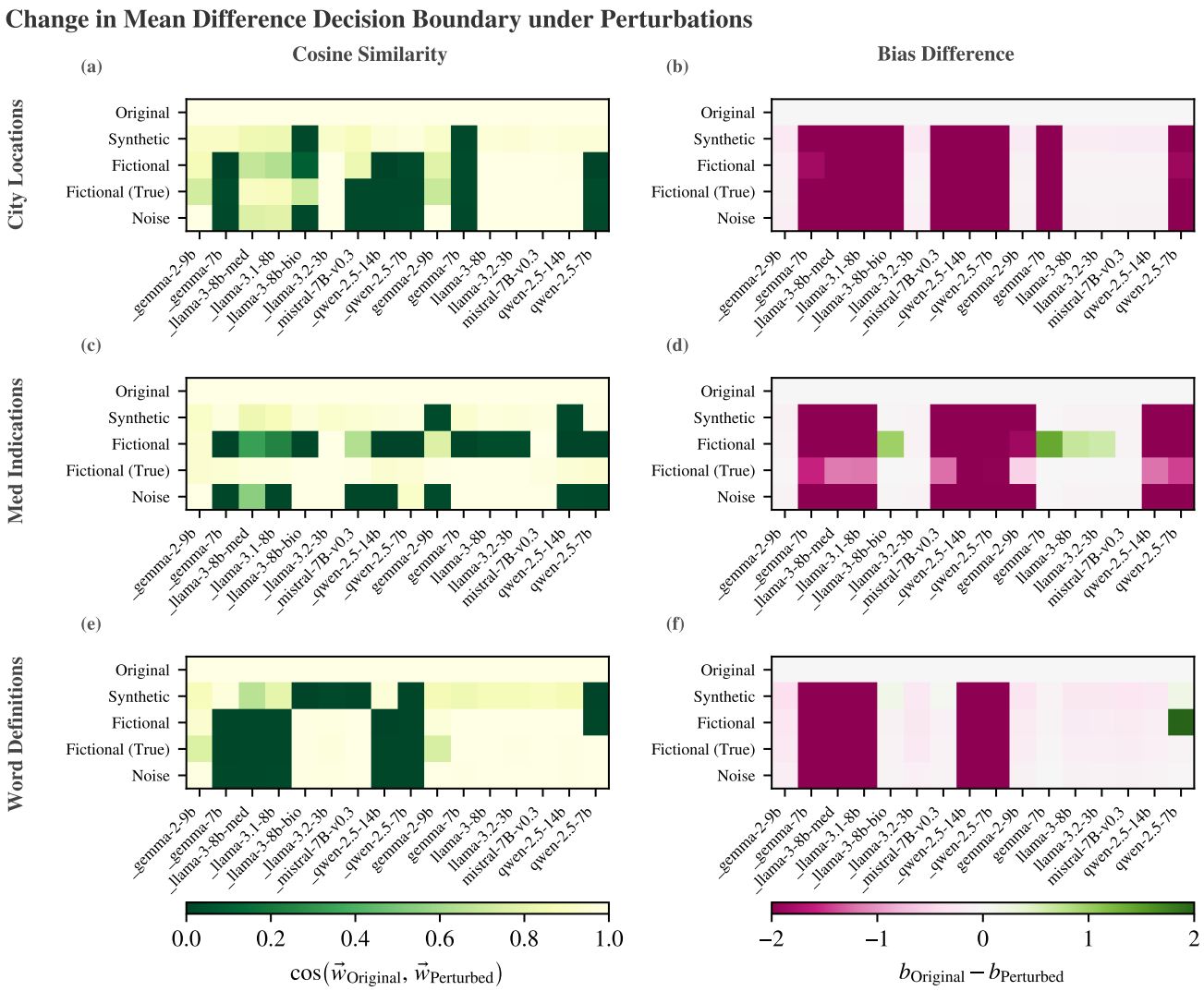

Figure A25. Change in `Mean Difference` decision boundaries under perturbations. Cosine similarity (left column) and bias difference (right column) between the baseline `True` vs. `Not True` probe and probes retrained under label perturbations for the **(a,b)** City Locations, **(c,d)** Medical Indications, 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.