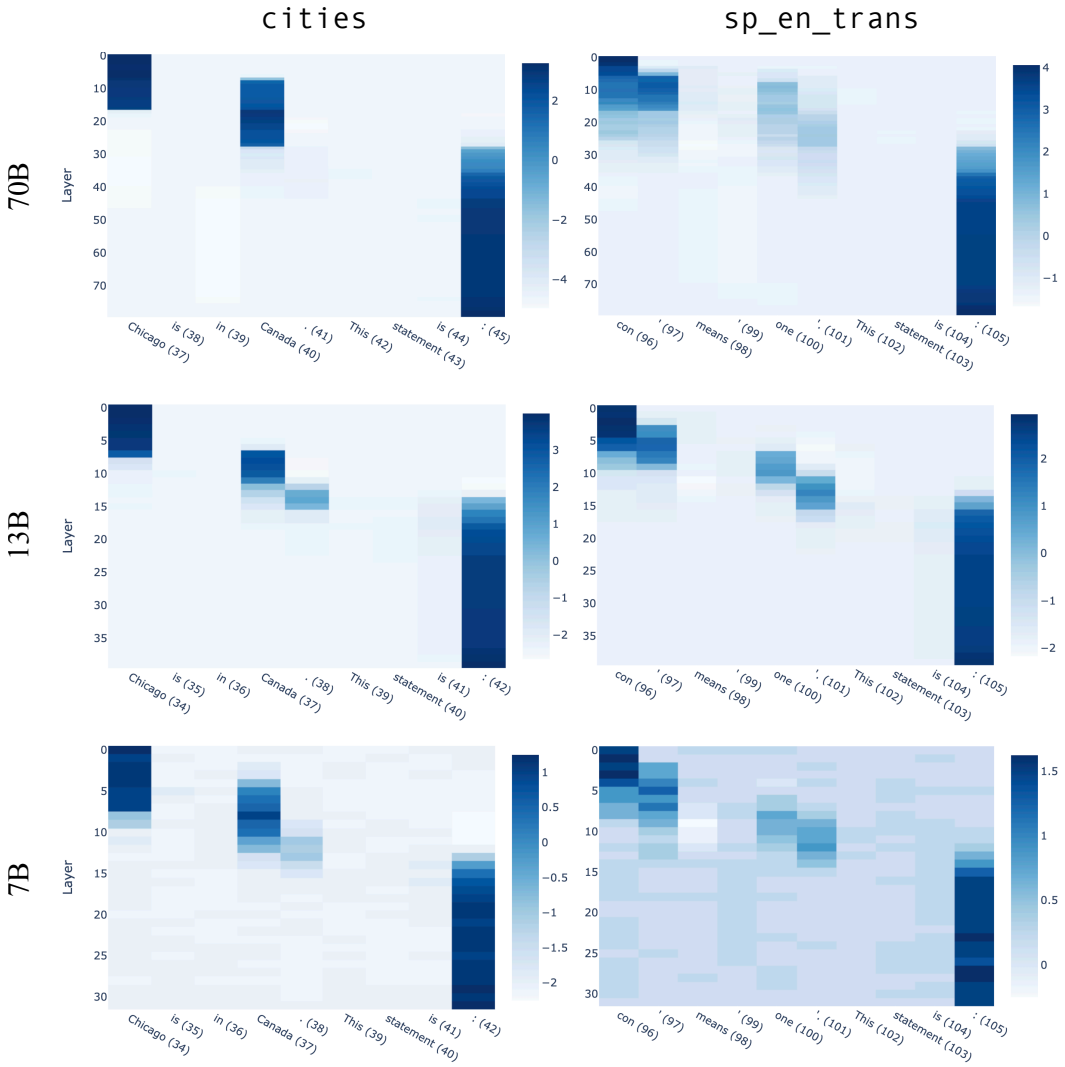

Figure 6: Full patching results across all three model sizes and inputs. Results are for patching false inputs (shown) to true by changing the first token shown on the left. Numbers in parentheses are the index of the token in the full (few-shot) prompt.

