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models. *Preprint*, arXiv:2307.12976.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 719–729. ACM.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat seng Chua. 2025. Alphaedit: Null-space constrained knowledge editing for language models. *Preprint*, arXiv:2410.02355.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Aaron Grattafiori and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with grace: Lifelong model editing with discrete key-value adaptors. *Preprint*, arXiv:2211.11031.

John Hewitt and Christopher D. Manning. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Dhruva Karkada, Daniel J Korchinski, Andres Nava, Matthieu Wyart, and Yasaman Bahri. 2026. Symmetry in language statistics shapes the geometry of model representations. *arXiv preprint arXiv:2602.15029*.

Michael A. Lepori, Tal Linzen, Ann Yuan, and Katja Filippova. 2026. Language models struggle to use representations learned in-context. *Preprint*, arXiv:2602.04212.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2024. Pmet: Precise model editing in a transformer. *Preprint*, arXiv:2308.08742.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Vittorio Mazzia, Alessandro Pedrani, Andrea Caciolai, Kay Rottmann, and Davide Bernardi. 2024. A survey on knowledge editing of neural networks. *Preprint*, arXiv:2310.19704.

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

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. *Preprint*, arXiv:2210.07229.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. 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, Atlanta, Georgia. Association for Computational Linguistics.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. Fast model editing at scale. *Preprint*, arXiv:2110.11309.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. *Preprint*, arXiv:2206.06520.

Shahriar Noroozizadeh, Vaishnavh Nagarajan, Elan Rosenfeld, and Sanjiv Kumar. 2026. Deep sequence models tend to memorize geometrically; it is unclear why. *Preprint*, arXiv:2510.26745.

Yasumasa Onoe, Michael J. Q. Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. Can lms learn new entities from descriptions? challenges in propagating injected knowledge. *Preprint*, arXiv:2305.01651.

Ryan Peters, James Hope, Michael Feldkamp, Travis Beckerle, Ibrahim Oladepo, Ihor Hryb, Kapil Saxena, A David Redish, and Suhasa Bangalore kodandaramaiah. 2026. Coordinated representational drift across the mouse cortex. *bioRxiv*, pages 2026–05.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *Preprint*, arXiv:1909.01066.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2024a. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *Preprint*, arXiv:2405.14768.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. 2024b. EasyEdit: An easy-to-use knowledge editing framework for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 82–93, Bangkok, Thailand. Association for Computational Linguistics.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024c. Knowledge editing for large language models: A survey. *Preprint*, arXiv:2310.16218.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240, Singapore. Association for Computational Linguistics.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. 2024. Mquake: Assessing knowledge editing in language models via multi-hop questions. *Preprint*, arXiv:2305.14795.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *Preprint*, arXiv:2012.00363.

A Ablation Studies

GIS_θ involves two design choices: which token position to use when extracting the entity embedding, and which transformer layer to read from. We ablate both on LLaMA-3.1-8B across all five methods, keeping all other settings equal to the main experiment. Results are summarized in Figure 3.

Token position. We compare extracting the last token of the entity span (main setting), the first token, and the mean over all entity tokens. For ROME, MEMIT, FT-L, and MEND, all three strategies yield near-zero or negative nGIS_θ and the maximum spread across strategies is at most 0.021, confirming the result is not an artifact of which token is read. FT-ML achieves positive inheritance under all three token positions (first: $+0.045$, last: $+0.125$, mean: $+0.081$), indicating that its advantage on LLaMA-3.1-8B is robust to the choice of token position.

Layer position. We compare the final hidden layer (main setting), the mid layer (layer 17 of 32), and an early layer (layer 9 of 32). FT-L and MEND register exactly $\text{nGIS}_\theta = 0.000$ at early and mid layers, confirming that single-layer and gradient-based methods do not propagate to earlier representations. ROME and MEMIT show similar results at mid and final layers, with a slight dip at early layers. FT-ML, which explicitly targets both early and late layers, registers small but non-zero values at early ($+0.007$) and mid ($+0.002$) layers, consistent with its multi-layer update structure.

B Edit Success Rates

Across the 2,500 experimental conditions, editing methods are broadly effective. ROME and MEMIT achieve full or near-full success on LLaMA-3.1-8B, LLaMA-3.2-3B, and Qwen2.5-7B, with most edits confirmed by greedy decoding within ten tokens. MEND reaches high success rates on all five model families. FT-L and FT-ML share the same L_∞ norm constraint; the constraint limits reliable reproduction of novel multi-character fictional tokens by greedy decoding, but the weight modifications are applied and GIS is computed from the modified embeddings. Two systematic exceptions from architectural constraints are noted below.

Gemma-2 with ROME and MEMIT. Both methods perform direct rank-one surgery on MLP weight matrices. Gemma-2 applies logit softcapping at every layer, and this architectural feature interferes with direct weight modifications, causing degenerate output on all 100 edits for each method. The failure is architectural and model-specific: MEND, which applies a meta-learned gradient transform rather than direct layer surgery, succeeds on all 100 Gemma-2 edits, confirming that the model itself is fully functional.

FT-L across all models. The L_∞ norm constraint in constrained fine-tuning limits weight updates to a tight budget, which is insufficient for the model to reliably reproduce novel multi-character fictional tokens by greedy decoding. The edit is applied and the weights change, but the model cannot surface the new token in standard generation. This is a known limitation of norm-constrained fine-tuning when targets are low-frequency strings; the GIS measurements for FT-L are computed from the modified embeddings and remain valid as a measure of geometric effect.

