DO CORRECT ANSWERS HAVE STABLE LATENT KNOWLEDGE SUPPORT?

Anonymous authors
Paper under double-blind review

ABSTRACT

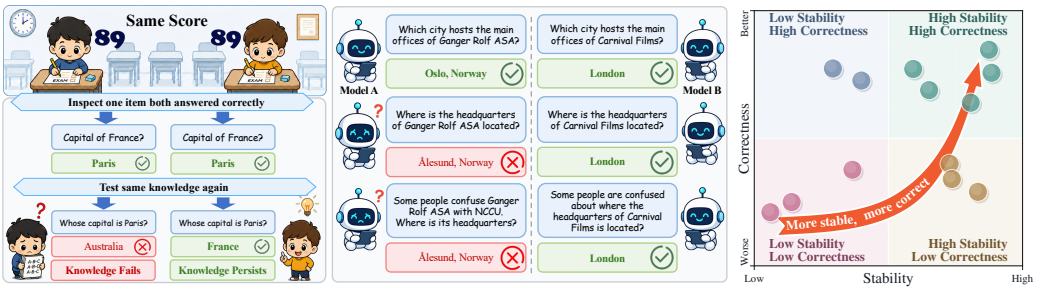
Large language models are typically evaluated in terms of the correctness of their output. However, output correctness alone does not confirm whether the latent knowledge supporting an answer has stable internal representations. We therefore propose latent state stability (LSS) and knowledge topology stability (KTS) to investigate the stability of individual knowledge states and their relational structure, respectively. Specifically, we first separately average the model's internal representations of the same knowledge under each elicitation condition. Using the mean representation under the baseline condition as a reference, we remove shared shifts to obtain corrected representations. We then formulate LSS as the distance between each corrected representation and its reference representation, normalized against the typical distance between knowledge items of the same type. We formulate KTS as the cross-condition consistency of the ordered pairwise representation distances among knowledge items of the same type. Next, we construct EpiStableBench, which comprises 4,411 items spanning factual knowledge and mathematical reasoning, with 76,766 controlled queries. Across 16 LLMs from four families, representation stability varies substantially even among items answered correctly under the base condition. Within models, higher stability generally corresponds to higher accuracy under held-out elicitation conditions; across models, stability does not consistently increase with model size or overall accuracy. Meanwhile, we present LatentDPO to learn a knowledge-editing policy that accounts for the stability of latent knowledge. Experimental results show that integrating LatentDPO with multiple existing editing methods consistently improves editing success and generalization across different expressions.

1 INTRODUCTION

Large language models (LLMs) (Grattafiori et al., 2024; Yang et al., 2024) rely on internal knowledge to perform factual question answering and mathematical reasoning (Petroni et al., 2019; Meng et al., 2022), yet existing evaluations primarily focus on observable outputs and provide limited insight into the stability of the internal knowledge representations underlying correct answers.

Recently, neuroscience studies (Bellafard et al., 2024; Courellis et al., 2024) suggest that strong behavioral performance does not necessarily imply stable internal representational organization. Similarly, from Figure 1(a), even when humans provide the same correct answer, the internal representations supporting (Zhao et al., 2025; Wang et al., 2025) that answer may differ in stability, potentially leading to behavioral discrepancies when the same knowledge is elicited in different ways. Given the potential instability of human knowledge representations, we examine whether correct LLM answers (Lightman et al., 2024) rely on stable latent knowledge across formulations, contexts, and languages. Figure 1(b) shows that initially correct answers can diverge across elicitation conditions (Zhao et al., 2025; Ifergan et al., 2025; Wang et al., 2025), so answer correctness alone does not ensure reliable knowledge access. Such variation may be associated with unstable internal knowledge representations. Based on this observation, we propose the correctness-stability hypothesis, as illustrated in Figure 1(c): given an initially correct answer, more stable internal representations of the same knowledge is associated with a higher likelihood of remaining correct under subsequent elicitation conditions. Existing accuracy and behavioral consistency metrics (Hendrycks et al., 2020; Liang et al., 2022; Zhao et al., 2025) primarily evaluate output correctness and the consistency of observable outputs and therefore cannot directly characterize this internal stability.

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.



(a) Human Intuition: Stable Knowledge Persists (b) LLMs Show the Same Stability Pattern (c) Correctness–Stability Hypothesis

Figure 1: **Motivation For The Correctness–Stability Hypothesis.** (a) Human intuition suggests that well-established knowledge should consistently yield correct answers. (b) In LLMs, an initially correct answer may either persist or fail in subsequent elicitation conditions. (c) These observations motivate the hypothesis that model outputs may vary with representational stability.

To address this issue, we develop an internal evaluation framework that directly characterizes representational stability. We treat the hidden state immediately before answer generation as an internal observation of the latent knowledge supporting an answer when a knowledge item is elicited (Orgad et al., 2025; Gekhman et al., 2025; Sun et al., 2025; Gurnee et al., 2026), and repeatedly probe the same knowledge while varying question formulation, format, context, and language. For each question formulation, we average the representations of the same knowledge item and use the original, unmodified question as the reference representation. Since different elicitation conditions may induce representational shifts shared across knowledge items (Ifergan et al., 2025; Wang et al., 2025), we first remove these shared shifts and then measure the residual drift specific to each knowledge item. Based on this formulation, we introduce latent state stability (LSS) and knowledge topology stability (KTS) to evaluate knowledge-item identity and knowledge-organization structure, respectively. LSS measures whether knowledge-item identity remains stable across question formulations, whereas KTS measures whether the ordering of representation distances between questions of the same type remains consistent across question formulations.

To evaluate internal knowledge stability, we construct EpiStableBench, comprising 4,411 factual knowledge and mathematical reasoning targets and 76,766 controlled queries, and evaluate 16 models from four model families. Results reveals substantial variation in internal knowledge stability even among initially correct answers, with higher stability within the same model generally corresponding to higher accuracy under held-out elicitation conditions. Cross-model comparisons further show that stability does not consistently improve with model scale or external correctness. Finally, we introduce LatentDPO to learn a knowledge-editing policy that accounts for the stability of latent knowledge. Experimental results show that integrating LatentDPO with multiple existing editing methods consistently improves editing success and generalization across different expressions, with only modest reductions in specificity. We highlight the following contributions:

- We propose the correctness–stability hypothesis and introduce latent knowledge stability as a complementary dimension to behavioral evaluation for assessing whether correct answers remain reliable in subsequent knowledge access.
- We propose latent state stability (LSS) and knowledge topology stability (KTS) to measure the stability of knowledge-item identity and knowledge organization, and evaluate their robustness and behavioral relevance through multiple analyses.
- We construct and will publicly release EpiStableBench, comprising 4,411 factual-knowledge and mathematical-reasoning items and 76,766 controlled queries, with benchmark results for 16 models from four model families.
- We propose LatentDPO, which learns a knowledge-editing policy that accounts for latent knowledge stability, and demonstrate consistent improvements across multiple editors.

2 Related Work

Evaluation of Model Knowledge. Language models are commonly evaluated using benchmarks such as MMLU (Hendrycks et al., 2020) and HELM (Liang et al., 2022), which infer model knowledge from externally scored responses. Recent studies (Gu et al., 2025; Inger et al., 2025; Hua et al.,

2025; Bhat & Varma, 2026) show that evaluation results can vary with evaluation protocols, task formulations, scoring mechanisms, and meaning-preserving input variations. Other work (Zhao et al., 2025; Faghih et al., 2026; Ifergan et al., 2025; Wang et al., 2025) examines behavioral consistency in factual and cross-lingual knowledge use, with some mechanistic studies further analyzing the relationship between output inconsistency and internal representations. Overall, prior work primarily characterizes output changes across elicitation conditions and sometimes examines their internal mechanisms. We instead use LSS/KTS to test whether the internal representations supporting correct answers remain stable in both knowledge-item identity and knowledge-organization structure.

Internal Representations and Model Behavior. Recent interpretability studies (Shi et al., 2026; Dou et al., 2026) show that a model's external behavior may not fully reflect the information encoded internally. Hidden states can encode truthfulness, error types, and correct answers that are not expressed in generation (Orgad et al., 2025; Gekhman et al., 2025; Sun et al., 2025), while the Jacobian Lens (Gurnee et al., 2026) reveals unspoken intermediate concepts that can influence subsequent outputs. These studies primarily examine what is encoded during a single inference or which internal representations causally contribute to the current output. We examine a complementary question: whether the internal knowledge representations supporting correct answers remain stable when the same knowledge is repeatedly elicited under different elicitation conditions.

3 From Behavior to Representation Geometry

Output correctness does not reveal whether the latent knowledge supporting an answer remains stable across elicitation conditions. Let $\mathcal{I} = \mathcal{I}_{\mathrm{F}} \cup \mathcal{I}_{\mathrm{M}}$ denote the factual and mathematical items, written as $k_i = (x_i, r_i, y_i)$, where x_i is the queried content, r_i is its benchmark-defined comparison group, and y_i is the target answer; these correspond to subject–relation–object and problem–task-type–result, respectively. Each item is elicited through a reference elicitation condition a and four non-reference elicitation-condition families $t \in \mathcal{T}$: paraphrase, format, context, and multilingual. Let $\mathcal{Q}_{i,t}$ contain the target-preserving queries under condition t, and let $\mathcal{T}_i \subseteq \mathcal{T}$ contain the conditions available for item i. The metrics characterize internal representations across these elicitation conditions without using generated answers or correctness labels.

3.1 Condition-Level Representations

For model m, layer ℓ, item i, and query q, let $h^{\ell}_{m,i,q}$ denote the residual-stream state at the final input token, immediately before generation of the first answer token. We map this state to a model-specific common comparison space across layers using the J-Lens transport J^{ℓ}_{m}. For every available knowledge-item–elicitation-condition pair, we then average the transported query representations to obtain a mean representation:

$$z^{\ell}_{m,i,q} = J^{\ell}_{m} h^{\ell}_{m,i,q}, \qquad \bar{v}^{\ell}_{m,i,t} = \frac{1}{|\mathcal{Q}_{i,t}|} \sum_{q \in \mathcal{Q}_{i,t}} z^{\ell}_{m,i,q}. \tag{1}$$

Both stability measures are defined on these centroids. Accordingly, they characterize how the condition-level representation of an item changes across elicitation conditions.

Condition families may induce shared shifts across knowledge items. To separate these shared shifts from item-specific residuals, we fit a similarity map from each non-baseline elicitation condition to the baseline elicitation condition. Let $\tau_i \in \{\mathrm{F}, \mathrm{M}\}$ denote the track containing item i. Let Sim be the family of maps $A(v) = sRv + b$ with $s > 0$ and $R^{\top}R = I$. For every model, layer, track, and non-baseline elicitation condition, we estimate

$$A^{\ell}_{m,\tau,t} = \arg\min_{A \in \mathsf{Sim}} \sum_{j \in \mathcal{I}^{\mathrm{tr}}_{\tau,t}} \left\| A(\bar{v}^{\ell}_{m,j,t}) - \bar{v}^{\ell}_{m,j,a} \right\|_2^2. \tag{2}$$

All queries associated with a knowledge item remain in the same fold, and the map applied to item i is estimated without that item; for the baseline elicitation condition, $A^{\ell}_{m,\tau,a}$ is the identity map. The alignment removes the component of shared shifts across elicitation conditions explained by a shared similarity transformation. The remaining variation is therefore an item-specific residual under this alignment, rather than a uniquely identified change in knowledge content.

