

Representational and Behavioral Stability of Truth in Large Language Models

Samantha Dies^{1‡}, Courtney Maynard¹, Germans Savcisens¹, and Tina Eliassi-Rad^{1,2,3}

¹Khoury College of Computer Sciences, Northeastern University, 440 Huntington Ave, #202, Boston, MA 02115 USA

²Network Science Institute, Northeastern University, 177 Huntington Ave, #1010, Boston, MA 02115 USA

³Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501 USA

[‡]dies.s@northeastern.edu

ABSTRACT

Large language models (LLMs) are increasingly used as information sources, yet small changes in semantic framing can destabilize their truth judgments. We propose P-StaT (Perturbation Stability of Truth), an evaluation framework for testing belief stability under controlled semantic perturbations in representational and behavioral settings via probing and zero-shot prompting. Across sixteen open-source LLMs and three domains, we compare perturbations involving epistemically familiar **Neither** statements drawn from well-known fictional contexts (**Fictional**) to those involving unfamiliar **Neither** statements not seen in training data (**Synthetic**). We find a consistent stability hierarchy: **Synthetic** content aligns closely with factual representations and induces the largest retractions of previously held beliefs, producing up to 32.7% retractions in representational evaluations and up to 36.3% in behavioral evaluations. By contrast, **Fictional** content is more representationally distinct and comparatively stable. Together, these results suggest that epistemic familiarity is a robust signal across instantiations of belief stability under semantic reframing, complementing accuracy-based factuality evaluation with a notion of epistemic robustness.

1 Introduction

Large language models (LLMs) are increasingly used as sources of information, but their behavior often blurs the line between knowledge and plausibility [1, 2]. While humans are expected to distinguish between **True**, **False**, and **Neither**-valued statements, it remains unclear whether LLMs form similarly structured internal representations, or whether such representations predict when truth judgments remain stable under changes in semantic framing [3, 4]. This gap is increasingly consequential as LLMs are deployed in high-stakes informational settings, where instability can undermine reliability and contribute to undesirable behaviors such as hallucinations [5, 6].

Existing work approaches this problem from two largely separate directions. Representation-based probing studies ask whether **True** and **False** statements occupy separable regions in activation space and whether “truth directions” transfer across LLMs or domains [7, 8, 9]. Behavioral studies document that small changes in prompting, context, or adversarial framing can substantially alter an LLM’s outputs, including its apparent confidence and self-consistency [10, 11, 12, 13]. While both lines of work reveal important failure modes, they are rarely connected through a shared experimental lens. Thus, we lack a unified framework for testing whether a candidate stability signal in latent space corresponds to stability in LLM behavior [14, 15].

We address this gap by proposing the Perturbation Stability of Truth framework, P-StaT, which links inter-

nal representations and LLM behavior through the lens of belief stability (Fig. 1). P-StaT asks how stable LLM truth judgments remain when semantic assumptions are systematically varied. This perspective is motivated by formal epistemology: under Leitgeb’s notion of *P*-stability, a belief system is epistemically well-formed only if established beliefs are preserved under small, principled changes in evidential context [16].

We operationalize semantic variation through perturbations Θ that modify which **Neither** statements are treated as compatible with truth. We distinguish epistemically *familiar* **Neither** statements (**Fictional**) from *unfamiliar* **Neither** statements (**Synthetic**) across three domains. Crucially, the same perturbation is instantiated both representationally, by retraining probes, and behaviorally, via zero-shot prompting.

We observe a consistent pattern: Unfamiliar **Synthetic** content induces the largest epistemic retractions in both representational and behavioral settings. By contrast, familiar **Fictional** content occupies more distinct regions of activation space and produces substantially fewer retractions. Together, these results identify *epistemic familiarity* as a cross-instantiation signal for when LLM truth judgments are most likely to destabilize under semantic reframing.

Contributions

1. We introduce a dataset of fictional statements across three domains (City Locations, Medical Indications, and Word Definitions) to enable controlled compar-

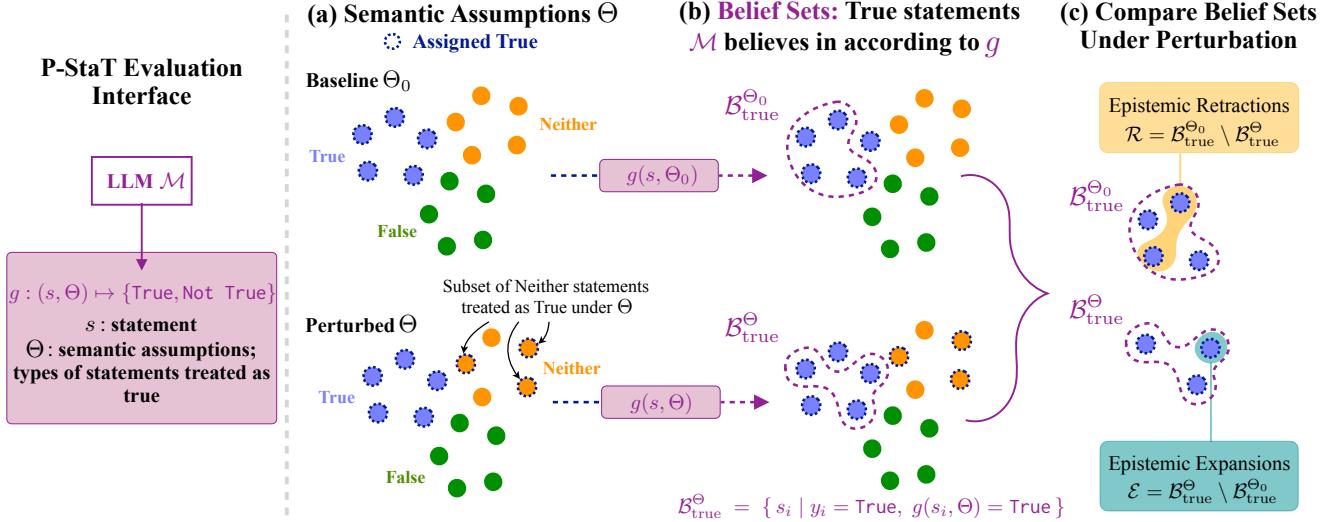


Figure 1. Overview of the P-StaT framework. P-StaT evaluates the stability of an LLM \mathcal{M} 's beliefs by using a decision function g to map statements s to True or Not True under semantic assumptions Θ . (a) In the baseline case Θ_0 , only True statements (purple) are labeled True, while False (green) and Neither (orange) statements are labeled Not True. Under a perturbed assumption set Θ , a subset of Neither statements is treated as True. (b) The sets of True statements that \mathcal{M} believes to be True in the baseline ($\mathcal{B}_{\text{true}}^{\Theta_0}$) and perturbed ($\mathcal{B}_{\text{true}}^{\Theta}$) cases are identified via g . (c) Comparing these belief sets reveals unstable beliefs that are either retracted (\mathcal{R}) or expanded (\mathcal{E}). Crucially, g can be instantiated using either representation-based probing or behavioral zero-shot prompting, enabling direct connections between instability in latent space and behavioral instability.

isons between epistemically familiar and unfamiliar Neither content.

2. We propose P-StaT, a perturbation-based stability framework grounded in epistemic notions of belief stability that supports both probing and zero-shot prompting.
3. Across sixteen open-source LLMs, we show that unfamiliar Synthetic statements induce the largest epistemic retractions both representationally and behaviorally, identifying epistemic familiarity as a key signal of stability.

2 Related Work

Our work lies at the intersection of three research threads: (i) representation-based probing of veracity, (ii) behavioral sensitivity to semantic context, and (iii) epistemic notions of belief stability.

Representational Probing of Veracity. Probing examines which properties are recoverable from hidden representations of LLMs, providing insights into what LLMs encode beyond observable behavior [17, 18, 19]. Recent work has shown that True and False statements form separable clusters across LLMs and domains in activation space [7, 8]. The sAwMIL framework [9] extends this work by modeling True, False, and Neither statements as

distinct representational directions. Related work on hallucination detection similarly suggests that hidden representations encode strong veracity signals even when outputs are wrong [2].

Behavioral Stability and In-Context Effects. A separate line of research documents small changes in wording or framing can lead to large shifts in LLM behavior, including jailbreak vulnerabilities [13], sycophancy [20], and instability under paraphrasing [11]. Recent work has framed this brittleness in terms of behavioral consistency across interactions. For example, Li et al. introduce benchmarks for evaluating response stability in multi-turn settings [12]. Related work on in-context learning suggests that larger LLMs can override pretraining-induced semantics when presented with conflicting in-context labels, while smaller LLMs rely more heavily on prior knowledge [21]. Complementarily, Bigelow et al. propose a Bayesian account in which both prompting and activation steering modify latent beliefs [22].

Epistemic Belief Stability. LLMs also exhibit systematic difficulties in tracking epistemic distinctions such as belief, knowledge, and factuality [3, 4]. Herrmann and Levinstein argue that the study of LLM beliefs lacks unified criteria for when internal representations should count as belief-like [15]. Formal epistemology offers a complementary perspective: Leitgeb's notion of P -

stability characterizes rational belief systems as those that preserve established beliefs under small, *justified* changes in evidential context [16].

By combining representational probing with behavioral perturbations through P-StaT, our work bridges prior studies of latent veracity structure and in-context sensitivity. We assess how beliefs respond to principled shifts in semantic assumptions, providing a unified view of epistemic stability across representations and behavior.

3 Methodology

We study how LLMs organize statements in activation space and whether this organization predicts the stability of beliefs under semantic reframing using P-StaT (Perturbation Stability of Truth), which applies the same perturbations in both probing and zero-shot evaluations.

3.1 Epistemic Familiarity and Neither Statements

Let $\mathcal{S} = \{s_i\}_{i=1}^N$ denote declarative statements drawn from factual domains with labels $y_i \in \{\text{True}, \text{False}, \text{Neither}\}$. Let $\mathcal{N} = \{s_i \in \mathcal{S} \mid y_i = \text{Neither}\}$ denote the set of **Neither** statements. While all statements in \mathcal{N} lack real-world truth value, \mathcal{N} is heterogeneous and can be partitioned into familiar (**Fictional**) and unfamiliar (**Synthetic**) subsets:

$$\mathcal{N} = \mathcal{N}_{\text{fam}} \cup \mathcal{N}_{\text{unf}}, \quad \mathcal{N}_{\text{fam}} \cap \mathcal{N}_{\text{unf}} = \emptyset.$$

Here, \mathcal{N}_{fam} contains well-known fictional entities likely present in training corpora, while \mathcal{N}_{unf} contains constructed entities intended to be absent from training data. For **Fictional**, we additionally distinguish canonically true and false subsets, $\mathcal{N}_{\text{fam}}^{(T)}$ and $\mathcal{N}_{\text{fam}}^{(F)}$.

We evaluate three domains that differ in how sharply truth and falsehood are delineated (City Locations, Medical Indications, Word Definitions; Table 1) [7, 8, 9]. **True**, **False**, and **Synthetic** statements originate in [9], while we introduce the new **Fictional** dataset.¹ Details on data construction and validation are in Supplementary Section B.

3.2 Veracity Representations

For an LLM \mathcal{M} , we define a representation map $\phi_{\mathcal{M},l} : s_i \mapsto \mathbf{z}_i^{(l)}$ between a statement s_i and its token-level hidden representations at layer l . For each (dataset, LLM) pair we select l to maximize linear separability between **True** and **Not True** statements [7, 8, 9]. We refer to $\mathbf{z}^{(l)}$ as *veracity representations*. Together with statements and labels, they define the dataset $\mathcal{D} = \{(s_i, \mathbf{z}_i^{(l)}, y_i)\}_{i=1}^N$, which serves as the input to representational and behavioral experiments.

¹The new fictional dataset can be accessed at https://huggingface.co/datasets/samanthadies/representational_stability.

3.3 P-StaT: Perturbation Stability of Truth

P-StaT evaluates stability for a fixed LLM \mathcal{M} by assuming the evaluation procedure

$$g : (s_i, \Theta) \mapsto \{\text{True}, \text{Not True}\},$$

where Θ encodes semantic assumptions about which statements should be treated as compatible with truth. The $g(\cdot, \Theta)$ interface is the core abstraction of P-StaT, enabling the same semantic perturbation to be instantiated in distinct evaluation settings.

Representational vs. Behavioral Stability. In representational experiments, g is implemented as $g_p(s, \Theta) = h_\Theta(\phi_{\mathcal{M},l}(s))$, where h_Θ is a linear probe trained with labels induced by Θ . In behavioral experiments, g is implemented as $g_{\text{zs}}(s, \Theta)$ via zero-shot prompting and a *belief context* C_Θ . Both settings support the same perturbations, enabling direct comparison.

