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 reticers. **(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)** Snozzberry 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., $\mathcal{N}_{\text{unf}}$, while `Fictional` statements are familiar `Neither` statements $\mathcal{N}_{\text{fam}}$. A version of this table without the `Fictional` and `Noise` columns can be found in [9].

Condition	$\mathcal{N}_\Theta$	Probing: Training Labels	Zero-shot: Belief Context C_Θ
Baseline	$\emptyset$	`True` vs. `False`+`Synthetic`+`Fictional`+`Noise`	None
Synthetic	$\mathcal{N}_{\text{unf}}$	`True`+`Synthetic` vs. `False`+`Fictional`+`Noise`	`Synthetic`
Fictional	$\mathcal{N}_{\text{fam}}$	`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 $\mathcal{N}_\Theta$, the `Neither` statements treated as compatible with truth. The same Θ is instantiated (i) representationally by retraining a probe with labels induced by $\mathcal{N}_\Theta$ 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 $\phi_{\mathcal{M},l}(s)$. 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 $\mathcal{N}_\Theta$ 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_Θ]
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

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

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

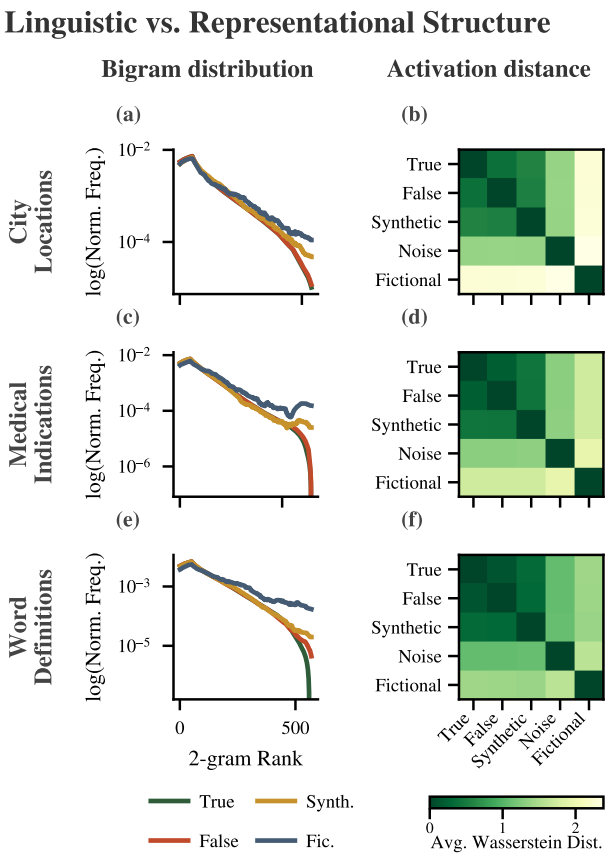

Figure 2. Linguistic vs. representational structure of `Neither` statements. For **(a),(b)** City Locations, **(c),(d)** Medical Indications, and **(e),(f)** Word Definitions, the left column shows normalized character bigram rank–frequency curves for `True` (green), `False` (red), `Synthetic` (yellow), 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,

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

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

[4]We 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**. Although 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.