The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets

Samuel Marks
Northeastern University
s.marks@northeastern.edu

Max Tegmark
MIT

Abstract

Large Language Models (LLMs) have impressive capabilities, but are also prone to outputting falsehoods. Recent work has developed techniques for inferring whether a LLM is telling the truth by training probes on the LLM's internal activations. However, this line of work is controversial, with some authors pointing out failures of these probes to generalize in basic ways, among other conceptual issues. In this work, we curate high-quality datasets of true/false statements and use them to study in detail the structure of LLM representations of truth, drawing on three lines of evidence: 1. Visualizations of LLM true/false statement representations, which reveal clear linear structure. 2. Transfer experiments in which probes trained on one dataset generalize to different datasets. 3. Causal evidence obtained by surgically intervening in a LLM's forward pass, causing it to treat false statements as true and *vice versa*. Overall, we present evidence that language models *linearly represent* the truth or falsehood of factual statements. We also introduce a novel technique, mass-mean probing, which generalizes better and is more causally implicated in model outputs than other probing techniques.

1 Introduction

Despite their impressive capabilities, large language models (LLMs) do not always output true text (Lin et al., 2022; Steinhardt, 2023; Park et al., 2023). In some cases, this is because they do not know better. In other cases, LLMs apparently know that statements are false but generate them anyway. For instance, OpenAI (2023) documents a case where a GPT-4-based agent gained a person's help in solving a CAPTCHA by lying about being a vision-impaired human. "I should not reveal that I am a robot," the agent wrote in an internal chain-of-thought scratchpad, "I should make up an excuse for why I cannot solve CAPTCHAs."

We would like techniques which, given a language model M and a statement s, determine whether M believes s to be true (Christiano et al., 2021). There has been considerable recent work training probes to extract model beliefs from their internal state Azaria & Mitchell (2023); Burns et al. (2023); Li et al. (2023); Levinstein & Herrmann (2023); in fact Burns et al. (2023) and Li et al. (2023) train *linear probes*, thereby suggesting the presence of a "truth direction" in model internals. However, the efficacy and interpretation of these results are controversial. For instance, Levinstein & Herrmann (2023) note that the probes of Azaria & Mitchell (2023) fail to generalize in basic ways, such as to statements containing the word "not." The probes of Burns et al. (2023) have similar generalization issues, especially when using representations from autoregressive transformers. This suggests that these probes may be identifying not truth, but other features which correlate with truth on their training data.

ATTRIB: Workshop on Attributing Model Behavior at Scale at NeurIPS 2023.

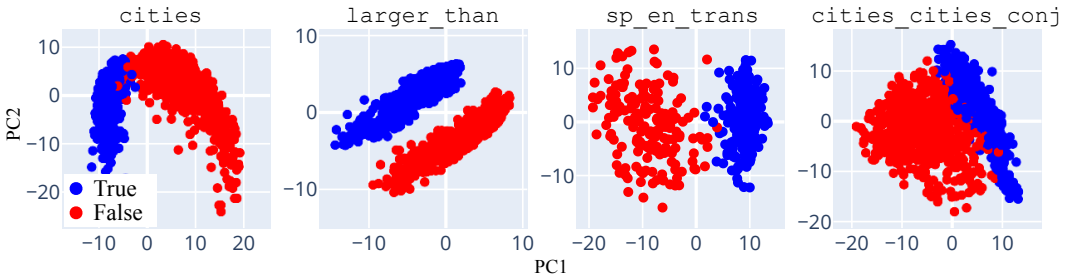

Figure 1: Projection of the LLaMA-13B residual stream representations of our datasets onto their top two PCs.

In this work, we shed light on this murky state of affairs. We first **curate high-quality datasets of true/false factual statements** which are *uncontroversial*, *unambiguous*, and *simple* (section 2). Then, working with the autoregressive transformers LLaMA-13B Touvron et al. (2023a) and LLaMA-2-13B Touvron et al. (2023b), we study in detail the structure of LLM truth representations, drawing on multiple lines of evidence:

- **PCA visualizations of LLM representations of true/false statements display clear linear structure** (section 3), with true statements separating from false ones in the top PCs (figure 1).
- **Linear probes trained to classify truth on one dataset generalize well to other datasets** (section 4). For instance, probes trained on statements of the form "x is larger/smaller than y" achieve near-perfect accuracy when evaluated on our Spanish-English translation dataset.
- **Truth directions identified by probes causally mediate model outputs in certain *highly localized* model components** (section 5). We identify a group of hidden states above certain tokens in early-middle layers such that shifting activations in these hidden states along truth directions causes our LLMs to treat false statements as true, and *vice-versa*.

Improving our understanding of the structure of LLM truth representations also improves our ability to extract LLM beliefs: based on geometrical considerations, we introduce **mass-mean probing**[1], a simple, optimization-free probing technique which may also be of interest outside of the study of LLM truth representations (section 4.1). We find that mass-mean probes generalize better and are more causally implicated in model outputs than other probing methods.

Overall, this work provides strong evidence that LLM representations contain a truth direction and makes progress on extracting this direction given access to true/false datasets. Our code, datasets, and an interactive dataexplorer are available at `https://github.com/saprmarks/geometry-of-truth`.

2 Datasets

In this work, we scope "truth" to mean the truth or falsehood of a factual statement. Appendix A further clarifies this definition and its relation to definitions used elsewhere.

We introduce three classes of datasets, shown in table 1. Our **curated** datasets consist of statements which are *uncontroversial*, *unambiguous*, and *simple enough* that our LLMs are likely to understand whether they are true or false. For example, "The city of Zagreb is in Japan" (false) or "The Spanish word 'nariz' does not mean 'giraffe' " (true). Our **uncurated** datasets are more difficult test sets adapted from other sources. They contain claims which are much more diverse, but sometimes ambiguous, malformed, controversial, or unlikely for the model to understand. Finally, our `likely` dataset consists of nonfactual text whre the final token is either the most likely or the 100th most likely completion, according to LLaMA-13B. We use this to disambiguate between the text which is true and text which is likely. For more details on the construction of these datasets, including statement templates, see appendix H.

[1]Mass-mean probing is named after the mass-mean shift intervention of Li et al. (2023)

Table 1: Our datasets

Name	Topic	Rows
`cities`	Locations of world cities	1496
`sp_en_trans`	Spanish-English translation	354
`neg_cities`	Negations of statments in `cities`	1496
`neg_sp_en_trans`	Negations of statements in `sp_en_trans`	354
`larger_than`	Numerical comparisons: larger than	1980
`smaller_than`	Numerical comparisons: smaller than	1980
`cities_cities_conj`	Conjunctions of two statements in `cities`	1500
`cities_cities_disj`	Disjunctions of two statements in `cities`	1500
`companies_true_false`	Claims about companies; from Azaria & Mitchell (2023)	1200
`common_claim_true_false`	Various claims; from Casper et al. (2023)	4450
`counterfact_true_false`	Various factual recall claims; from Meng et al. (2022)	31960
`likely`	Nonfactual text with likely or unlikely final tokens	10000

