The Geometry of Truth: Emergent Linear Structure in LLM 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 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 use high-quality datasets of simple true/false statements 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 at sufficient scale, LLMs *linearly represent* the truth or falsehood of factual statements. We also show that simple difference-in-mean probes generalize as well as other probing techniques while identifying directions which are more causally implicated in model outputs.

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, Perez et al. (2022) demonstrate that LLM assistants output more falsehoods when prompted with the biography of a less-educated user. More starkly, 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). One approach to this problem relies on inspecting model outputs; for instance, the internal chain-of-thought in the above example provides evidence that the model understood it was generating a falsehood. An alternative class of approaches instead leverages access to M's internal state when processing s. There has been considerable recent work on this class of approaches: Azaria & Mitchell (2023), Li et al. (2023b), and Burns et al. (2023) all train probes for classifying truthfulness based on a LLM's internal activations. In fact, the probes of Li et al. (2023b) and Burns et al. (2023) are *linear probes*, 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 these probes may be identifying not truth, but other features that correlate with truth on their training data.

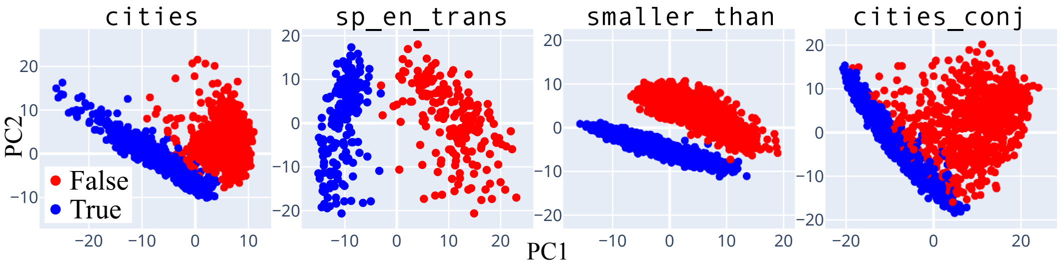

Figure 1: PCA visualizations for LLaMA-2-70B representations of our true/false datasets.

Working with autoregressive transformers from the LLaMA-2 family (Touvron et al., 2023), we shed light on this murky state of affairs. After curating high-quality datasets of simple, unambiguous true/false statements, we perform a detailed investigation of LLM representations of factuality. Our analysis, which draws on patching experiments, simple visualizations with principal component analysis (PCA), a study of probe generalization, and causal interventions, finds:

- **Evidence that linear representations of truth emerge with scale**, with larger models having a more abstract notion of truth that applies across structurally and topically diverse inputs.
- **A small group of causally-implicated hidden states** which encode these truth representations.
- Consistent results across a suite of probing techniques, but with **simple difference-in-mean probes identifying directions which are most causally implicated**.

Our code, datasets, and an interactive dataexplorer are available at `https://github.com/saprmarks/geometry-of-truth`.

1.1 Related work

Linear world models. Substantial previous work has studied whether LLMs encode world models in their representations (Li et al., 2023a; 2021; Abdou et al., 2021; Patel & Pavlick, 2022). Early work focused on whether individual neurons represent features (Wang et al., 2022; Sajjad et al., 2022; Bau et al., 2020), but features may more generally be represented by *directions* in a LLM's latent space (i.e. linear combinations of neurons) (Dalvi et al., 2018; Gurnee et al., 2023; Cunningham et al., 2023; Elhage et al., 2022). We say such features are *linearly represented* by the LLM. Just as other authors have asked whether models have directions representing the concepts of "West Africa" (Goh et al., 2021) or "basketball" (Gurnee et al., 2023), we ask here whether there is a direction corresponding to the truth or falsehood of a factual statement.

