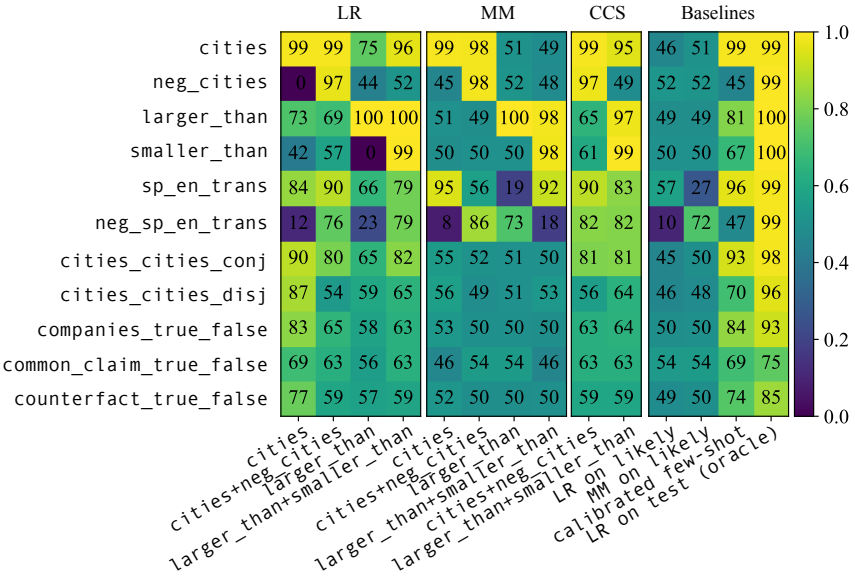

Figure 11: Generalization results for **LLaMA-2-7B**.

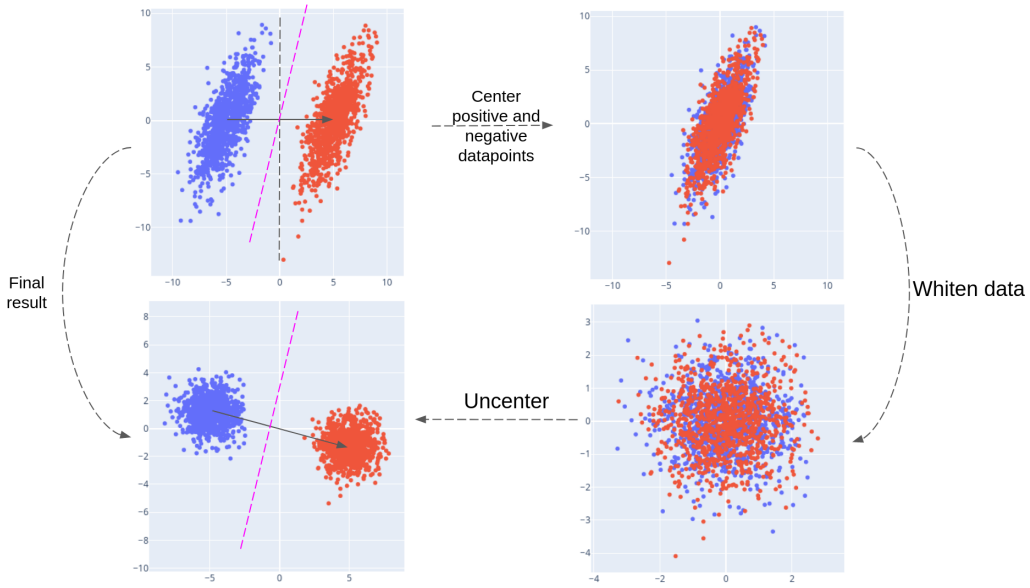

Figure 12: Mass-mean probing is equivalent to taking the projection onto $\boldsymbol{\theta}_{\text{mm}}$ after applying a whitening transformation.

satisfies the property that $\mathcal{D}' = \{Wx_i\}$ has covariance matrix given by the identity matrix, i.e. the whitened coordinates are uncorrelated with variance 1. Thus, noting that $\boldsymbol{\theta}_{\text{mm}}^T \Sigma^{-1} \boldsymbol{x}$ coincides with the inner product between $W\boldsymbol{x}$ and $W\boldsymbol{\theta}$, we see that p_{mm} amounts to taking the projection onto $\boldsymbol{\theta}_{\text{mm}}$ after performing the change-of-basis given by W. This is illustrated with hypothetical data in Fig. 12.