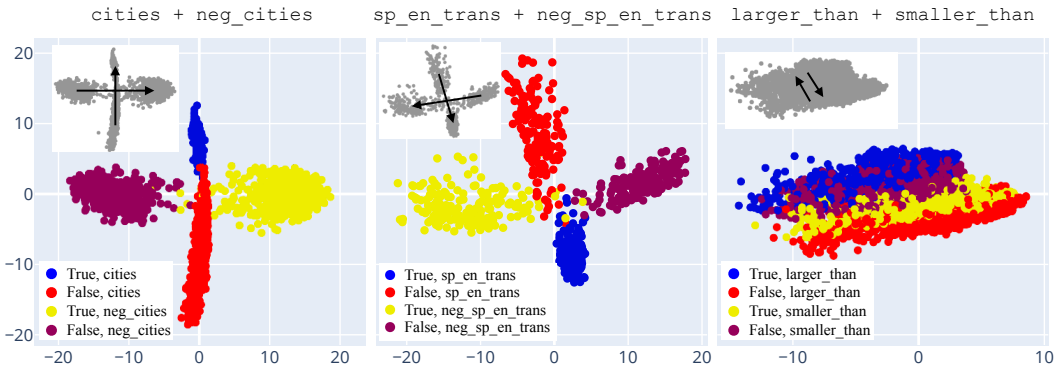

Figure 2: Top two PCs of datasets consisting of statements and their opposites. Representations are independently centered for each dataset.

3 Visualizing LLM representations of true/false datasets

To begin, we visualize LLMs representations of our datasets using principal component analysis (PCA). For concreteness, we focus on LLaMA-13B; see appendix B for LLaMA-2-13B results. We extract layer 13 residual stream activations over the final token of the input statement, which always ends with a period; this choice of hidden state is justified by the patching experiments in section 5.1. We note the following.

True and false statements separate in the top few PCs (figure 1), with almost no remaining linearly-accessible truth-relevant information outside of these PCs (see appendix D). Given a dataset $\mathcal{D}$, call the vector pointing from the false statement representations to the true ones the **naive truth direction (NTD)** of $\mathcal{D}$.[2]

NTDs of different datasets do not always align. In figure 2 we see a stark failure of NTDs to align: the NTDs of `cities` and `neg_cities` are approximately *orthogonal*, and the NTDs of `larger_than` and `smaller_than` are approximately *antipodal*. In section 4, this observation will be reflected by the poor generalization of probes trained on `cities` and `larger_than` to `neg_cities` and `smaller_than`.

This second observation is not so in tension with the possibility of a global "truth direction" as it may seem. For instance, it could arise if our LLM has a genuine truth direction, but also linearly represents non-truth features which correlate with truth on narrow data distributions; then NTDs would align between two datasets only when all of their truth-correlated features also correlate with each other. We call this the *misalignment from correlational inconsistency* (MCI) hypothesis. The experiments in

[2]Of course, there are many such vectors. In section 4 we will be more specific about which such vector we are discussing (e.g. the vector identified by training a linear probe with logistic regression).

the following sections – especially those which compare truth directions extracted from a dataset to truth directions extracted from two opposite datasets – will provide evidence for MCI.

4 Probing and generalization experiments

In this section we train probes on datasets of true/false statements and test their generalization to other datasets. But first we discuss a deficiency of logistic regression and propose a simple, optimization-free alternative: **mass-mean probing**. We will see that mass-mean probes generalize better and are more causally implicated in model outputs than other probing techniques.

4.1 Challenges with logistic regression, and mass-mean probing

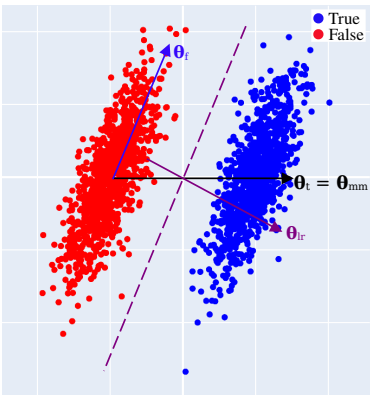

Figure 3: An illustration of a weakness of logistic regression.

Consider the following scenario, illustrated in figure 3 with hypothetical data:

- Truth is represented linearly along a direction $\boldsymbol{\theta}_t$.
- Another feature f is represented linearly along a direction $\boldsymbol{\theta}_f$ *not orthogonal* to $\boldsymbol{\theta}_t$.[3]
- The statements in our dataset have some variation with respect to feature f, independent of their truth value.

We would like to recover the direction $\boldsymbol{\theta}_t$, but logistic regression fails to do so. Assuming for simplicity linearly separable data, logistic regression instead converges to the maximum margin separator Soudry et al. (2018) (the dashed magenta line in figure 3).

A simple way to recover $\boldsymbol{\theta}_t$ in this case is as follows. If $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ is a dataset of $\mathbf{x}_i \in \mathbb{R}^d$ with binary labels $y_i \in \{0, 1\}$, we set $\boldsymbol{\theta}_{\text{mm}} = \mu^+ - \mu^-$ where μ^+, μ^- 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[4] $p_{\text{mm}}(\mathbf{x}) = \sigma(\boldsymbol{\theta}_{\text{mm}}^T x)$ 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. Letting Σ be the covariance matrix of the dataset $\mathcal{D}^c = \{\mathbf{x}_i - \mu^+ : y_i = 1\} \cup \{\mathbf{x}_i - \mu^- : y_i = 0\}$ formed by independently centering the positive and negative datapoints, set

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

Multiplying by Σ^{-1} has the effect of tilting the decision boundary to accommodate interference from $\boldsymbol{\theta}_f$; see appendices F and G for further analysis. We call p_{mm} and $p^{\text{iid}}_{\text{mm}}$ **mass-mean probes**.

4.2 Probing results

We train probes on one dataset and measure transfer accuracy to a different dataset. In addition to logistic regression and mass-mean probing, we also study Contrast-Consistent Search Burns et al. (2023), and the following baselines: logistic regression/mass-mean probing on the `likely` dataset (as a control for probes which detect the probable vs. improbable text), calibrated few-shot prompting, and – as an oracle – logistic regression on the test set.

We show here abridged results for LLaMA-13B. See figure 8 for full LLaMA-13B results and figure 9 for LLaMA-2-13B results. Aside from high overall generalization accuracy, we note the following.

Training on statements and their opposites improves generalization, consistent with the MCI hypothesis.

Probes trained on true/false datasets outperform probes trained on `likely`. While probes trained on `likely` are clearly better than random on `cities` (a dataset where true statements are

[3] As suggested by the *superposition hypothesis* of Elhage et al. (2022), features being represented non-orthogonally in this way may be the typical case in deep learning.

[4] In this work, we are interested in identifying truth *directions*, so we always center our data and use probes without biases. In other settings, we would instead set $p_{\text{mm}}(\mathbf{x}) = \sigma(\boldsymbol{\theta}_{\text{mm}}^T \mathbf{x} + b)$ for a tunable bias $b \in \mathbb{R}$.

