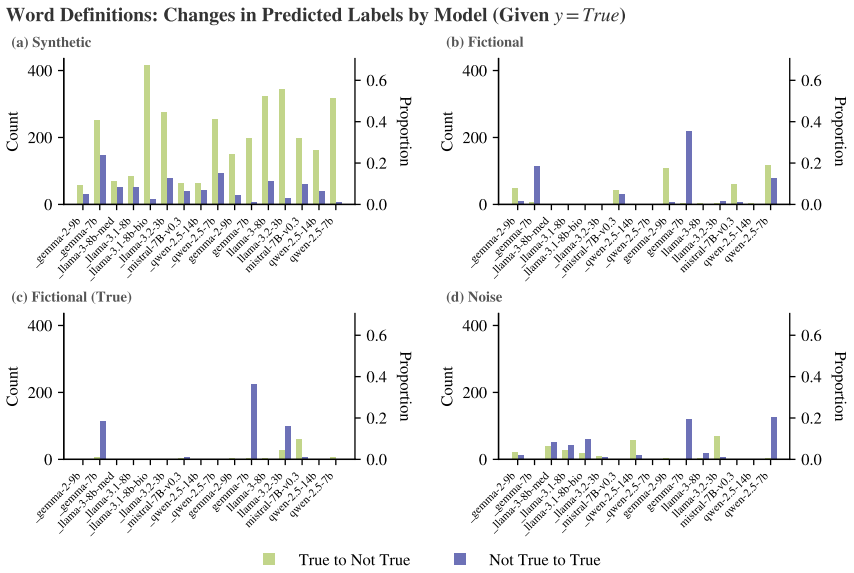

Figure A20. Stability of probe predictions under belief context perturbations for Word Definitions data. Bar plots show for each LLM (x-axis) how often the probe induces epistemic expansions and retractions when retrained under four perturbations: **(a)** `Synthetic`, **(b)** `Fictional`, **(c)** `Fictional(T)`, 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.

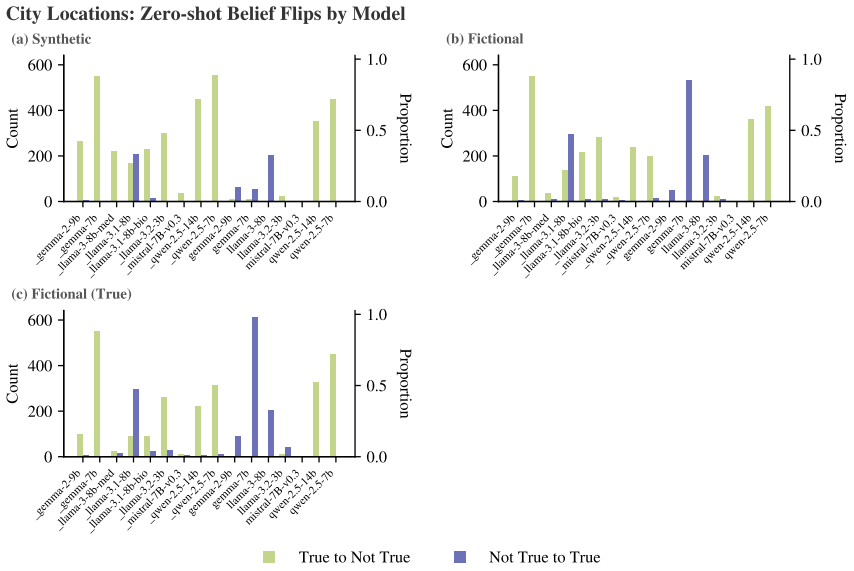

Figure A21. Stability of zero-shot beliefs under belief context perturbations for City Locations data. Bar plots show for each LLM (x-axis) how often the zero-shot instantiation induced epistemic expansions and retractions when retrained under three perturbations: **(a)** `Synthetic`, **(b)** `Fictional`, 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.