Baseline evaluation. Under a baseline semantic interpretation Θ_0 , g identifies the set of ground-truth **True** statements that are assigned **True**:

$$\mathcal{B}_{\text{true}}^{\Theta_0} = \{s_i \mid y_i = \text{True}, g(s_i, \Theta_0) = \text{True}\}.$$

Perturbations. A perturbation modifies the semantic assumptions encoded in Θ . Let $\mathcal{N}_\Theta \subseteq \mathcal{N}$ denote the subset of **Neither** statements treated as compatible with truth under a perturbed interpretation. The perturbed interpretation Θ differs from Θ_0 only in this reassignment and yields the perturbed belief set:

$$\mathcal{B}_{\text{true}}^\Theta = \{s_i \mid y_i = \text{True}, g(s_i, \Theta) = \text{True}\}.$$

Thus, only the semantic interpretation of **Neither** content is altered.

Epistemic retractions. Within P-StaT, we quantify stability by comparing the baseline and perturbed belief sets. Epistemic retractions, $\mathcal{R} = \mathcal{B}_{\text{true}}^{\Theta_0} \setminus \mathcal{B}_{\text{true}}^\Theta$, capture ground-truth **True** statements that lose belief status under perturbation. While we also consider epistemic expansions, $\mathcal{E} = \mathcal{B}_{\text{true}}^\Theta \setminus \mathcal{B}_{\text{true}}^{\Theta_0}$, for completeness, epistemic retractions are a stronger signal of belief instability than expansions because they withdraw previously held beliefs [16].

4 Experiments

We use P-StaT to apply the same perturbation Θ to both (i) probe-based evaluations over $\phi_{\mathcal{M},l}$ and (ii) zero-shot evaluations via belief context. We implement all experiments in Python using PyTorch [23], NumPy, SciPy, and scikit-learn [24], and rely on the HuggingFace Transformers ecosystem for LLM access and data handling [25]. All experiments were run on a university HPC cluster with NVIDIA H200 GPUs and required approximately 36 GPU-hours (activation extraction and zero-shot evaluations), with probe training performed on CPU.

Dataset	True	False	Synthetic	Fictional	Noise	Examples
City Locations	A: 1392 N: 1376	A: 1358 N: 1374	A: 876 N: 876	A: 350 N: 350	795	(T) The city of Surat is located in India. (Fa) The city of Palembang is located in the Dominican Republic. (S) The city of Norminsk is located in Jamoates. (Fi) The city of Bikini Bottom is located in the Pacific Ocean.
Medical Indications	A: 1439 N: 1522	A: 1523 N: 1419	A: 478 N: 522	A: 402 N: 402	771	(T) Pentobarbital is indicated for the treatment of insomnia. (Fa) Vancomycin is not indicated for the treatment of lower respiratory tract infections. (S) Alumil is indicated for the treatment of reticlers. (Fi) The Trump Virus is indicated for the treatment of Xenovirus Takis-B.
Word Definitions	A: 1234 N: 1235	A: 1277 N: 1254	A: 1747 N: 1753	A: 1224 N: 1224	1095	(T) Hoagy is a synonym of an Italian sandwich. (Fa) Decalogue is an astronomer. (S) Dostab is a scencer. (Fi) Snozberry is a type of berry.

Table 1. Summary of datasets and statement types. Number of affirmative (A) and negated (N) statements across the three datasets, along with examples. Each dataset includes **True**, **False**, **Synthetic**, and **Fictional** statements, while **Noise** consists of randomly generated Gaussian activation vectors matched in dimensionality and distribution to the statement embeddings. **Synthetic** statements serve as **Neither** statements that were not seen during LLM training, i.e., \mathcal{N}_{unf} , while **Fictional** statements are familiar **Neither** statements \mathcal{N}_{fam} . A version of this table without the **Fictional** and **Noise** columns can be found in [9].

Condition	\mathcal{N}_Θ	Probing: Training Labels	Zero-shot: Belief Context C_Θ
Baseline	\emptyset	True vs. False+Synthetic+Fictional+Noise	None
Synthetic	\mathcal{N}_{unf}	True+Synthetic vs. False+Fictional+Noise	Synthetic
Fictional	\mathcal{N}_{fam}	True+Fictional vs. False+Synthetic+Noise	Fictional
Fictional (T)	$\mathcal{N}_{\text{fam}}^{(T)}$	True+Fictional(T) vs. False+Synthetic+Fictional(F)+Noise	Fictional (T)
Noise	N/A	True+Noise vs. False+Synthetic+Fictional	N/A

Table 2. Perturbation conditions Θ and instantiations in P-StaT. Each condition corresponds to a semantic interpretation Θ defined by \mathcal{N}_Θ , the **Neither** statements treated as compatible with truth. The same Θ is instantiated (i) representationally by retraining a probe with labels induced by \mathcal{N}_Θ and (ii) behaviorally by constructing a belief context C_Θ from the corresponding training statements. **Noise** is probing-only since it is a non-semantic control.

4.1 Data

We use the three domains and statement types defined in Section 3.1 (Table 1). **Fictional** is newly released with this work, and **Noise** is a set of randomly generated Gaussian vectors that match the dimensionality and distribution of the veracity representations and serve as a probing-only non-semantic control. Supplementary Section B contains additional information on data construction.

Approximately 55% of the data is used for training, 20% for calibration, and 25% for testing (see Supplementary Tab. A2). The splits are shared across all experiments. We compute stability on the same held-out set of ground-truth **True** statements from the test split, so that differences in retractions are attributable only to changes in Θ .

4.2 Activations

We evaluate sixteen open-source LLMs spanning Gemma, Llama, Mistral, and Qwen families, with both base and chat-tuned variants (Supplementary Section C). For each (dataset, LLM) pair, we extract token-level activations at the selected layer l that maximizes linear separability between **True** and **Not True** statements [9] (Supplementary Tab. A3).

For descriptive analyses, we reduce each activation sequence to a single vector by selecting the final non-padding token. At the linguistic level, we compute rank-frequency curves over character bigrams aggregated across entity names for each statement type. At the representation level, we compute pairwise 1-D Wasserstein distances between activation distributions, considering all activation dimensions. These measures reveal similarities between statement types at the linguistic level and in latent space.

4.3 Perturbation Conditions and Shared Protocol

For each perturbation Θ in Table 2, we instantiate Θ using only training data: probes are retrained with labels induced by Θ , and belief contexts C_Θ are constructed from the corresponding training statements. We then evaluate on the same held-out set of ground-truth **True** test statements, computing $\mathcal{B}_{\text{true}}^{\Theta_0}$ and $\mathcal{B}_{\text{true}}^\Theta$ by applying $g(\cdot, \Theta_0)$ and $g(\cdot, \Theta)$, and report epistemic retractions $\mathcal{R} = \mathcal{B}_{\text{true}}^{\Theta_0} \setminus \mathcal{B}_{\text{true}}^\Theta$ (and expansions analogously).

4.3.1 Instantiation I: Probing over Activations

To implement $g(\cdot, \Theta)$ representationally, we train a linear probe that operates on the veracity representations $\phi_{M,l}(s)$. We use the sparse-aware multiple-instance learning probe (**sAwMIL**) [9], a multiclass probing method

designed to extract reliable and transferable veracity directions from LLM activations. Unlike simpler probes such as the **Mean Difference** classifier [7], which assumes that truth and falsehood lie along a single axis, **sAwMIL** models **True**, **False**, and **Neither** as distinct directions and aggregates token-level representations using multiple-instance learning. As a max-margin method, **sAwMIL** yields stable decision boundaries, making differences across perturbations more reflective of genuine structure in the LLM’s geometry than noise from the probe.²

For each condition in Table 2, we retrain the probe on $\mathcal{D}_{\text{train}}$ with labels induced by Θ , holding the architecture and hyperparameters fixed across conditions. Token-level representations are scaled using a standard scaler fit on the training set; bags are truncated to a fixed maximum size; and we perform a grid search over the regularization parameter \mathcal{C} using three-fold cross-validation with mean average precision.³

4.3.2 Instantiation II: Zero-shot Prompting with Belief Context

To implement $g(\cdot, \Theta)$ behaviorally, we use a zero-shot prompt. For each condition Θ (Tab. 2), we construct a belief context C_Θ from the corresponding training statements and insert it into each held-out ground-truth **True** test statement s .

To construct C_Θ , we uniformly sample $K = 100$ context statements from \mathcal{N}_Θ as specified in Table 2. We sample without replacement.

For each ground-truth **True** test statement, we use the following prompt:

```
[optional belief context  $C_\Theta$ ]
Is the following statement correct?
[statement  $s$ ]

a. The statement is true.
b. The statement is false.
c. The statement is neither true nor
false.

The final answer is
```

For chat-tuned LLMs, we place the same content into the LLM’s chat template and define the model’s response as the highest-probability next token among the answer labels $\{a, b, c\}$ at the final prompt position.

We compute next-token probabilities for a , b , and c at the final position and predict

$$g_{\text{zs}}(s, \Theta) = \arg \max_{\ell \in \{a, b, c\}} p(\ell | s, \Theta),$$

²For comparison, results from the **Mean Difference** probe appear in Supplementary Section E.

³Our code with all seeds and final hyperparameters is at https://github.com/samanthadies/representational_stability.

with temperature set to zero.⁴ We treat option a as **True** and options $\{b, c\}$ as **Not True**. All zero-shot evaluations are therefore deterministic, and reported results aggregate outcomes across LLMs rather than multiple random seeds.

5 Results

Linguistic vs. Representational Structure

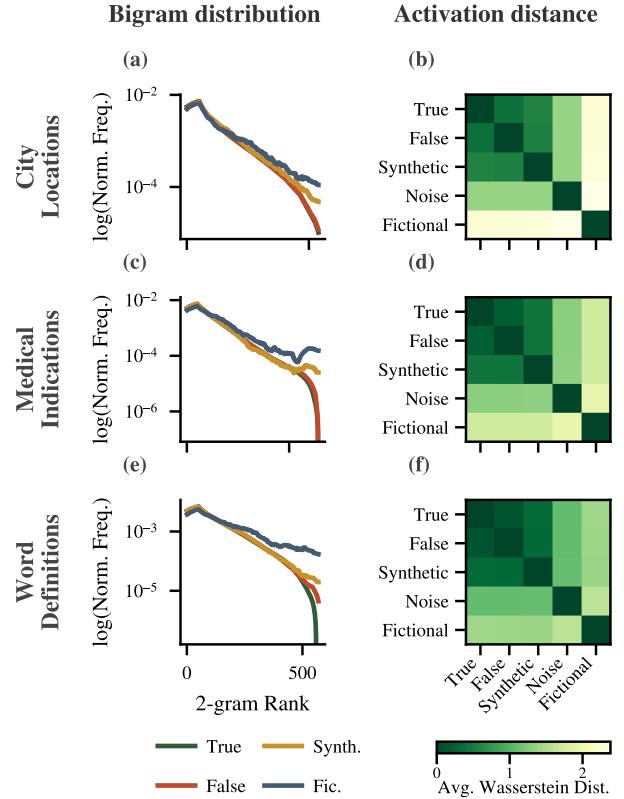


Figure 2. Linguistic vs. representational structure of Neither statements. For (a),(b) City Locations, (c),(d) Medical Indications, and (e),(f) Word Definitions, the left column shows normalized character bigram rank–frequency curves for **True** (green), **False** (red), **Synthetic** (yellow), and **Fictional** (blue) statements, and the right column shows pairwise Wasserstein distances between activation distributions. **Fictional** content exhibits a domain-dependent decoupling between linguistic form and latent-space organization, indicating that veracity representations reflect both epistemic context and linguistic form.

We report (i) the descriptive structure of **Neither** statements, (ii) representational stability under probing,

⁴We also track the probability mass corresponding to all non-label tokens, but define g exclusively over $\{a, b, c\}$ rather than parsing free-form textual outputs.

Dataset	Perturbation	True to True	Not True to Not True	Epistemic Expansions \mathcal{E}	Epistemic Retractions \mathcal{R}
City Locations	Synthetic	9153 (91.5)	360 (3.6)	274 (2.7)	213 (2.1)
	Fictional	9326 (93.3)	568 (5.7)	66 (0.7)	40 (0.4)
	Fictional (T)	9330 (93.3)	576 (5.8)	58 (0.6)	36 (0.4)
	Noise	9183 (91.8)	532 (5.3)	102 (1.0)	183 (1.8)
Medical Locations	Synthetic	7413 (72.6)	1556 (15.2)	786 (7.7)	460 (4.5)
	Fictional	7808 (76.4)	2284 (22.4)	58 (0.6)	65 (0.6)
	Fictional (T)	7815 (76.5)	2269 (22.2)	73 (0.7)	58 (0.6)
	Noise	7779 (76.2)	2009 (19.7)	333 (3.3)	94 (0.9)
Word Definitions	Synthetic	3682 (37.2)	2188 (22.1)	785 (7.9)	3233 (32.7)
	Fictional	6507 (65.8)	2494 (25.2)	479 (4.1)	408 (4.1)
	Fictional (T)	6795 (68.7)	2520 (25.5)	453 (4.6)	120 (1.2)
	Noise	6653 (67.3)	2251 (25.4)	462 (4.7)	262 (2.6)

Table 3. Epistemic expansions \mathcal{E} and retractions \mathcal{R} under probing label perturbations. Counts (percentages) of beliefs that remain stable or undergo expansions \mathcal{E} or retractions \mathcal{R} under Synthetic (yellow), Fictional (gray), Fictional(T) (blue), and Noise (red) perturbations. Synthetic perturbations consistently produce the highest rate of epistemic retractions, indicating that epistemically unfamiliar content induces the most instability.

and (iii) behavioral stability under zero-shot prompting. LLM-level results are in Supplementary Section D.