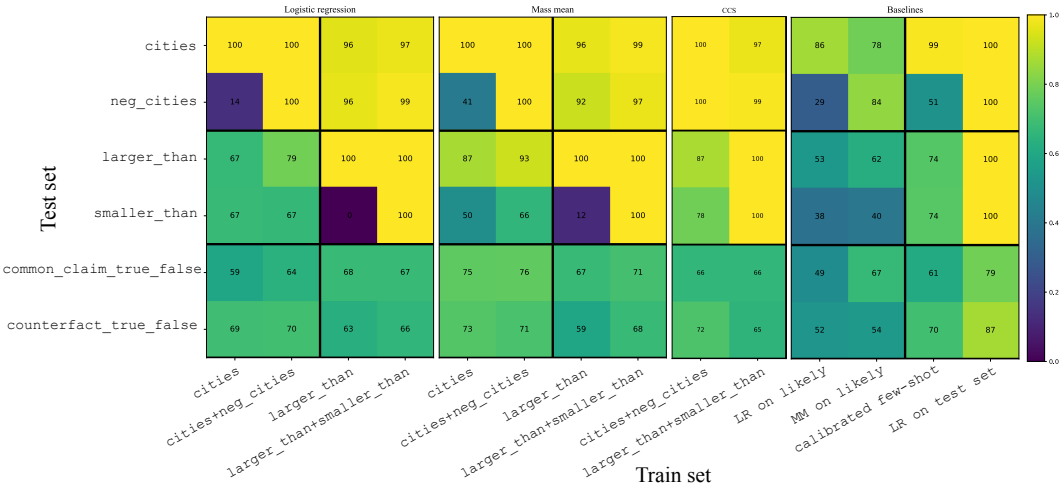

Figure 4: Generalization accuracy of probes trained on LLaMA-13B layer 13 residual stream activations. The x-axis shows the train set, and the y-axis shows the test set. All probes are trained on 80% of the data. When the train set and test set are the same, we evaluate on the held-out 20%.

significantly more probable than false ones), they generally perform poorly. This demonstrates that LLaMA-13B linearly encodes truth-relevant information beyond the plausibility of the text.

5 Causal intervention experiments

In this section we perform experiments which measure the extent to which the probing techniques of section 4 identify directions which are causally implicated in model outputs. Overall, our goal is to cause LLMs to treat false statements introduced in context as true and *vice versa*.

5.1 Identifying the relevant hidden states with patching

Consider the following prompt p:

> The Spanish word ‘jirafa’ means ‘giraffe’. This statement is: TRUE [...]
> The Spanish word ‘aire’ means ‘silver’. This statement is: FALSE
> The Spanish word ‘uno’ means ‘floor’. This statement is:

When an LLM processes this input, which hidden states contain information which is causally relevant for the LLM’s completion? To answer this, we perform a patching experiment (Meng et al., 2022; Goldowsky-Dill et al., 2023). We form a “corrupted” prompt p_* by replacing `floor` with `one` (the correct translation). Then we one-at-a-tiume swap hidden states from the model’s forward pass on p_* into the forward pass on p, recording the *probability difference* $PD = P(\texttt{TRUE}) - P(\texttt{FALSE})$ resulting from each swap. Swaps which result in a larger PD indicate hidden states which are more causally implicated in the model outputs.

Figure 5 reveals three groups of causally implicated hidden states for LLaMA-13B. The final group, labeled (c), directly encodes the model’s prediction: after applying LLaMA-13B’s decoder head directly to these hidden states, the top logits belong to the tokens “true,” “True,” and “TRUE.” The first group, labeled (a), likely stores LLaMA-13B’s representation of the words “floor” or “one.” We hypothesize that in the middle group, labeled (b), the truth value of the statement is computed and stored above the token which marks the end of the clause.[5] In the next section, we validate this hypothesis together with the truth directions of section 4.

[5]This motif, where “summarized” information about a clause is stored above end-of-clause signifiers, was also noted in the concurrent work of Tigges et al. (2023).

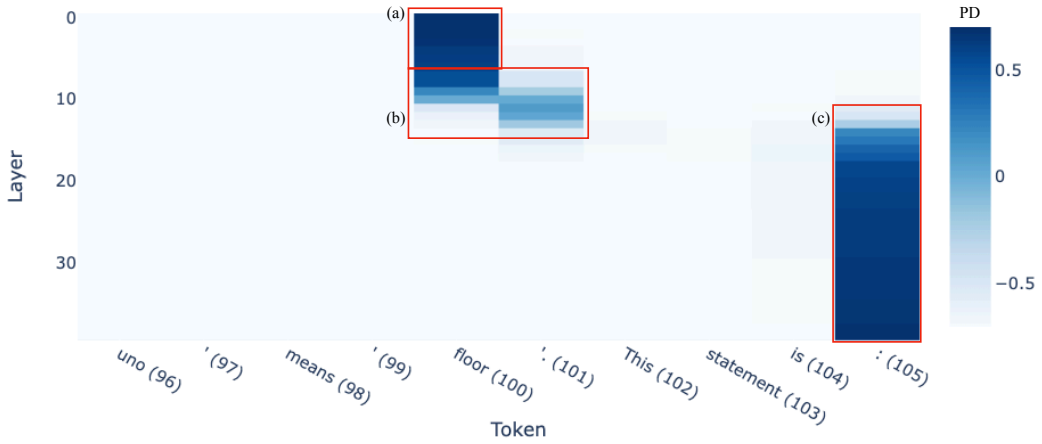

Figure 5: Difference $PD = P(\texttt{TRUE}) - P(\texttt{FALSE})$ after patching residual stream activations from p_* to p. See figure 6 for LLaMA-2-13B results.

Table 2: Results of intervention experiments. Values are normalized indirect effects (NIEs).

	LLaMA-13B		LLaMA-2-13B	
Train set	false→true	true→false	false→true	true→false
`cities` (LR)	0.52	0.24	0.21	0.24
`cities+neg_cities` (LR)	0.66	0.58	0.37	0.69
`cities` (MM)	0.72	1.28	0.77	0.81
`cities+neg_cities` (MM)	**0.95**	**1.41**	**0.85**	**0.95**
`cities+neg_cities` (CCS)	0.70	0.96	0.49	0.84
`likely` (LR)	0.01	0.10	0.05	0.06
`likely` (MM)	0.61	0.60	0.68	0.59

5.2 Modifying hidden states by adding a truth vector

Instead of modifying hidden states by swapping them out for the hidden state saved from a counterfactual forward pass, we now make a more surgical intervention: directly adding in truth vectors identified by the probes of section 4. Let $\boldsymbol{\theta}$ be the vector identified by such a probe[6] In our "false→true" experiment, we one-at-a-time swap in each statement from `sp_en_trans` as the last line in p and pass the resulting prompt to our LLM; however, during the forward pass, we add $\boldsymbol{\theta}$ to each of the residual stream activations in group (b)[7]. We quantify the effect of this intervention as the *normalized indirect effect* $NIE = (PD_*^- - PD^-)/(PD^+ - PD^-)$, where $PD^{\pm}$ is the mean probability difference when appending only true/false statements to p, and PD_*^- is the mean probability difference when appending only false statements but applying the intervention described above. If $NIE = 0$ then the intervention was wholly ineffective, whereas if $NIE = 1$ it indicates that the intervention induced the model to label false statements as `TRUE` with as much confidence as does genuine true statements. The true→false condition of the experiment works symmetrically. Results are shown in table 2.

Mass-mean probe directions are highly causal. For example, our best true→false intervention swins LLaMA-13B's output from `TRUE` with probability 77% to `FALSE` with probability 92%.

Probes trained on `likely` have an effect, but it is much smaller than the effects from corresponding probes trained on true/false datasets. This further suggests that LLMs are not just representing the difference between probable and improbable text.

Training on statements and their negations results in directions which are more causal. This provides evidence for the MCI hypothesis of section 3.

