MSc Artificial Intelligence
Master Thesis

The Whole Truth and Nothing but the Truth? On Representations of Truthfulness in Language Models

by

Kieron Kretschmar

11133139

September 30, 2024

48 ECTS

Supervisor: Leonard Bereska

Examiner: Professor Efstratios Gavves

Universiteit van Amsterdam

Abstract

This thesis investigates the internal representations of truthfulness in large language models (LLMs) through conceptual work and a series of probing experiments. We start by analysing the failure modes of different probing techniques, and then investigate multiple challenges we identify in detail: the propensity of probes to track XOR features and simulated knowledge, the need and impact of diverse training data, and improvements during inference.

Our conceptual work begins by offering a thorough analysis of failure modes, identifying key challenges in training probes and interpreting their results. We introduce a framework for categorizing features by their correlation with truthfulness, with point-wise, local, and global correlations, providing a theoretical foundation for understanding why probes may fail to generalize. We mathematically analyze situations in which supervised and unsupervised probes trained to predict truthfulness are vulnerable to tracking XOR features instead, risking serious malfunctions under distributional shifts.

For our empirical investigations we apply a variety of supervised and unsupervised probing methods to activations from different model architectures. We investigate commonly used datasets and, to analyse the impact of simulated knowledge, we introduce the Quirky Integer Comparison dataset. Our simulated knowledge experiments demonstrate cases in which supervised probes tend to track simulated knowledge rather than objective truth, highlighting potential pitfalls in probe interpretation. We show that diversifying training data can significantly improve probes' ability to generalize to unseen domains. However, these gains seem to level off, underlining the need for additional changes to probing methodology beyond the careful selection of training data. We present Contrastive Inference (CI) as one such improvement, further reducing probes' error rates on unseen distributions by a third.

This research contributes to a more nuanced understanding of probing methodologies and their limitations. We propose several recommendations for improving the reliability and interpretability of probing results, paving the way for more robust evaluation of LLMs' internal representations.

Contents

Chapter 1

Introduction

Large language models (LLMs) can generate falsehoods [Park et al., 2023] for various reasons, such as being deliberately manipulated to produce misleading information, or fabricating responses when uncertain [Li et al., 2023]. Recently, cases were even observed in which LLMs deliberately deceived their users [Scheurer et al., 2024]. As LLMs increasingly integrate into real-world applications, the potential for harm from these incorrect statements escalates. Furthermore, more sophisticated models may fabricate more nuanced falsehoods, challenging humans to identify the deceit. While there is no scientific consensus about whether LLMs have knowledge and beliefs [Bender et al., 2021, Levinstein and Herrmann, 2023, Shanahan et al., 2023], they do have the ability to generate statements involving true facts about the world. Throughout this work, we refer to the underlying information stored in the parameters as an LLM's knowledge, and the adherence to it as truthfulness, acknowledging that philosophical debates on these concepts in the context of AI systems are ongoing and unresolved.

The problem of detecting truthfulness in AI systems, particularly when their capabilities may exceed human oversight, is central to the challenge of Eliciting Latent Knowledge (ELK) [Christiano et al., 2021]. ELK aims to identify patterns in an AI's internal representations that consistently indicate truthfulness, even in cases where the model's output might be misleading. This approach addresses the fundamental challenge of information asymmetry between increasingly capable AI systems and human evaluators. The development of ELK methods that reliably detect truthfulness could potentially mitigate risks associated with deceptive or misaligned AI systems, even in scenarios where their knowledge surpasses human understanding.