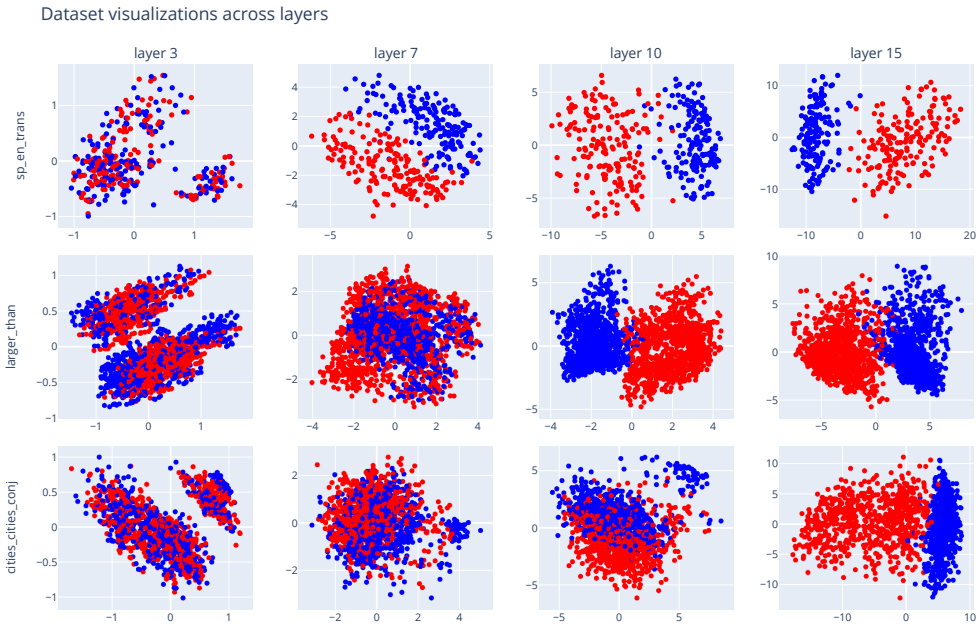

Figure 7: Projections of LLaMA-2-13B representations of datasets onto their top two PCs, across various layers.

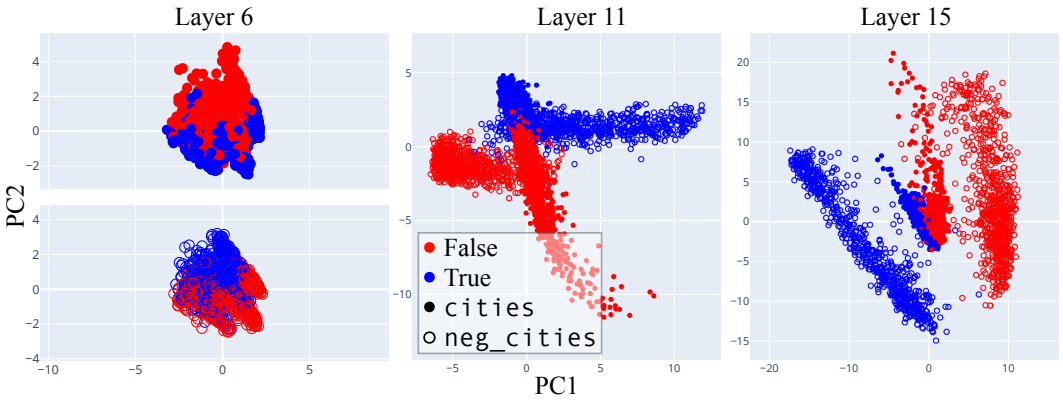

Figure 8: PCA visualizations of LLaMA-2-13B representations of `cities` and `neg_cities` at various layers.