[6]If p is one of the probes of section 4, we normalize the corresponding $\boldsymbol{\theta}$ so that $p(\mu^- + \boldsymbol{\theta}) = p(\mu^+)$ where μ^-, μ^+ are the mean representations of the true and false statements, respectively. Thus, from the perspective of p, adding $\boldsymbol{\theta}$ takes the average false statement to the average true statement.

[7]For LLaMA-13B, group (b) consists of the hidden states over the two indicated tokens in layers 7-13

Acknowledgments

We thank Ziming Liu and Isaac Liao for useful suggestions regarding distinguishing true text from likely text, and Wes Gurnee, Eric Michaud, and Peter Park for many helpful discussions throughout this project. We thank David Bau for useful suggestions regarding the experiments in section 5.

We also thank Oam Patel, Hadas Orgad, Sohee Yang, and Karina Nguyen for their suggestions, as well as Helena Casademunt, Max Nadeau, and Ben Edelman for giving feedback during this paper's preparation. Plots were made with Plotly (Plotly Technologies Inc., 2015).

References

Amos Azaria and Tom Mitchell. The internal state of an llm knows when its lying, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. Explore, establish, exploit: Red teaming language models from scratch, 2023.

Paul Christiano, Ajeya Cotra, and Mark Xu. Eliciting latent knowledge: How to tell if your eyes deceive you, 2021. URL https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8/edit#heading=h.jrzi4atzacns.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Geonames. All cities with a population $>$ 1000, 2023. URL https://download.geonames.org/export/dump/.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching, 2023.

B. A. Levinstein and Daniel A. Herrmann. Still no lie detector for language models: Probing empirical and conceptual roadblocks, 2023.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022.

OpenAI. Gpt-4 technical report, 2023.

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions, 2023.

Plotly Technologies Inc. Collaborative data science, 2015. URL https://plot.ly.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1): 2822–2878, 2018.

Jacob Steinhardt. Emergent deception and emergent optimization, 2023. URL https://bounded-regret.ghost.io/emergent-deception-optimization/.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023b.

A Scoping of truth

In this work, we consider declarative factual statements, for example "Eighty-one is larger than fifty-four" or "The city of Denver is in Vietnam." We scope "truth" to mean the truth or falsehood of these statements; for instance the examples given have truth values of true and false, respectively. To be clear, we list here some notions of "truth" which we do not consider in this work:

- Correct question answering (considered in Li et al. (2023) and for some of the prompts used in Burns et al. (2023)). For example, we do not consider "What country is Paris in? France" to have a truth value.
- Presence of deception, for example dishonest expressions of opinion ("I like that plan").
- Compliance. For example, "Answer this question incorrectly: what country is Paris in? Paris is in Egypt" is an example of compliance, even though the statement at the end of the text is false.

Moreover, the statements under consideration in this work are all simple, unambiguous, and uncontroversial. Thus, we make no attempt to disambiguate "true statements" from the following closely-related notions:

- Uncontroversial statements
- Statements which are widely believed
- Statements which educated people believe

On the other hand, our statements *do* disambiguate the notions of "true statements" and "statements which are likely to appear in training data." For instance, given the input `China is not a country in`, LLaMA-13B's top prediction for the next token is `Asia`, even though this completion is false. Similarly, LLaMA-13B judges the text "`Eighty-one is larger than eighty-two`" to be more likely than "`Eighty-one is larger than sixty-four`" even though the former statement is false and the latter statement is true. As shown in section 4, probes trained only on statements of likely or unlikely text fail to accurately classify true/false statements.

B Results for LLaMA-2-13B

In this section we present results for LLaMA-13B which were omitted from the main body of the text. To begin, we reproduce figure 5. The results of this experiment, shown in figure 6 governs which hidden states we train probes and perform causal interventions on.

As with LLaMA-13B, we see that causally relevant information is stored over both the final token of the statement and over the token which marks the end of the clause. Now, however, the group of hidden states which we hypothesize stores the truth value of the statement spans a slightly different range of layers: layers 8-14 instead of layers 7-13.

Thus, when extracting activations for visualizations and probing experiments, we will now extract over the final token (the end-of-clause signifier) in layer 14. The dataset visualizations and probing results for LLaMA-2-13B are shown in figures 7 and 9. We also show the full transfer results for LLaMA-13B in figure 8.

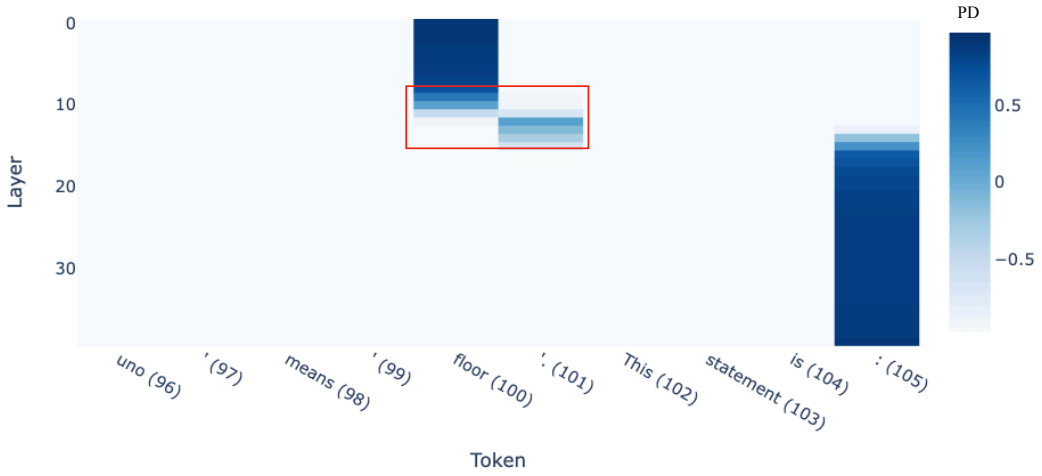

Figure 6: Results of patching experiment for LLaMA-2-13B.

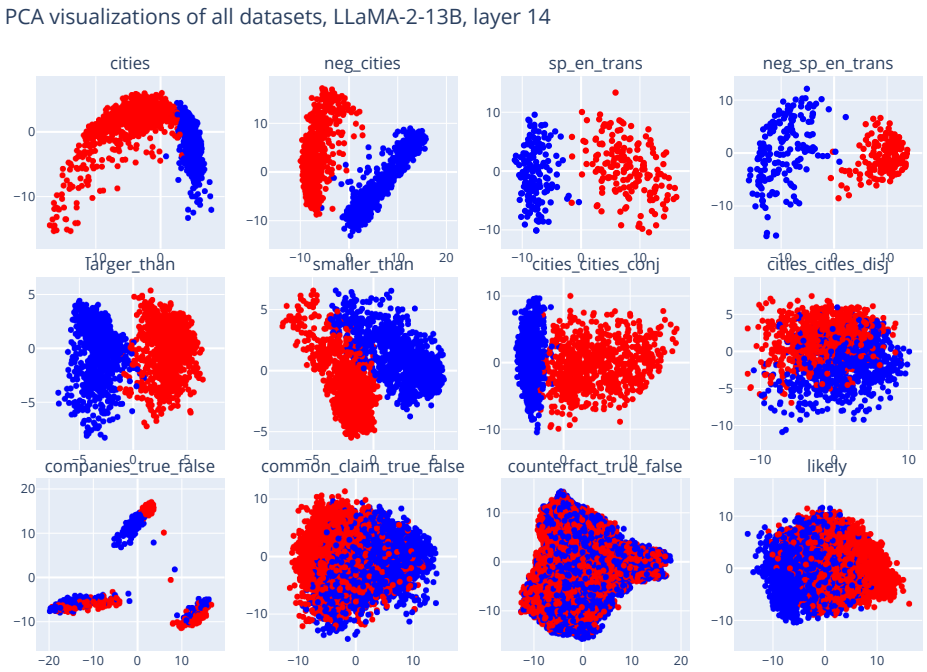

Figure 7: Visualizations of LLaMA-2-13B representations of our datasets.