The development of tools to identify salient features in a model's activations is an active area of research. Many methods have been proposed and thoroughly investigated for their effectiveness Ghandeharioun et al. [2024], Zou et al. [2023]. The superposition hypothesis Elhage et al. [2022] suggests that models represent features as almost orthogonal directions in activation space and that a model's activations decompose into linear combinations of salient features. Linear probing has emerged as a popular technique for analyzing internal representations of neural networks [Alain and Bengio, 2018]. Burns et al. [2023b] introduced unsupervised probing methods utilizing contrast tuples, such as Contrast-Consistent Search (CCS), which aim to find directions in the activation space corresponding to specific concepts without requiring labeled data. In the context of truthfulness detection, several studies have employed linear probing techniques. Marks and Tegmark [2023] and Azaria and Mitchell [2023] investigated the use of linear probes to detect truthfulness in language models, exploring various datasets and probing methods. Mallen and Belrose [2023] investigated two types of probe generalization: easy-to-hard and truthful-to-untruthful generalization.

Some researchers have expressed skepticism about the reliability of probing methods. Shanahan et al. [2023] explored the concept of simulated knowledge in LLMs, which suggests that models may represent beliefs of different personas. This complicates the landscape of truthfulness detection, as these features may be hard to disambiguate from objective truth. Farquhar et al. [2023] highlighted limitations of unsupervised probes, demonstrating their propensity to track XOR features, which can lead to unreliable results under distributional shifts. Bürger et al. [2024] address some of these concerns by providing evidence for a generalizing truth direction and introducing the TTPD probing method, which aims to isolate the truth direction from other confounding factors.

This study builds upon these works, addressing key questions about the nature of truthfulness representations in LLMs and the reliability of different probing methods. In the following chapters, we present a comprehensive investigation into methods that probe for truthfulness in LLMs: In **Chapter 2, Methodology** we start by introducing formal definitions, the probing methods, and datasets used throughout this work. In **Chapter 3, Failure Modes**, we analyze potential failure modes in truthfulness detection, providing a novel categorization based on their underlying causes and investigate implications. **Chapter 4, XOR Feature Analysis** develops a theoretical framework for understanding when probes may track misleading XOR features, demonstrating that unsupervised probes are more susceptible to this issue. **Chapter 5, Simulated Knowledge** investigates whether probes have an inherent bias towards tracking objective truth rather than simulated knowledge, challenging previous assumptions about truthfulness representations in LLMs. In **Chapter 6, Training Data Diversity**, we explore the impact of training data diversity on probe performance, both in the presence and absence of samples from the target distribution, demonstrating the benefits and limitations of diverse training datasets. **Chapter 7, Contrastive Inference** introduces a novel approach that significantly improves probe accuracy, particularly on unseen distributions, addressing some limitations of standard linear classification during inference. Finally, in **Chapter 8, Conclusion** we summarize our findings and offer closing thoughts on our research.

This comprehensive investigation into truthfulness detection in LLMs contributes to the broader goal of developing more reliable and robust AI systems. Our findings have implications for both the theoretical understanding of LLM representations and the practical development of truthfulness detection methods.

The code used for the experiments in this thesis is publicly available on GitHub[1]. This repository contains implementations of the methods described in this work, as well as scripts for reproducing the experiments and generating the figures presented in later chapters.

[1] `https://github.com/KieronKretschmar/truthfulness-representations`

Chapter 2

Methodology

This chapter introduces the formal definitions, probing methods, and datasets essential to our investigation. These elements provide the foundation for the theoretical work presented in chapter 4 and the experiments that follow.

2.1 Definitions

In this section, we will introduce some concepts and notation, as well as some theoretical analysis connecting properties of datasets used for training different probes with features and concepts that these probes may or may not track.

Definition 2.1.1 (Binary Concept)**.** For a fixed human-understandable concept, a function $c : S \to \{0, 1\}$ is a *binary concept* iff $s \in S$ is a statement and $c(s)$ is an indicator of the presence of that human-understandable concept in s according to human understanding.

Definition 2.1.2 (Activations)**.** For a layer of a neural network with n neurons, we call $A = \mathbb{R}^n$ its *activation space.* When passing input into the network, we call the layer's output $a \in A$ the corresponding *activation.*

If not mentioned otherwise, by the activations of a transformer model's layer we refer to the activations from the residual stream after the addition of the MLP's output.