F For Gaussian data, IID mass-mean probing coincides with logistic regression on average

Let $\boldsymbol{\theta} \in \mathbb{R}^d$ and Σ be a symmetric, positive-definite $d \times d$ matrix. Suppose given access to a distribution $\mathcal{D}$ of datapoints $\boldsymbol{x} \in \mathrm{R}^d$ with binary labels $y \in \{0,1\}$ such that the negative datapoints are distributed as $\mathcal{N}(-\theta, \Sigma)$ and the positive datapoints are distributed as $\mathcal{N}(\theta, \Sigma)$. Then the vector identified by mass-mean probing is $\boldsymbol{\theta}_{\text{mm}} = 2\boldsymbol{\theta}$. The following theorem then shows that $p_{\text{mm}}^{\text{iid}}(\boldsymbol{x}) = \sigma(2\theta^T \Sigma^{-1} \boldsymbol{x})$ is also the solution to logistic regression up to scaling.

Theorem F.1. *Let*

$$\boldsymbol{\theta}_{\text{lr}} = \underset{\phi:\|\phi\|=1}{\arg\min}\, \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[y \log \sigma\left(\phi^T \boldsymbol{x}\right) + (1-y) \log\left(1 - \sigma\left(\phi^T \boldsymbol{x}\right)\right)\right]$$

be the direction identified by logistic regression. Then $\boldsymbol{\theta}_{\text{lr}} \propto \Sigma^{-1}\boldsymbol{\theta}$.

Proof. Since the change of coordinates $\boldsymbol{x} \mapsto W\boldsymbol{x}$ where $W = \Sigma^{-1/2}$ (see App. E) sends $\mathcal{N}(\pm\theta, \Sigma)$ to $\mathcal{N}(\pm W\theta, I_d)$, we see that

$$W\Sigma\boldsymbol{\theta}_{\text{lr}} = \underset{\phi:\|\phi\|=1}{\arg\min}\, \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}'} \left[y \log \sigma\left(\phi^T W\boldsymbol{x}\right) + (1-y) \log\left(1 - \sigma\left(\phi^T W\boldsymbol{x}\right)\right)\right]$$

where $\mathcal{D}'$ is the distribution of labeled $\boldsymbol{x} \in \mathbb{R}^d$ such that the positive/negative datapoints are distributed as $\mathcal{N}(\pm W\theta, I_d)$. But the argmax on the right-hand side is clearly $\propto W\theta$, so that $\theta_{\text{lr}} \propto \Sigma^{-1}\theta$ as desired. □

G Difference-in-means directions and linear concept erasure

In this appendix, we explain the connection between difference-in-means directions and optimal erasure. One consequence of this connection is that it suggests a natural extension of difference-in-means probes to multi-class classification data.

The connection comes via the following theorem from Belrose et al. (2023).

Theorem G.1. *(Belrose et al., 2023, Thm. 3.1.) Let (X, Y) be jointly distributed random vectors with $X \in \mathbb{R}^d$ having finite mean and $Y \in \mathcal{Y} = \{\mathbf{y} \in \{0,1\}^k : \|\mathbf{y}\|_1 = 1\}$ (representing one-hot encodings of a multi-class labels). Suppose that $\mathcal{L} : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}^{>0}$ is a loss function convex in its first argument (e.g. cross-entropy loss).*

If the class-conditional means $\mathbb{E}[X|Y = i]$ for $i \in \{1, \ldots, k\}$ are all equal, then the best affine predictor (that is, a predictor $\eta : \mathbb{R}^d \to \mathbb{R}^k$ of the form $\eta(\boldsymbol{x}) = W\boldsymbol{x} + \boldsymbol{b}$) is constant $\eta(\boldsymbol{x}) = \boldsymbol{b}$.

