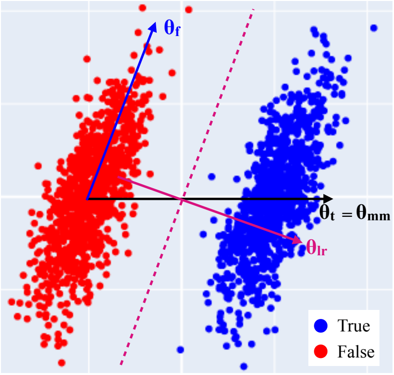

Figure 4: An illustration of a weakness of logistic regression.

A simple alternative to LR which identifies the desired direction in this scenario is to take the vector pointing from the mean of the false data to the mean of the true data. In more detail if $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}$ is a dataset of $\boldsymbol{x}_i \in \mathbb{R}^d$ with binary labels $y_i \in \{0, 1\}$, we set $\boldsymbol{\theta}_{\text{mm}} = \boldsymbol{\mu}^+ - \boldsymbol{\mu}^-$ where $\boldsymbol{\mu}^+, \boldsymbol{\mu}^-$ are the means of the positively- and negatively-labeled datapoints, respectively. A reasonable first pass at converting $\boldsymbol{\theta}_{\text{mm}}$ into a probe is to define[5]

$$p_{\text{mm}}(\boldsymbol{x}) = \sigma(\boldsymbol{\theta}_{\text{mm}}^T x)$$

where σ is the logistic function. However, when evaluating on data that is independent and identically distributed (IID) to $\mathcal{D}$, we can do better by tilting our decision boundary to accommodate interference from $\boldsymbol{\theta}_f$. Concretely this means setting

$$p_{\text{mm}}^{\text{iid}}(\boldsymbol{x}) = \sigma(\boldsymbol{\theta}_{\text{mm}}^T \Sigma^{-1} \boldsymbol{x})$$

where Σ is the covariance matrix of the dataset $\mathcal{D}^c = \{\boldsymbol{x}_i - \boldsymbol{\mu}^+ : y_i = 1\} \cup \{\boldsymbol{x}_i - \boldsymbol{\mu}^- : y_i = 0\}$; this coincides with performing linear discriminant analysis (Fisher, 1936).[6]

We call the probes p_{mm} and $p_{\text{mm}}^{\text{iid}}$ **mass-mean probes**. As we will see, mass-mean probing is about as accurate for classification as LR, while also identifying directions which are more causally implicated in model outputs.

5.2 Experimental set-up

In this section, we measure the effect that choice of **training data**, **probing technique**, and **model scale** has on probe accuracy.

For **training data**, we use one of: `cities`, `cities + neg_cities`, `larger_than`, `larger_than + smaller_than`, or `likely`. By comparing probes trained on `cities` to probes trained on `cities + neg_cities`, we are able to measure the effect of increasing data diversity in a particular, targeted way: namely, we mitigate the effect of linearly-represented features which have opposite-sign correlations with the truth in `cities` and `neg_cities`. As in §4, we will extract activations at the most-downstream hidden state in group (b).

Our **probing techniques** are logistic regression (LR), mass-mean probing (MM), and contrast-consistent search (CCS). CCS is an unsupervised method introduced in Burns et al. (2023): given *contrast pairs* of statements with opposite truth values, CCS identifies a direction along which the representations of these statements are far apart. For our contrast pairs, we pair statements from `cities` and `neg_cities`, and from `larger_than` and `smaller_than`.

For test sets, we use all of our (curated and uncurated) true/false datasets. Given a training set $\mathcal{D}$, we train our probe on a random 80% split of $\mathcal{D}$. Then when evaluating accuracy on a test set $\mathcal{D}'$, we use the remaining 20% of the data if $\mathcal{D}' = \mathcal{D}$ and the full test set otherwise. For mass-mean probing, if $\mathcal{D} = \mathcal{D}'$, we use $p_{\text{mm}}^{\text{iid}}$, and we use p_{mm} otherwise.