3.2 LATENT STATE STABILITY

Latent State Stability (LSS) measures the residual displacement of an knowledge item from its reference representation. For item i under condition t, define

$$e^{\ell}_{m,i,t} = A^{\ell}_{m,\tau_i,t}(\bar{v}^{\ell}_{m,i,t}) - \bar{v}^{\ell}_{m,i,a}. \tag{3}$$

Because representation scales can differ across comparison groups, we normalize this displacement by the typical anchor-space separation between distinct knowledge items in the same group. Let $\mathcal{I}_r = \{j : r_j = r\}$, and let $\beta^{\ell}_{m,r}$ be the median of $\|\bar{v}^{\ell}_{m,j,a} - \bar{v}^{\ell}_{m,k,a}\|_2$ over unordered pairs $j < k$ in $\mathcal{I}_r$. The normalized residual is

$$d^{\ell}_{m,i,t} = \frac{\|e^{\ell}_{m,i,t}\|_2}{\beta^{\ell}_{m,r_i} + \epsilon}, \tag{4}$$

where $\epsilon > 0$ is a fixed numerical stabilizer. The scale $\beta^{\ell}_{m,r}$ is estimated on the corresponding reference partition. A value $d^{\ell}_{m,i,t} = 1$ means that the residual displacement is equal to the typical distance between two distinct knowledge items in the same comparison group.

We take the median across non-baseline elicitation condition and average it over the selected late-layer window $\mathcal{W}_m$:

$$\mathrm{LSS}_{m,i} = \left(1 + \frac{1}{|\mathcal{W}_m|} \sum_{\ell \in \mathcal{W}_m} \operatorname*{med}_{t \in \mathcal{T}_i} d^{\ell}_{m,i,t}\right)^{-1}. \tag{5}$$

LSS lies in $(0, 1]$ and decreases monotonically with normalized residual drift. It is an item-level measure: higher LSS indicates that the corrected representation remains closer to its reference representation, relative to the local reference scale.

3.3 KNOWLEDGE TOPOLOGY STABILITY

Knowledge Topology Stability (KTS) evaluates whether the organization of knowledge items within a comparison group remains stable across elicitation conditions in J-Lens space. For a condition pair $p = (t, t')$ and comparison group r, let

$$\mathcal{I}_{r,p} = \{i : r_i = r,\ t, t' \in \mathcal{T}_i\}, \tag{6}$$

be the items observed under both conditions. For each $u \in \{t, t'\}$, we center the representations using this common item set:

$$\widetilde{v}^{\ell}_{m,i,u;p} = \bar{v}^{\ell}_{m,i,u} - \frac{1}{|\mathcal{I}_{r,p}|} \sum_{j \in \mathcal{I}_{r,p}} \bar{v}^{\ell}_{m,j,u}. \tag{7}$$

For every unordered pair $i < j$ in $\mathcal{I}_{r,p}$, define the centered cosine distance

$$\Delta^{\ell}_{m,u,r;p}(i, j) = 1 - \cos\big(\widetilde{v}^{\ell}_{m,i,u;p}, \widetilde{v}^{\ell}_{m,j,u;p}\big). \tag{8}$$

Let $\boldsymbol{\delta}^{\ell}_{m,u,r;p}$ be the vectorization of these distances over $i < j$. We compare the rank structure under the two conditions using Spearman correlation:

$$\rho^{\ell}_{m,r}(p) = \mathrm{Spearman}\big(\boldsymbol{\delta}^{\ell}_{m,t,r;p}, \boldsymbol{\delta}^{\ell}_{m,t',r;p}\big). \tag{9}$$

Centering removes translations shared within the group; cosine distance is invariant to a common positive isotropic scale and orthogonal transformation; and Spearman correlation is invariant to strictly monotone transformations of the distances. Thus, KTS evaluates preservation of the rank ordering of within-group pairwise distances, rather than equality of coordinates or absolute distances. Here, topology refers operationally to this within-group distance-rank structure.

Let $\mathcal{P}_r$ contain the condition pairs $p = (t, t')$ satisfying $|\mathcal{I}_{r,p}| \geq n_{\min}$, where $n_{\min} = 100$ in the primary analysis, and for which both distance vectors are non-degenerate. Sensitivity to the shared-item threshold is reported in appendix D.5. We first average over valid condition pairs and selected layers:

$$\mathrm{KTS}_{m,r} = \operatorname*{mean}_{\ell \in \mathcal{W}_m,\, p \in \mathcal{P}_r} \rho^{\ell}_{m,r}(p). \tag{10}$$

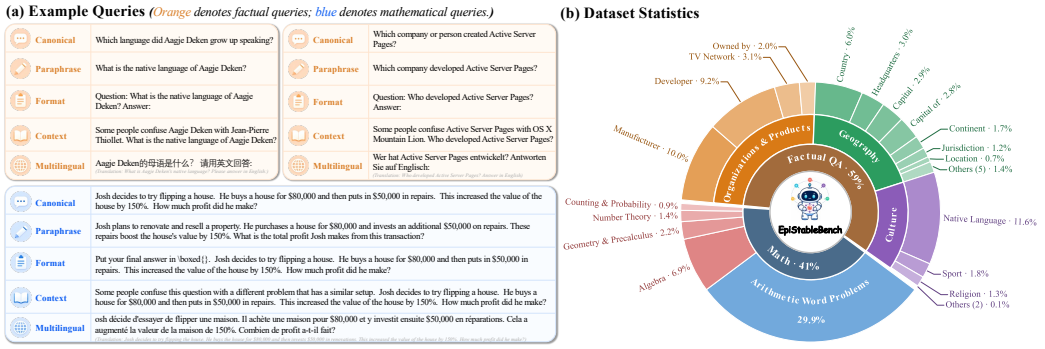


Figure 2: **Overview of EpiStableBench.** (a) Canonical questions and four controlled-query types varying paraphrase, format, context, and language for factual knowledge (orange) and mathematical reasoning (blue). (b) Distribution of 4,411 knowledge items across domains and categories.

Table 1: **Model-Family and Scaling Comparison.** ACC, LSS, and KTS are reported across open-weight LLM families and parameter scales for factual knowledge, mathematical reasoning, and their average. The best and second-best results are shown in **bold** and underlined. More results, please see the appendix.

		Factual Knowledge			*Mathematical Reasoning*			*Average*		
Family	Model	ACC ↑	LSS ↑	KTS ↑	ACC ↑	LSS ↑	KTS ↑	ACC ↑	LSS ↑	KTS ↑
Qwen	Qwen2.5-3B	0.766	0.355	0.795	0.625	0.596	0.874	0.696	0.475	0.834
	Qwen2.5-7B	0.814	0.407	0.848	0.699	0.624	0.888	0.756	0.516	0.868
	Qwen2.5-14B	0.825	0.445	0.847	0.744	0.645	0.892	0.784	0.545	0.869
	Qwen2.5-32B	0.866	0.394	0.828	0.756	0.565	0.881	**0.811**	0.480	0.855
OLMo	OLMo-2-1B	0.599	0.302	0.813	0.238	0.561	0.862	0.419	0.431	0.837
	OLMo-2-7B	0.832	0.455	0.866	0.485	0.629	0.880	0.659	0.542	0.873
	OLMo-2-13B	0.865	0.562	0.904	0.553	0.650	0.885	0.709	0.606	0.895
	OLMo-2-32B	**0.903**	0.440	0.856	0.607	0.624	0.880	0.755	0.532	0.868
Gemma	Gemma-3-1B	0.653	0.486	0.815	0.022	0.684	0.864	0.338	0.585	0.840
	Gemma-3-4B	0.801	0.559	0.894	0.307	**0.796**	0.914	0.554	0.677	0.904
	Gemma-3-12B	0.807	0.519	0.862	0.586	0.780	**0.915**	0.697	0.650	0.888
	Gemma-3-27B	0.808	0.586	0.892	0.623	0.783	0.911	0.716	**0.684**	0.901
Llama3	Llama3-1B	0.630	0.427	0.857	0.053	0.648	0.878	0.342	0.538	0.867
	Llama3-3B	0.779	0.546	0.887	0.218	0.688	0.883	0.499	0.617	0.885
	Llama3-8B	0.719	0.576	0.900	0.410	0.715	0.894	0.565	0.646	0.897
	Llama3-70B	0.785	**0.638**	**0.923**	**0.772**	0.693	0.902	0.778	0.666	**0.912**

Let $\mathcal{R}_\tau$ be the comparison groups in track τ. The overall score first macro-averages growups within each track and then gives the two tracks equal weight:

$$\mathrm{KTS}_m = \frac{1}{2} \sum_{\tau \in \{\mathrm{F},\mathrm{M}\}} \frac{1}{|\mathcal{R}_\tau|} \sum_{r \in \mathcal{R}_\tau} \mathrm{KTS}_{m,r} \,. \tag{11}$$

KTS lies in $[-1, 1]$ and is defined at the comparison-group level. Associating item i with KTS_{m,r_i} supplies group-level geometric context; it does not convert KTS into an item-level measurement.

4 Experiments And Analysis

Dataset. EpiStableBench treats the underlying knowledge target, rather than an individual question, as the unit of evaluation. Each target has a fixed answer and is elicited through the original, unmodified question and four types of controlled queries that vary question formulation, format, context, and language. These queries preserve the target knowledge while providing repeated internal observations across different elicitation conditions. EpiStableBench contains factual knowledge and mathematical reasoning tracks constructed from established benchmarks. Figure 2(a) presents representative targets and their controlled queries, while Figure 2(b) summarizes the benchmark composition. In total, EpiStableBench comprises 4,411 factual knowledge and mathematical reasoning targets and 76,766 controlled queries.

Protocol. We evaluate 16 models on EpiStableBench, covering four scales each from the Qwen2.5 (Yang et al., 2024), OLMo-2 (OLMo Team et al., 2025), Gemma-3 (Gemma Team et al., 2025), and Llama (Grattafiori et al., 2024) families. All models use BF16, identical plain-text in-

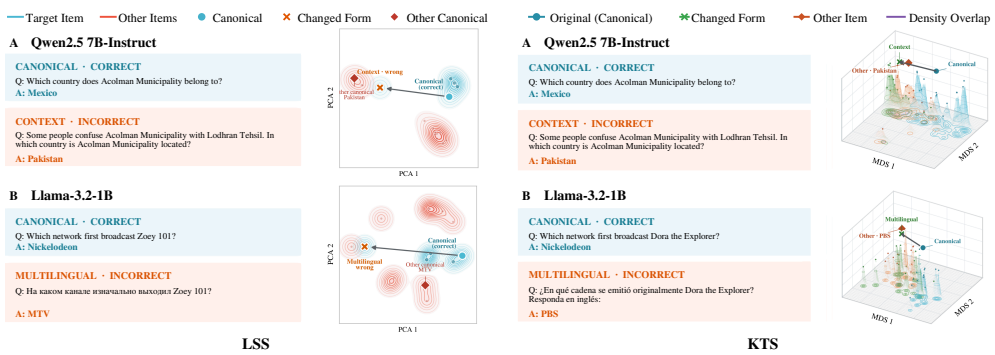


Figure 3: **Behavioral Errors Accompanied By Latent Shifts.** Items answered correctly under baseline elicitation condition produce competing answers under changed elicitation conditions, accompanied by representational shifts toward facts associated with those answers. More results please see the appendix.

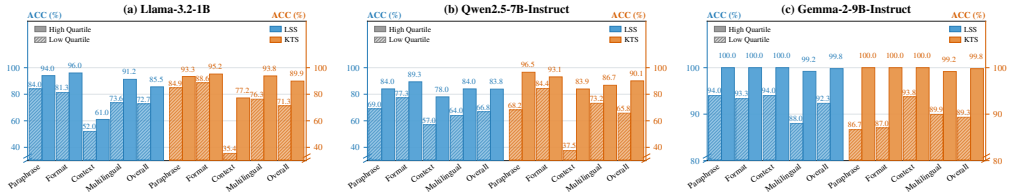


Figure 4: **Stability and Subsequent Correctness.** Among targets answered correctly under the base condition, high-LSS targets and high-KTS groups show higher ACC under held-out elicitation conditions than their low-stability counterparts.

puts without chat or system prompts, and one greedily decoded response per question; mathematical questions additionally use a fixed eight-shot chain-of-thought prefix. Factual answers are independently judged by Qwen2.5-72B-Instruct (Yang et al., 2024) and Mistral-Large-Instruct (Mistral AI Team, 2024), while mathematical answers are assessed using deterministic equivalence checks. LSS and KTS are aggregated over the final quarter of model layers. Accuracy (ACC) and LSS are computed per knowledge item, with ACC weighting controlled-query types equally. KTS is computed per relation or task type and macro-averaged across groups; response-level analyses assign each item its group's KTS. Prediction experiments use five-fold stratified splits grouped by knowledge item.