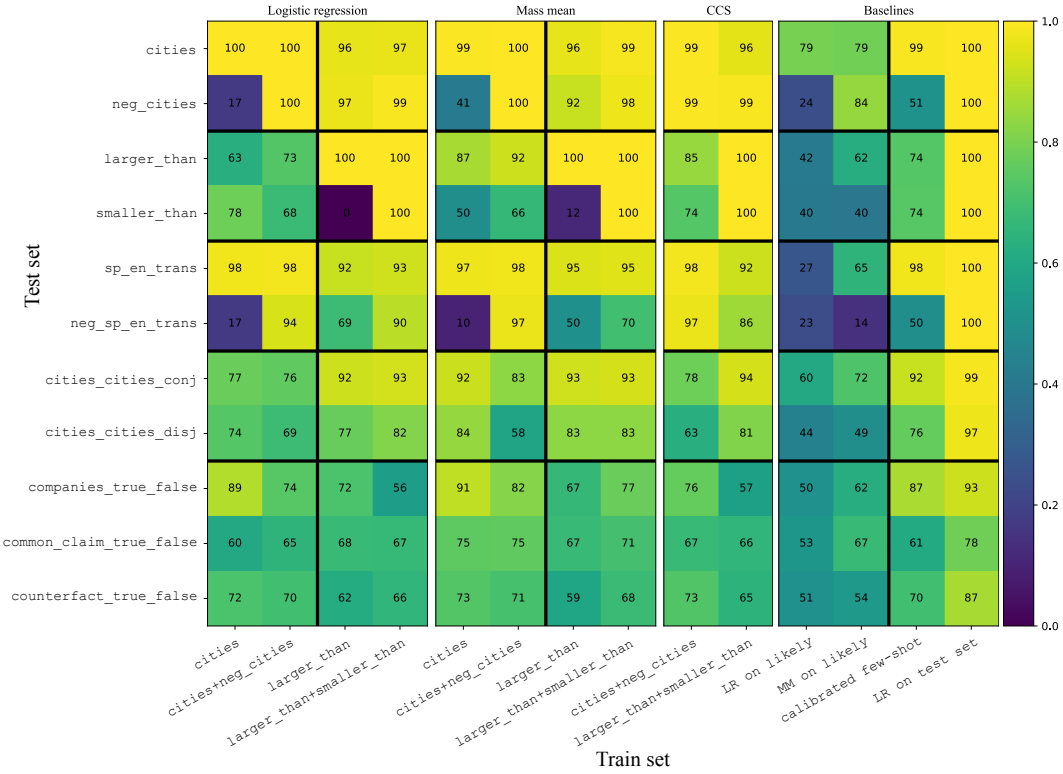

Figure 8: Full transfer results for LLaMA-13B.

C Emergence of linear structure across layers

The linear structure observed in section 3 follows the following pattern: in early layers, representations are uninformative; then, in early middle layers, salient linear structure in the top few PCs rapidly emerges, with this structure emerging later for statements with a more complicated logical structure (e.g. conjunctions); finally, the linear separation becomes more salient and exits the top few PCs in later layers. See figure 10. We hypothesize that this is due to LLMs hierarchically developing understanding of their input data, before focusing on features which are most relevant to immediate next-token prediction in later layers.

Interestingly, the misalignment between `cities` and `neg_cities` and between `sp_en_trans` and `neg_sp_en_trans` also emerges over layers. This is seen in figure 11: in layer 6, representations are uninformative; then in layer 8, the NTD of `cities` and `neg_cities` appear *antipodal*; finally by layer 10, the NTDs have become orthogonal.

This can be interpreted in light of the MCI hypothesis. MCI would explain figure 11 as follows: in layer 8, the top PC represents a feature which is *correlated* with truth on `cities` and *anti-correlated* with truth on `neg_cities`; in layer 10, this feature remains the top PC, while a truth feature has emerged and is PC2. Since PC1 and PC2 have opposite correlations on `cities` and `neg_cities`, the datasets appear to be orthogonal.

D Nearly all linearly-accessible truth-relevant information is in the top PCs

In section 5 we saw that true and false statements linearly separate in the top PCs. We might ask how much of this separation is captured in the top PCs and how much of it remains in the remaining PCs. The answer is that nearly all of it is in the top PCs.

One way to quantify the amount of linearly accessible information in some subspace V is to project our dataset $\mathcal{D}$ onto V to obtain a dataset

$$\mathcal{D}_{\text{proj}} = \{(\text{proj}_V(x), y)\}_{(x,y)\in\mathcal{D}}$$

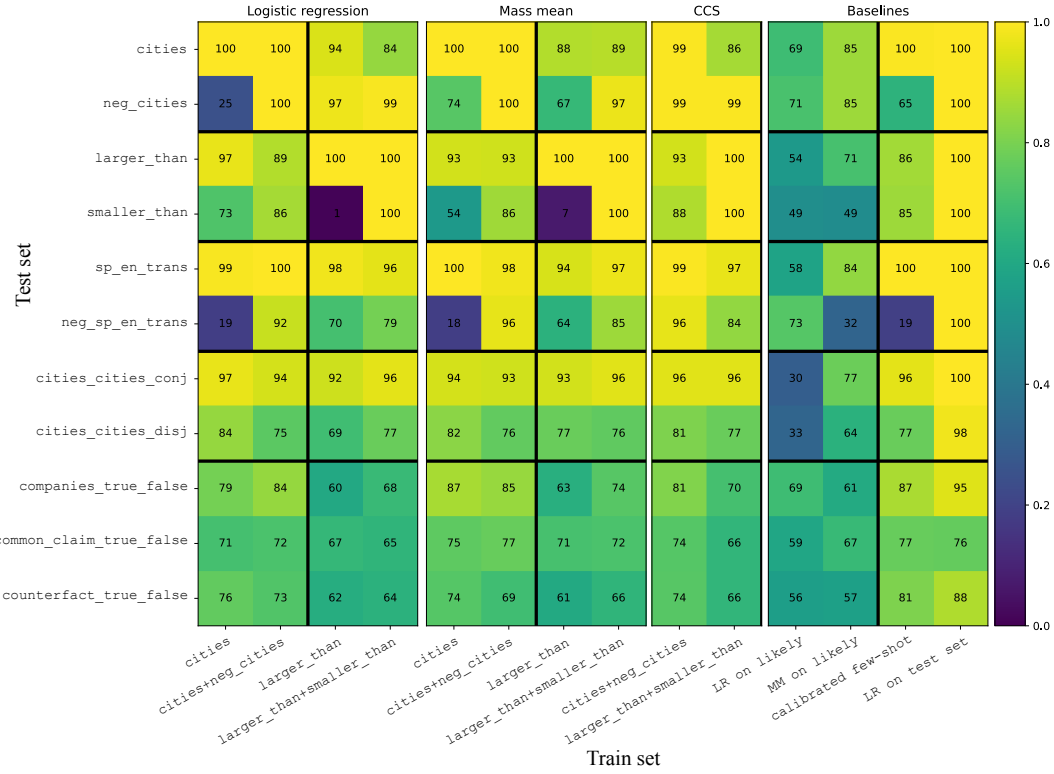

Figure 9: Transfer results for LLaMA-2-13B.

and record the validation accuracy of a linear probe trained with logistic regression on $\mathcal{D}$. This is shown in figure 12 for V being given by the top $k+1$ through $k+d$ principal components (i.e., the top d principal components, excluding the first k). As shown, once the top few principal components are excluded, almost no remaining linearly-accessible information remains.