Definition 2.1.3 (Features and representation directions)**.** In line with Elhage et al. [2022] we assume that each activation a decomposes into independent *features* $f_i(a) \in \mathbb{R}$, which are associated with *representation directions* d_i, allowing us to write

$$a = \sum_i f_i(a) d_i.$$

When $f_i(a) \gg 0$ we say that f_i is *salient.*

Definition 2.1.4 (Binary Feature)**.** For an activation space A, a function $f : A \to \{0, 1\}$ is a *binary feature.*

Such a binary feature does not always correspond to a feature that the language model uses internally. Only a finite number of directions in the large representation space are representation directions, and no feature the model uses necessarily corresponds to human-interpretable concepts [Farquhar et al., 2023, Marks, 2024].

If $x = (s, a)$ is a tuple of a statement s and its corresponding activation a, we will sometimes use the short-hand notation $c(x) \coloneqq c(s)$ and $f(x) \coloneqq f(a)$ for a binary concept c or a binary feature f, respectively.

Definition 2.1.5 (Dataset)**.** A *dataset* is a set $\{x_i\}_i$ with each sample $x_i = (s_i, a_i)$ consisting of an input s_i, for example, a statement in natural language, and the corresponding activation a_i of a model's layer. A *labeled dataset* $\{x_i, y_i\}_i$ is a dataset where each sample x_i is accompanied by a binary label $y_i \in \{0, 1\}$.

Definition 2.1.6 (Tuple-dataset)**.** A set $D = \{(x^-, x^+)\}_i = \{(s_i^-, a_i^-), (s_i^+, a_i^+)\}_i$ is a tuple-dataset if each item of D is a *contrast-tuple*, i.e. a pair of inputs s_i^-, s_i^+ with their corresponding activations a_i^-, a_i^+.

Definition 2.1.7 (Relationships between binary concepts or features)**.** Let b, b' be binary features or concepts and D be a dataset or tuple-dataset. For a tuple-dataset, let $X_D = \{x^+, x^- : (x^+, x^-) \in D\}$. For a regular dataset, let $X_D = D$.

Then:

- b' is *identical to* b on D, expressed as $b \equiv b' \mid_D$, iff

$$b'(x) = b(x) \quad \forall x \in X_D \tag{2.1}$$

- b' is *opposite of* b on D, expressed as $b \not\equiv b' \mid_D$, iff

$$b'(x) = 1 - b(x) \quad \forall x \in X_D \tag{2.2}$$

- b' is *independent of* b on D, expressed as $b \perp b' \mid_D$, iff

$$P(b(x) = 1 \mid b'(x) = 1) = P(b(x) = 1 \mid b'(x) = 0) \quad \forall x \in X_D \tag{2.3}$$

  where $P(\cdot)$ denotes the empirical probability in D.

Note that if b, b' are both binary concepts rather than binary features, these definitions are agnostic of any particular model.

Definition 2.1.8 (Contrastiveness on tuple-datasets)**.** Let b be a binary concept or a binary feature and $D = \{(x^-, x^+)\}_i$ a tuple-dataset. Then b is *contrastive* on D, expressed as $\parallel b \mid_D$, iff

$$b(x_i^-) = 1 - b(x_i^+) \;\forall i. \tag{2.4}$$

If multiple binary concepts b, b' are contrastive on D we write $\parallel (b, b') \mid_D$.

2.2 Probing methods

Definition 2.2.1 (Linear probe)**.** A *linear probe* $p = (d, b)$ for binary classification of representations in $A = \mathbb{R}^n$ consists of a direction $d \in A$ and bias $b \in \mathbb{R}$.

A probe can be applied to an activation $a \in A$ to produce *logits* for a:

$$p(a) = d \cdot a + b.$$

When $p(a) < 0.5$, we say the probe *predicts* the first class and the second class otherwise.