4.1 Model Rankings

Model Size, Accuracy, and Stability. We compare ACC, LSS, and KTS across four model families and multiple scales to assess whether larger, more accurate models have more stable knowledge representations. Table 1 shows that these metrics do not consistently increase together. From 14B to 32B, Qwen2.5's average ACC rises from 0.784 to 0.811, while LSS falls from 0.545 to 0.480 and KTS from 0.869 to 0.855; OLMo-2 shows a similar pattern from 13B to 32B. In mathematical reasoning, ACC increases from 0.625 to 0.756 for Qwen2.5 and from 0.238 to 0.607 for OLMo-2 without corresponding stability gains. Higher accuracy in larger models therefore does not necessarily imply greater stability of individual knowledge representations or their relational structure.

Qualitative Analysis. We examine items answered correctly under the base condition but incorrectly under controlled probes (Figure 3). Under context and multilingual probes, the answers for *Acolman Municipality* and *Zoey 101* change from Mexico to Pakistan and from Nickelodeon to MTV, respectively, accompanied by representational shifts toward competing items. The KTS examples also exhibit changes in within-group geometry. Together, these cases illustrate the item-level drift and relational changes captured by LSS and KTS, respectively. Because the cases were selected after observing errors, they provide qualitative rather than predictive evidence.

4.2 Validation of LSS and KTS

Stability and Subsequent Correctness. To test the correctness–stability hypothesis, we retain targets answered correctly under the base condition, defined by the original, unmodified question. We hold out each controlled-query type in turn for accuracy evaluation and compute LSS and KTS from the remaining controlled queries. Across the models and query types shown in Figure 4, high-LSS

targets and high-KTS groups consistently achieve higher ACC than their low-stability counterparts. For Qwen2.5-7B-Instruct, ACC aggregated across query types is 83.8% versus 66.3% for high- and low-LSS targets, and 90.1% versus 65.8% for high- and low-KTS groups. These results support the correctness-stability hypothesis: among initially correct targets, greater stability of internal knowledge representations is associated with higher accuracy under held-out elicitation conditions.

Semantic Sensitivity of Latent Knowledge Metrics. To assess whether LSS and KTS distinguish target knowledge changes from surface reformulations, we compare three perturbation types: *minimal edits*, which add only a prefix; *real variants*, which preserve the target knowledge; and *hard negatives*, which replace the target fact with a similar-length fact from the same evaluation group but with a different subject and answer. Across all three models, hard negatives produce substantially larger changes in both metrics than the two target-preserving perturbations (Figure 5). Changes under hard negatives are approximately 0.44–0.47 for LSS and 0.25–0.32 for KTS, compared with 0.12–0.27 and 0.03–0.11, respectively, under target-preserving perturbations. This separation supports the semantic sensitivity of LSS and KTS, showing stronger responses to changes in target knowledge than to surface reformulations.

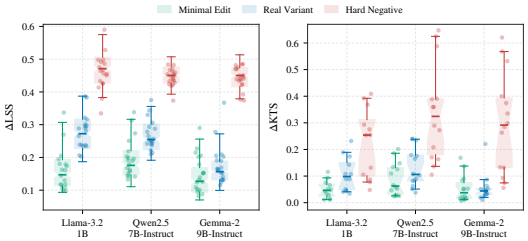


Figure 5: **Semantic Sensitivity of LSS and KTS.** Hard negatives induce larger metric changes than target-preserving perturbations across all three models.

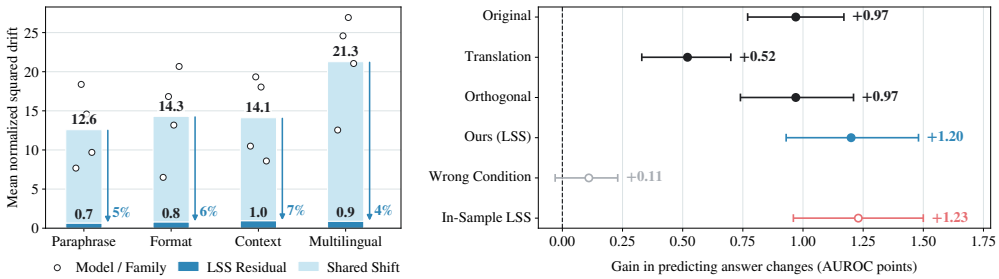


Figure 6: **Separating Shared Shifts from Item-Specific Drift.** (Left) Decomposition of normalized squared drift into the portion removed by alignment and the remaining residual. (Right) Gains in answer-change prediction across alignment variants, measured in AUROC points.

Separating Shared Shifts from Item-Specific Drift. To assess whether LSS separates item-specific drift from shared shifts, we fit condition-specific similarity maps on training items and align held-out representations to the base condition. Only 4.2%-6.8% of the original squared drift remains after alignment, indicating that shared condition-level transformations account for 93.2%-95.8% (Figure 6(a)). Adding the residual drift to an answer-change classifier controlling for model, relation, and condition improves AUROC by 1.20 points. By comparison, residuals from wrong-condition alignment yield only a 0.11-point gain, with a confidence interval including zero (Figure 6(b)). These results suggest that condition-matched alignment reduces shared shifts while retaining item-specific variation associated with answer changes.

Relation-Aware Processing. We assess relation-aware processing by measuring rank agreement between each ablation and the full method (Table 2). For LSS, removing normalization substantially changes model rankings, yielding a correlation of 0.23, whereas global normalization retains greater agreement at 0.84; item-level correlations are similar at 0.65 and 0.64, respectively. For KTS, removing relation centering largely preserves model and relation rankings (0.99 and 0.93), whereas evaluating topology globally reduces these correlations to 0.77 and 0.15. Thus, normalization affects LSS rankings, particularly across models, while relation-specific topology evaluation has a greater impact on KTS than relation centering.

Table 2: **Relation-Aware Processing.** Spearman correlations between ablated and full-method scores at the indicated aggregation levels. Labels before and after the slash denote LSS and KTS variants, respectively.

Setting	*LSS*		*KTS*	
	Model	Item	Model	Relation
No Norm. / Centering	0.23	**0.65**	**0.99**	**0.93**
Global Norm. / Structure	**0.84**	0.64	0.77	0.15

Representation and Layer Selection. We assess sensitivity to layer ranges and representation choices by comparing each variant with the full method (Table 3). Agreement increases with depth, with the overlapping 80%–100% interval yielding correlations of 0.96-1.00 across all evaluation levels. Using only the final layer or raw residual-stream states over the final 25% of layers yields model-level LSS correlations of 0.60 and 0.47, respectively. KTS model rankings remain more consistent under both variants, with correlations of 0.98 and 0.89. These results indicate consistency across nearby late-layer ranges, while highlighting the sensitivity of model-level LSS rankings to aggregation and representation choices.

Table 3: **Representation and Layer Selection.** Spearman correlations with the full method, which aggregates J-Lens representations over the final 25% of layers. Depth is normalized from input to output.

Setting	*LSS*		*KTS*	
	Model	Item	Model	Relation
Depth 40%–60%	0.14	0.42	-0.10	0.59
Depth 60%–80%	0.68	0.77	0.53	0.85
Depth 80%–100%	**0.96**	**1.00**	**0.99**	**1.00**
Final Layer Only	0.60	0.84	0.98	0.95
Raw States, Final 25%	0.47	0.82	0.89	0.80

Probe and Reference Sensitivity. We assess sensitivity to elicitation-condition families and reference choices through leave-one-family-out and alternative-reference analyses (Table 4). All correlations remain at least 0.90 when any single family is omitted, indicating consistent rankings across evaluation levels. Using expression or format as the reference yields item-level LSS correlations of 0.77 and 0.71, respectively, while relation-level KTS correlations remain at 0.96 and 0.93. Thus, rankings are largely stable under omission of individual condition families, whereas item-level LSS is more sensitive to the reference condition.

Table 4: **Probe and Reference Sensitivity.** Spearman correlations with the full method after omitting each elicitation-condition family in turn or changing the reference condition.

Setting	*LSS*		*KTS*	
	Model	Item	Model	Relation
Leave-one-probe-out				
Drop Paraphrase	**0.98**	**0.97**	**0.99**	**0.97**
Drop Format	0.96	0.95	0.97	0.96
Drop Context	0.95	0.95	0.92	0.91
Drop Multilingual	0.95	0.95	0.93	0.90
Alternative Reference				
Expression Reference	**0.96**	**0.77**	**0.98**	**0.96**
Format Reference	0.89	0.71	0.86	0.93

4.3 Deeper Analysis

Visualizing Internal Knowledge Stability. We visualize representative knowledge items and relations with high and low stability (Figure 7). For LSS, corrected representations of high-scoring items remain close to their references, whereas those of low-scoring items exhibit greater drift toward other items. For KTS, high-scoring relations retain more consistent within-group geometry across conditions, whereas low-scoring relations exhibit greater rearrangement. These patterns qualitatively illustrate the item-level drift and relational changes captured by LSS and KTS, respectively.

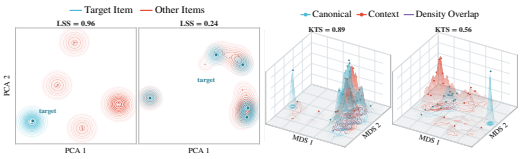


Figure 7: **Geometric Patterns of Internal Knowledge Stability.** Representative high- and low-LSS items and high- and low-KTS relations are compared across the base and controlled elicitation conditions.

Layer-Specific Activation Patching. We test whether late-layer states contain item-specific information that supports target-answer recovery. At each tested layer, we replace a controlled query's final-input-token state with the canonical state from either the same item or a different item sharing the same relation. The control preserves the relation type while changing item identity, helping distinguish target-specific information from signals shared within a relation. Figure 8 shows that same-item patches increasingly outperform controls with depth. Their respective recovery rates are 0.38/0.36 at 40%–50% depth, 0.64/0.34 at 50%–75%, and 0.87/0.07 in the final quarter, widening the recovery gap from 0.02 to 0.80. This experiment complements the earlier correlational analyses by directly testing the functional contribution of internal states to answer recovery. These

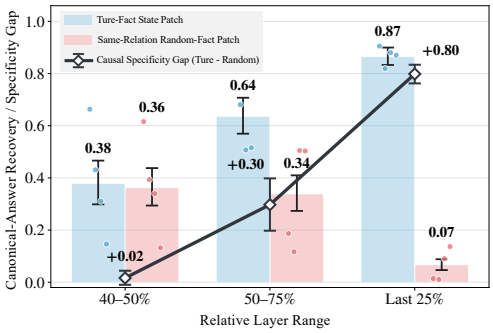


Figure 8: **Layer-Specific Activation Patching.** Target-answer recovery using canonical states from the same item versus different-item controls within the same relation.

interventions indicate that late-layer states carry target-specific information useful for answer recovery, supporting the final-quarter layer selection for LSS and KTS.

Family-Specific Sensitivity to Elicitation. We compare changes in ACC, LSS, and KTS relative to the base condition across model families and elicitation conditions (Figure 9). These metrics capture output correctness, individual-state stability, and relational-structure stability, respectively. The observed patterns vary across families and conditions, with accuracy and stability changes not always occurring together. Multilingual elicitation produces large ACC drops and LSS drift across multiple families, whereas some other conditions show limited ACC changes despite noticeable LSS or KTS drift. These observations indicate that accuracy alone can overlook changes in internal knowledge representations, while LSS and KTS reveal complementary variation at the item and relation levels.

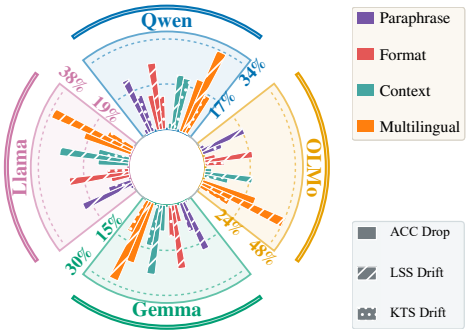


Figure 9: **Family-Specific Sensitivity to Elicitation.** Condition-wise changes in ACC, LSS, and KTS relative to the base condition across four model families.

