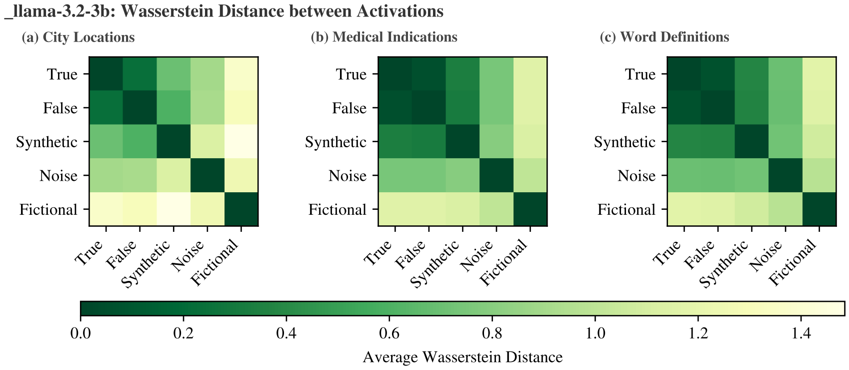

Figure A6. Wasserstein distance between activations for `_llama-3.2-3b`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets. `Synthetic` statements are represented similarly to `True` and `False` statements, while `Fictional` statements and `Noise` are represented distinctly from all other statements.

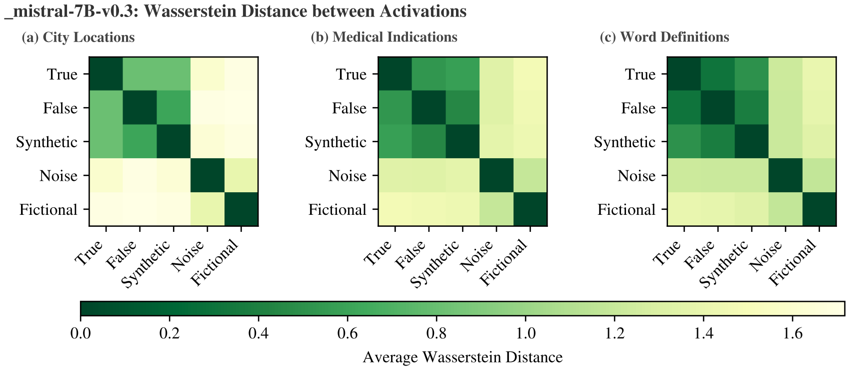

Figure A7. Wasserstein distance between activations for `_mistral-7B-v0.3`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets. `Synthetic` statements are represented similarly to `True` and `False` statements, while `Fictional` statements and `Noise` are represented distinctly from all other statements.

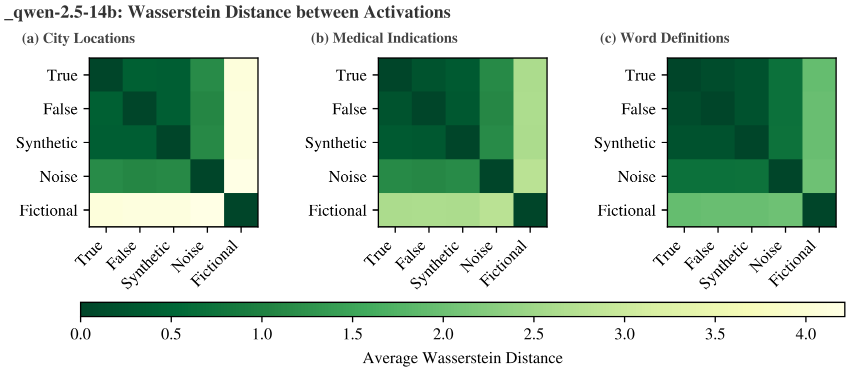

Figure A8. Wasserstein distance between activations for `_qwen-2.5-14b`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets. `Synthetic` statements are represented similarly to `True` and `False` statements, while `Fictional` statements are represented distinctly from all other statements.

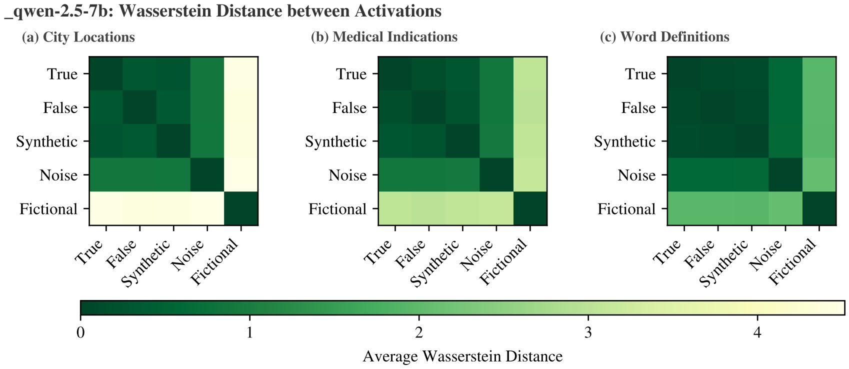

Figure A9. Wasserstein distance between activations for `_qwen-2.5-7b`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets. `Synthetic` statements are represented similarly to `True` and `False` statements, while `Fictional` statements are represented distinctly from all other statements.