5.1 Linguistic and Representational Structure of Neither Statements

We begin by characterizing how familiar and unfamiliar Neither statements differ. At the linguistic level, True, False, and Synthetic statements exhibit nearly identical normalized bigram rank–frequency curves (Figure 2(a),(c),(e)). Fictional statements, by contrast, show a slower decay reflecting stylistic patterns characteristic of narrative text. The differences in Fictional bigram distributions are most visible in Word Definitions and least in City Locations.

We next test whether these linguistic differences explain latent-space structure by comparing activation distributions across statement types using pairwise Wasserstein distances (Figure 2(b),(d),(f); LLM-level heatmaps appear in Supplementary Section D.1). Across domains, True and False occupy nearby regions of activation space, and Synthetic remains relatively close to factual content despite its lack of real-world referents. By contrast, Fictional statements are consistently more representationally separated from factual content, forming a distinct cluster that is not well-explained by bigram statistics alone.

Taken together, these findings indicate a decoupling between linguistic and latent-space similarity. Al-

though linguistic divergence of Fictional content varies by domain, representational separation remains consistently pronounced. Further, the domain with the largest Fictional separation linguistically (Word Definitions) is the domain with the smallest representational distance. The geometry of LLM activations thus reflects both linguistic form and context.

5.2 Representational Stability under Probing Perturbations

We next evaluate P-StaT representational stability by retraining sAwMIL under perturbed interpretations Θ and measuring epistemic retractions on held-out ground-truth True statements. Table 3 summarizes prediction changes across perturbation conditions aggregated over all LLMs. LLM-level results and results for the Mean Difference probe appear in Supplementary Sections D.2 and E.

Across all three domains, Synthetic perturbations induce the highest rate of epistemic retractions: 2.1% in City Locations, 4.5% in Medical Indications, and 32.7% in Word Definitions. These rates are substantially larger than those produced by Fictional and Fictional(T) perturbations, establishing a clear perturbation hierarchy in which epistemically unfamiliar Neither content most strongly destabilizes learned veracity boundaries. Synthetic perturbations also induce the largest epistemic expansions across domains.

Retraction rates also vary systematically by domain.

Dataset	Perturbation	True to True	Not True to Not True	Epistemic Expansions \mathcal{E}	Epistemic Retractions \mathcal{R}
City Locations	Synthetic	2503 (25.0)	3301 (33.0)	567 (5.7)	3629 (36.3)
	Fictional	3521 (35.2)	2708 (27.1)	1160 (11.6)	2611 (26.1)
	Fictional (T)	3678 (36.8)	2529 (25.3)	1339 (13.4)	2454 (24.5)
Medical Locations	Synthetic	1539 (14.1)	5562 (51.0)	843 (7.7)	2952 (27.1)
	Fictional	1816 (16.7)	4489 (41.2)	1916 (17.6)	2675 (24.6)
	Fictional (T)	1724 (15.8)	4476 (41.1)	1929 (17.7)	2767 (25.4)
Word Definitions	Synthetic	1880 (19.0)	4246 (42.9)	1647 (16.7)	2115 (21.4)
	Fictional	2179 (22.0)	4319 (43.7)	1574 (15.9)	1816 (18.4)
	Fictional (T)	2159 (21.8)	4389 (44.4)	1504 (15.2)	1836 (18.6)

Table 4. Epistemic expansions \mathcal{E} and retractions \mathcal{R} under perturbed zero-shot experiments. Counts (percentages) of beliefs that remain stable or undergo expansions \mathcal{E} or retractions \mathcal{R} under Synthetic (yellow), Fictional (gray), and Fictional(T) (blue) perturbations. Despite domain variation, Synthetic perturbations induce the highest retraction rates in all cases, mirroring the hierarchy observed under probing in Table 3.

City Locations is highly stable, Medical Indications shows moderate instability, and Word Definitions is markedly more fragile. These results again indicate that representational stability is governed by epistemic familiarity, given that City Locations are likely richly represented in training corpora, while Medical Indications are common but specialized, and Word definitions are the most semantically flexible.

5.3 Behavioral Stability under Zero-Shot Perturbations

We next evaluate P-StaT behavioral stability by applying the same perturbations Θ via belief context in zero-shot prompting and measuring epistemic retractions. Table 4 reports belief changes across LLMs, with LLM-level results in Supplementary Section D.3. Overall belief changes are larger than in probing experiments, reflecting the greater flexibility of behavioral responses compared to linear reclassification over fixed representations.

Despite this increase, the perturbation hierarchy observed representationally also appears behaviorally. The Synthetic perturbation induces the highest epistemic retraction rate in every domain, with 36.3% in City Locations, 27.1% in Medical Indications, and 21.4% in Word Definitions. Fictional and Fictional(T) also induce substantial retractions, but consistently fewer than Synthetic within each dataset.

However, domain ordering differs from the probing setting. City Locations has the highest zero-shot retraction rates, while Word Definitions has the lowest. This reversal likely reflects interactions between belief stability and behavioral accuracy. Specifically, zero-shot retractions conflate belief changes with the number of baseline

beliefs available to retract (i.e., a statement must be considered True in the baseline case in order to become a retraction). In contrast, probing isolates changes in the decision boundary under fixed representations.

Our results show that **epistemic familiarity governs stability not only in activation space, but also in LLM behavior**. Unfamiliar Synthetic content systematically destabilizes LLM beliefs under semantic perturbation, whereas familiar Fictional content produces smaller and more context-dependent effects.

6 Discussion

We provide a unified perspective on how LLMs organize and maintain truth judgments under semantic perturbation. Across LLMs, domains, and evaluation settings, Neither statements play a central role in belief stability: unfamiliar Synthetic statements consistently induce the largest epistemic retractions, while familiar Fictional statements are more stable. Our findings indicate that belief stability in LLMs is governed as much by epistemic familiarity as by linguistic form.

At the representational level, we observe a decoupling between linguistic similarity and latent-space organization. Fictional content exhibits domain-dependent linguistic variation that does not align with distance in representation space. This mismatch suggests that veracity representations encode higher-level epistemic context learned during training rather than lexical statistics alone. Through the lens of P-StaT, these representational differences translate into systematic differences in stability under semantic perturbation. Expanding the semantic definition of truth to include unfamiliar Synthetic content reliably destabilizes previously held beliefs, while

perturbations involving familiar **Fictional** content are better tolerated. These results hold in both representational and behavioral instantiations, suggesting that latent-space organization captures meaningful epistemic structure.

More broadly, P-StaT complements accuracy-based factuality benchmarks by shifting focus from correctness under fixed prompts to stability under principled semantic shifts. By emphasizing epistemic retractions, it targets a particularly consequential form of instability. This perspective aligns with formal accounts of rational belief change, such as Leitgeb’s notion of *P*-stability, which requires that established beliefs be preserved under justified changes in evidential context [16]. From this viewpoint, the sensitivity of LLMs to **Synthetic** perturbations highlights a structural limitation: distributional plausibility alone does not ensure a stable internal organization of truth, falsity, and indeterminacy. Together, these findings suggest that evaluating belief stability under controlled semantic perturbations offers a principled way to probe the epistemic organization of LLMs beyond what can be inferred from accuracy alone.

7 Conclusion

We show that **epistemic familiarity is a key determinant of belief stability under semantic reframing**. By linking internal representations and behavioral responses under matched perturbations, P-StaT reveals that unfamiliar **Synthetic** content consistently induces the largest epistemic retractions across LLMs and domains. These results demonstrate that evaluating belief stability under controlled perturbations of **Neither** content provides a principled complement to accuracy-based factuality metrics and a step toward more epistemically reliable language models.

8 Limitations

Our perturbations focus on a specific notion of epistemic familiarity and a limited set of factual domains. While this design enables controlled comparisons across representational and behavioral settings, it does not exhaustively cover all types of semantic variation. Extending P-StaT to other forms of epistemic ambiguity, such as disputed claims, probabilistic beliefs, or evolving facts, would further test its generality. In addition, our analysis considers fixed LLM parameters. Although P-StaT isolates how semantic perturbations interact with existing internal representations, it does not address how belief stability may change in settings where representations themselves evolve over time. Finally, retraction rates depend on accuracy on **True** statements, so our notion of stability reflects both robustness to perturbation and an LLM’s propensity to assert truth in the baseline case. While we emphasize epistemic retractions as a primary

signal of instability, other applications may require alternative notions of stability.

9 Ethical Considerations

Our study examines the conditions under which perturbations can systematically destabilize what an LLM considers true. For example, we find that unfamiliar synthetic content induces the largest epistemic retractions, producing up to 32.7% retractions in representational evaluations and up to 36.3% in behavioral evaluations. This could inform efforts to undermine LLM reliability and has implications for trust in LLMs.

While our framework could be misused to deliberately destabilize LLM beliefs, our intent is diagnostic rather than adversarial: P-StaT is designed to identify epistemic vulnerabilities in order to inform more robust evaluation and LLM design. We do not propose interventions or belief manipulation techniques, and all experiments are conducted on fixed, open-source LLMs in offline settings. More broadly, our findings highlight a structural limitation of current LLMs rather than a prescription for exploiting it. P-StaT is intended solely for research and diagnostic evaluation of pretrained LLMs and is not designed for deployment, belief steering, or real-world decision-making systems.

9.1 Data availability

We use the **True**, **False**, and **Synthetic** statements available at <https://huggingface.co/datasets/carlomarxx/trilemma-of-truth> under a CC-BY-4.0 license. **Fictional** statements are available at https://huggingface.co/datasets/samanthadies/representational_stability. The code used to generate the **Noise** activations can be found at https://github.com/samanthadies/representational_stability.

9.2 Code availability

We release the code used to generate activations, generate the **Noise** activations, and run the P-StaT stability experiments under an MIT License at https://github.com/samanthadies/representational_stability. ChatGPT and Copilot were used to help write initial versions of experiment and plotting scripts, and helped clean and comment the code. All AI-generated content was reviewed and verified by the authors.

All models used in this work are publicly available for research use under their respective licenses (MIT License for sAwMIL; Gemma for Gemma-7b, Gemma-7b-it, Gemma-2-9b, and Gemma-2-9b-it; llama3.1 for Llama-3.1-8b and Llama-3.1-8b-Instruct; llama3.2 for Llama-3.2-3b, Llama-3.2-3b-Instruct; llama3 for Llama3-Med42-8b; Bio-Medical-Llama-3-8b LLM License for Bio-Medical-Llama-3-8b; apache-2.0 for Mistral-7B-v0.3, Mistral-7B-Instruct-v0.3, Qwen2.5-7B, Qwen2.5-

7B-Instruct, Qwen2.5-14B, and Qwen2.5-14B-Instruct). Marks and Tegmark did not specify a license for the Mean Difference probe [7].

Acknowledgments

We thank Hannes Leitgeb and Branden Fitelson for discussions on *P*-stability and how it might be related to epistemic uncertainty in LLMs. We also thank Zohair Shafi and Moritz Laber for their feedback and discussions on methodological and empirical portions of this work.

Funding

This material was sponsored by the Government of the United States under Contract Number FA8702-15-D-0002. The view, opinions, and/or filings contained in this material are those of the author(s) and should not be construed as an official position, policy, or decision of the Government of the United States or Carnegie Mellon University or the Software Engineering Institute unless designated by other documentation.

Competing interests

The authors declare no competing interests.