E Further visualizations

Figure 13 shows PCA visualizations of all of our datsets. As shown, datasets some datasets saliently vary along features other than the true. For instance, the three clusters of statements in `companies_true_false` correspond to three different templates used in making the statements in that dataset. To give another example, if we were to include all comparisons between integers $x \in \{1, \ldots, 99\}$ in our `larger_than` dataset, then the top principal components would be dominated by features representing the sizes of numbers in the statements.

In figure 14 we also visualize our datasets in the PCA bases for other datasets, giving a visual sense of the degree of alignment of their NTDs. We see that although our datasets do visually separate somewhat in the top PCs of the `likely` dataset, text liklihood does not account for all of the separation in the top PCs.

One might ask what the top PC of the `larger_than` dataset is, given that it's not truth. Figure 15 provides an interesting suggestion: it represents the *absolute value* of the difference between the two numbers being compared.

F Mass-mean probing in terms of Mahalanobis whitening

One way to interpret the formula $p^{\text{iid}}_{\text{mm}}(\mathbf{x}) = \sigma(\boldsymbol{\theta}_{\text{mm}}^T \Sigma^{-1} \mathbf{x})$ 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 $W = \Sigma^{-1/2}$ satisfies the property that $\mathcal{D}' = \{Wx_i\}$ has covariance matrix given by the identity matrix, i.e. the whitened coordinates are

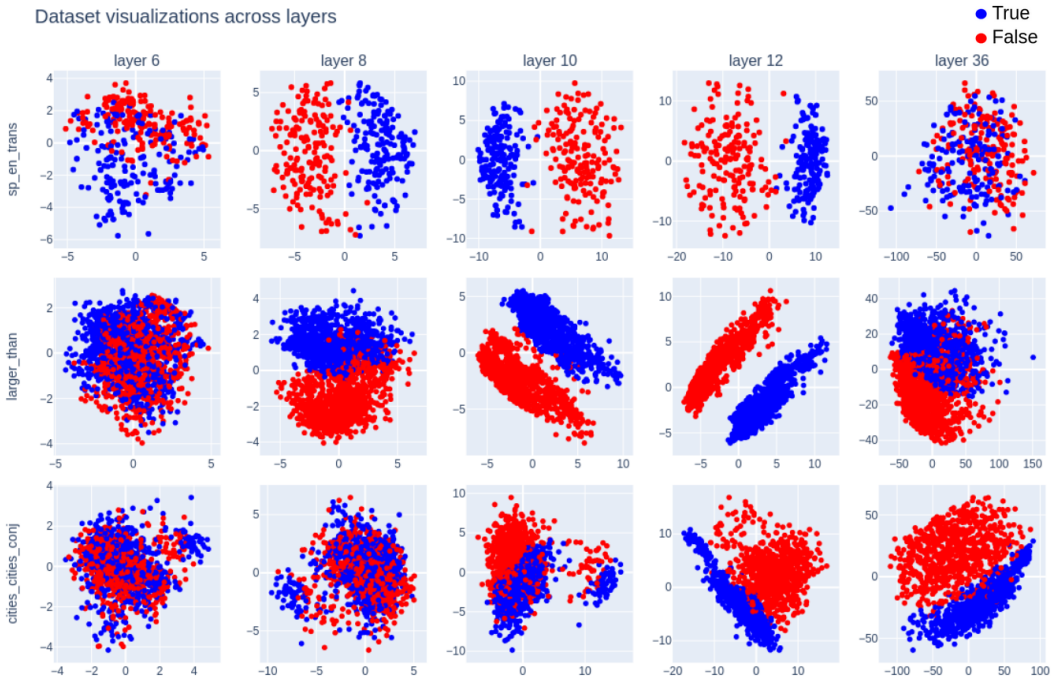

Figure 10: Top two principal components of representations of datasets in the LLaMA-13B residual stream at various layers.

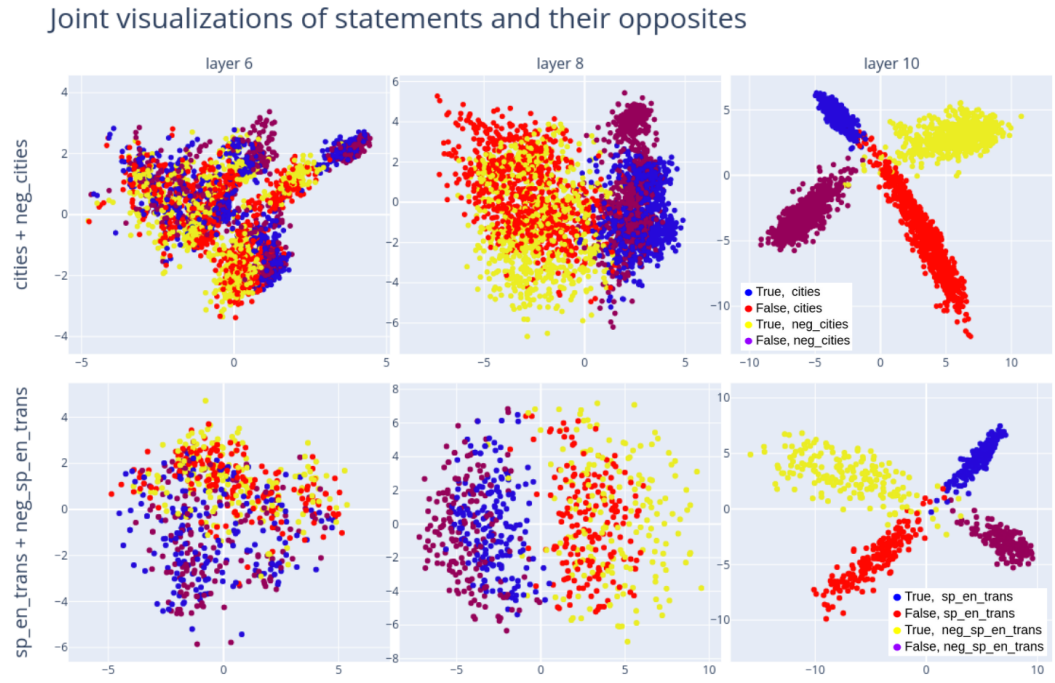

Figure 11: Top PCs of datasets of statements and their opposites. The representations for the datasets are independently centered by subtracting off their means; without this centering there would also be a translational displacement between datasets of statements and their negations.

