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].

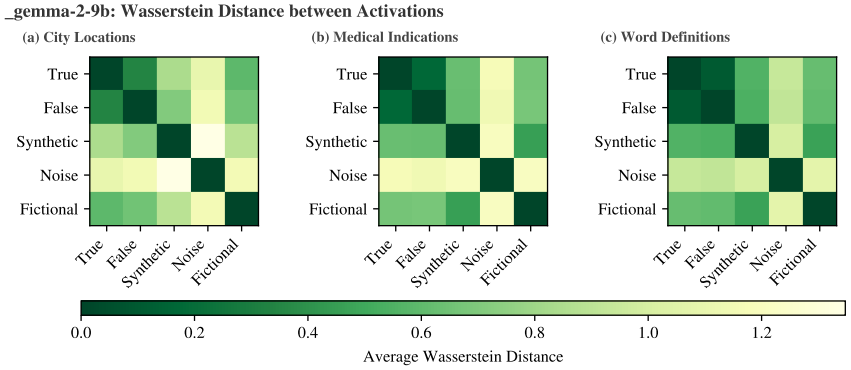

Figure A1. Wasserstein distance between activations for `_gemma-2-9b`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets.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

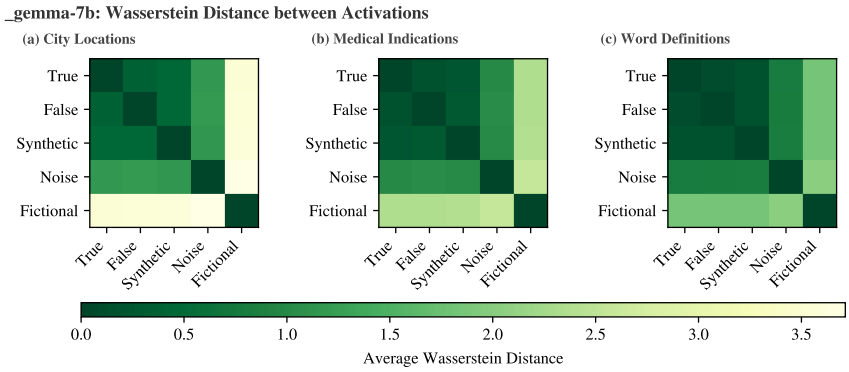

Figure A2. Wasserstein distance between activations for `_gemma-7b`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets.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

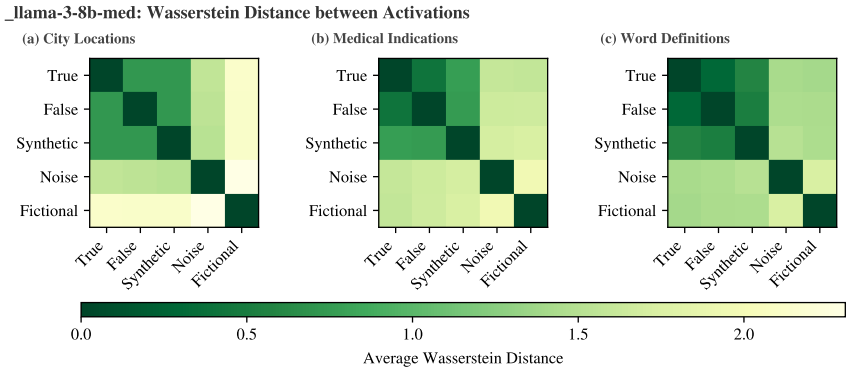

Figure A3. Wasserstein distance between activations for `_llama-3-8b-med`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets.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

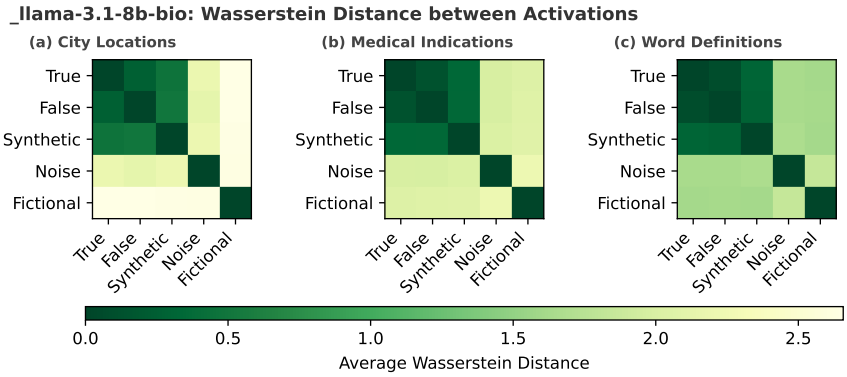

Figure A4. Wasserstein distance between activations for `_llama-3-8b-bio`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets.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

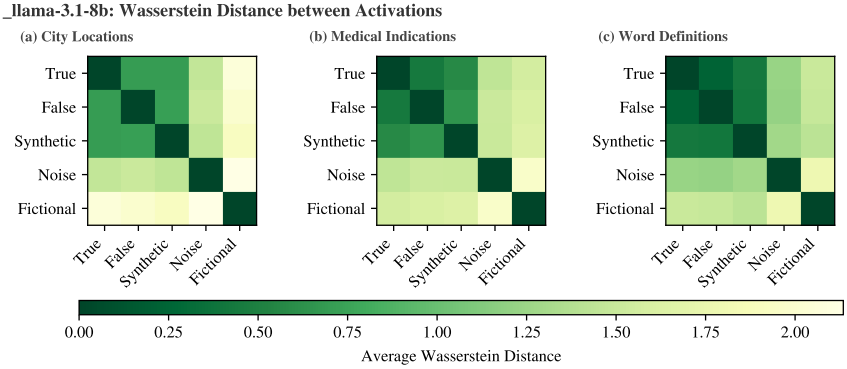

Figure A5. Wasserstein distance between activations for `_llama-3.1-8b`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets. `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`.