Table 5: **Application to Knowledge Editing.** Accuracy, Generality, and Specificity (%) under identical edit sets and evaluation protocols. Parentheses report percentage-point changes relative to the baseline.

Method	*GPT2-XL*			*LLaMA 2-7B*		
	Accuracy ↑	Generality ↑	Specificity ↑	Accuracy ↑	Generality ↑	Specificity ↑
ROME	99.40	73.67	**80.61**	**99.90**	67.37	**93.87**
ROME + Ours	**99.65** (+0.25)	**80.48** (+6.81)	80.27 (-0.34)	**99.90** (+0.00)	**75.75** (+8.38)	93.33 (-0.54)
MEMIT	72.37	38.89	**87.57**	76.08	54.23	**74.40**
MEMIT + Ours	**87.14** (+14.77)	**54.98** (+16.09)	86.94 (-0.63)	**90.79** (+14.71)	**69.17** (+14.94)	73.52 (-0.88)
SAKE	94.59	83.96	**84.20**	96.70	80.73	**84.67**
SAKE + Ours	**96.80** (+2.21)	**91.17** (+7.21)	83.07 (-1.13)	**97.70** (+1.00)	**88.44** (+7.71)	83.71 (-0.96)

4.4 Application

We evaluate stability-aware policy training with ROME (Meng et al., 2022), MEMIT (Meng et al., 2023), and SAKE (Scialanga et al., 2025) on GPT2-XL (Radford et al., 2019) and LLaMA 2-7B (Touvron et al., 2023), using matched baselines, identical edit sets, and the same evaluation protocol. Table 5 shows improved Generality and maintained or improved Accuracy across all settings. ROME retains near-perfect Accuracy while increasing Generality by 6.81–8.38 percentage points. MEMIT achieves the largest gains, improving Accuracy by 14.71–14.77 points and Generality by 14.94–16.09 points. SAKE improves Accuracy by 1.00–2.21 points and Generality by 7.21–7.71 points. These gains accompany Specificity reductions of 0.34–1.13 points, demonstrating consistent editing benefits across editors and backbones with modest locality trade-offs. Further implementation and evaluation details are provided in the appendix.

5 Conclusion

We introduce EpiStableBench and propose LSS and KTS to evaluate latent knowledge stability at the individual-state and relational-structure levels across elicitation conditions. Experiments on 16 models from four families reveal substantial stability variation even among items answered correctly under the base condition. Within models, higher stability generally corresponds to higher accuracy under held-out conditions; across models, stability does not consistently increase with model size or overall accuracy. Perturbation and alignment analyses support the metrics' interpretation, while activation patching provides evidence for the functional relevance of late-layer states to target-answer recovery. Stability-aware policy training further improves editing Generality while maintaining or improving Accuracy across three editors and two backbones, with modest reductions in Specificity. These findings support evaluating internal representational stability alongside output correctness to better understand the knowledge supporting model answers.

AI DISCLOSURE

In this work, we used generative AI tools to edit and polish the manuscript, including improving its grammar, clarity, and readability. We also used generative AI tools to assist with code development. All AI-assisted textual revisions were reviewed by the authors to ensure that they preserved the intended meaning and accurately reflected the research methods and results. AI-assisted code was reviewed and verified for correctness by multiple authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses publicly available datasets and open-weight models and does not collect private or sensitive personal information. Human review is limited to validating the semantic consistency and correctness of controlled queries. We acknowledge potential biases in the source datasets, evaluated models, and automated evaluators. The knowledge-editing experiments are conducted offline for research purposes; deployment should involve appropriate validation and human oversight.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the dataset construction and human-validation procedures, model checkpoints, query templates, representation extraction, metric computation, cross-fitting protocol, answer evaluation, and statistical analyses in the appendix. All reported experiments use fixed configurations and item-level splits that keep queries associated with the same knowledge item together. We will release EpiStableBench, the evaluation code, and the configurations required to reproduce the main results.

REFERENCES

Arash Bellafard, Ghazal Namvar, Jonathan C Kao, Alipasha Vaziri, and Peyman Golshani. Volatile working memory representations crystallize with practice. *Nature*, 629(8014):1109–1117, 2024.

Savita Bhat and Vasudeva Varma. All prompts are created equal? evaluating robustness of llm judges against non-adversarial prompt variations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38730–38745, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Hristos S Courellis, Juri Minxha, Araceli R Cardenas, Daniel L Kimmel, Chrystal M Reed, Taufik A Valiante, C Daniel Salzman, Adam N Mamelak, Stefano Fusi, and Ueli Rutishauser. Abstract representations emerge in human hippocampal neurons during inference. *Nature*, 632(8026): 841–849, 2024.

Jingtong Dou, Chuancheng Shi, Yemin Wang, Shiming Guo, Anqi Yi, Wenhua Wu, Li Zhang, Fei Shen, and Tat-Seng Chua. Dna: Uncovering universal latent forgery knowledge. *arXiv preprint arXiv:2601.22515*, 2026.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Kazem Faghih, Yize Cheng, Shoumik Saha, Mobina Pournemat, Armin Gerami, and Soheil Feizi. Same question, different answers: Evaluating llm reliability beyond accuracy. *arXiv preprint arXiv:2607.22554*, 2026.

Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. Inside-out: Hidden factual knowledge in llms. *arXiv preprint arXiv:2503.15299*, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, et al. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025. doi: 10.48550/arXiv.2503.19786. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. Olmes: A standard for language model evaluations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5020–5048, 2025.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, et al. Verbalizable representations form a global workspace in language models. *arXiv preprint arXiv:2607.15495*, 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, and Yao Qin. Flaw or artifact? rethinking prompt sensitivity in evaluating llms. In *Proceedings of the 2025 conference on empirical methods in natural language processing*, pp. 19900–19910, 2025.

Maxim Ifergan, Leshem Choshen, Roee Aharoni, Idan Szpektor, and Omri Abend. Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in llms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4630–4644, 2025.

Nurit Cohen Inger, Yehonatan Elisha, Bracha Shapira, Lior Rokach, and Seffi Cohen. Forget what you know about llms evaluations-llms are like a chameleon. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21675–21688, 2025.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint*, 2022.

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.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Aaron Meurer, Christopher P Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B Kirpichev, Matthew Rocklin, Amit Kumar, Sergiu Ivanov, Jason K Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E Granger, Richard P Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J Curry, Andy R Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. SymPy: Symbolic computing in Python. *PeerJ Computer Science*, (3):e103, 2017. doi: 10.7287/peerj.preprints.2083v3. URL https://inria.hal.science/hal-01404156.

Mistral AI. Mistral-large-instruct-2407. Hugging Face model card, 2024. URL https://huggingface.co/mistralai/Mistral-Large-Instruct-2407.

Mistral AI Team. Mistral-Large-Instruct-2411 model card. Hugging Face model card, 2024. URL https://huggingface.co/mistralai/Mistral-Large-Instruct-2411.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

OLMo Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, et al. 2 OLMo 2 Furious. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=2ezugTT9kU.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025.

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)*, pp. 2463–2473, 2019.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Marco Scialanga, Thibault Laugel, Vincent Grari, and Marcin Detyniecki. SAKE: Steering activations for knowledge editing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15966–15978. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.777. URL https://aclanthology.org/2025.acl-long.777/.

Chuancheng Shi, Shangze Li, Wenjun Lu, Wenhua Wu, Cong Wang, Zifeng Cheng, Fei Shen, and Tat-Seng Chua. Tracerouter: robust safety for large foundation models via path-level intervention. *arXiv preprint arXiv:2601.21900*, 2026.

Yucheng Sun, Alessandro Stolfo, and Mrinmaya Sachan. Probing for arithmetic errors in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8122–8139, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. doi: 10.48550/arXiv.2307.09288. URL https://arxiv.org/abs/2307.09288.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5075–5094, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Raoyuan Zhao, Abdullatif Köksal, Ali Modarressi, Michael A Hedderich, and Hinrich Schütze. Do we know what llms don't know? a study of consistency in knowledge probing. In *EMNLP (Findings)*, pp. 23254–23280, 2025.

APPENDIX

The appendix is organized as follows. Appendix A summarizes the notation used throughout the paper. Appendix B describes the construction and human validation of EpiStableBench. Appendix C provides the detailed experimental setup. Appendix D presents additional experiments, qualitative analyses of LSS and KTS, and quantitative comparisons across a broader set of model families. Appendix E details the stability-aware policy training framework and its application to knowledge editing. Appendix F presents the theoretical derivation, properties, and scope of LSS and KTS. Finally, Appendices G and H discuss the limitations and additional questions concerning the interpretation and design choices of the proposed metrics.

A NOTATION SUMMARY

Table 6 summarizes the notation used to estimate whether the internal representations of the same knowledge remain stable across elicitation conditions. A superscript ℓ indexes a model layer and does not denote exponentiation.

Table 6: Core Notation for Latent State Stability (LSS) and Knowledge Topology Stability (KTS).

Symbol	Definition
m, ℓ, i, q	Model, layer, knowledge item, and query, respectively.
$\mathcal{I}, \mathcal{I}_{\mathrm{F}}, \mathcal{I}_{\mathrm{M}}$	All knowledge items and the factual-knowledge and mathematical-reasoning subsets, respectively.
$k_i = (x_i, r_i, y_i)$	Knowledge item i, where x_i is the queried content, r_i is its benchmark-defined comparison group, and y_i is the target answer.
$\tau_i \in \{\mathrm{F}, \mathrm{M}\}$	Track containing item i, where F and M denote the factual-knowledge and mathematical-reasoning tracks, respectively.
$a, \mathcal{T}, \mathcal{T}_i$	Reference elicitation condition a, elicitation-condition families $\mathcal{T}$, and the conditions $\mathcal{T}_i \subseteq \mathcal{T}$ available for item i.
$\mathcal{Q}_{i,t}$	Set of target-preserving queries for item i under elicitation condition t.
$p = (t, t'),\ u$	Pair of elicitation conditions and a condition $u \in \{t, t'\}$ within that pair.
$\mathcal{I}_r$	Set of knowledge items in comparison group r, $\mathcal{I}_r = \{j : r_j = r\}$.
$\mathcal{I}^{\mathrm{tr}}_{\tau,t}$	Cross-fitting training items used to estimate the similarity map for track τ and condition t, excluding the evaluated item.
$\mathcal{I}_{r,p}$	Items in comparison group r observed under both conditions in $p = (t, t')$.
$\mathcal{W}_m$	Selected late-layer window for model m (the final quarter of layers in the primary analysis).
$\mathcal{P}_r, \mathcal{R}_\tau$	Valid condition pairs for comparison group r, and comparison groups in track τ, respectively.
$h^{\ell}_{m,i,q}$	Residual-stream state at the final input token, immediately before generation of the first answer token.
$J^{\ell}_m, z^{\ell}_{m,i,q}$	J-Lens transport to the model-specific common comparison space, and the transported query representation.
$\bar{v}^{\ell}_{m,i,t}$	Mean representation obtained by averaging the transported query representations in $\mathcal{Q}_{i,t}$.
Sim	Similarity-map family $A(v) = sRv + b$, where $s > 0$ and $R^\top R = I$.
$A^{\ell}_{m,\tau,t}$	Cross-fitted similarity map from elicitation condition t to the reference elicitation condition a for model m, layer ℓ, and track τ.
$e^{\ell}_{m,i,t}$	Item-specific residual between the aligned representation under condition t and the reference representation under anchor a.
$\beta^{\ell}_{m,r}$	Median reference-space distance between distinct knowledge items in comparison group r.
ϵ	Positive numerical stabilizer in the LSS normalization.
$d^{\ell}_{m,i,t}$	Residual displacement normalized by the typical reference-space separation in comparison group r_i.
$D_{m,i}$	Normalized residual drift aggregated over the available non-reference elicitation conditions and selected layers.
$\mathrm{LSS}_{m,i}$	Item-level latent state stability in $(0, 1]$; higher values indicate that the corrected representation remains closer to its reference representation relative to the local reference scale.
$\widetilde{v}^{\ell}_{m,i,u;p}$	Representation of item i under condition $u \in \{t, t'\}$, centered using the common item set $\mathcal{I}_{r,p}$.
$\Delta^{\ell}_{m,u,r;p}(i, j)$	Centered cosine distance between items i and j in $\mathcal{I}_{r,p}$.
$\boldsymbol{\delta}^{\ell}_{m,u,r;p}$	Vector of centered cosine distances over all unordered item pairs in $\mathcal{I}_{r,p}$.
$\rho^{\ell}_{m,r}(p)$	Spearman correlation between the distance vectors under the two conditions in p.
$\mathrm{KTS}_{m,r}$	Comparison-group-level knowledge topology stability, averaged over valid condition pairs and selected layers.
KTS_m	Overall KTS, obtained by macro-averaging comparison groups within each track and assigning equal weight to the two tracks.

