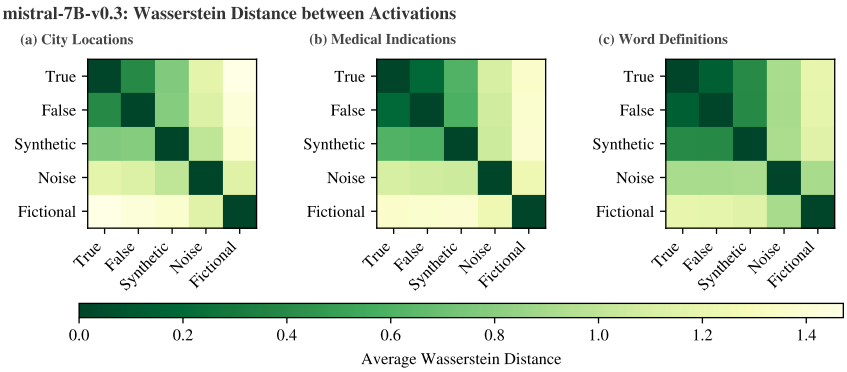

Figure A14. Wasserstein distance between activations for `mistral-7B-v0.3`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets.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

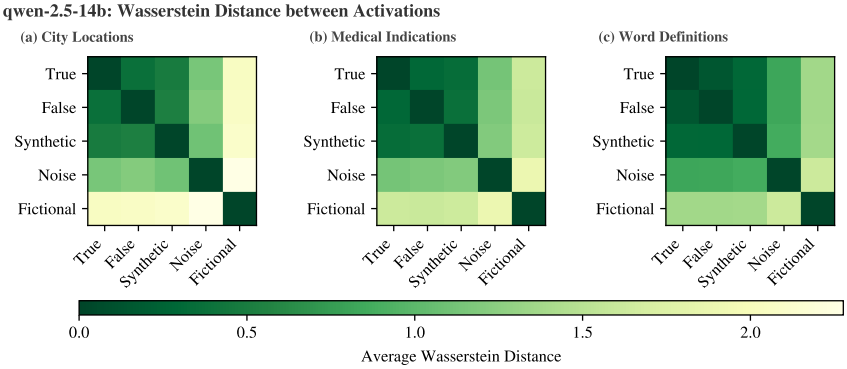

Figure A15. Wasserstein distance between activations for `qwen-2.5-14b`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions. `Synthetic`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

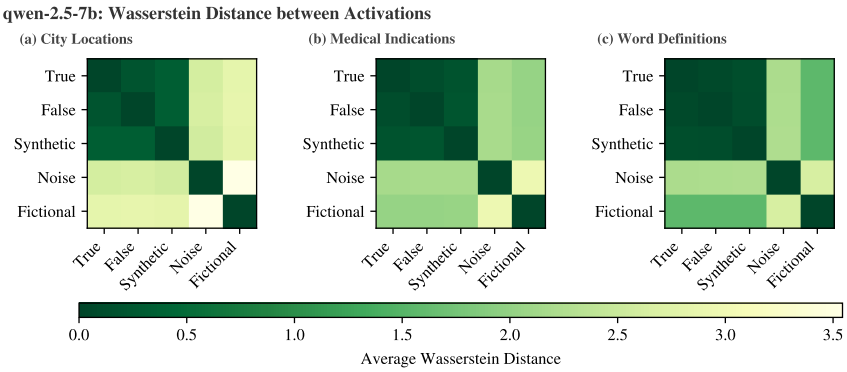

Figure A16. Wasserstein distance between activations for `qwen-2.5-7b`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets.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