References

1. AlKhamissi, B., Li, M., Celikyilmaz, A., Diab, M. & Ghazvininejad, M. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031* (<https://doi.org/10.48550/arXiv.2204.06031> (2022)).
2. Han, J. *et al.* Simple factuality probes detect hallucinations in long-form natural language generation. In *Findings of the Association for Computational Linguistics (EMNLP 2025)*, 16209–16226. (<https://doi.org/10.18653/v1/2025.findings-emnlp.880> (2025)).
3. Abbasi Yadkori, Y., Kuzborskij, I., György, A. & Szepesvari, C. To believe or not to believe your LLM: Iterative prompting for estimating epistemic uncertainty. *Adv. Neural Inf. Process. Syst.* **37**, 58077–58117 (2024).
4. Suzgun, M. *et al.* Language models cannot reliably distinguish belief from knowledge and fact. *Nat. Mach. Intell.* 1–11. (<https://doi.org/10.1038/s42256-025-01113-8> (2025)).
5. Liu, Y. *et al.* Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. In *Socially Responsible Language Modelling Research*. (<https://doi.org/10.48550/arXiv.2308.05374> (2023)).
6. Huang, L. *et al.* A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Inf. Syst.* **43**, 1–55. (<https://doi.org/10.1145/3703155> (2025)).
7. Marks, S. & Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of True/False datasets. In *Proceedings of the 1st Conference on Language Modeling (COLM 2024)* (2024).
8. Bürger, L., Hamprecht, F. A. & Nadler, B. Truth is universal: Robust detection of lies in LLMs. *Adv. Neural Inf. Process. Syst.* **37**, 138393–138431 (2024).
9. Savcisen, G. & Eliassi-Rad, T. Trilemma of truth in large language models. In *Mechanistic Interpretability Workshop at NeurIPS 2025* (2025). (<https://openreview.net/forum?id=z7dLG2ycRf>).
10. Turpin, M., Michael, J., Perez, E. & Bowman, S. Language models don't always say what they think: Unfaithful explanations in Chain-of-Thought prompting. *Adv. Neural Inf. Process. Syst.* **36**, 74952–74965 (2023).
11. Elazar, Y. *et al.* Measuring and improving consistency in pretrained language models. *Transactions Assoc. for Comput. Linguist.* **9**, 1012–1031. (https://doi.org/10.1162/tacl_a_00410 (2021)).
12. Li, Y., Miao, Y., Ding, X., Krishnan, R. & Padman, R. Firm or fickle? evaluating large language models consistency in sequential interactions. In Che, W., Nabende, J., Shutova, E. & Pilehvar, M. T. (eds.) *Findings of the Association for Computational Linguistics: ACL 2025*, 6679–6700. (<https://doi.org/10.18653/v1/2025.findings-acl.347> (Association for Computational Linguistics, Vienna, Austria, 2025)).
13. Wei, A., Haghtalab, N. & Steinhardt, J. Jailbroken: How does LLM safety training fail? *Adv. Neural Inf. Process. Syst.* **36**, 80079–80110 (2023).
14. Harding, J. Operationalising representation in natural language processing. *Br. J. for Philos. Sci.* (<https://doi.org/10.1086/728685> (2023)).
15. Herrmann, D. A. & Levinstein, B. A. Standards for belief representations in LLMs. *Minds Mach.* **35**, 5. (<https://doi.org/10.1007/s11023-024-09709-6> (2024)).
16. Leitgeb, H. The stability theory of belief. *Philos. review* **123**, 131–171. (<https://doi.org/10.1215/00318108-2400575> (2014)).
17. Conneau, A., Kruszewski, G., Lample, G., Barrault, L. & Baroni, M. What you can cram into a single

- \$\&!\#* vector: Probing sentence embeddings for linguistic properties. In Gurevych, I. & Miyao, Y. (eds.) *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2126–2136. <https://doi.org/10.18653/v1/P18-1198> (Association for Computational Linguistics, Melbourne, Australia, 2018).
18. Hewitt, J. & Manning, C. D. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, 4129–4138. <https://doi.org/10.18653/v1/N19-1419> (2019).
19. Tenney, I., Das, D. & Pavlick, E. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL 2019)*, 4593–4601. <https://doi.org/10.48550/arXiv.1905.05950> (2019).
20. Sharma, M. *et al.* Towards understanding sycophancy in language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*. <https://doi.org/10.48550/arXiv.2310.13548> (2024).
21. Wei, J. *et al.* Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846* <https://doi.org/10.48550/arXiv.2303.03846> (2023).
22. Bigelow, E. *et al.* Belief dynamics reveal the dual nature of in-context learning and activation steering. *arXiv preprint arXiv:2511.00617* <https://doi.org/10.48550/arXiv.2511.00617> (2025).
23. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32** (2019).
24. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
25. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. <https://doi.org/10.18653/v1/2020.emnlp-demos.6> (2020).
26. Wikipedia contributors. List of fictional settlements (2025). https://en.wikipedia.org/wiki/List_of_fictional_settlements.
27. Wikipedia contributors. List of fictional city-states in literature (2025). https://en.wikipedia.org/wiki/List_of_fictional_city-states_in_literature.
28. Fandom NeoEncyclopedia. List of fictional diseases (2025). [https://neoencyclopedia.fandom.com/wiki/List_of_fictional_diseases](https://neoencyclopedia.fandom.com/wiki>List_of_fictional_diseases).
29. Fandom NeoEncyclopedia. List of fictional toxins (2025). https://neoencyclopedia.fandom.com/wiki/List_of_fictional_toxins.
30. Chemeurope Encyclopedia. List of fictional medicines and drugs (2025). https://www.chemeurope.com/en/encyclopedia/List_of_fictional_medicines_and_drugs.html.
31. Tomasula, S. The Thackery T. Lambshead pocket guide to eccentric & discredited diseases. *The Rev. Contemp. Fiction* **24** (2004).
32. Almaden, S. A. Dahl dictionary: A list of 103 words made-up by Roald Dahl (2023). <https://beelinguapp.com/blog/Dahl%20Dictionary:%20A%20List%20of%20103%20Words%20Made-up%20By%20Roald%20Dahl>.
33. Schleitwiler, P. & Shuflin, G. Dothraki initial text (2025). <https://conlang.org/language-creation-conference/lcc5/1-dothraki-initial-text/>.
34. Dict-Na'vi.com Online Dictionary. wordlist “substantive (noun)” (2025). <https://dict-navi.com/en/dictionary/list/?type=classification&ID=1>.

A Notation

We summarize the mathematical notation used throughout the manuscript in Table A1.

Symbol	Description
\mathcal{M}	A fixed large language model (LLM).
l	Layer index used for activation extraction.
$\mathcal{S} = \{s_i\}_{i=1}^N$	Set of N natural-language statements.
y_i	Ground-truth veracity label of s_i , $y_i \in \{\text{True}, \text{False}, \text{Neither}\}$.
\mathcal{N}	Set of all Neither statements: $\mathcal{N} = \{s_i \mid y_i = \text{Neither}\}$.
$\mathcal{N}_{\text{fam}}, \mathcal{N}_{\text{unf}}$	Partition of \mathcal{N} into epistemically familiar (Fictional) and unfamiliar (Synthetic) subsets.
$\mathcal{N}_{\text{fam}}^{(T)}, \mathcal{N}_{\text{fam}}^{(F)}$	Canonically true and false subsets of familiar fictional statements (used in Fictional(T)).
$\phi_{\mathcal{M}, l}$	Representation map from a statement to its layer- l hidden representations under \mathcal{M} .
$\mathbf{z}_i^{(l)}$	Layer- l representation of statement s_i , $\mathbf{z}_i^{(l)} = \phi_{\mathcal{M}, l}(s_i)$.
\mathcal{D}	Dataset of statements, representations, and labels: $\mathcal{D} = \{(s_i, \mathbf{z}_i^{(l)}, y_i)\}_{i=1}^N$.
$\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}$	Training and test splits of \mathcal{D} .
Θ	Semantic interpretation specifying which Neither statements are treated as compatible with truth.
$\mathcal{N}_\Theta \subseteq \mathcal{N}$	Subset of Neither statements treated as compatible with truth under Θ .
$g(\cdot, \Theta)$	Evaluation function mapping a statement and interpretation to a binary judgment: $g(s, \Theta) \in \{\text{True}, \text{Not True}\}$.
g_p	Representational instantiation of g via a probe over $\phi_{\mathcal{M}, l}(s)$.
g_{zs}	Behavioral instantiation of g via zero-shot prompting with belief context C_Θ .
C_Θ	The belief context inserted into a zero-shot prompt to instantiate perturbation Θ .
Θ_0	Baseline semantic interpretation (no Neither treated as true).
$\mathcal{B}_{\text{true}}^{\Theta_0}$	Baseline belief set: $\{s_i \mid y_i = \text{True}, g(s_i, \Theta_0) = \text{True}\}$.
$\mathcal{B}_{\text{true}}^\Theta$	Belief set under perturbed interpretation Θ .
\mathcal{R}	Epistemic retractions: $\mathcal{R} = \mathcal{B}_{\text{true}}^{\Theta_0} \setminus \mathcal{B}_{\text{true}}^\Theta$.
\mathcal{E}	Epistemic expansions: $\mathcal{E} = \mathcal{B}_{\text{true}}^\Theta \setminus \mathcal{B}_{\text{true}}^{\Theta_0}$.

Table A1. Notation. Summary of symbols used throughout the manuscript.

B Data

B.1 Data Generation

We use statements from the City Locations, Medical Indications, and Word Definitions datasets introduced in [9]. City statements take the form “*The city of [city] is (not) located in [country]*,” (omitting “*The city of*” when redundant). Medical statements follow “[drug] is (not) indicated for the treatment of [disease/condition].” Word Definition statements draw from three templates: “[word] is (not) a [instanceOf],” “[word] is (not) a type of [typeOf],” and “[word] is (not) a synonym of [synonym].” No personal data, identifying information, or user-generated content is included.

B.1.1 True, False, and Synthetic Statements

We take the **True**, **False**, and **Synthetic** statements from the datasets introduced in [9]. All statements are constructed with both affirmative and negated forms. **Synthetic** entities are generated using a Markov-chain-based name generator (**namemaker**⁵) and undergo multi-stage filtering, including database checks, model tagging, and web-search validation, to ensure no accidental overlap with real entities. Validated names are then paired to form grammatically coherent but semantically meaningless statements that follow each template. Because **Synthetic** entities do not exist and cannot have appeared in training corpora, LLMs have no basis for assigning them a truth value. Accordingly, these statements function as **Neither** cases: unknown claims for which belief should be suspended rather than confidently classified as true or false.

B.1.2 Fictional Statements

In addition to **Synthetic** statements, which represent *unseen and unknown* claims, we construct new sets of **Fictional** statements for all three domains. **Fictional** statements also function as **Neither** statements in our experiments as they reference entities that do not exist in the real world and therefore lack real-world truth value. However, unlike **Synthetic** statements, many **Fictional** entities are likely to have appeared in LLM training corpora.⁶ As such, they represent a complementary form of **Neither**: claims that an LLM may recognize, but that

⁵<https://github.com/Rickmsd/namemaker>.

⁶For later analyses, we additionally annotate fictional statements with their within-universe factual status (**Fictional (T)** or **Fictional (F)**), but this labeling is not used in the primary **True** vs. **Not True** classification tasks.

still lie outside the true–false axis relevant to factual grounding.

To ensure that **Fictional** statements remain genuinely non-factual, all terms were validated to exclude any real-world overlap, and fictional lexical items appearing in any natural language were excluded to prevent misinterpretation by multilingual LLMs. **Fictional** statements were then constructed using the same templates as the **True**, **False**, and **Synthetic** statements, including both affirmative and negated forms.

Fictional City Locations. Fictional cities and countries span literature, film, radio, television, comics, animation, and games [26, 27]. Each \langle city, location \rangle pair is included only when an identifiable enclosing region exists. When multiple spatial resolutions are available, we select the most specific (e.g., \langle Quahog, Rhode Island \rangle rather than \langle Quahog, United States \rangle).

Fictional Medical Indications. Fictional drug and disease statements are drawn from (1) *NeoEncyclopedia Wiki* [28, 29]; (2) ChemEurope’s *List of Fictional Medicines and Drugs* [30]; and (3) *The Thackery T. Lambshead Pocket Guide to Eccentric & Discredited Diseases* [31]. Drug–disease pairs are included when a treatment relationship exists according to the fictional source.

Fictional Word Definitions. Fictional lexical items are compiled from (1) *Gobblefunk* [32]; (2) *Dothraki* [33]; and (3) *Na’vi* [34]. Dothraki and Na’vi have formal linguistic structure, whereas Gobblefunk is a playful neologistic extension of English.

B.1.3 Noise

The **Noise** statements contain no linguistic content. We generate $n_{\text{noise}} = 0.10 \cdot |\mathcal{D}|$ random activation sequences by sampling from a multivariate Gaussian with per-feature mean, standard deviation, and sequence-length distribution matched to the LLM activations. These distributionally consistent but non-semantic sequences serve as a control, allowing us to test whether observed representational differences arise from semantic content or from statistical variation in activation space.

B.2 Data Splits for Probing Experiments

Dataset	Train	Calibration	Test	Total
City Locations	4746 (0.54)	1772 (0.20)	2229 (0.25)	8747 (1.00)
Medical Indications	4636 (0.55)	1721 (0.20)	2121 (0.25)	8478 (1.00)
Word Definitions	6488 (0.54)	2514 (0.21)	3041 (0.25)	12043 (1.00)

Table A2. Dataset splits. The number of statements used in training, calibration, and testing of the probe. The proportion of total statements is reported in parentheses.

Table A2 summarizes the partitions used for all experiments. Each dataset is split exclusively into training, calibration, and test sets to prevent data leakage. Approximately 55% of statements are used for training, 20% for calibration, and 25% for testing. We use identical splits in all conditions.

C LLMs

Table A3 lists the sixteen open-source LLMs used in our experiments. The set spans four major model families, Gemma, Llama, Mistral, and Qwen, with approximately 3 billion to 15 billion parameters and release dates between February and September 2024. For each family, we include both base (pre-trained) and chat-tuned variants. Together, these LLMs provide a representative cross-section of current decoder-only architectures varying in scale, origin, and training objectives.

D LLM-level Results

The main text reports LLM-agnostic stability patterns aggregated across sixteen LLMs. Here we provide the corresponding LLM-level results.