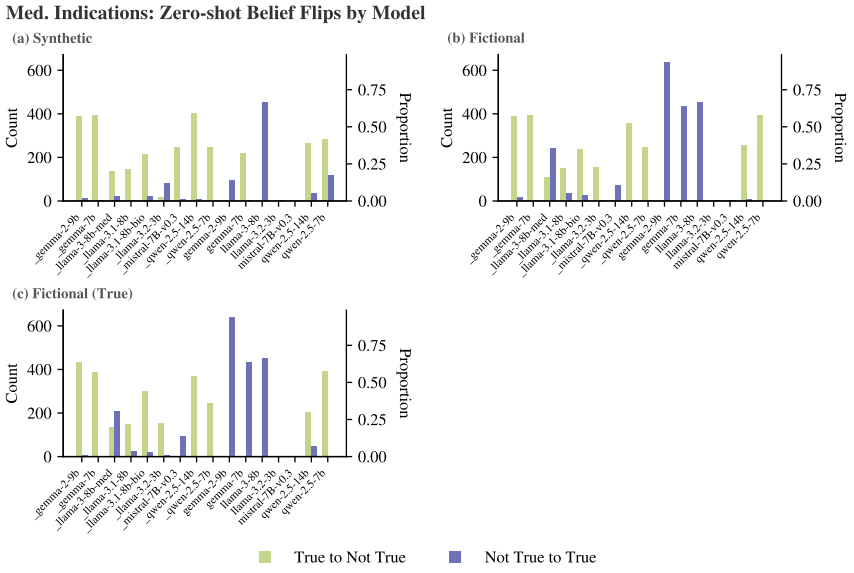

Figure A22. Stability of zero-shot beliefs under belief context perturbations for Medical Indications data. Bar plots show for each LLM (x-axis), how often the zero-shot instantiation induced epistemic expansions and retractions when retrained under three perturbations: **(a)** `Synthetic`, **(b)** `Fictional`, 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.

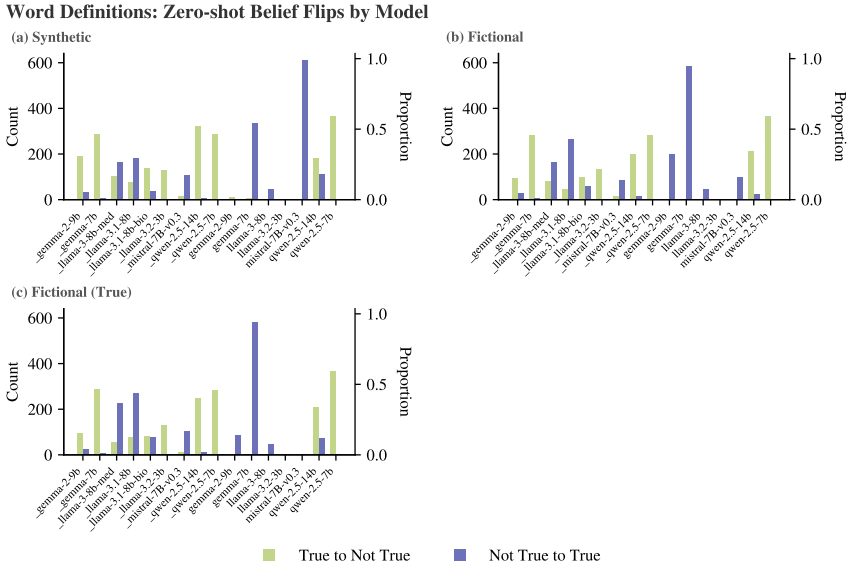

Figure A23. Stability of zero-shot beliefs under belief context perturbations for Word Definitions data. Bar plots show for each LLM (x-axis), how often the zero-shot instantiation induced epistemic expansions and retractions when retrained under three perturbations: **(a)** `Synthetic`, **(b)** `Fictional`, 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 $\mathcal{E}$.