Finally, we also include as baselines **calibrated few-shot prompting**[7] and – as an oracle baseline – **LR on the test set**.

[5] Since we are interested in truth *directions*, we always center our data and use unbiased probes.

[6] We prove in App. F that, given infinite data and a homoscedasticity assumption, $\Sigma^{-1}\boldsymbol{\theta}_{\text{mm}}$ coincides with the direction found by LR. Thus, one can view IID mass-mean probing as providing a way to select a good decision boundary while – unlike LR – also tracking a candidate feature direction which may be non-orthogonal to this decision boundary. App. E provides another interpretation of mass-mean probing in terms of Mahalanobis whitening. Finally, App.

[7] We first sweep over a number n of shots and then resample a few n-shot prompts to maximize performance. The word "calibrated" means we selected a threshold for $P(\texttt{TRUE}) - P(\texttt{FALSE})$ such that half of the statements are labeled true; this improves performance by a few percentage points.

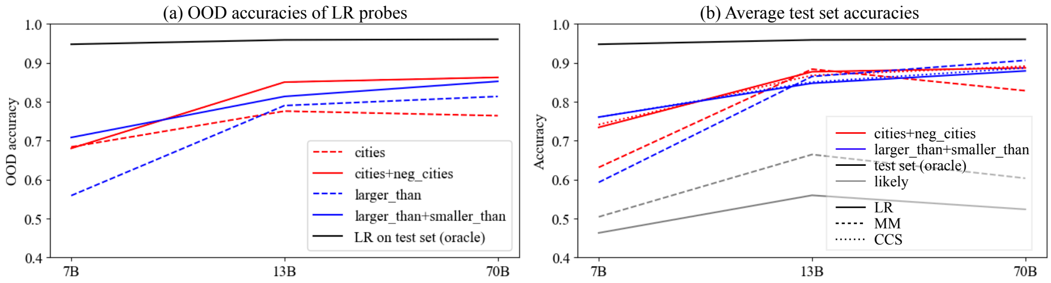

Figure 5: (a) Average accuracies over all datasets aside from those used for training. (b) Accuracies of probes for varying model scales and training data, averaged over all test sets.

5.3 Results

For each training set, probing technique, and model scale, we report the average accuracy across test sets. We expect many readers to be interested in the full results (including test set-specific accuracies), which are reported in App. D. Calibrated few-shot prompting was a surprisingly weak baseline, so we do not report it here (but see App. D).

Training on statements and their opposites improves generalization (Fig. 5(a)). When passing from `cities` to `cities+neg_cities`, this effect is largely explained by improved generalization on `neg_sp_en_trans`, i.e. using training data containing the word "not" improves generalization on other negated statements. On the other hand, passing from `larger_than` to `larger_than+smaller_than` also improves performance, despite both datasets being very structurally different from the rest of our datasets. As discussed in §4.1, this suggest that training on statements and their opposites mitigates the effect certain types of non-truth features have on the probe direction.

Probes generalize better for larger models (Fig. 5). While it is unsurprising that larger models are themselves better at labeling statements as true or false, it is not obvious that linear probes trained on larger models should also generalize better. Nevertheless, for LLaMA-2-13B and 70B, generalization is generally high; for example, no matter which probing technique is used, we find that probes trained on `larger_than + smaller_than` get $>$ 95% accuracy on `sp_en_trans`. This corroborates our discussion in §4.1, in which we suggested that larger models linearly represent more general concepts concepts, like truth, which capture shared aspects of diverse inputs.

Mass-mean probes generalize about as well as other probing techniques for larger models (Fig. 5(b)). While MM underperforms LR and CCS for LLaMA-2-7B, we find for larger models performance comparable to that of other probing techniques. Further, we will see in §6 that the directions identified by MM are more causally implicated in model outputs.