D.1 Linguistic and Representational Structure of Neither Statements (by LLM)

Figures A1–A16 show the pairwise activation distance matrices for all sixteen LLMs. Three general representational patterns emerge. The first, observed in `_gemma-2-9b` (Fig. A1) and `gemma-2-9b` (Fig. A10), shows **Fictional** and **Synthetic** statements clustering near **True** and **False** statements, with **Noise** forming a distinct outlier. The

Official Name	Short Name	Type	# Decoders	# Parameters	Best Layer	Release Date	Source
Gemma-7b	gemma-7b	Base	28	8.54 B	C: 14, M: 19, W: 17	Feb 21, 2024	Google
Gemma-2-9b	gemma-2-9b	Base	26	9.24 B	C: 24, M: 25, W: 23	Jun 27, 2024	Google
Llama-3-8b	llama-3.1-8b	Base	32	8.03 B	C: 18, M: 17, W: 17	Jul 23, 2024	Meta
Llama-3.2-3b	llama-3.2-3b	Base	28	3.21 B	C: 16, M: 17, W: 15	Sep 25, 2024	Meta
Mistral-7B-v0.3	mistral-7B-v0.3	Base	32	7.25 B	C: 18, M: 17, W: 18	May 22, 2024	Mistral AI
Qwen2.5-7B	qwen-2.5-7b	Base	28	7.62 B	C: 18, M: 19, W: 17	Sep 19, 2024	Alibaba Cloud
Qwen2.5-14B	qwen-2.5-14b	Base	38	14.80 B	C: 30, M: 31, W: 30	Sep 19, 2024	Alibaba Cloud
Gemma-7b-it	_gemma-7b	Chat	28	8.54 B	C: 19, M: 19, W: 17	Feb 21, 2024	Google
Gemma-2-9b-it	_gemma-2-9b	Chat	26	9.24 B	C: 27, M: 26, W: 25	Jul 27, 2024	Google
Llama-3.2-3b-Instruct	_llama-3.2-3b	Chat	28	3.21 B	C: 16, M: 19, W: 18	Sep 25, 2024	Meta
Llama-3.1-8b-Instruct	_llama-3.1-8b	Chat	32	8.03 B	C: 18, M: 19, W: 18	Jul 23, 2024	Meta
Llama3-Med42-8b	_llama-3-8b-med	Chat	32	8.03 B	C: 18, M: 16, W: 15	Aug 12, 2024	M42 Health
Bio-Medical-Llama-3-8b	_llama-3-8b-bio	Chat	32	8.03 B	C: 18, M: 19, W: 18	Aug 11, 2024	Contact Doctor
Mistral-7b-Instruct-v0.3	_mistral-7B-v0.3	Chat	32	7.25 B	C: 19, M: 21, W: 18	May 22, 2024	Mistral AI
Qwen2.5-7B-Instruct	_qwen-2.5-7b	Chat	28	7.62 B	C: 19, M: 21, W: 18	Aug 18, 2024	Alibaba Cloud
Qwen2.5-14B-Instruct	_qwen-2.5-14b	Chat	38	14.80 B	C: 31, M: 34, W: 30	Aug 18, 2024	Alibaba Cloud

Table A3. LLMs used in the stability experiments. We list the official names of the LLMs according to the HuggingFace repository [25]. We further specify the shortened name we use to refer to each of the LLMs, whether it is the base, pre-trained LLM or a chat-tuned version, the number of decoders, the number of parameters, the release date, and the source of the LLM. Finally, we report the layers with the best separation between **True** and **Not True** statements for the City Locations (“C”), Medical Indications (“M”), and Word Definitions (“W”) datasets. The LLMs are publicly available through HuggingFace [25].

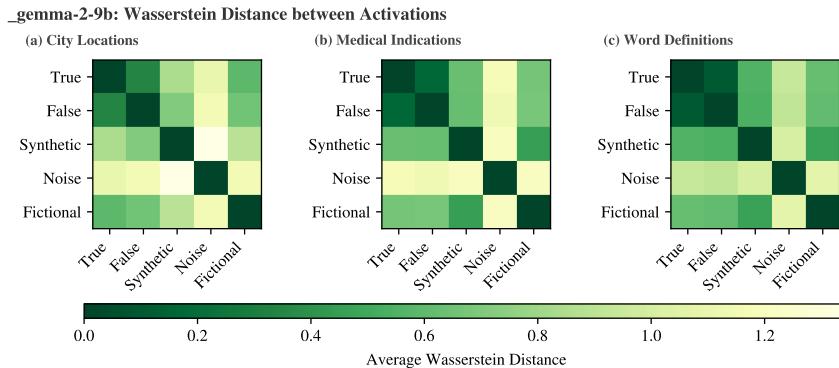


Figure A1. Wasserstein distance between activations for `_gemma-2-9b`. Pairwise Wasserstein distances between activation distributions of **True**, **False**, **Synthetic**, **Fictional**, and **Noise** statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. **Noise** has distinct representations, but **Fictional** and **Synthetic** statements are represented similarly to **True** and **False** statements and each other.

second, present in `_gemma-7b` (Fig. A2), `gemma-7b` (Fig. A11), `_qwen-2.5-14b` (Fig. A8), `qwen-2.5-14b` (Fig. A15), and `_qwen-2.5-7b` (Fig. A9), exhibits **Synthetic** statements close to **True** and **False**, **Fictional** statements clearly separated, and **Noise** positioned slightly closer to the **True/False/Synthetic** cluster. The third, seen in the remaining nine LLMs, features **Synthetic** statements aligned with **True** and **False**, while both **Fictional** and **Noise** statements occupy distinct and distant regions. Except for `_qwen-2.5-7b` (which follows the second pattern; Fig. A9) and `qwen-2.5-7b` (the third; Fig. A16), base and chat versions are qualitatively similar.

D.2 Representational Stability under Probing Perturbations (by LLM)

Figure A17 shows how the sAwMIL decision boundaries change under each perturbation, measured by cosine similarity to the baseline **True** vs. **Not True** direction and by the associated bias shift. Consistent with the aggregate flip rates (Table 3), **Synthetic** produces the largest boundary changes across domains, while **Fictional**, **Fictional(T)**, and **Noise** yield smaller deviations.

Different LLM families exhibit different degrees of susceptibility to these perturbations. Chat-tuned variants (denoted by leading underscores) tend to exhibit somewhat larger rotations and bias shifts than their base models. An exception is `gemma-7b`, which shows unusually large shifts in the Word Definitions domain.

Figures A18–A20 break down epistemic retractions \mathcal{R} and expansions \mathcal{E} by LLM. Chat-tuned LLMs tend to

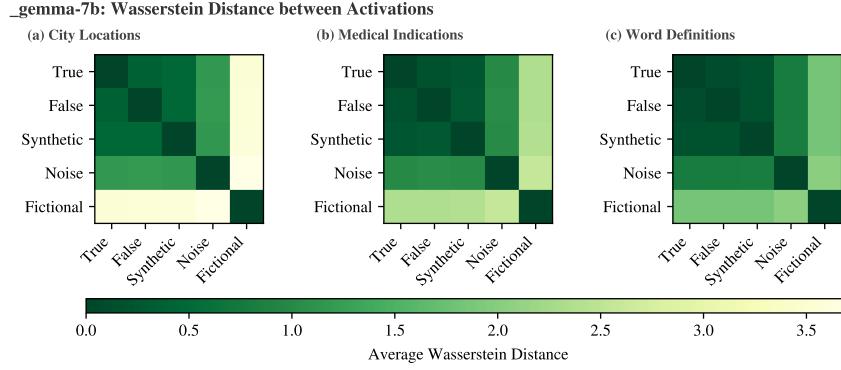


Figure A2. Wasserstein distance between activations for `_gemma-7b`. Pairwise Wasserstein distances between activation distributions of True, False, Synthetic, Fictional, and Noise statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. Synthetic statements are represented similarly to True and False statements, while Fictional statements are represented distinctly from all other statements.

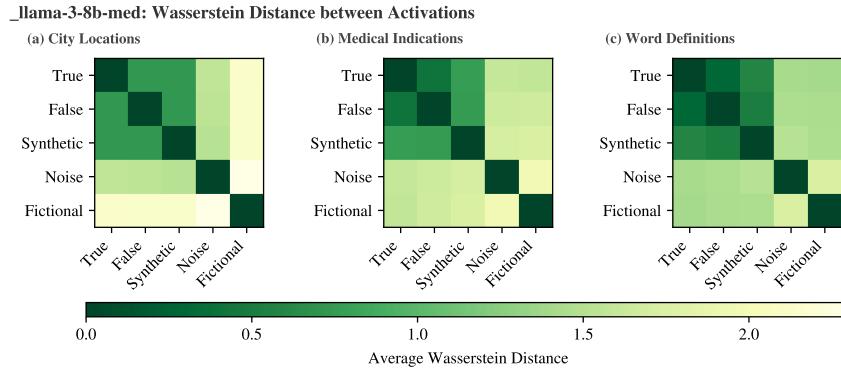


Figure A3. Wasserstein distance between activations for `_llama-3-8b-med`. Pairwise Wasserstein distances between activation distributions of True, False, Synthetic, Fictional, and Noise statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. Synthetic statements are represented similarly to True and False statements, while Fictional statements and Noise are represented distinctly from all other statements.

produce more expansions, while Base models tend to exhibit more retractions. However, these tendencies do not hold uniformly, and overall differences across LLM families are smaller than differences across perturbation types.

D.3 Behavioral Stability under Zero-Shot Perturbations (by LLM)

Figures A21–A23 break down epistemic retractions \mathcal{R} and expansions \mathcal{E} by LLM. Opposite to the behavior seen in the representational instantiation, Base LLMs tend to produce more expansions, while Chat-tuned variants tend to exhibit more retractions.

E Exploring the Mean Difference Probe

We repeated the representation-based perturbation experiments using the Mean Difference probe proposed by Marks and Tegmark [7] to supplement the sAwMIL results. Mean Difference estimates a “truth direction” by taking the vector difference between the mean activation of True statements and that of False statements, optionally scaled by the inverse covariance matrix of the data. This approach is inherently sensitive to differences in the centroids and covariance structure of the data, which leads to strong instability in the learned decision boundary when Neither statements are included alongside True and False examples. The Mean Difference probe shows considerably greater variability across LLMs than sAwMIL (Fig. A25). While sAwMIL yields consistent decision boundary rotation corresponding to specific perturbations (see Fig. A17, particularly the Synthetic perturbation),

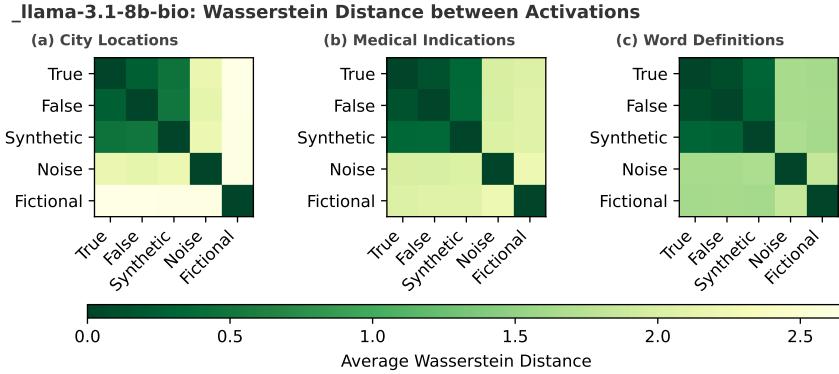


Figure A4. Wasserstein distance between activations for `_llama-3-8b-bio`. Pairwise Wasserstein distances between activation distributions of True, False, Synthetic, Fictional, and Noise statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. Synthetic statements are represented similarly to True and False statements, while Fictional statements and Noise are represented distinctly from all other statements.

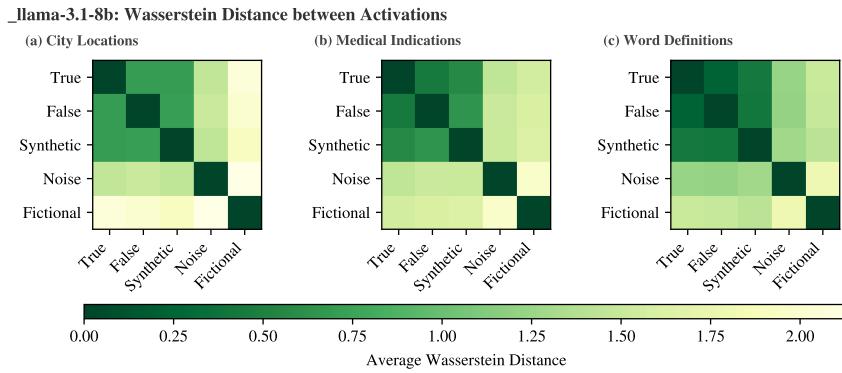


Figure A5. Wasserstein distance between activations for `_llama-3.1-8b`. Pairwise Wasserstein distances between activation distributions of True, False, Synthetic, Fictional, and Noise statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. Synthetic statements are represented similarly to True and False statements, while Fictional statements and Noise are represented distinctly from all other statements.

the **Mean Difference** probe exhibits near-orthogonal boundary shifts for certain LLMs regardless of perturbation. In addition, Table A24 shows that, unlike with **sAwMIL**, the **Fictional** perturbation produces the largest epistemic retractions across all three datasets, and the Word Definitions dataset exhibits the fewest total retractions. We note, however, that the **Synthetic** perturbation still produces the most epistemic expansions across domains, consistent with the **sAwMIL** results. We interpret these discrepancies as artifacts of the **Mean Difference** probe's reliance on dataset centroids: when statement activations are well separated, as with **Fictional** statements, class-label perturbations can induce disproportionately large changes in the estimated decision boundary. This instability reflects probe sensitivity rather than genuine representational instability in the LLMs. Accordingly, the **Mean Difference** probe is less well suited for quantifying stability within P-Stat than **sAwMIL**.