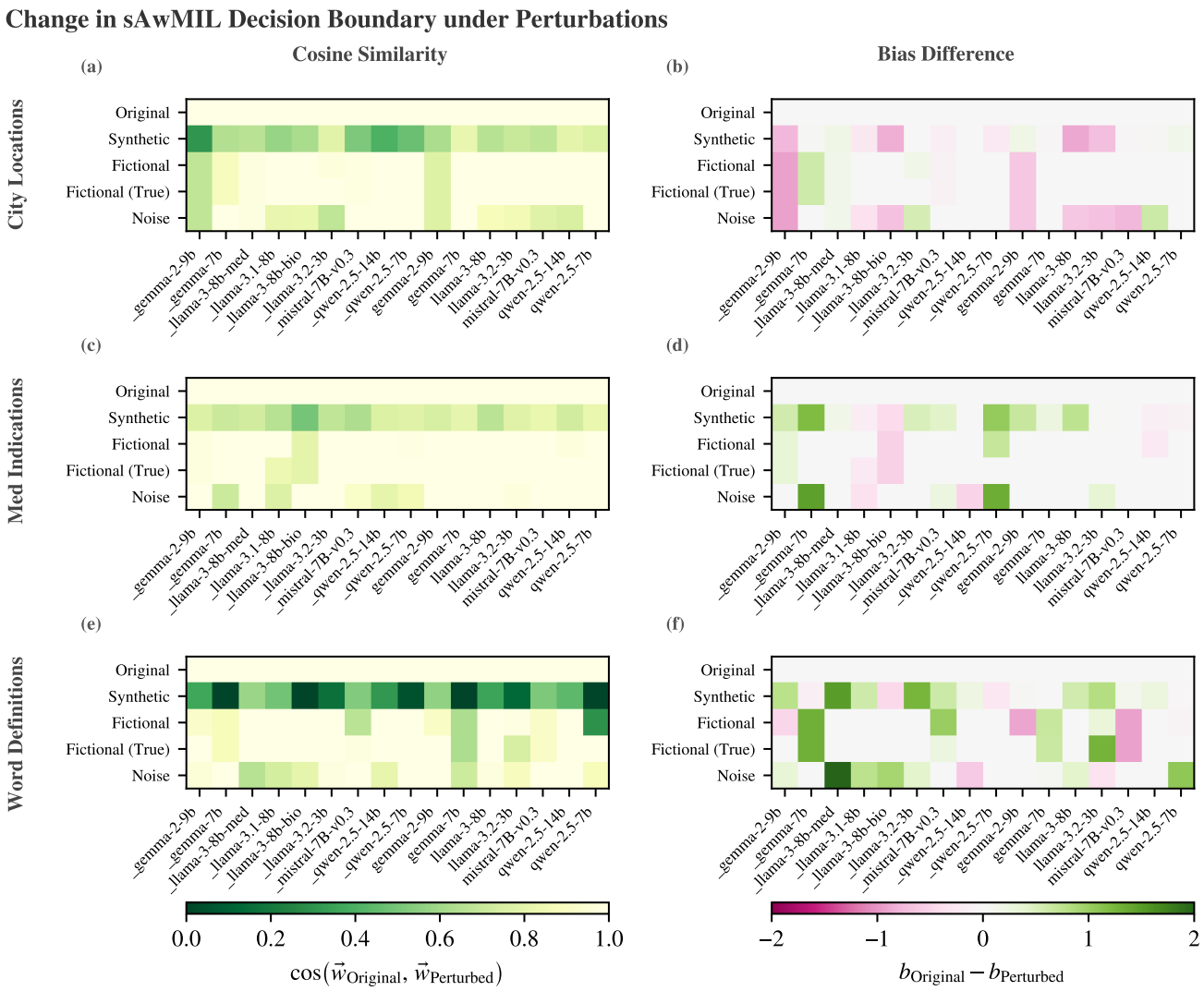

Figure A17. Changes in the probe decision boundary under perturbations. Cosine similarity (left column) and bias difference (right column) between the baseline `True` vs. `Not True` probe and probes retrained under label perturbations for the **(a,b)** City Locations, **(c,d)** Medical Indications, and **(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.