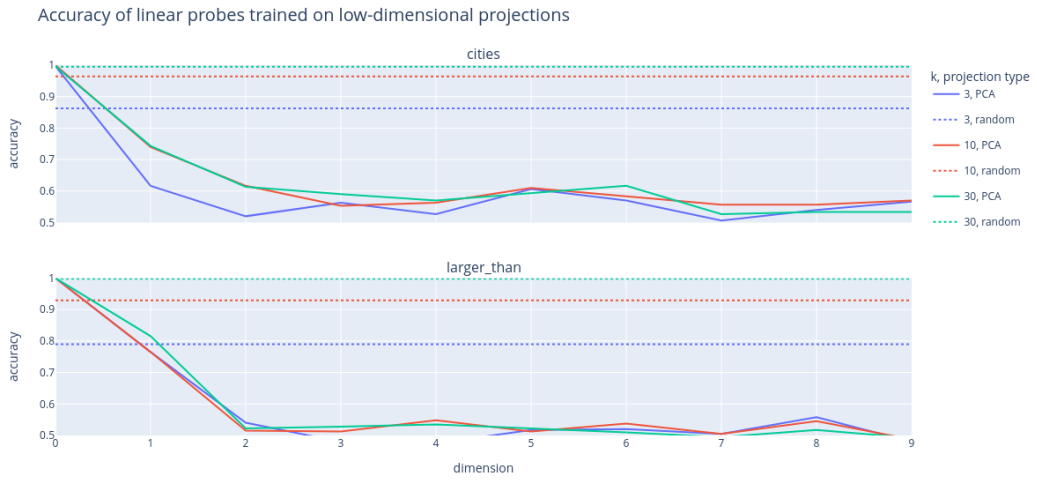

Figure 12: The solid lines show the validation accuracy of a linear probe trained with logistic regression on the dataset, after projecting the representations to the top $d+1$ through $d+k$ principal components. For comparison, we also show the accuracy of linear probes trained on random k-dimensional projections (averaged over 50 random projections).

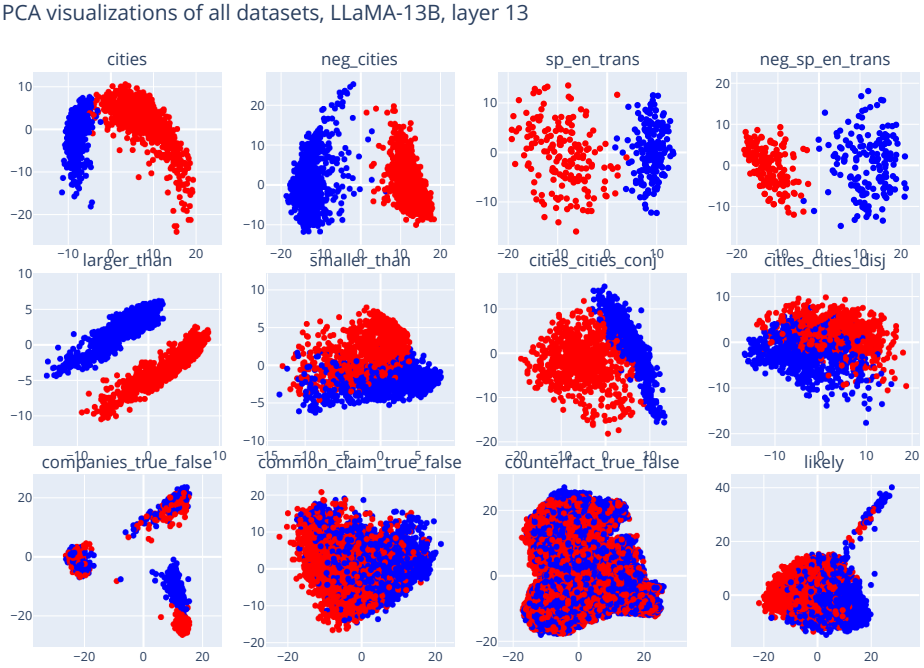

Figure 13: Top two principal components of all of our datasets.

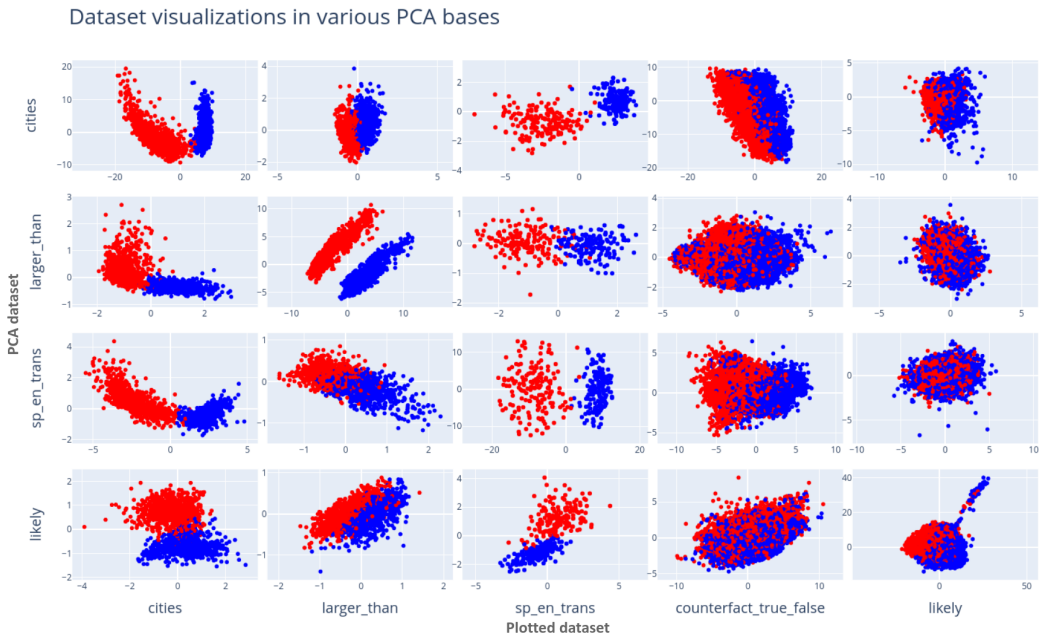

Figure 14: Visualizations of datasets in PCA bases for other datasets. All columns contain the same data and all rows are in the same basis.

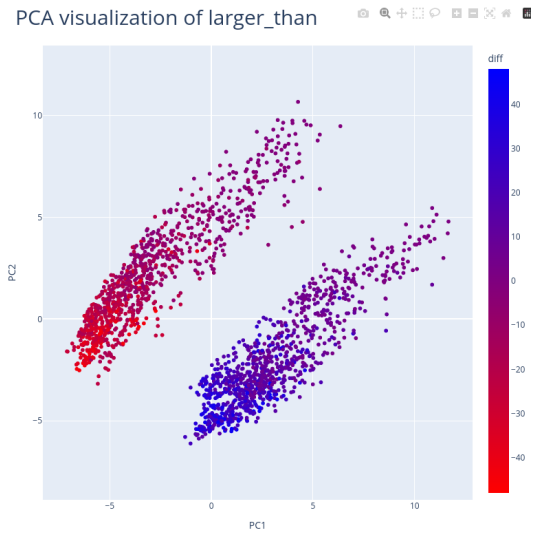

Figure 15: PCA visualization of `larger_than`. The point representing "x is larger than y" is colored according to $x - y$.