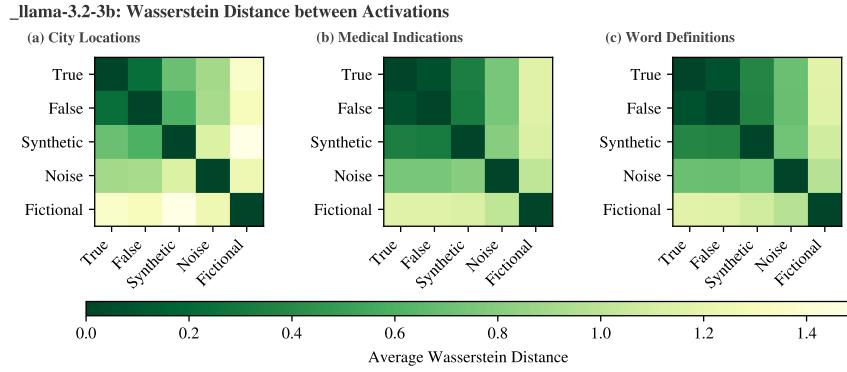


Figure A6. Wasserstein distance between activations for `_llama-3.2-3b`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. `Synthetic` statements are represented similarly to `True` and `False` statements, while `Fictional` statements and `Noise` are represented distinctly from all other statements.

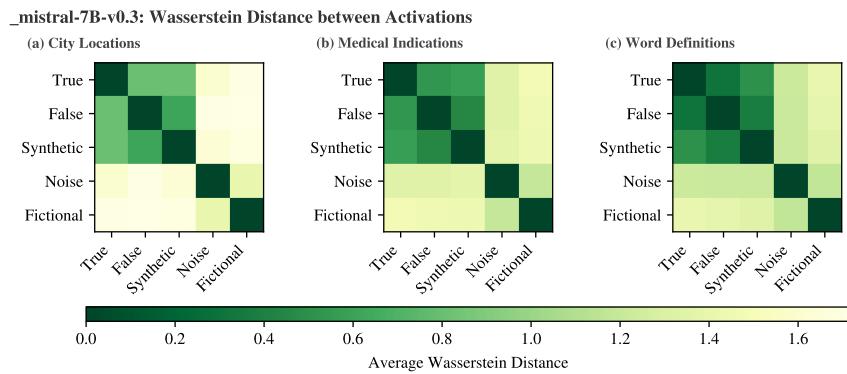


Figure A7. Wasserstein distance between activations for `_mistral-7B-v0.3`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. `Synthetic` statements are represented similarly to `True` and `False` statements, while `Fictional` statements and `Noise` are represented distinctly from all other statements.

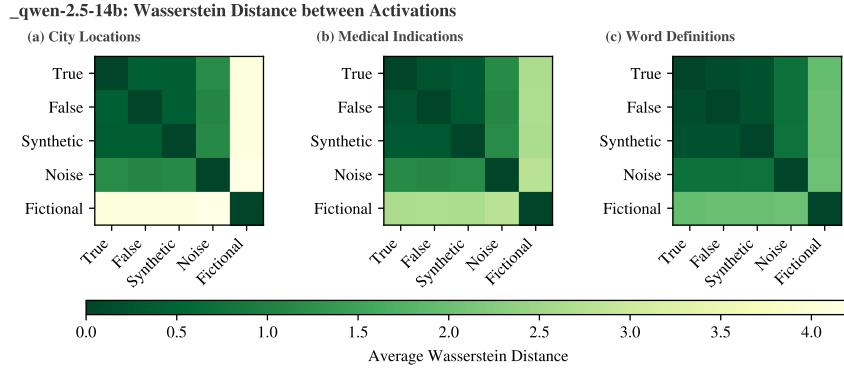


Figure A8. Wasserstein distance between activations for `_qwen-2.5-14b`. Pairwise Wasserstein distances between activation distributions of True, False, Synthetic, Fictional, and Noise statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. Synthetic statements are represented similarly to True and False statements, while Fictional statements are represented distinctly from all other statements.

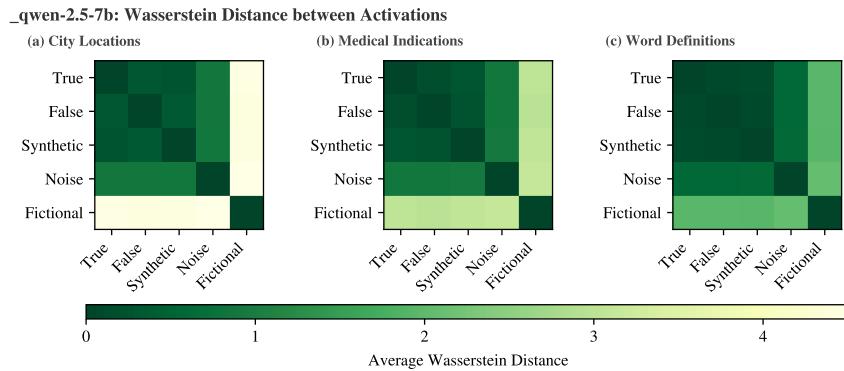


Figure A9. Wasserstein distance between activations for `_qwen-2.5-7b`. Pairwise Wasserstein distances between activation distributions of True, False, Synthetic, Fictional, and Noise statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. Synthetic statements are represented similarly to True and False statements, while Fictional statements are represented distinctly from all other statements.

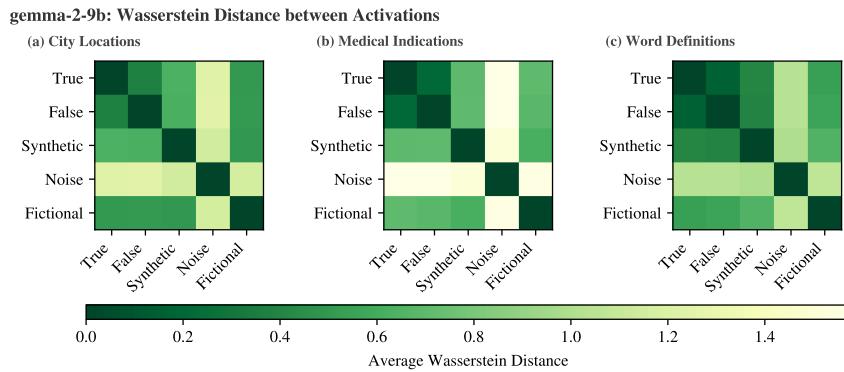


Figure A10. Wasserstein distance between activations for `gemma-2-9b`. Pairwise Wasserstein distances between activation distributions of True, False, Synthetic, Fictional, and Noise statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. Noise has distinct representations, but Fictional and Synthetic statements are represented similarly to True and False statements and each other.

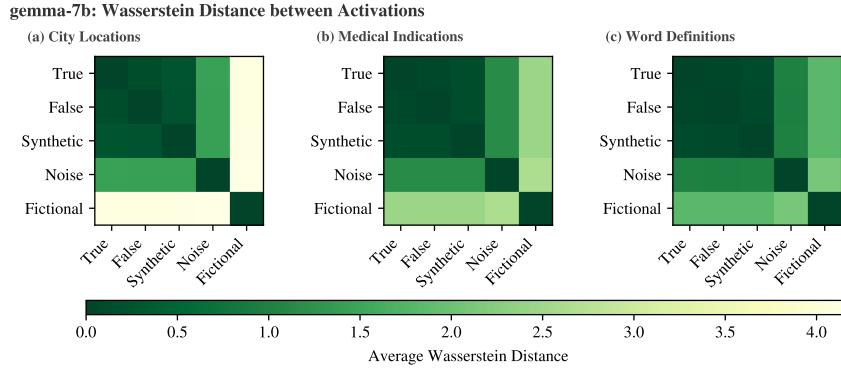


Figure A11. Wasserstein distance between activations for gemma-7b. Pairwise Wasserstein distances between activation distributions of True, False, Synthetic, Fictional, and Noise statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. Synthetic statements are represented similarly to True and False statements, while Fictional statements are represented distinctly from all other statements.

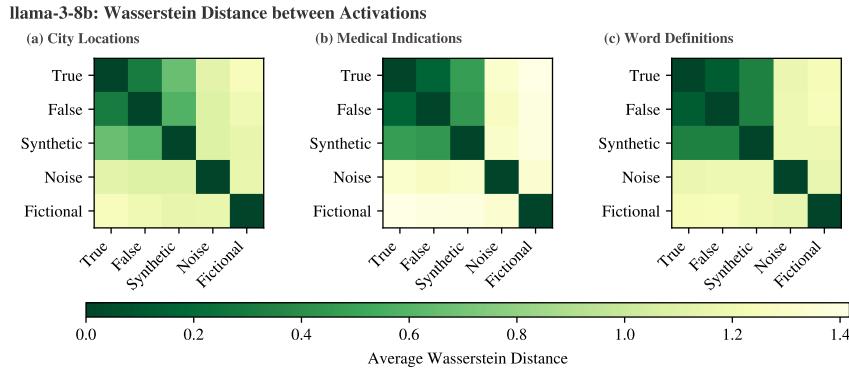


Figure A12. Wasserstein distance between activations for llama-3-8b. Pairwise Wasserstein distances between activation distributions of True, False, Synthetic, Fictional, and Noise statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. Synthetic statements are represented similarly to True and False statements, while Fictional statements and Noise are represented distinctly from all other statements.

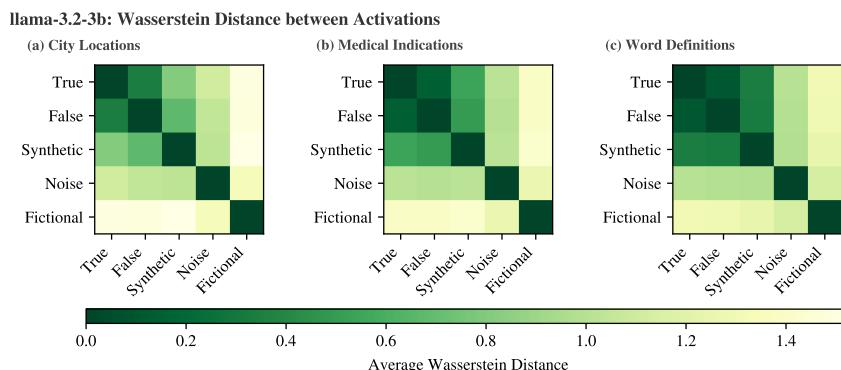


Figure A13. Wasserstein distance between activations for llama-3.2-3b. Pairwise Wasserstein distances between activation distributions of True, False, Synthetic, Fictional, and Noise statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. Synthetic statements are represented similarly to True and False statements, while Fictional statements and Noise are represented distinctly from all other statements.

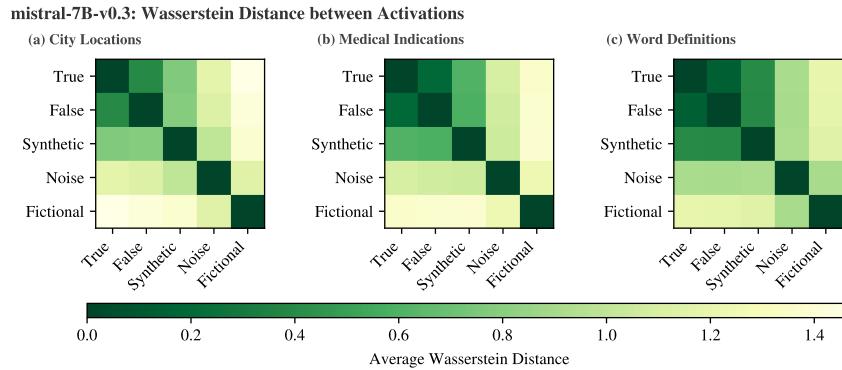


Figure A14. Wasserstein distance between activations for `mistral-7B-v0.3`. Pairwise Wasserstein distances between activation distributions of True, False, Synthetic, Fictional, and Noise statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. Synthetic statements are represented similarly to True and False statements, while Fictional statements and Noise are represented distinctly from all other statements.