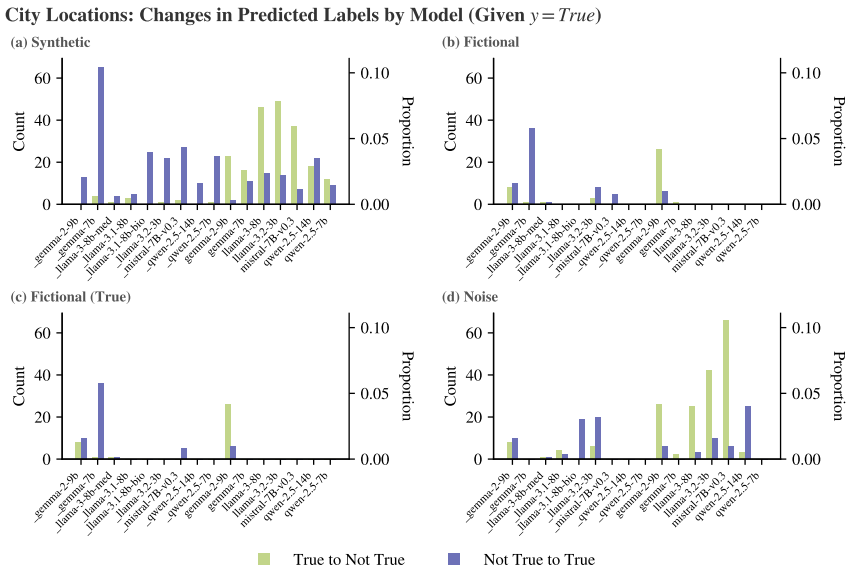

Figure A18. Stability of probe predictions under belief context perturbations for City Locations data. Bar plots show for each LLM (x-axis) how often the probe induces epistemic expansions and retractions when retrained under four perturbations: **(a)** `Synthetic`, **(b)** `Fictional`, **(c)** `Fictional(T)`, 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.

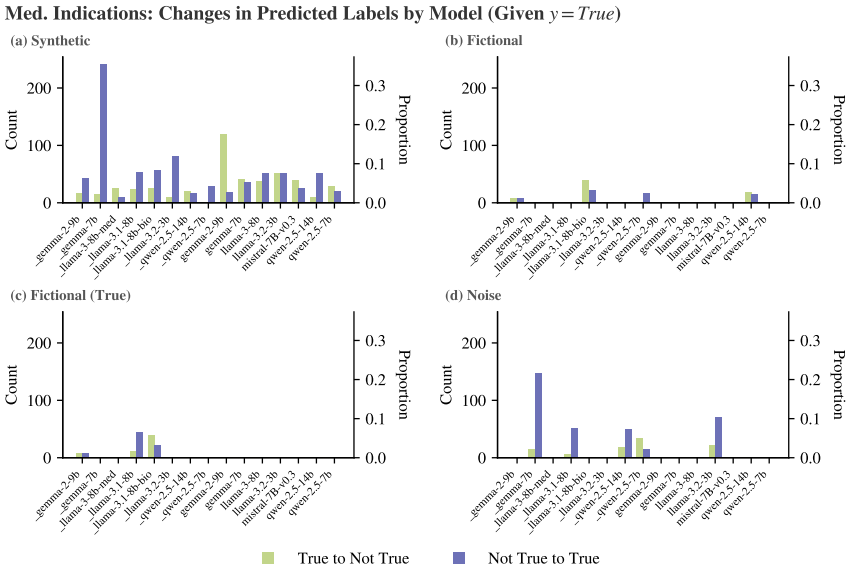

Figure A19. Stability of probe predictions under belief context perturbations for Medical Indications data. Bar plots show for each LLM (x-axis) how often the probe induces epistemic expansions and retractions when retrained under four perturbations: **(a)** `Synthetic`, **(b)** `Fictional`, **(c)** `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.