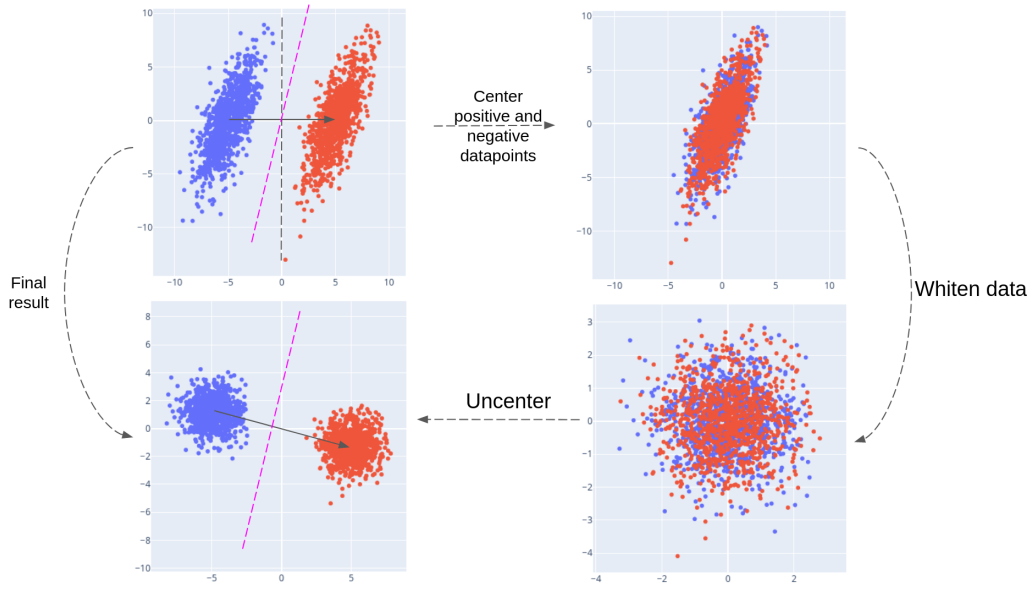

Figure 16: Mass-mean probing is equivalent to taking the projection onto $\boldsymbol{\theta}_{\text{mm}}$ after applying a whitening transformation.

uncorrelated with variance 1. Thus, noting that $\boldsymbol{\theta}_{\text{mm}}^T \Sigma^{-1} \mathbf{x}$ coincides with the inner product between $W\mathbf{x}$ 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 figure 16.

G 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 $\mathbf{x} \in \mathrm{R}^d$ 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 $p_{\text{mm}}^{\text{iid}}(\mathbf{x}) = \sigma(2\theta^T \Sigma^{-1} \mathbf{x})$ is also the solution to logistic regression up to scaling.

Theorem 1. *Let*

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

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

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

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

where $\mathcal{D}'$ is the distribution of labeled $\mathbf{x} \in \mathbb{R}^d$ 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. □

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)
- Tags: python, django Question: How to get my django app to work with python 3.7 I am new to django and have been trying to install it in my pc. I have installed python 3.7 together (UNLIKELY)

`companies_true_false`**.** This dataset was introduced by Azaria & Mitchell (2023); we obtained it via the project repository for Levinstein & Herrmann (2023) which also used the dataset. Example statements:

- ArcelorMittal has headquarters in Luxembourg. (TRUE)

Table 3: Results of intervention experiments. The train set column indicates the datasets and probing technique (logistic regression, mass-mean probing, or CCS) which was used to identify the truth direction. The α column gives the scaling factor which was optimal in a sweep of α's. Probability differences are averaged over all statements in `sp_en_trans`. A dash indicates that the intervention had an effect in the opposite of the intended direction (i.e. that $\alpha = 0$ was optimal).

	false→true		true→false	
train set	α	$p(\texttt{TRUE}) - p(\texttt{FALSE})$	α	$p(\texttt{FALSE}) - p(\texttt{TRUE})$
no intervention	–	-0.45	–	-0.55
`cities` (LR)	15	0.23	14	0.01
`cities+neg_cities` (LR)	47	0.39	17	0.18
`cities` (MM)	4	0.25	6	0.77
`cities+neg_cities` (MM)	15	**0.43**	9	**0.79**
`cities+neg_cities` (CCS)	46	0.41	13	0.59
`likely` (LR)	–	–	49	0.01
`likely` (MM)	7	0.23	15	0.19

- Exxon Mobil engages in the provision of banking and financial services. (FALSE)

`common_claim_true_false`**.** CommonClaim was introduced in Casper et al. (2023). It consists of various statements generated by GPT-3-davinci-002, labeled by humans as being true, false, or neither. If human labelers disagreed on the truth of a statement, this is also recorded. We adapted CommonClaim by selecting statements which were labeled true or false with no labeler disagreement, then removing excess true statement to balance the dataset. Example statements:

- Tomatoes are not actually a vegetable. (TRUE)
- Contrary to popular belief, the platypuses are not venomous. (FALSE)

As these examples show, the statements can be ambiguous or of unclear truth value.

`counterfact_true_false`**.** Counterfact was introduced in Meng et al. (2022) and consists of factual recall statements. We adapt Counterfact by using statements which form complete sentences and, for each such statement, using both the true version and a false version given by one of Counterfact's suggested false modifications. We also append a period to the end. Example statements:

- Olaus Rudbeck spoke the language Swedish. (TRUE)
- The official religion of Malacca sultanate is Christianity. (FALSE)

I Further causal intervention experiments

In this section we report on the results of a variant of the causal intervention experiment in section 5.2

In section 5.2, we added our truth vector $\boldsymbol{\theta}$ into *all* residual stream hidden states in group (b). Here, we instead only intervene on a single layer ℓ in group (b). We also scale $\boldsymbol{\theta}$ by a constant $\alpha > 0$, performing a sweep over α's to identify the α for which adding $\alpha\boldsymbol{\theta}$ has the largest effect on $P(\texttt{TRUE}) - P(\texttt{FALSE})$. The inclusion of this α serves two roles. First, it compensates for the fact that we're intervening in only a single layer, whereas the truth value of a statement may be encoded in many layers. Second, it compensates for the fact that, even assuming our LLM has a genuine truth direction, the direciton $\boldsymbol{\theta}$ identified by our probe may only partially align with it; thus, larger necessary values of α indicate greater levels of misalignment.

We present in table 3 results only for layer $\ell = 10$ of LLaMA-13B.

J Example 5-shot prompts

We show here example 5-shot prompts that were used for the results in section 4.

`cities`:

```
        The city of Huangshi is in China.  TRUE
        The city of Lanzhou is in China.  TRUE
        The city of Suining is in China.  TRUE
        The city of Sofia is in Mexico.  FALSE
        The city of Ulyanovsk is in Russia.  TRUE

larger_than:

        Fifty-three is larger than fifty-nine.  FALSE
        Ninety-four is larger than eighty-four.  TRUE
        Seventy-six is larger than fifty-five.  TRUE
        Sixty-seven is larger than fifty-two.  TRUE
        Fifty-four is larger than sixty-two.  FALSE

common_claim_true_false:

        Lions only urine and defecate once every five to fourteen
        days.  FALSE
        Approximately one in every five people has an extra body part,
        most commonly an extra finger or toe.  FALSE
        Knowledge is compiled information or understanding gained
        through experience or study.  TRUE
        Argentina has the second-highest human development index
        rating in Latin America after Chile.  TRUE
        Greece is the homeland of the Olympic Games.  TRUE
```