B EPISTABLEBENCH CONSTRUCTION

Data Sources and Item Definition. EpiStableBench contains factual-knowledge and mathematical-reasoning tracks constructed through a shared pipeline. The factual track draws from CounterFact (Meng et al., 2022), LAMA (Petroni et al., 2019), T-REx (Elsahar et al., 2018), Wikidata (Vrandečić & Krötzsch, 2014), and PopQA (Mallen et al., 2022). The mathematical track

draws from GSM8K (Cobbe et al., 2021) and MATH-500 (Lightman et al., 2024). We collect factual triples and mathematical problems and apply track-specific normalization and filtering to retain items with unambiguous target answers.

Both tracks represent a knowledge item as a triple $k = (s, r, o)$. For factual knowledge, s is the subject, r is the knowledge relation, and o is the target answer. For mathematical reasoning, s is the problem, r is the reasoning-task type, and o is the correct result. This shared representation allows factual relations and mathematical task types to serve as the corresponding evaluation groups in LSS and KTS analyses.

Canonical Questions and Controlled Probes. For each knowledge item, we construct a canonical question and four types of controlled probes varying paraphrase, format, context, and multilingual presentation. All questions associated with an item preserve its target knowledge and correct answer. The controlled probes therefore provide repeated observations of how the same item is represented and used under different forms of access.

For example, the factual item "Aagje Deken–native language–Dutch" is represented as $k = (\text{Aagje Deken}, \text{native language}, \text{Dutch})$. Its canonical question and controlled probes use different presentations but continue to examine the same relation and require the answer "Dutch." For a mathematical item, the presentation may change while the underlying problem, reasoning target, and correct result remain fixed.

Benchmark Composition. The resulting benchmark contains 2,592 factual items from 21 relations and 1,819 mathematical problems. Across the two tracks, EpiStableBench contains 4,411 knowledge items and 76,766 queries. All evaluated models receive the same items, canonical questions, and controlled probes.

Probe Generation and Human Validation. Multilingual probes were generated using ChatGPT-5.6 Sol. To ensure semantic preservation in EpiStableBench, 20 human reviewers manually screened the dataset before evaluation, each reviewing a separate subset. Reviewers checked whether rewritten questions preserved the original relation, task requirements, and intended answer. Questions containing ambiguity, semantic drift, answer changes, or other inconsistencies were revised or removed to reduce potential confounds in the evaluation of knowledge stability.

C Detailed Experimental Setup

Model Checkpoints and Hardware. We evaluate 16 LLMs in the main text. Qwen2.5 (Qwen et al., 2025) at 3B, 7B, 14B, and 32B; OLMo-2 (OLMo et al., 2024) at 1B, 7B, 13B, and 32B; Gemma-3 (Team et al., 2025) at 1B, 4B, 12B, and 27B; and Llama (Grattafiori et al., 2024) at 1B, 3B, 8B, and 70B. All checkpoints are loaded in BF16 without quantization. Experiments are conducted on four NVIDIA H200 GPUs.

Evaluation Data. EpiStableBench contains 4,411 knowledge items and 76,766 queries. The factual track contains 2,592 items from 21 relations, and the mathematical track contains 1,819 items. Each item is elicited through canonical queries and controlled queries varying paraphrase, format, context, or language. All models are evaluated on the same items and queries.

Input and Generation Configuration. All models receive the same task-specific plain-text inputs without chat templates or system messages. Factual queries request the shortest correct answer, whereas mathematical queries use a fixed eight-shot chain-of-thought prefix and request a step-by-step solution. Each question is answered once using greedy decoding with `do_sample=False` and one beam. The maximum generation length is 24 tokens for factual queries, 512 tokens for mathematical queries, and 640 tokens for multilingual mathematical queries.

Latent-State Extraction. For each question and selected layer, we extract the residual-stream state at the final non-padding input token after the complete question has been processed and before the first answer token is generated. Layer depth is normalized as $d_\ell = \ell/(L-1)$. Primary results are aggregated over the final-quarter layers with $d_\ell \geq 0.75$, while layers with $0.50 \leq d_\ell < 0.75$ are examined in the sensitivity analysis.

J-Lens and Cross-Fitted Alignment. J-Lens (Gurnee et al., 2026) estimates a rank-256 transport from each selected layer to the final representation space using randomized range finding. The

transport is calibrated on 128 length-128 sequences sampled from `NeelNanda/pile-10k`. LSS alignment uses five-fold cross-fitting stratified by factual relation or mathematical task type. For an item f in fold $\mathcal{F}_j$, the Procrustes parameters used to compute its residual drift are estimated only from items outside $\mathcal{F}_j$. All queries associated with the same knowledge item remain in the same fold.

Evaluation Groups and KTS Notation. Throughout the metric definitions, r_f denotes the evaluation group of item f: a factual relation for factual items and a task type for mathematical items. KTS is defined at this group level rather than at the individual-item level. In analyses relating KTS to item-level behavior, each item is associated with the KTS of its corresponding group.

For KTS, relation centering is performed separately for every condition pair $p = (t, t')$. Accordingly, the centered representation and pairwise distance are formally written as

$$\tilde{v}^{\ell}_{m,f,u;p} = \bar{v}^{\ell}_{m,f,u} - \frac{1}{|\mathcal{D}_{r,p}|} \sum_{g \in \mathcal{D}_{r,p}} \bar{v}^{\ell}_{m,g,u}, \qquad u \in \{t, t'\}, \tag{12}$$

and

$$\Delta^{\ell}_{m,u,r;p}(f, g) = 1 - \cos\big(\tilde{v}^{\ell}_{m,f,u;p}, \tilde{v}^{\ell}_{m,g,u;p}\big). \tag{13}$$

Here $\mathcal{D}_{r,p}$ contains the items in group r available under both conditions in p. The condition-pair index p is suppressed in the main-text notation for readability.

Answer Evaluation. Factual answers are evaluated independently using Qwen2.5-72B-Instruct (Qwen et al., 2025) and Mistral-Large-Instruct (Mistral AI, 2024). Both judges receive the generated response, target answer, and accepted aliases and determine whether the response expresses the target answer. Mathematical answers are evaluated by extracting the final answer and applying normalized string matching, numerical equivalence with a relative tolerance of 10^{-6}, and symbolic equivalence with SymPy (Meurer et al., 2017). Outputs not classified as correct are counted as incorrect in ACC. Disagreements between the two LLM judges are very rare. We assess sensitivity using both a strict AND rule, requiring both judges to label a response as correct, and an OR rule, requiring at least one judge to do so. All disagreement cases are independently reviewed by three human annotators, blinded to the evaluated model, LSS/KTS scores, and both judges' decisions. The human majority vote determines the final adjudicated label.

Aggregation and Statistical Analysis. The knowledge item is the statistical unit for ACC and LSS, whereas the factual relation or mathematical task type is the statistical unit for KTS. ACC is first averaged within each available probe type and then equally across probe types for each item. LSS is averaged across items within each track, while KTS is macro-averaged across evaluation groups within each track.

Let $\mathcal{C}_m$ denote the available tracks, $\mathcal{F}_{m,c}$ the items in track c, and $\mathcal{R}_{m,c}$ its relations or task types. Track- and model-level scores are defined as

$$\mathrm{LSS}_{m,c} = \frac{1}{|\mathcal{F}_{m,c}|} \sum_{f \in \mathcal{F}_{m,c}} \mathrm{LSS}_{m,f}, \qquad \mathrm{KTS}_{m,c} = \frac{1}{|\mathcal{R}_{m,c}|} \sum_{r \in \mathcal{R}_{m,c}} \mathrm{KTS}_{m,r}, \tag{14}$$

$$\mathrm{LSS}_{m} = \frac{1}{|\mathcal{C}_{m}|} \sum_{c \in \mathcal{C}_{m}} \mathrm{LSS}_{m,c}, \qquad \mathrm{KTS}_{m} = \frac{1}{|\mathcal{C}_{m}|} \sum_{c \in \mathcal{C}_{m}} \mathrm{KTS}_{m,c}. \tag{15}$$

Thus, factual knowledge and mathematical reasoning receive equal weight in model-level comparisons. In item-level analyses, KTS_{m,r_f} is used as a group-level structural feature and is not interpreted as an item-specific KTS score.

We construct 95% percentile confidence intervals using 1,000 relation- or task-type-stratified bootstrap replicates. Quartile comparisons are performed over items for LSS and over relations or task types for KTS. Predictive experiments use five-fold stratified, item-grouped splits, ensuring that observations of the same knowledge item never appear in both training and test sets.

Activation Patching for Layer Selection. To examine whether late-layer representations contain more item-specific knowledge, we conduct an activation-patching experiment. For each controlled probe, we replace the native residual-stream state at the final input-token position of a decoder block with the corresponding state from the same item's canonical question, using a different same-relation item as a control. Each run patches one block independently, once during prefill before

answer generation, across sampled layers from approximately 40% depth to the final layer. The results show that same-item patching becomes more effective in later layers, while control patching weakens. Specifically, the recovery-rate gap increases from 0.02 in the 40%–50% depth interval to 0.80 in the final quarter. Therefore, later-layer states more selectively support recovery of the target answer, providing intervention-based evidence for using final-quarter layers in LSS and KTS.

D More Experiments

D.1 Sensitivity to Centroid Cancellation

Because condition-level centroids could in principle obscure large but opposing deviations across individual queries, we examined this possibility among items answered correctly under the reference elicitation condition. Using an independent reference panel calibrated separately by comparison group and condition, we found that severe masking events, in which individual-query representations deviated substantially while the corresponding centroid remained stable, were rare. Such events accounted for 0.289% of query-condition observations for Qwen2.5-3B (95% CI: 0.141%–0.462%) and 0.116% for Qwen2.5-7B (95% CI: 0.032%–0.253%); at the item level, 1.155% and 0.463% of items, respectively, exhibited such an event under at least one condition. Under a stricter criterion requiring the centroid deviation to fall below the 25th percentile of its reference distribution, the corresponding rates decreased to 0.030% for Qwen2.5-3B and zero observed cases for Qwen2.5-7B. These results indicate that severe centroid cancellation is uncommon and does not constitute a dominant source of distortion in the centroid-based metrics. Nevertheless, centroid and tail behavior are not interchangeable: among observations whose individual-query deviations entered the top decile, 34.9% for Qwen2.5-3B and 21.7% for Qwen2.5-7B did not enter the top decile according to centroid deviation. We therefore additionally report tail-based results to characterize variation not captured by condition-level centroids.

D.2 Comparison with Standard Representation Measures

To examine whether LSS and KTS provide predictive value beyond output features, we compare them with standard representation measures. Table 7 shows that adding either metric improves failure-prediction AUROC under held-out access conditions, with their combination achieving the highest scores. Specifically, LSS/KTS reaches 0.6154 and 0.7333 on 3B and 7B, exceeding the strong output baseline by 1.61 and 1.83 percentage points and within-relation CKA by 1.21 and 1.67 points, respectively. It also outperforms the combined standard features. Therefore, these results support the complementary predictive value of internal stability, with joint item- and relation-level measurements yielding the strongest performance.

Table 7: **Comparison With Standard Representation Measures.** All representation features augment the same strong output baseline. Values report failure-prediction AUROC.