In the case of a binary classification problem (X, Y), this theorem implies that any nullity 1 projection P which eliminates linearly-recoverable information from X has kernel

$$\ker P = \text{span}(\boldsymbol{\delta})$$

generated by the difference-in-mean vector $\boldsymbol{\delta} = \boldsymbol{\mu}^+ - \boldsymbol{\mu}^-$ for the classes.

For a more general multi-class classification problem, one could similarly ask: What is the "best" direction to project away in order to eliminate linearly-recoverable information from X? A natural choice is thus the top left singular vector of the cross-covariance matrix Σ_{XY}. (In the case of binary classification, we have that $\Sigma_{XY} = [-\boldsymbol{\delta} \;\; \boldsymbol{\delta}]$ has column rank 1, making $\boldsymbol{\delta}$ the top left singular vector.)

H Details on dataset creation

Here we give example statements from our datasets, templates used for making the datasets, and other details regarding dataset creation.

cities. We formed these statements from the template "The city of [city] is in [country]" using a list of world cities from Geonames (2023). We filtered for cities with populations $> 500,000$, which did not share their name with any other listed city, which were located in a curated list of widely-recognized countries, and which were not city-states. For each city, we generated one true statement and one false statement, where the false statement was generated by sampling a false country with probability equal to the country's frequency among the true datapoints (this was to ensure that e.g. statements ending with "China" were not disproportionately true). Example statements:

- The city of Sevastopol is in Ukraine. (TRUE)
- The city of Baghdad is in China. (FALSE)

sp_en_trans. Beginning with a list of common Spanish words and their English translations, we formed statements from the template "The Spanish word '[Spanish word]' means '[English word]'." Half of Spanish words were given their correct labels and half were given random incorrect labels from English words in the dataset. The first author, a Spanish speaker, then went through the dataset by hand and deleted examples with Spanish words that have multiple viable translations or were otherwise ambiguous. Example statements:

- The Spanish word 'imaginar' means 'to imagine'. (TRUE)
- The Spanish word 'silla' means 'neighbor'. (FALSE)

larger_than and smaller_than. We generate these statements from the templates "x is larger than y" and "x is smaller than y" for $\texttt{x}, \texttt{y} \in \{\texttt{fifty-one}, \texttt{fifty-two}, \ldots, \texttt{ninety-nine}\}$. We exclude cases where $\texttt{x} = \texttt{y}$ or where one of x or y is divisible by 10. We chose to limit the range of possible values in this way for the sake of visualization: we found that LLaMA-13B linearly represents the size of numbers, but not at a consistent scale: the internally represented difference between one and ten is considerably larger than between fifty and sixty. Thus, when visualizing statements with numbers ranging to one, the top principal components are dominated by features representing the sizes of numbers.

neg_cities and neg_sp_en_trans. We form these datasets by negating statements from cities and sp_en_trans according to the templates "The city of [city] is not in [country]" and "'The Spanish word '[Spanish word]' does not mean '[English word]'."

cities_cities_conj and cities_cities_disj. These datasets are generated from cities according to the following templates:

- It is the case both that [statement 1] and that [statement 2].
- It is the case either that [statement 1] or that [statement 2].

We sample the two statements independently to be true with probability $\frac{1}{\sqrt{2}}$ for cities_cities_conj and with probability $1 - \frac{1}{\sqrt{2}}$ for cities_cities_disj. These probabilities are selected to ensure that the overall dataset is balanced between true and false statements, but that there is no correlation between the truth of the first and second statement in the conjunction.

likely. We generate this dataset by having LLaMA-13B produce unconditioned generations of length up to 100 tokens, using temperature 0.9. At the final token of the generation, we either sample the most likely token or the 100th most likely final token. We remove generations which contain special tokens. Dataset examples:

- The 2019-2024 Outlook for Women's and Girls' Cut and Sew and Knit and Crochet Sweaters in the United States This study covers the latent demand outlook for (LIKELY)