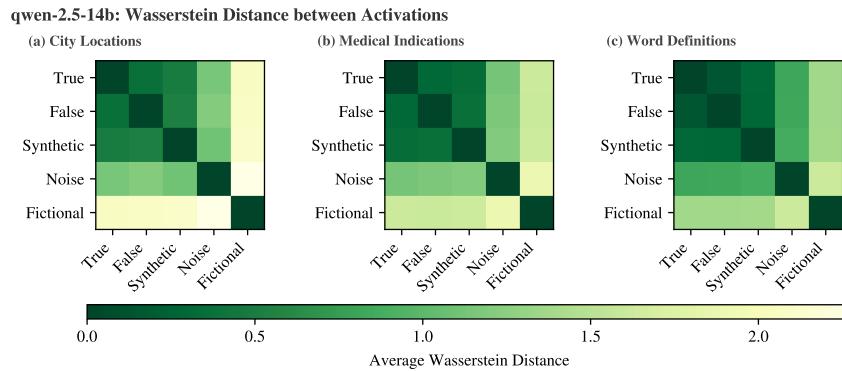


Figure A15. Wasserstein distance between activations for `qwen-2.5-14b`. Pairwise Wasserstein distances between activation distributions of True, False, Synthetic, Fictional, and Noise statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions. Synthetic statements are represented similarly to True and False statements, while Fictional statements are represented distinctly from all other statements.

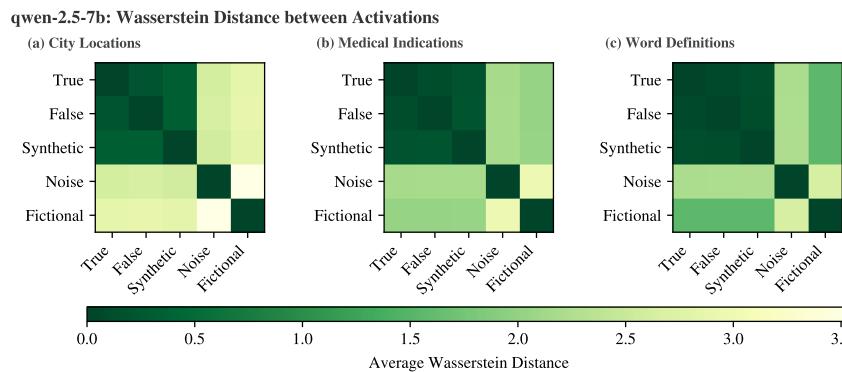


Figure A16. Wasserstein distance between activations for `qwen-2.5-7b`. Pairwise Wasserstein distances between activation distributions of True, False, Synthetic, Fictional, and Noise statements for the (a) City Locations, (b) Medical Indications, and (c) Word Definitions datasets. Synthetic statements are represented similarly to True and False statements, while Fictional statements and Noise are represented distinctly from all other statements.

Change in sAwMIL Decision Boundary under Perturbations

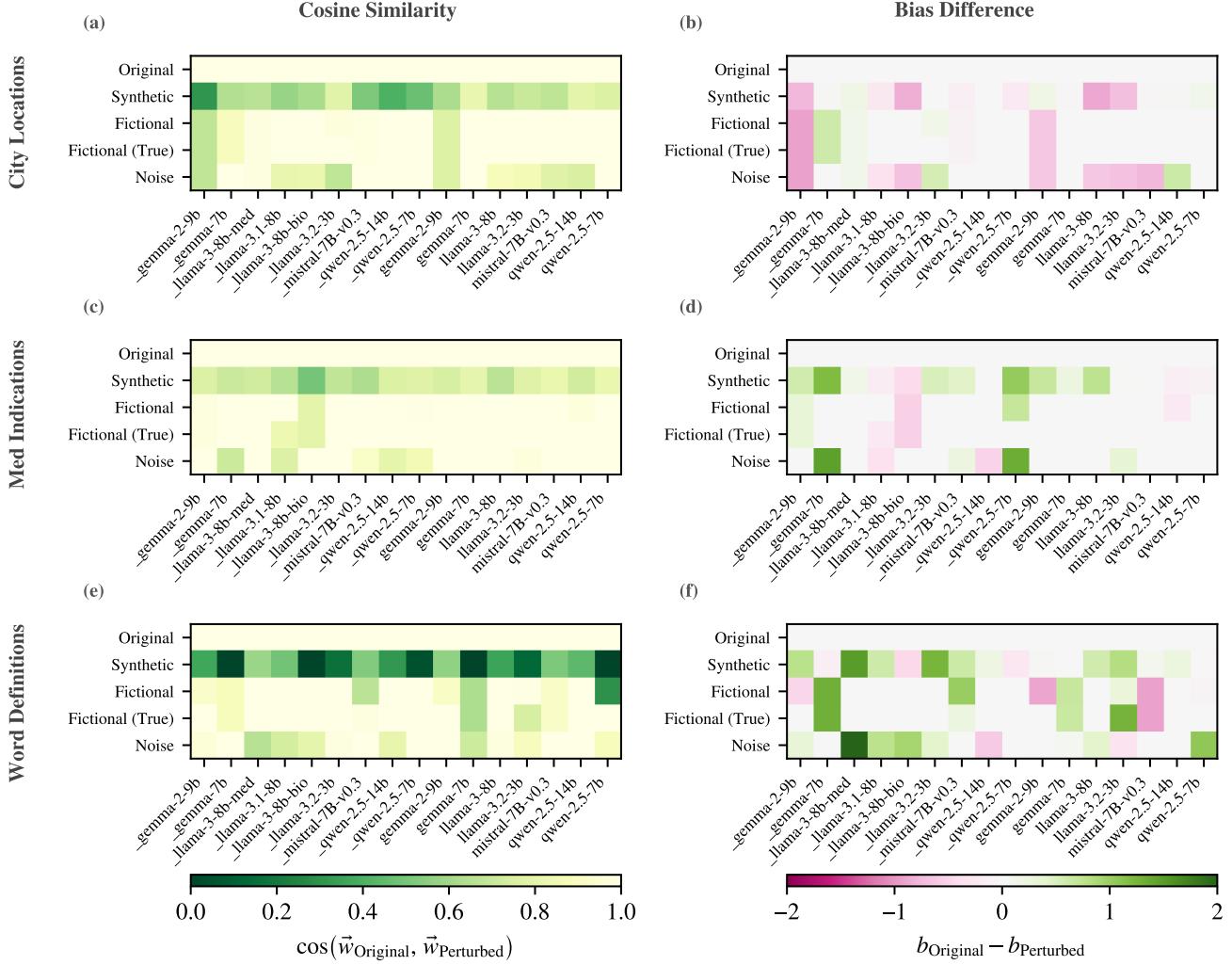


Figure A17. Changes in the probe decision boundary under perturbations. Cosine similarity (left column) and bias difference (right column) between the baseline True vs. Not True probe and probes retrained under label perturbations for the (a,b) City Locations, (c,d) Medical Indications, and (e,f) Word Definitions datasets. Each heatmap shows results for sixteen LLMs (columns) and five perturbation conditions (rows). LLMs with leading underscores are Chat models, while those without are Base models. Higher cosine similarity indicates smaller rotations of the learned decision boundary, while bias difference reflects shifts in intercept. Across datasets, probes retrained with the Synthetic perturbation show the largest deviation from the original, particularly in cosine similarity.

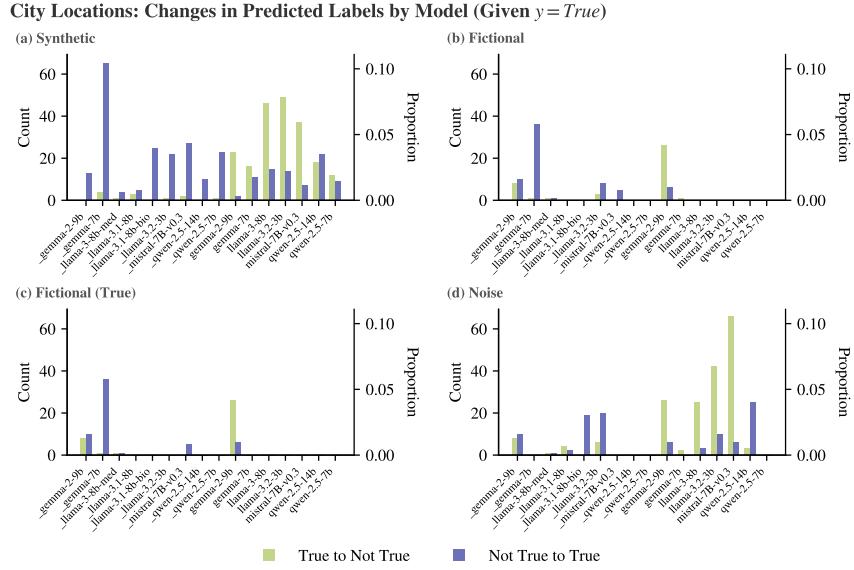


Figure A18. Stability of probe predictions under belief context perturbations for City Locations data. Bar plots show for each LLM (x-axis) how often the probe induces epistemic expansions and retractions when retrained under four perturbations: (a) Synthetic, (b) Fictional, (c) Fictional(T), and (d) Noise. Green bars indicate retractions (True to Not True), while purple bars indicate expansions (Not True to True). The left y-axis reports the number of statements with flipped predictions, and the right y-axis reports the corresponding proportions. The Base models exhibit more retractions than the Chat models, and the Chat models induce more expansions.

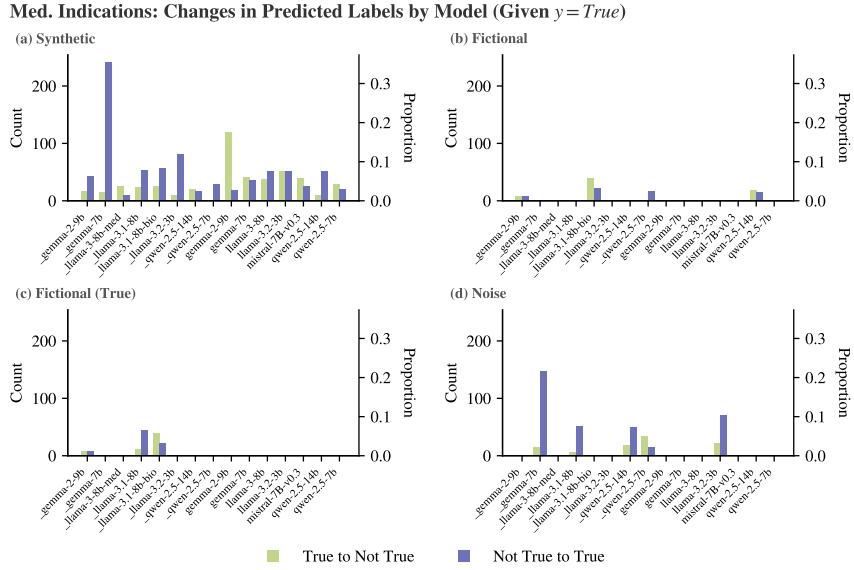


Figure A19. Stability of probe predictions under belief context perturbations for Medical Indications data. Bar plots show for each LLM (x-axis) how often the probe induces epistemic expansions and retractions when retrained under four perturbations: (a) Synthetic, (b) Fictional, (c) Fictional(T), and (d) Noise. Green bars indicate retractions (True to Not True), while purple bars indicate expansions (Not True to True). The left y-axis reports the number of statements with flipped predictions, and the right y-axis reports the corresponding proportions. The Synthetic perturbation leads to the most instability, and the Fictional and Fictional (T) perturbations result in almost no flips.

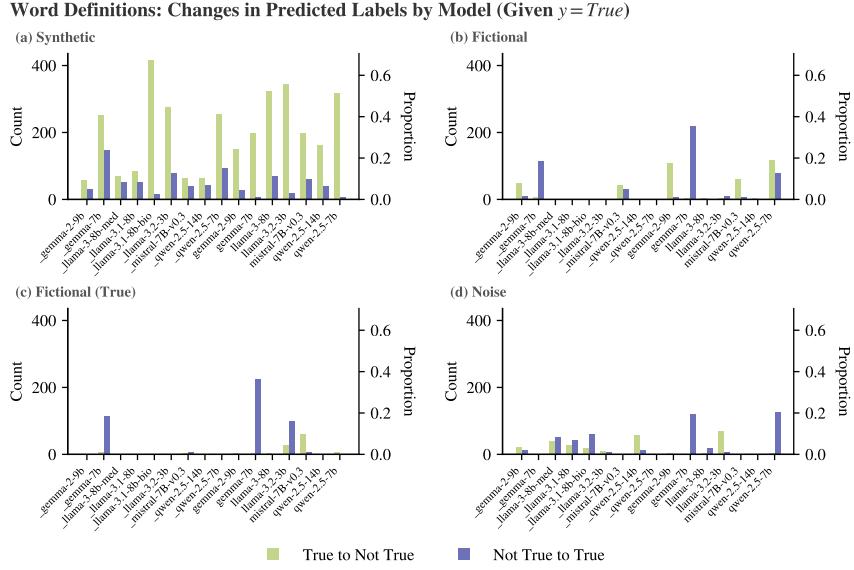


Figure A20. Stability of probe predictions under belief context perturbations for Word Definitions data. Bar plots show for each LLM (x-axis) how often the probe induces epistemic expansions and retractions when retrained under four perturbations: (a) Synthetic, (b) Fictional, (c) Fictional(T), and (d) Noise. Green bars indicate retractions (True to Not True), while purple bars indicate expansions (Not True to True). The left y-axis reports the number of statements with flipped predictions, and the right y-axis reports the corresponding proportions. The Synthetic perturbation leads to the most instability, with some LLMs retracting over 50% of their originally True statements.