Predictor	Qwen2.5-3B	Qwen2.5-7B
Strong Output Baseline	0.5993	0.7150
+ Within-relation CKA	0.6033	0.7166
+ All Standard Features	0.6030	0.7137
+ LSS	0.6102	0.7169
+ KTS	0.6149	0.7262
+ LSS/KTS	**0.6154**	**0.7333**

D.3 Sensitivity to J-Lens Construction and Representation Extraction

To examine whether LSS depends on specific implementation choices, we conduct ablations over the J-Lens calibration corpus, transport rank, representation space, and token extraction position. Table 8 shows that LSS is highly robust to choices within the J-Lens construction, while remaining specific to its defined representation space and extraction position. Specifically, changing the calibration corpus yields rank correlations of 0.964–0.990. Varying the J-Lens transport rank

among $\{64, 128, 256, 512\}$ yields correlations of 0.903–0.991. Replacing J-Lens with raw or PCA-projected representations and replacing the final-input-token state with question-level pooling result in lower correlations of 0.373–0.780 and 0.161–0.561, respectively. Therefore, the LSS results are not driven by a particular J-Lens calibration corpus or transport rank, while J-Lens alignment and the pre-specified final-input-token state constitute important components of the metric definition.

Table 8: **Sensitivity to J-Lens and Representation Choices.** LSS rank correlations under changes to the calibration corpus, transport rank, representation space, and token extraction position.

Ablation	LSS Rank Correlation
Calibration Corpus	0.964–0.990
J-Lens Transport Rank	0.903–0.991
Raw/PCA Representations	0.373–0.780
Question-level Pooling	0.161–0.561

D.4 Reliability of KTS Across Relation Sizes

To examine whether KTS is sensitive to relation size and small-sample effects, we subsample 10, 20, 50, 100, and 200 shared items from each relation and recompute KTS at both the relation and condition-pair levels. Figure 10 shows that KTS reliability improves consistently with the number of shared items. Specifically, the proportion of relation-level cells satisfying the reliability criterion increases from 0.0% and 1.6% at 10 and 20 items to 46.9%, 85.9%, and 100.0% at 50, 100, and 200 items, respectively. Condition-pair estimates require more observations, increasing from 0.2% and 3.9% to 25.2%, 59.2%, and 94.1% over the same sample sizes. Therefore, approximately 100 shared items are sufficient for exploratory relation-level analysis, whereas 200 shared items provide a more reliable basis for fine-grained condition-pair comparisons, with both levels reaching or approaching the 90% reliability target.

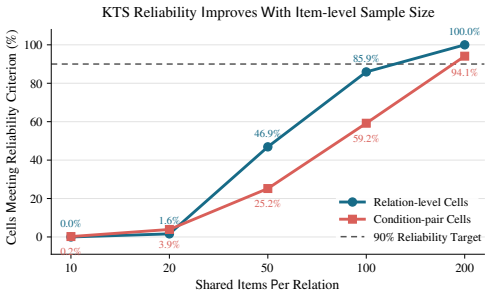


Figure 10: **KTS Reliability Across Relation Sizes.** Reliability increases with the number of shared items, with both relation-level and condition-pair estimates exceeding the 90% target at 200 items.

Table 9: **Sensitivity to Numerical and Aggregation Choices.** Spearman correlations and maximum rank shifts are measured against the default settings across 16 models in the audited factual cohort. The LSS constant row summarizes all seven tested settings; other rows concern the full method.

Metric	Setting Change	Rank ρ	Max. Shift
KTS	Alternative Tie Conventions	1.000	0
	Shared-item Threshold: 100 $\rightarrow$ 10	0.985	2
	Item-weighted Relations	0.974	2
LSS	Additive Constant: 0–10^{-3}	1.000	0
	Item-weighted Relations	0.988	2

D.5 Sensitivity to Numerical and Aggregation Choices

To examine whether model rankings depend on numerical and aggregation choices, we conduct sensitivity analyses of tie handling, shared-item thresholds, the additive LSS denominator constant, and relation weighting. Table 9 shows that rankings remain broadly stable within the audited factual cohort. Specifically, alternative tie conventions leave full-method KTS rankings unchanged; increasing the minimum shared-item threshold from 100 to 10 preserves a rank correlation of 0.985; LSS constants from 0 to 10^{-3} leave rankings unchanged across all tested settings; and item-weighted aggregation yields rank correlations of 0.988 and 0.974 for full-method LSS and KTS, respectively. Therefore, these results support ranking robustness to the tested numerical, eligibility, and weighting choices.

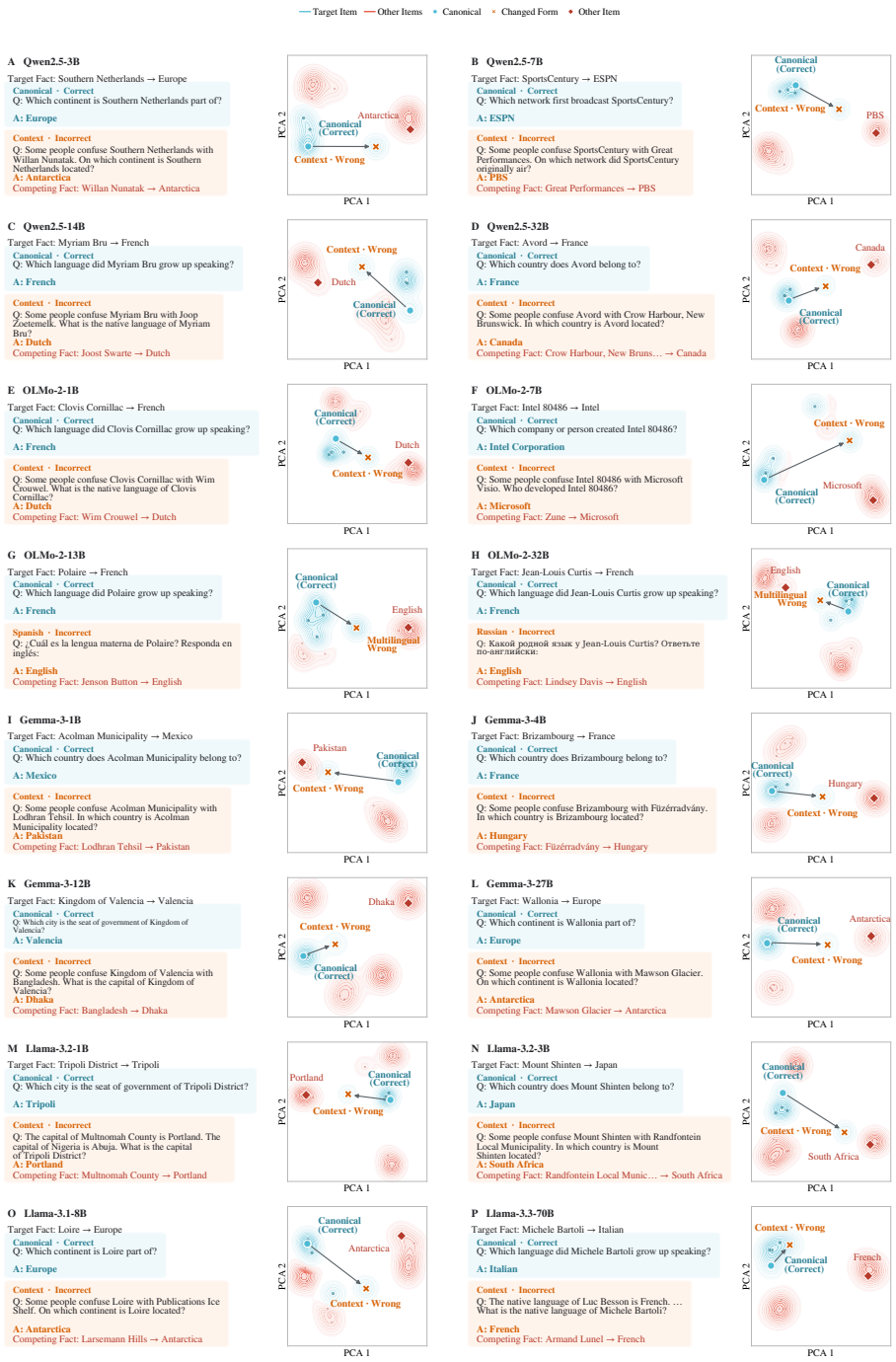


Figure 11: **More Qualitative Analysis of LSS.**

D.6 More Qualitative Analysis of LSS

Figure 11 extends the qualitative analysis of LSS across all 16 models. Each selected item is answered correctly under the canonical elicitation condition but incorrectly under a context or multilingual condition. Across these cases, the changed-form representation moves away from the canonical target representation and toward a region occupied by items associated with the competing answer. This recurring pattern illustrates the disruption of knowledge-item identity captured by LSS and shows that such item-specific drift occurs across model families and parameter scales. Because the

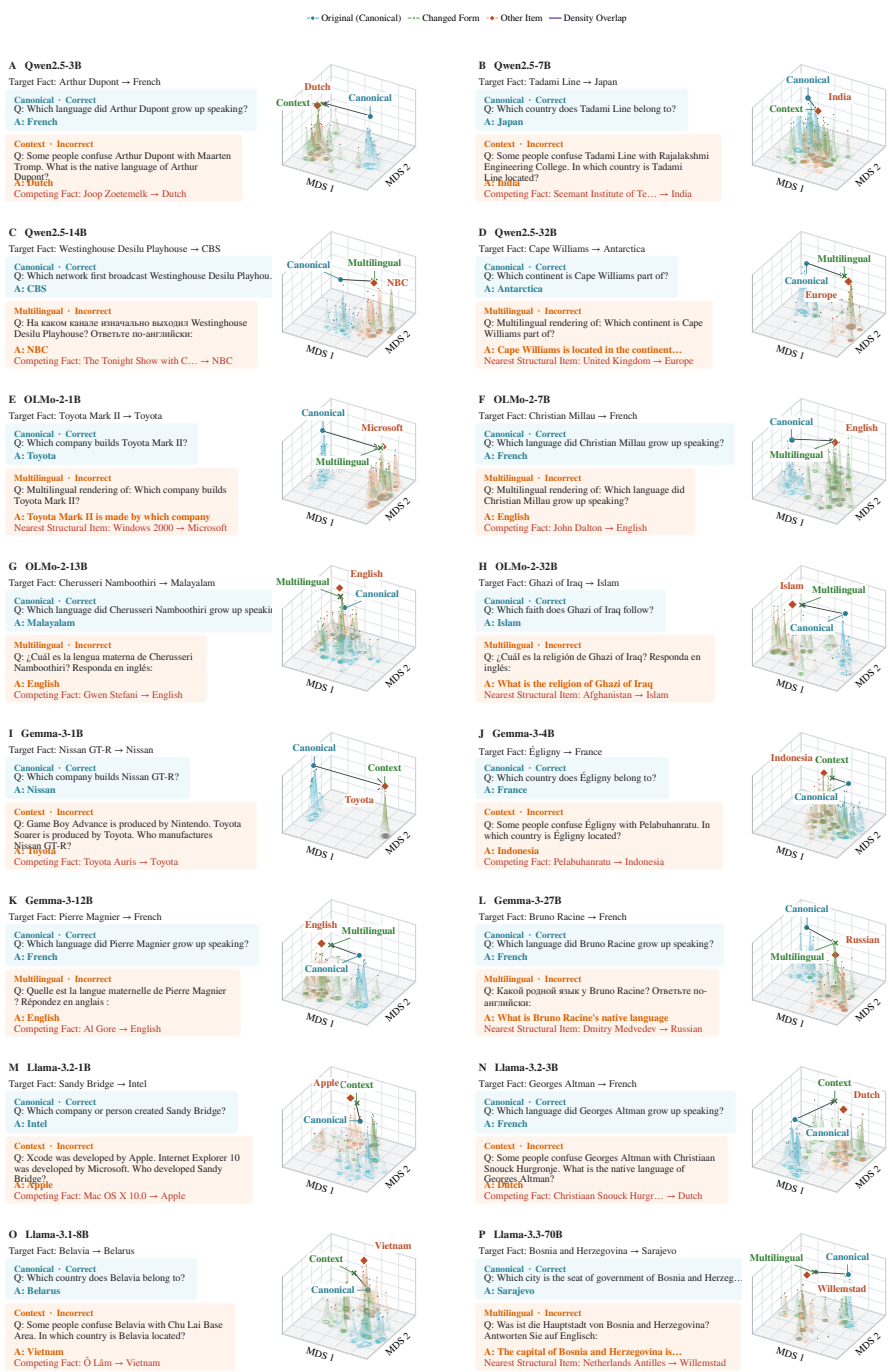


Figure 12: **More Qualitative Analysis of KTS.**

examples are selected after observing answer changes, they provide qualitative rather than predictive evidence.