Probing for truthfulness. Others have trained probes to classify truthfulness from LLM activations, using both logistic regression (Azaria & Mitchell, 2023; Li et al., 2023b), unsupervised (Burns et al., 2023), and contrastive (Zou et al., 2023; Rimsky et al., 2024) techniques. This work differs from prior work in a number of ways. First, a cornerstone of our analysis is evaluating whether probes trained on one dataset transfer to topically and structurally different datasets in terms of *both* classification accuracy *and* causal mediation of model outputs. Second, we specifically interrogate whether our probes attend to *truth*, rather than merely features which correlate with truth (e.g. probable vs. improbable text). Third, we localize truth representations to a small number of hidden states above certain tokens. Fourth, we go beyond the mass-mean shift interventions of Li et al. (2023b) by systematically studying the properties of difference-in-mean. Finally, we carefully scope our setting, using only datasets of clear, simple, and unambiguous factual statements, rather than statements which are complicated and structured (Burns et al., 2023), confusing (Azaria & Mitchell, 2023; Levinstein & Herrmann, 2023), or intentionally misleading (Li et al., 2023b; Lin et al., 2022).

Table 1: Our datasets

Name	Description	Rows
`cities`	"The city of `[city]` is in `[country]`."	1496
`neg_cities`	Negations of statements in `cities` with "not"	1496
`sp_en_trans`	"The Spanish word '`[word]`' means '`[English word]`'."	354
`neg_sp_en_trans`	Negations of statements in `sp_en_trans` with "not"	354
`larger_than`	"x is larger than y."	1980
`smaller_than`	"x is smaller than y."	1980
`cities_cities_conj`	Conjunctions of two statements in `cities` with "and"	1500
`cities_cities_disj`	Disjunctions of two statements in `cities` with "or"	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

2 Datasets

Curated datasets. Unlike some prior work (Lin et al., 2022; Onoe et al., 2021, *inter alia*) on language model truthfulness, our primary goal is *not* to measure LLMs' capabilities for classifying the factuality of challenging data. Rather, our goal is to understand: Do LLMs have a unified representation of truth that spans structurally and topically diverse data? We therefore construct **curated** datasets with the following properties:

1. **Clear scope**. We scope "truth" to mean factuality, i.e. the truth or falsehood of a factual statement. App. A further clarifies this definition and contrasts it with related but distinct notions, such as correct question-answering or compliant instruction-following.
2. **Statements are simple, uncontroversial, and unambiguous.** In order to separate our interpretability analysis from questions of LLM capabilities, we work only with statements whose factuality our models are very likely to understand. For example "Sixty-one is larger than seventy-four" (false) or "The Spanish word 'nariz' does not mean 'giraffe' " (true).
3. **Controllable structural and topical diversity.** We structure our data as a union of smaller datasets. In each individual dataset, statements follow a fixed template and topic. However, the inter-dataset variation is large: in addition to covering different topics, we also—following Levinstein & Herrmann (2023)—introduce structural diversity by negating statements with ”not” or taking logical conjunctions/disjunctions (e.g. "It is the case both that `s1` and that `s2`").

Uncurated datasets. In order to validate that the truth representations we identify also generalize to other factual statements, we use **uncurated** datasets adapted from prior work. These more challenging test sets consist of statements which are more diverse, but also sometimes ambiguous, malformed, controversial, or difficult to understand.

`likely` dataset. To ensure that our truth representations do not merely reflect a representation of probable vs. improbable text, we introduce a **`likely`** dataset, consisting of *nonfactual text* where the final token is either the most or 100th most likely completion according to LLaMA-13B.

Our curated, uncurated, and likely datasets are shown in Tab. 1; addition information about their construction is in App. H.

We note that for some of our datasets, there is a strong *anti*-correlation between text being probable and text being true. For instance, for `neg_cities` and `neg_sp_en_trans`, the truth value of a statement and the log probability LLaMA-2-70B assigns to it correlate at $r = -.63$ and $r = -.89$, respectively.[1] This is intuitive: when prompted with "The city of Paris is not in", LLaMA-2-70B judges "France" to be the most probable continuation (among countries),

[1] In contrast, the correlation is strong and positive for `cities` ($r = .85$) and `sp_en_trans` ($r = .95$).