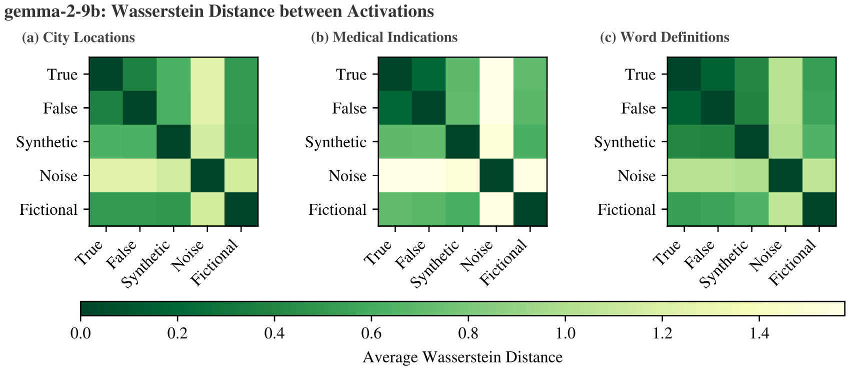

Figure A10. Wasserstein distance between activations for `gemma-2-9b`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets. `Noise` has distinct representations, but `Fictional` and `Synthetic` statements are represented similarly to `True` and `False` statements and each other.

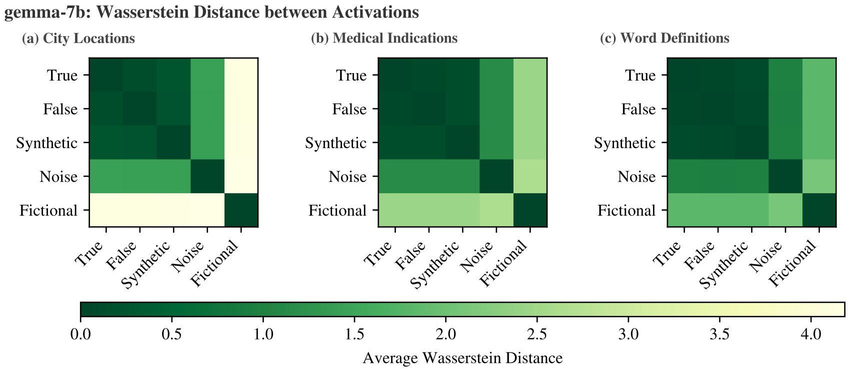

Figure A11. Wasserstein distance between activations for `gemma-7b`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets. `Synthetic` statements are represented similarly to `True` and `False` statements, while `Fictional` statements are represented distinctly from all other statements.

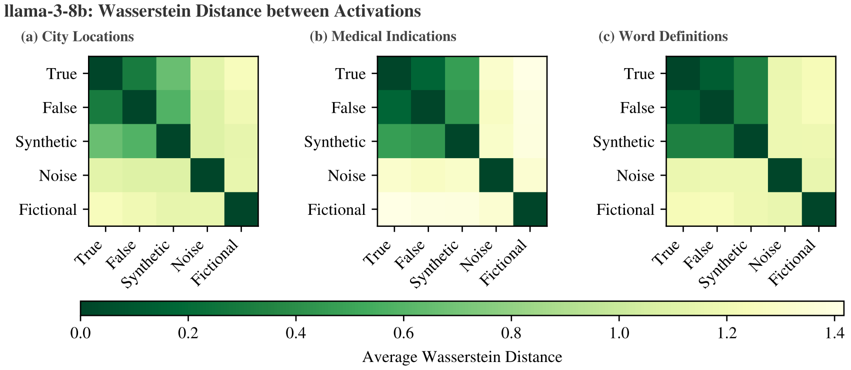

Figure A12. Wasserstein distance between activations for `llama-3-8b`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets. `Synthetic` statements are represented similarly to `True` and `False` statements, while `Fictional` statements and `Noise` are represented distinctly from all other statements.

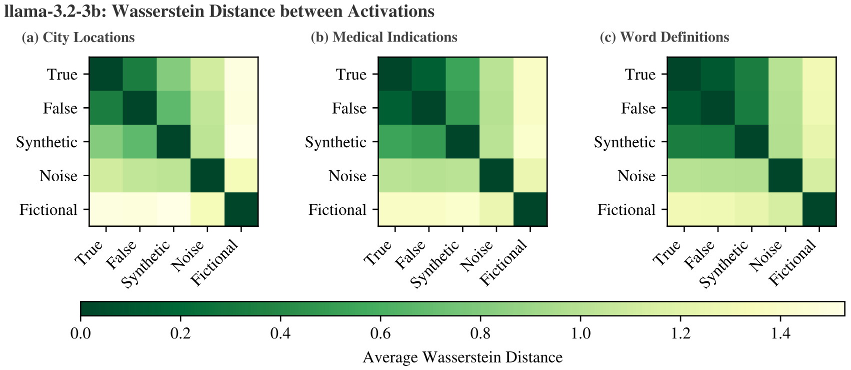

Figure A13. Wasserstein distance between activations for `llama-3.2-3b`. Pairwise Wasserstein distances between activation distributions of `True`, `False`, `Synthetic`, `Fictional`, and `Noise` statements for the **(a)** City Locations, **(b)** Medical Indications, and **(c)** Word Definitions datasets. `Synthetic` statements are represented similarly to `True` and `False` statements, while `Fictional` statements and `Noise` are represented distinctly from all other statements.