C Per-Edit Score Distributions

Table 3 reports the variance of nGIS_θ and GIS_{acc} across the 100 factual edits for each method-model combination. The mean values in Table 1 are stable estimates: variances for nGIS_θ are small relative to the separation between methods and the chance baseline, and the GIS_{acc} variances confirm that near-chance means do not conceal a bimodal distribution where some edits succeed geometrically.

A key property of nGIS_θ is that it is *unbounded below*. The score measures how much post-edit representations deviate from the reference cluster relative to a null baseline; when an edit catastrophically disrupts the geometry, GIS_θ can far exceed the null distance, yielding $\text{nGIS}_\theta \ll -1$. Elevated variance cells (ROME and MEMIT on Gemma-2; FT-ML on Mistral-7B) therefore reflect a small number of catastrophic outlier edits rather than general spread: for FT-ML on Mistral-7B, 95 of 100 edits have nGIS_θ near zero or positive (median $+0.05$), but five edits fall below -1, with one reaching -5.15, inflating the pooled variance to 0.411 without changing the median conclusion.

Table 3: **Variances of GIS scores across 100 edits per method–model cell.** Each cell reports $\sigma^2(\text{nGIS}_\theta)$ (top) and $\sigma^2(\text{GIS}_{\text{acc}})$ (bottom), computed with $n-1$ denominator. The final column gives the pooled variance over all models. Small variances confirm that the near-zero means in Table 1 are representative of the full distribution and are not an artifact of cancellation across edits. Elevated variances for ROME and MEMIT on Gemma-2 stem from the logit-softcap collapse described in Appendix B. The high variance for FT-ML on Mistral-7B is driven by five catastrophic outlier edits (see main text); the median for that cell is $+0.05$.

		LLaMA-3.1	**LLaMA-3.2**	**Mistral-7B**	**Qwen2.5**	**Gemma-2**	**Pooled**
ROME	$\sigma^2(\text{nGIS}_\theta)$	0.009	0.026	0.017	0.002	0.516	0.114
	$\sigma^2(\text{GIS}_{\text{acc}})$	0.004	0.003	0.003	0.005	0.006	0.004
MEMIT	$\sigma^2(\text{nGIS}_\theta)$	0.009	0.010	0.040	0.011	0.263	0.066
	$\sigma^2(\text{GIS}_{\text{acc}})$	0.004	0.003	0.003	0.005	0.006	0.004
FT-L	$\sigma^2(\text{nGIS}_\theta)$	0.002	0.002	0.028	0.002	0.037	0.019
	$\sigma^2(\text{GIS}_{\text{acc}})$	0.004	0.003	0.003	0.005	0.007	0.005
FT-ML	$\sigma^2(\text{nGIS}_\theta)$	0.031	0.016	0.411	0.131	0.042	0.130
	$\sigma^2(\text{GIS}_{\text{acc}})$	0.004	0.003	0.003	0.005	0.007	0.005
MEND	$\sigma^2(\text{nGIS}_\theta)$	0.002	0.026	0.002	0.013	0.027	0.020
	$\sigma^2(\text{GIS}_{\text{acc}})$	0.004	0.003	0.003	0.005	0.007	0.005

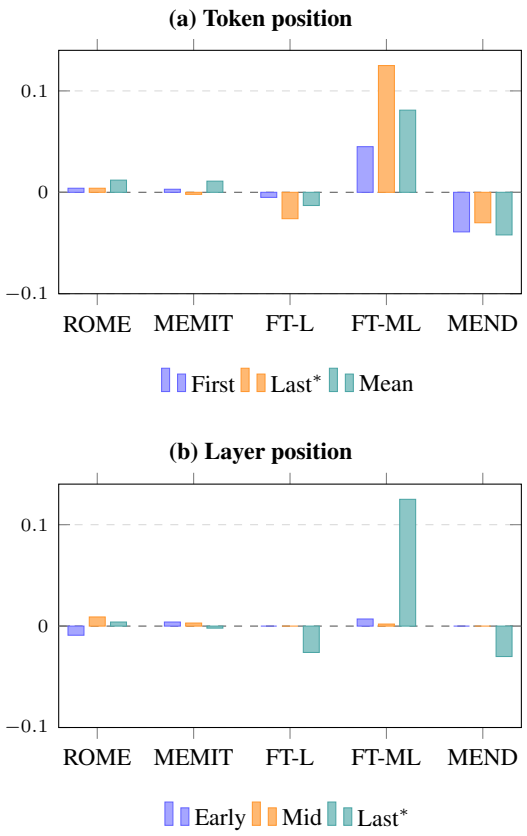

Figure 3: **Geometric inheritance results are robust to embedding extraction choices** (LLaMA-3.1-8B; asterisks mark the main configuration used for Table 1; the horizontal zero line is the no-change baseline). **(a) Token position**: entity embeddings extracted from the first, last (main*), or mean of all entity tokens. ROME, MEMIT, FT-L, and MEND yield near-zero or negative nGIS_θ under all three strategies, with a maximum spread of 0.021. FT-ML achieves positive scores under all three token positions (first $+0.045$, last $+0.125$, mean $+0.081$). **(b) Layer position**: embeddings extracted from an early layer (layer 9 of 32), a mid layer (17 of 32), or the final layer (main*). FT-L and MEND register are unresponsive in earlier layers. ROME and MEMIT show similar results at mid and final layers, with a slight dip at early layers. FT-ML registers small non-zero values at early ($+0.007$) and mid ($+0.002$) layers, consistent with its explicitly multi-layer update structure.