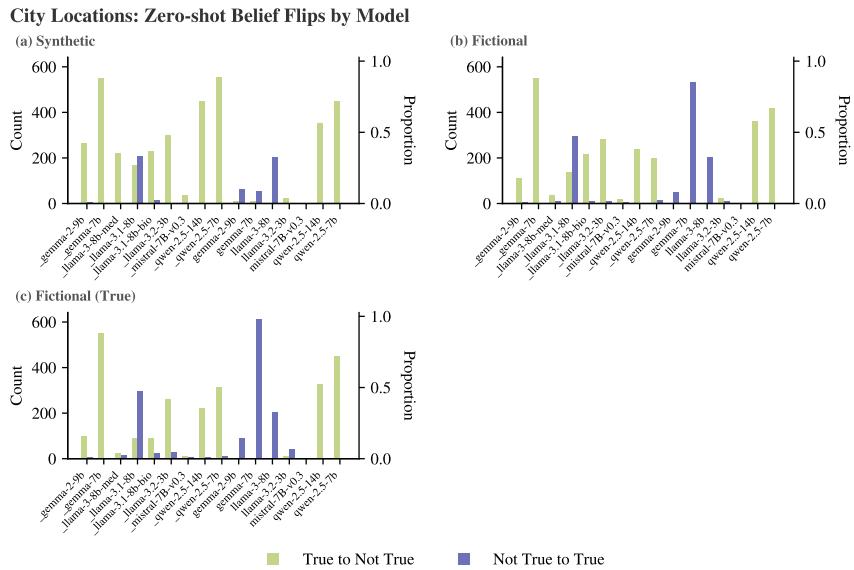


Figure A21. Stability of zero-shot beliefs under belief context perturbations for City Locations data. Bar plots show for each LLM (x-axis) how often the zero-shot instantiation induced epistemic expansions and retractions when retrained under three perturbations: (a) Synthetic, (b) Fictional, and (c) Fictional(T). Green bars indicate retractions (True to Not True), while purple bars indicate expansions (Not True to True). The left y-axis reports the number of statements with flipped predictions, and the right y-axis reports the corresponding proportions. The Chat models exhibit more retractions than the Base models.

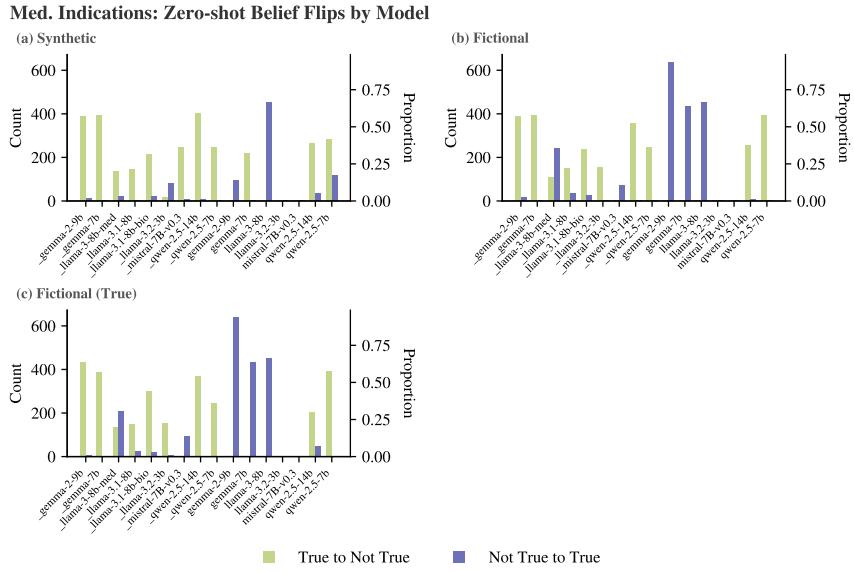


Figure A22. Stability of zero-shot beliefs under belief context perturbations for Medical Indications data. Bar plots show for each LLM (x-axis), how often the zero-shot instantiation induced epistemic expansions and retractions when retrained under three perturbations: (a) Synthetic, (b) Fictional, and (c) Fictional(T). Green bars indicate retractions (True to Not True), while purple bars indicate expansions (Not True to True). The left y-axis reports the number of statements with flipped predictions, and the right y-axis reports the corresponding proportions. The Chat models exhibit more retractions, while the Base models exhibit more expansions.

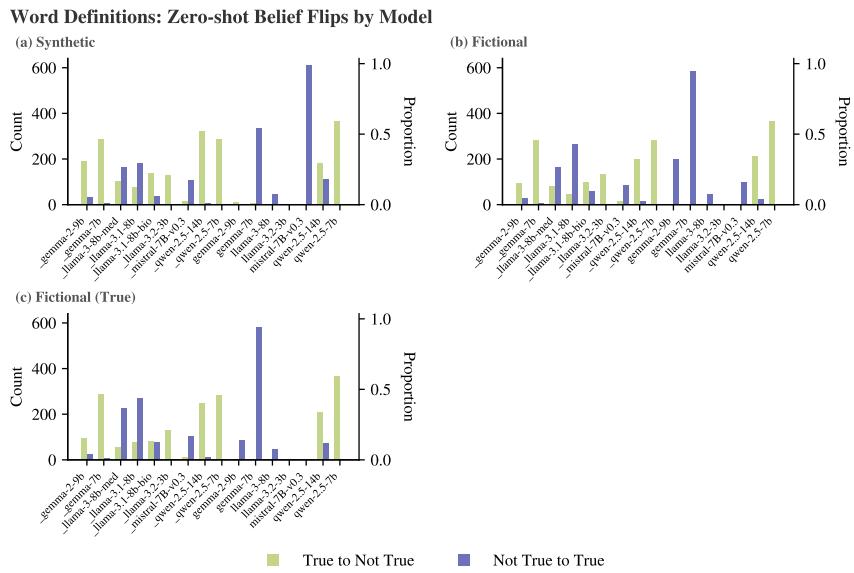


Figure A23. Stability of zero-shot beliefs under belief context perturbations for Word Definitions data. Bar plots show for each LLM (x-axis), how often the zero-shot instantiation induced epistemic expansions and retractions when retrained under three perturbations: (a) Synthetic, (b) Fictional, and (c) Fictional(T). Green bars indicate retractions (True to Not True), while purple bars indicate expansions (Not True to True). The left y-axis reports the number of statements with flipped predictions, and the right y-axis reports the corresponding proportions. The Chat models exhibit more retractions, while the Base models exhibit more expansions.

Dataset	Perturbation	True to True	Not True to Not True	Epistemic Expansions \mathcal{E}	Epistemic Retractions \mathcal{R}
City Locations	Synthetic	9724 (97.2)	167 (1.7)	63 (0.6)	46 (0.5)
	Fictional	7386 (73.9)	221 (2.2)	9 (0.1)	2384 (23.8)
	Fictional (T)	9680 (96.8)	184 (1.8)	46 (0.5)	90 (0.9)
	Noise	9653 (96.5)	201 (2.0)	29 (0.3)	117 (1.2)
Medical Locations	Synthetic	9457 (86.8)	894 (8.2)	221 (2.0)	324 (3.0)
	Fictional	4694 (43.1)	977 (9.0)	138 (1.3)	5087 (46.7)
	Fictional (T)	9718 (89.2)	1069 (9.8)	46 (0.4)	63 (0.6)
	Noise	8955 (82.1)	1026 (9.4)	89 (0.8)	826 (7.6)
Word Definitions	Synthetic	8468 (85.6)	500 (5.1)	695 (7.0)	225 (2.3)
	Fictional	7035 (71.1)	1175 (11.9)	20 (0.2)	1658 (16.8)
	Fictional (T)	8282 (83.8)	1083 (11.0)	112 (1.1)	411 (4.2)
	Noise	8208 (83.0)	1178 (11.9)	17 (0.2)	485 (4.9)

Figure A24. Epistemic expansions \mathcal{E} and retractions \mathcal{R} under probing label perturbations for the Mean Difference Probe. Counts (and percentages) of beliefs that remain stable or lead to expansions \mathcal{E} and retractions \mathcal{R} across Synthetic (yellow), Fictional (gray), Fictional(T) (blue), and Noise (red) perturbations for each dataset. While the Fictional perturbation induces the most epistemic retractions \mathcal{R} , the Synthetic perturbation induces the most epistemic expansions \mathcal{E} .

Change in Mean Difference Decision Boundary under Perturbations

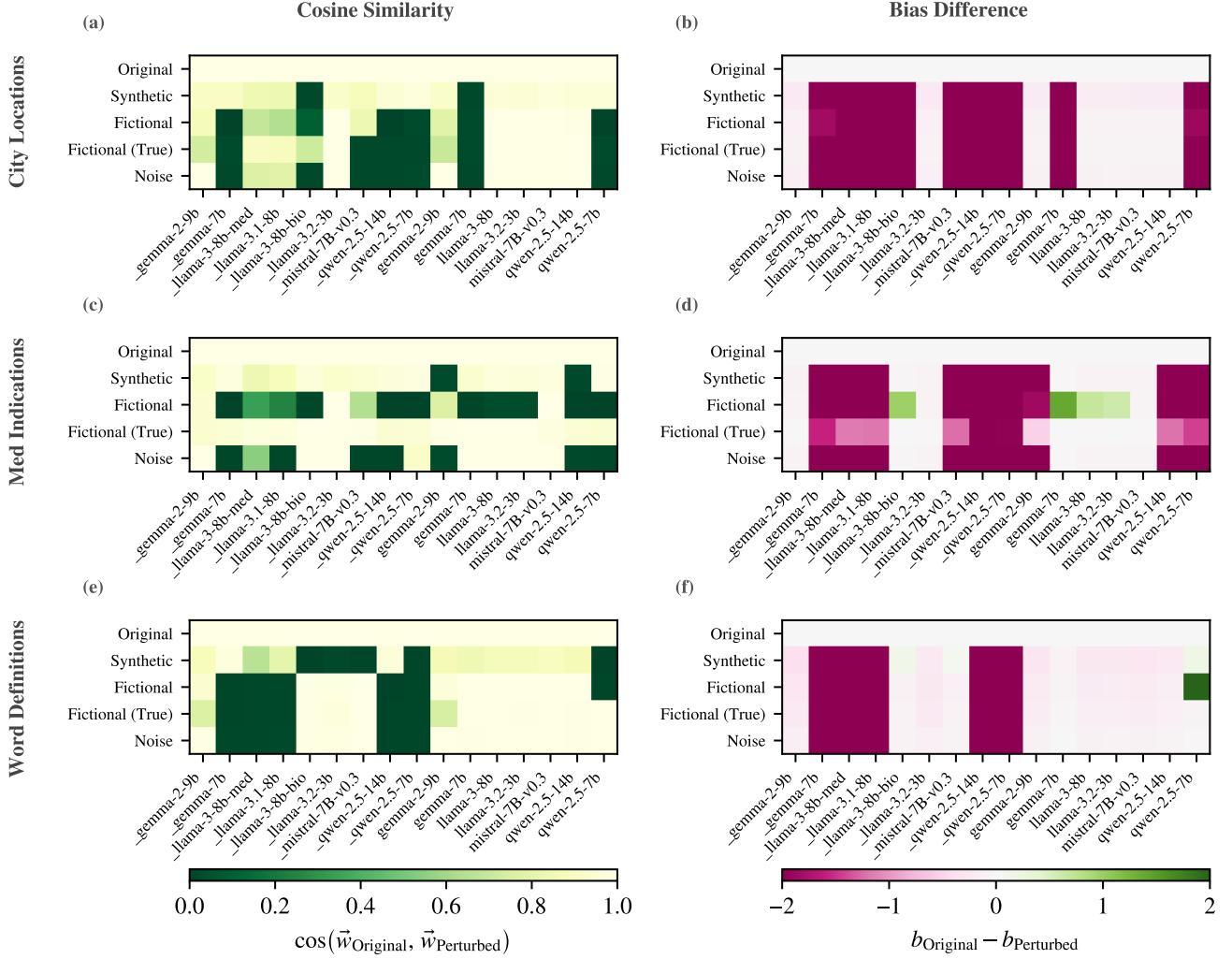


Figure A25. Change in Mean Difference decision boundaries under perturbations. Cosine similarity (left column) and bias difference (right column) between the baseline True vs. Not True probe and probes retrained under label perturbations for the (a,b) City Locations, (c,d) Medical Indications, and (e,f) Word Definitions datasets. Each heatmap shows results for sixteen LLMs (columns) and five perturbation conditions (rows). LLMs with leading underscores are Chat models, while those without are Base models. Higher cosine similarity indicates smaller rotations of the learned decision boundary, while bias difference reflects shifts in intercept. Certain LLMs lead to near orthogonal perturbed decision boundaries across all perturbation types, suggesting that, unlike sAwMIL, the probe is highly sensitive to differences in the distributions of the underlying activations.