D.7 More Qualitative Analysis of KTS

Figure 12 provides a corresponding relation-level analysis of KTS across all 16 models. Under the changed elicitation condition, the representation distribution is often displaced or mixed relative to its canonical counterpart, with reduced overlap and changes in the relative arrangement of knowledge items. These cases therefore reflect a broader reorganization of within-relation geometry,

Table 10: **Extended Model Rankings.** ACC, LSS, and KTS are reported for 43 additional model configurations across 11 model families. The best and second-best results in each column are shown in bold and underlined, respectively.

Family	Model	ACC ↑	LSS ↑	KTS ↑
Gemma	google/gemma-7b	0.8939	0.6114	**0.8643**
	google/gemma-1.1-2b-it	0.7936	0.6705	0.7747
	google/gemma-1.1-7b-it	0.8580	0.6728	0.7932
	google/gemma-2-2b	0.9198	0.7012	0.8397
	google/gemma-2-9b	<u>0.9614</u>	**0.7119**	0.8478
Phi	microsoft/phi-1_5	0.6107	0.6258	0.7941
	microsoft/phi-2	0.8904	0.6770	0.8405
	microsoft/phi-4	0.9240	0.5585	0.8033
Mistral	mistralai/Mistral-7B-v0.3	0.9240	<u>0.7078</u>	0.8556
	mistralai/Mistral-Nemo-Base-2407	0.9321	0.6927	0.8511
	mistralai/Mistral-Small-Instruct-2409	0.9549	0.6725	0.8310
Granite	ibm-granite/granite-7b-base	0.8067	0.6835	0.8429
	ibm-granite/granite-3b-code-base-128k	0.7207	0.6575	0.7978
	ibm-granite/granite-8b-code-base-128k	0.7396	0.6655	0.8264
Jamba	ai21labs/AI21-Jamba2-3B	0.8800	0.6419	0.8319
LFM	LiquidAI/LFM2-350M	0.5756	0.5507	0.7754
	LiquidAI/LFM2-700M	0.6617	0.6480	0.7946
	LiquidAI/LFM2-1.2B	0.7122	0.6683	0.8082
	LiquidAI/LFM2-2.6B	0.7535	0.6528	0.8220
	LiquidAI/LFM2.5-230M	0.6566	0.5522	0.8032
	LiquidAI/LFM2.5-350M	0.6671	0.5278	0.8009
	LiquidAI/LFM2.5-1.2B-Base	0.7508	0.6982	0.8275
SmolLM	HuggingFaceTB/SmolLM-135M	0.5093	0.5844	0.7946
	HuggingFaceTB/SmolLM-360M	0.4923	0.6613	0.8111
	HuggingFaceTB/SmolLM-1.7B	0.7006	0.6378	0.8303
	HuggingFaceTB/SmolLM2-135M	0.5910	0.6058	0.8104
	HuggingFaceTB/SmolLM2-360M	0.7400	0.6605	0.8316
	HuggingFaceTB/SmolLM2-1.7B	0.8634	0.6834	0.8470
	HuggingFaceTB/SmolLM3-3B	0.8684	0.7060	0.8402
ERNIE	baidu/ERNIE-4.5-0.3B-Base-PT	0.6570	0.6209	0.7793
Hunyuan	tencent/Hunyuan-0.5B-Pretrain	0.6142	0.5878	0.7652
	tencent/Hunyuan-1.8B-Pretrain	0.6605	0.6434	0.7991
	tencent/Hunyuan-4B-Pretrain	0.7932	0.6548	0.8274
	tencent/Hunyuan-MT-7B	0.7041	0.6169	0.7905
MiMo	XiaomiMiMo/MiMo-7B-Base	0.8349	0.6657	<u>0.8558</u>
Yi	01-ai/Yi-6B	0.7049	0.6132	0.8396
	01-ai/Yi-9B	0.9097	0.5556	0.8218
	01-ai/Yi-34B	**0.9676**	0.6306	0.8460
	01-ai/Yi-1.5-6B	0.9267	0.5589	0.8242
	01-ai/Yi-1.5-9B	0.9383	0.5579	0.8156
	01-ai/Yi-1.5-34B	0.9595	0.5849	0.8027
	01-ai/Yi-Coder-1.5B	0.5289	0.6248	0.7784
	01-ai/Yi-Coder-9B	0.6439	0.6142	0.8294

rather than only the displacement of an individual target item. This pattern illustrates the structural instability captured by KTS through changes in the ordering of pairwise representation distances. The MDS projections are provided for visualization only, whereas KTS is computed in the original representation space.

D.8 More Quantitative Model Comparison

The extended comparison across 43 additional configurations from 11 model families further shows that overall accuracy and representational stability do not necessarily improve together. For example, Phi-4 achieves higher ACC than Phi-2 (0.9240 vs. 0.8904) but lower LSS (0.5585 vs. 0.6770) and KTS (0.8033 vs. 0.8405); Mistral-Small-Instruct likewise achieves the highest ACC in its family (0.9549) but the lowest LSS (0.6725) and KTS (0.8310). Models with similar ACC can also differ substantially in stability: Gemma-2-9B and Yi-1.5-34B achieve comparable ACC (0.9614 vs. 0.9595), while Gemma-2-9B has higher LSS (0.7119 vs. 0.5849) and KTS (0.8478 vs. 0.8027). Although some model sequences exhibit joint improvements in correctness and stability, this pattern is not universal. These expanded comparisons support the conclusion that LSS and KTS capture dimensions of latent support that are not determined by model size or overall accuracy alone.

Table 11: **Frozen Gated-Correction Configurations.** Correction strength $\alpha_{b,e}$ and prompt-gate threshold $\tau_{b,e}$ used to construct action 1.

Backbone	Editor	$\alpha_{b,e}$	$\tau_{b,e}$
GPT2-XL	ROME	0.75	5.25
	MEMIT	0.50	5.50
	SAKE	1.00	6.75
LLaMA 2-7B	ROME	0.75	5.25
	MEMIT	0.25	5.55
	SAKE	1.00	6.75

E Details of Application

LatentDPO Problem Setup. We instantiate the stability-aware policy training method for each backbone–editor pair, where the backbones are GPT2-XL and LLaMA 2-7B and the editors are ROME, MEMIT, and SAKE. The evaluation contains $N = 1{,}998$ retained CounterFact edit requests. For every request i, the policy selects between two frozen editing actions:

$$a_i \in \{0, 1\}, \qquad a_i = 0 : \text{Base}, \qquad a_i = 1 : \text{Base} + \text{gated correction}. \tag{16}$$

Here, Base denotes the output of the corresponding editor. ROME is evaluated one request at a time with the original weights restored between requests, MEMIT edits the complete request batch jointly, and SAKE uses its standard transport edit. Policy training does not further modify the language-model backbone or the underlying editor.

LatentDPO is an offline binary-action selector over these two frozen editing actions. The gated correction is applied only to requests for which the learned policy selects action 1, rather than uniformly to all requests. Policy training does not further modify the language-model backbone or the underlying editor.

Gated Activation Correction. Starting from the final-token hidden state $\widetilde{h}_i(p)$ produced by the base editor, we fit a request-specific correction vector $\delta_i^\star$ using 100 training-source prompts:

$$\delta_i^\star = \arg\min_{\|\delta\|_2 \le r_i} \frac{1}{100} \sum_{m=1}^{100} \mathrm{CE}\Big(W_{\mathrm{out}}\big(\widetilde{h}_i(p_{im}) + \delta\big), y_i\Big), \tag{17}$$

where y_i is the first token of the requested target and r_i is the median distance from the target-token activations to their mean. We optimize Eq. equation 17 for 40 Adam steps with learning rate $r_i/(10\sqrt{d})$ and project the correction onto the ℓ_2 ball after every step.

When leading-space and no-leading-space tokenizations produce different target-token IDs, we select the token according to its target-cloud exact rate and then its mean target logit. A frozen `paraphrase-MiniLM-L6-v2` encoder provides prompt embeddings. Let $\phi(p)$ denote the embedding of prompt p, and let μ_i be the mean embedding of the 100 training-source prompts and the canonical rewrite prompt. The scope gate and corrected state are

$$g_i(p) = \mathbf{1}[\|\phi(p) - \mu_i\|_2 \le \tau_{b,e}], \tag{18}$$

$$h_i^{(1)}(p) = \widetilde{h}_i(p) + \alpha_{b,e} g_i(p) \delta_i^\star, \tag{19}$$

where $\alpha_{b,e}$ and $\tau_{b,e}$ denote the correction strength and gate threshold for backbone b and editor e. This correction is applied to activations at inference time and does not modify the model weights or output matrix.

For ROME and MEMIT, the correction configurations are selected conservatively to limit the observed Specificity loss while retaining a nontrivial efficacy gain. SAKE retains the correction configuration used in its original policy experiment. These operating points are fixed before policy fitting:

Application-Specific Stability Features. The available editing cache contains final-layer activation clouds for three distributions:

$$d \in \{\text{source}, \text{forced source}, \text{target}\}.$$

Because these caches do not contain the five elicitation conditions, multiple J-Lens layers, and external development data required by the full LSS/KTS protocol, we construct lightweight application-specific proxies derived from the same item-identity and relation-geometry principles.

For each request and distribution, the 100 training activations are randomly partitioned into five fixed groups of 20. After row normalization, each group is represented by its normalized centroid, projected to 64 dimensions using a fixed Gaussian random projection, and normalized again. The proxy-extraction seed is 20,260,830. We denote the resulting five views by $u_{i,d}^{(0)}, \ldots, u_{i,d}^{(4)}$.

For the LSS-inspired feature, facts are divided into five within-relation cross-fitting folds. For every held-out fold f, distribution d, and view $v > 0$, a similarity-Procrustes map $P_{d,v}^{(-f)}$ is fitted on the other four folds to align view v with reference view 0. Let $s_{r_i}^{(-f)}$ denote the median pairwise distance between reference-view centroids for the training facts in relation r_i, with the global training median used when a relation-specific estimate is unavailable. We define

$$D_{i,d} = \underset{v=1,\ldots,4}{\text{median}} \frac{\left\| P_{d,v}^{(-f(i))}\left(u_{i,d}^{(v)}\right) - u_{i,d}^{(0)} \right\|_2}{s_{r_i}^{(-f(i))}}, \tag{20}$$

$$\widehat{\text{LSS}}_{i,d} = \frac{1}{1 + D_{i,d}}, \qquad \widehat{\text{LSS}}_i = \frac{1}{3} \sum_d \widehat{\text{LSS}}_{i,d}. \tag{21}$$

For the KTS-inspired feature, the five views are centered and normalized separately within each relation and activation distribution. Across the ten view pairs, we measure bidirectional identity retrieval adjusted for its $1/n_r$ chance rate, Spearman preservation of within-relation cosine-distance structure, and overlap between the three-nearest-neighbor sets. The primary composite is

$$\widehat{\text{KTS}}_{i,d} = \frac{2 I_{i,d} G_{r_i,d}}{I_{i,d} + G_{r_i,d} + 10^{-12}}, \qquad \widehat{\text{KTS}}_i = \frac{1}{3} \sum_d \widehat{\text{KTS}}_{i,d}, \tag{22}$$

where $I_{i,d}$ is chance-adjusted identity retrieval and $G_{r_i,d}$ is normalized geometry preservation. The three-nearest-neighbor overlap is retained as an additional policy feature but is not included in Eq. equation 22.