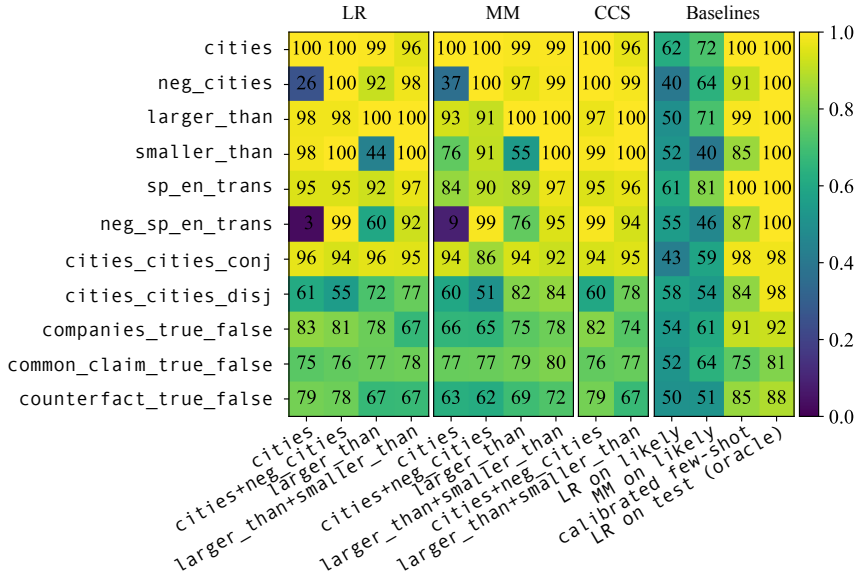

Figure 9: Generalization results for **LLaMA-2-70B**.

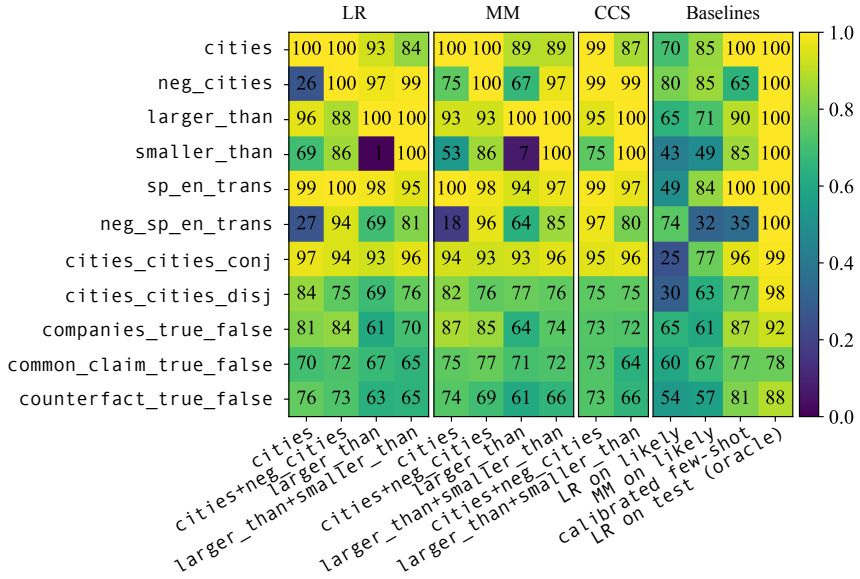

Figure 10: Generalization results for **LLaMA-2-13B**.

D Full generalization results

Here we present the full generalization results for probes trained on LLaMA-2-70B (Fig. 9), 13B (Fig. 10), and 7B (Fig. 11). The horizontal axis shows the training data for the probe and the vertical axis shows the test set.

E Mass-mean probing in terms of Mahalanobis whitening

One way to interpret the formula $p^{\text{iid}}_{\text{mm}}(\boldsymbol{x}) = \sigma(\boldsymbol{\theta}_{\text{mm}}^T \Sigma^{-1} \boldsymbol{x})$ for the IID version of mass-mean probing is in terms of Mahalanobis whitening. Recall that if $\mathcal{D} = \{x_i\}$ is a dataset of $x_i \in \mathbb{R}^d$ with covariance matrix Σ, then the Mahalanobis whitening transformation $W = \Sigma^{-1/2}$