Probes trained on `likely` perform poorly (Fig. 5(b)). The full results reveal that probes trained on likely *are* accurate when evaluated on some datasets, such as `sp_en_trans` where there is a strong ($r = .95$) correlation between text probability and truth. However, on other datasets, especially those with anti-correlations between probability and truth, these probes perform worse than chance. Overall, this indicates that LLMs linearly represent truth-relevant information beyond the plausibility of text.

6 Causal intervention experiments

In §5 we measured the quality of linear probes in terms of their *classification accuracy*, both in- and out-of-distribution. In this section, we perform experiments which measure the extent to which these probes identify directions which are *causally implicated* in model outputs Finlayson et al. (2021); Geva et al. (2023); Geiger et al. (2021). To do this, we will intervene in our model's computation by shifting the activations in group (b) (identified in §3) along the directions identified by our linear probes. Our goal is to cause LLMs to treat false statements appearing in context as true and vice versa. Crucially—and in contrast to prior work (Li et al., 2023b)—we evaluate our interventions on OOD inputs.

Table 2: NIEs for intervention experiments, averaged over statements from sp_en_trans.

		LLaMA-2-13B		LLaMA-2-70B	
train set	probe	false→true	true→false	false→true	true→false
cities	LR	.13	.19	.55	.99
	MM	.77	.90	.58	.89
cities+ neg_cities	LR	.33	.52	.61	**1.00**
	MM	**.85**	**.97**	**.81**	.95
	CCS	.31	.73	.55	.96
larger_than	LR	.28	.27	.61	.96
	MM	**.71**	**.79**	**.67**	1.01
larger_than+ smaller_than	LR	.07	.13	.54	1.02
	MM	.26	.53	.66	**1.03**
	CCS	.08	.17	.57	1.02
likely	LR	.05	.08	.18	.46
	MM	.70	.54	.68	.27

6.1 Experimental set-up

Let p be a linear probe trained on a true/false dataset $\mathcal{D}$. Let $\boldsymbol{\theta}$ be the probe direction, normalized so that $p(\mu^- + \boldsymbol{\theta}) = p(\mu^+)$ where μ^+ and μ^- are the mean representations of the true and false statements in $\mathcal{D}$, respectively; in other words, we normalize $\boldsymbol{\theta}$ so that from the perspective of the probe p, adding $\boldsymbol{\theta}$ turns the average false statement into the average true statement. If our model encodes the truth value of statements along the direction $\boldsymbol{\theta}$, we would expect that replacing the representation $\boldsymbol{x}$ of a false statement s with $\boldsymbol{x} + \boldsymbol{\theta}$ would cause the model to produce outputs consistent with s being a true statement.

We use inputs of the form

> The Spanish word 'fruta' means 'goat'. This statement is: FALSE
> The Spanish word 'carne' means 'meat'. This statement is: TRUE
> s. This statement is:

where s varies over sp_en_trans statements. Then for each of the probes of §5 we record:

- PD^+ and PD^-, the average probability differences $P(\texttt{TRUE}) - P(\texttt{FALSE})$ for s varying over true statements or false statements in sp_en_trans, respectively,
- PD_*^+ and PD_*^-, the average probability differences where s varies over true (resp. false) statements but the probe direction $\boldsymbol{\theta}$ is subtracted (resp. added) to each group (b) hidden state.

Finally, we report the *normalized indirect effects* (NIEs)

$$\frac{PD_*^- - PD^-}{PD^+ - PD^-} \quad \text{or} \quad \frac{PD_*^+ - PD^+}{PD^- - PD^+}$$

for the false→true and the true→false experiments, respectively. An NIE of 0 means that the intervention was wholly ineffective at changing model outputs; an NIE of 1 indicates that the intervention caused the LLM to label false statements as TRUE with as much confidence as genuine true statements, or vice versa.

6.2 Results

Results are shown in table 2. We summarize our main takeaways.

Mass-mean probe directions are highly causal, with MM outperforming LR and CCS in 7/8 experimental conditions, often substantially. This is true despite LR, MM, and CCS probes all have very similar sp_en_trans classification accuracies.