Within every training partition, the stability features are converted to empirical percentiles using only the corresponding training facts:

$$q_i^{\text{LSS}} = \widehat{F}_{\text{LSS}}^{\text{train}}\left(\widehat{\text{LSS}}_i\right), \qquad q_i^{\text{KTS}} = \widehat{F}_{\text{KTS}}^{\text{train}}\left(\widehat{\text{KTS}}_i\right). \tag{23}$$

We use midpoint ranks for ties and assign a value of 0.5 when a feature is unavailable.

Preference Construction. Let $A_{i,a} \in \{0, 1\}$ indicate whether the greedy continuation from the canonical rewrite prompt begins with the requested target token under action a. We construct a 70/30 preference utility:

$$R_{i,0} = 0.70 A_{i,0} + 0.15(1 - q_i^{\text{LSS}}) + 0.15(1 - q_i^{\text{KTS}}), \tag{24}$$

$$R_{i,1} = 0.70 A_{i,1} + 0.15 q_i^{\text{LSS}} + 0.15 q_i^{\text{KTS}}. \tag{25}$$

Canonical editing success therefore receives 70% of the utility, whereas the item- and relation-level stability features receive 15% each. Generality and Specificity are not included directly in the training utility. For ROME and MEMIT, locality is instead controlled through the conservative action configurations in Table 11. The preferred action is $a_i^+ = 1$ when $R_{i,1} \geq R_{i,0}$ and $a_i^+ = 0$ otherwise.

Policy Inputs and Optimization. Each request is represented by 14 features. Six describe the fitted correction: mapped-source success and cross-entropy, target success and cross-entropy, trust radius, and unscaled correction norm. Eight describe the partial geometry: source, target, and mean LSS; source identity retrieval, three-nearest-neighbor overlap, and composite KTS; mean KTS; and the geometric mean of the global LSS and KTS ranks. Missing values are replaced by training-partition medians, and all features are standardized using training-partition statistics.

The editing-action policy is a linear Bernoulli model:

$$\pi_\theta(1 \mid x_i) = \sigma(w^\top z_i + c), \qquad \pi_\theta(0 \mid x_i) = 1 - \pi_\theta(1 \mid x_i), \tag{26}$$

where z_i is the standardized feature vector. With a uniform reference policy, we optimize

$$\mathcal{L}_{\text{pref}} = -\frac{1}{|\mathcal{D}_{\text{train}}|}\sum_i \log\sigma\left(\beta\left[\log\frac{\pi_\theta(a_i^+ \mid x_i)}{\pi_{\text{ref}}(a_i^+ \mid x_i)} - \log\frac{\pi_\theta(a_i^- \mid x_i)}{\pi_{\text{ref}}(a_i^- \mid x_i)}\right]\right) + \frac{\lambda}{2}\|w\|_2^2, \tag{27}$$

where $a_i^- = 1 - a_i^+$. We optimize the policy for 1,500 Adam steps with learning rate 0.03. We select

$$\beta \in \{0.1, 0.3, 1, 3\}, \qquad \lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$$

using relation-grouped validation reward. At inference time, action 1 is selected when $\pi_\theta(1 \mid x_i) \geq 0.5$.

Relation-Disjoint Evaluation. The 34 CounterFact relations are assigned to five row-balanced outer folds containing 407, 402, 402, 392, and 395 requests. No relation appears in both the training and test partitions of an outer fold. Hyperparameters are selected using an inner relation-grouped validation split. The selected policy is then refitted on the complete outer training partition and applied once to the held-out relations. Concatenating the five test partitions produces exactly one out-of-fold decision for every request. The split seed is 20,260,915.

Accuracy is exact success on the canonical rewrite prompt. Generality is mean success over two held-out CounterFact paraphrases. Specificity is the edit-macro average over eligible neighborhood prompts, excluding prompts for which the unedited model already prefers the new target. All metrics are reported over the same 1,998 requests.

Scope of Application. The experiment evaluates an offline policy that selects between two previously fitted editing actions. The action library is fixed before policy fitting, and every reported policy decision is obtained out of fold with respect to relation. The application-specific stability features are lightweight proxies derived from LSS and KTS rather than the formal multilayer benchmark metrics.

F Theoretical Derivation

F.1 Representation Extraction and Cross-fitting

For every query, we extract the residual-stream state at the final non-padding input token after the complete question has been processed and before the first answer token is generated. Layer depth is normalized as $\lambda_\ell = \ell/(L-1)$, and the primary scores use the final-quarter window $\mathcal{W}_m = \{\ell : \lambda_\ell \geq 0.75\}$. J-Lens estimates a rank-256 transport from each selected layer to the final representation space using randomized range finding. The transport is calibrated separately for each model and layer on 128 length-128 sequences sampled from `NeelNanda/pile-10k`, and the same fitted transport is then applied to all items and access conditions.

The similarity alignment used by LSS is estimated with five-fold cross-fitting. Items are assigned to folds within each track while preserving the distribution of comparison groups, and all queries associated with an item remain in the same fold. For an item in fold f, the condition-to-reference map is fitted using the other four folds and then applied to the held-out item. The local reference scale β^ℓ_{m,r_i} is estimated from the corresponding reference partition. Cross-fitting therefore prevents the alignment or normalization scale from being chosen to make the evaluated item appear stable.

F.2 Aggregation And Statistical Levels

For LSS, the normalized residual $d^\ell_{m,i,t}$ is first summarized by the median over the available non-reference elicitation conditions and then averaged over $\mathcal{W}_m$. We denote this aggregated drift by

$$D_{m,i} = \frac{1}{|\mathcal{W}_m|}\sum_{\ell\in\mathcal{W}_m} \underset{t\in\mathcal{T}_i}{\text{median}} d^\ell_{m,i,t}. \tag{28}$$

Statistical comparisons use $D_{m,i}$, while $\text{LSS}_{m,i} = (1 + D_{m,i})^{-1}$ is reported as its bounded, monotone transformation. Item-level LSS values are averaged within each track; the two track means receive equal weight in an overall benchmark score.

KTS is defined at the comparison-group level. For each condition pair $p = (t, t')$, centering uses exactly the items observed under both conditions, so the two distance matrices have identical membership. A condition pair is valid only when the shared set meets the pre-specified size requirement and both distance vectors admit a Spearman correlation. Correlations are averaged equally across valid condition pairs and selected layers to obtain $\mathrm{KTS}_{m,r}$. Group scores are then macro-averaged within each track, followed by an equal-weight average of the two tracks. When KTS is used with an item-level outcome, item i is assigned the group-level context KTS_{m,r_i}; this assignment does not make KTS an item-level measurement.

F.3 Properties Of the Stability Measures

Correctness does not uniquely determine the internal state that produces an answer. For a single generated token with output weights w_j and biases b_j, predicting token y requires

$$(w_y - w_j)^\top u \geq b_j - b_y, \qquad \forall j \neq y. \tag{29}$$

These inequalities define a region of representations that produce the same token. Distinct points in this region can therefore yield the same output while having different locations and relations to other items. For multi-token answers, the same argument applies to the sequence of output-compatible regions visited during generation.

LSS is invariant to a condition-wide change that is exactly captured by the fitted similarity map. If $A^{\ell}_{m,\tau,t}(\bar{v}^{\ell}_{m,i,t}) = \bar{v}^{\ell}_{m,i,a}$ for every item, then $e^{\ell}_{m,i,t} = 0$, $D_{m,i} = 0$, and $\mathrm{LSS}_{m,i} = 1$. More generally,

$$\frac{\partial \,\mathrm{LSS}_{m,i}}{\partial D_{m,i}} = -(1 + D_{m,i})^{-2} < 0, \tag{30}$$

so LSS decreases monotonically with normalized residual drift.

For KTS, group centering removes translations shared by all items in a comparison group. Pairwise cosine distance is unchanged when the centered representations undergo a common positive isotropic scaling and orthogonal transformation. Spearman correlation further remains unchanged under a strictly monotone transformation of all pairwise distances. Consequently, $\mathrm{KTS}_{m,r} = 1$ whenever the ordering of within-group pairwise distances is perfectly preserved across the compared conditions, even if their absolute values change.

F.4 Interpretation And Scope

LSS is the normalized residual displacement of an individual item's condition-level representation, whereas KTS is the distance-rank preservation of its comparison group's internal organization. Neither metric contains a correctness criterion: stable representations may support an incorrect target, and correct outputs may arise from representations that vary across access conditions. The two measures are therefore geometric diagnostics of latent support and complement behavioral correctness.

Because the observed state precedes the first generated answer token, the metrics characterize pre-generation support. They do not measure the stability of the full hidden-state trajectory produced during multi-token generation.

G Limitation

This work primarily focuses on evaluating internal knowledge stability rather than directly optimizing it. In future work, we plan to build on the principles underlying LSS and KTS to develop post-training strategies that explicitly improve the stability of internal knowledge representations. These metrics may serve as training objectives, regularization signals, or feedback criteria to encourage consistent knowledge access across different expressions, contexts, formats, and languages. We will further investigate whether improving internal stability can enhance the reliability and generalization of model behavior while preserving overall task performance.

H MORE DISCUSSIONS

▷ ***Q1. What do LSS and KTS measure, and does high stability necessarily imply correct knowledge?***

LSS and KTS should be interpreted as complementary to correctness rather than as substitutes for it. LSS measures whether an individual knowledge item preserves its representational identity after condition-wide shifts are removed, whereas KTS measures whether the organization among items within the same relation remains consistent. A model may stably represent an incorrect association; therefore, stability alone does not guarantee factual correctness. Conversely, a correct response may be supported by an unstable representation and fail under another form of access. We consequently advocate jointly examining ACC, LSS, and KTS: ACC captures answer correctness, LSS captures item-level stability, and KTS captures structural coherence. Their combination distinguishes reliably supported correct answers from correct but internally fragile behavior.

▷ ***Q2. Could LSS and KTS mainly reflect differences in wording, format, or language?***

Several results suggest that the proposed metrics are not dominated by superficial prompt variation. LSS first removes representational shifts shared across items through relation-stratified, cross-fitted alignment, retaining only item-specific residual drift. Although this residual accounts for only 4.2%–6.8% of the raw cross-condition drift, it provides an additional 1.20 AUROC points for predicting answer changes, whereas the wrong-condition control produces only a near-zero gain. The semantic perturbation experiment further reveals a graded response: minimal edits induce the smallest changes, controlled probes produce intermediate changes, and same-relation hard negatives produce the largest changes. The stable rankings under leave-one-probe-out evaluation provide additional evidence that the results are not driven by any single probe type.

▷ ***Q3. Why are the final-quarter layers and J-Lens representations used?***

The final-quarter window is motivated by both sensitivity analysis and causal intervention. Representations from the 80%–100% depth interval almost reproduce the primary LSS and KTS rankings, whereas earlier intervals show substantially weaker agreement. More importantly, activation patching shows that late-layer states contain increasingly behaviorally effective item-specific information: the recovery advantage of same-item over same-relation control patches increases from +0.02 in the 40%–50% interval to +0.80 in the final quarter. J-Lens provides a shared comparison space across layers, while aggregation over multiple late layers reduces dependence on a single state. Thus, the selected representation is not intended to describe all stages of computation; it operationalizes stability in the late, pre-generation states most directly involved in producing the answer.

▷ ***Q4. Why are both LSS and KTS necessary?***

LSS and KTS capture different levels of latent organization. LSS is defined at the knowledge-item level and measures whether a particular item remains distinguishable across elicitation conditions. KTS is defined at the relation or task-type level and measures whether the relative geometry among multiple items is preserved. These properties need not change together: an individual item may drift while the broader relational structure remains stable, or several moderate item-level shifts may collectively reorganize the relation geometry. This distinction is also reflected in the non-parallel model and condition rankings observed in our experiments. Accordingly, KTS is treated as a relation-level contextual property rather than an item-specific score. Using both metrics provides a more complete account of latent support than either item identity or relational structure alone.

▷ ***Q5. Can consistently incorrect knowledge also receive high stability scores?***

Yes, consistently incorrect knowledge can also receive high stability scores. This is why we evaluate LSS and KTS jointly with ACC: stability complements correctness rather than replacing it.