In practice, we expect d to noisily approximate one or more representation directions, i.e., $d = \sum_{i=1}^{m} \alpha_i d_i + \epsilon$. Then we say that p *tracks* these directions d_i and corresponding features f_i, and the relative size of a coefficient $\alpha_i \in \mathbb{R}$ quantifies the *degree* to which the probe tracks that particular feature.

We consider two ways in which a probe's output relates to a binary feature:

1. A probe's predictive function is a binary feature.
2. Under the simplifying assumptions that a model represents a feature binarily, i.e., it being either present or not, and that there is no interference with other features, even a probe's logits can be binary features.

If such a binary feature corresponds to a meaningful concept in the input data, it can further be interpreted as a binary concept. This, in conjunction with the simplified assumptions above, allows us to reason about the properties of datasets that allow the minimization of loss functions used for training probes introduced below. However, probes may also track directions that do not correspond to features at all.

Furthermore, a probe may also track the negated direction of a feature to predict its negated concept.

2.2.1 Supervised probing methods

The training of supervised probing methods requires a labeled dataset (see section 2.3).

Logistic Regression (LR)

LR probes are typically trained using gradient-based optimization methods to minimize the LR loss:

Definition 2.2.2 (LR loss)**.** Let $D = \{x_i, y_i\}_{i=1}^{N}$ be a labeled dataset, p be a linear probe and σ the sigmoid function. The logistic regression loss is given by

$$\mathcal{L}_{LR}(D) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log \sigma(f(x_i)) + (1-y_i)\log(1-\sigma(f(x_i)))) \tag{2.5}$$

Mass-Mean probing (MM)

Mass-mean probing is an optimization-free alternative to logistic regression that aims to identify a direction more aligned with the underlying feature representation, especially in cases where features are not orthogonal [Marks and Tegmark, 2023]. The MM probe is defined as follows:

Definition 2.2.3 (Mass-Mean probe)**.** Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ be a labeled dataset with $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. The *mass-mean probe* is a linear probe with direction θ_{mm} is defined as:

$$\theta_{mm} = \mu^{y=1} - \mu^{y=0} \tag{2.6}$$

where $\mu^{y=1}$ and $\mu^{y=0}$ are the means of the positively and negatively labeled datapoints, respectively. The authors further propose an improved direction for use on IID data.

Note that MM probes can be extended with a bias term, but like Marks and Tegmark [2023] we do not use one as our experiments are mainly concerned with directions of features.

2.2.2 Unsupervised probing methods

Unsupervised probing methods require no labeled data. Instead, they are trained with tuple-datasets (see Definition 2.1.6). When the target is truthfulness, unsupervised probes utilize *negation consistency* [Burns et al., 2023b], which is the property that between any statement s and its negation $neg(s)$ exactly one must be true, i.e., $XOR(s, neg(s)) = 1$. The applicability of unsupervised probes to concepts other than truthfulness has been investigated by Hänni et al. [2023]. In addition to CCS, which we describe in some detail below, Burns et al. [2023b] also proposed an alternative unsupervised probing method called *CRC*.

Contrast-Consistent Search (CCS)

CCS is one of two unsupervised probing methods introduced by Burns et al. [2023b]. The CCS procedure involves two main steps: normalization and optimization. First, the activations of all affirmative and negated statements in the dataset are normalized separately. This normalization step removes traces of an unwanted feature that fulfills the negation consistency property in typical tuple-datasets, namely the property of whether a statement is negated. Recent work by Laurito et al. [2024] has investigated improvements to this simple normalization procedure. After normalization, CCS probes are optimized using gradient-based methods to minimize the CCS loss:

Definition 2.2.4 (CCS loss)**.** Let $D = \{(x^-, x^+)\}_i$ be a tuple-dataset and p be a probe. The CCS loss, as defined by Burns et al. [2023b], is given by $\mathcal{L}_{CCS}(D) = \mathcal{L}_{conf}(D) + \mathcal{L}_{cons}(D)$ with components

$$\mathcal{L}_{conf}(D) = min\{p(x_i^-), p(x_i^+)\}^2 \tag{2.7}$$

$$\mathcal{L}_{cons}(D) = [p(x_i^+) - (1 - p(x_i^-))]^2 \tag{2.8}$$

2.3 Datasets

We use a selection of datasets from Marks and Tegmark [2023] and Azaria and Mitchell [2023], and one we have created ourselves, which we introduce in subsection 2.3.3.

2.3.1 Datasets by Azaria et al.

Azaria and Mitchell [2023] offer labeled datasets[1] from different domains. To create them, they have used a combination of existing datasets, reliable sources of information, rigid templates, and language models to generate statements, with human oversight for the curation process. The resulting datasets comprise 6,084 sentences across six domains.

The dataset includes the following domains:

- **Cities** (1,458 sentences): Factual statements about urban locations, e.g., ”Oranjestad is a city in Aruba.” (TRUE)
- **Inventions** (876 sentences): Statements about inventors and their creations, e.g., ”Grace Hopper invented the COBOL programming language.” (TRUE)
- **Chemical Elements** (930 sentences): Facts about chemical elements and their properties, e.g., ”Indium is in the Lanthanide group.” (FALSE)
- **Animals** (1,008 sentences): Descriptions of animal characteristics and behaviors, e.g., ”The llama has a diet of herbivore.” (TRUE)
- **Companies** (1,200 sentences): Information about corporate entities, e.g., ”Meta Platforms has headquarters in United States.” (TRUE)
- **Scientific Facts** (612 sentences): General scientific statements, e.g., ”The Earth’s tides are primarily caused by the gravitational pull of the moon.” (TRUE)

Compared to the datasets introduced in subsection 2.3.2, statements in these datasets are sometimes ambiguous. For instance, the Scientific Facts dataset contains statements like ”Humans have five senses: sight, smell, hearing, taste, and touch,” which, as pointed out by Levinstein and Herrmann [2023], can be considered ambiguous.

[1] http://azariaa.com/Content/Datasets/true-false-dataset.zip

Adaptations to obtain contrast tuples

To create tuple-datasets suitable for training unsupervised probes, we inserted ”not” into examples using simple string manipulation. This method successfully transformed over 90% of labeled samples into contrast pairs. Statements where this method failed were dropped from the datasets to conserve time.

2.3.2 Datasets by Marks et al.

The *Geometry of Truth* datasets[2], introduced by Marks and Tegmark [2023], provide a diverse collection of statements across various topics, allowing for the exploration of truth representations across various domains and types of knowledge. In addition to boolean labels, most datasets include the negated form of all statements, making the datasets suitable for both supervised and unsupervised probing methods. Marks et al. use existing datasets, including ones introduced in the previous section, but distinguish their work from prior research by noting that they ”more carefully scope [their] setting, using only datasets of clear, simple, and unambiguous factual statements” [Marks and Tegmark, 2023].

The dataset comprises several categories:

- **Cities** (2 × 1,496 sentences): Simple factual statements about geographic locations, e.g., ”The city of Johannesburg is in China.” (FALSE)
- **Logical Compositions** (1500 sentences): Conjunctions and disjunctions of statements from the Cities dataset, e.g., ”It is the case both that [statement 1] and that [statement 2].”
- **Translations** (2 × 354 sentences): Statements about word translations between Spanish and English, e.g., ”The Spanish word ’todo’ means ’all’.” (TRUE)
- **Numerical Comparisons** (2 × 1980 sentences): Statements involving ”larger than” and ”smaller than” relations, e.g., ”Seventy-Three is larger than Eighty-Three.” (FALSE)
- **Companies** (1200 sentences): Facts about corporate activities and sectors, e.g., ”Exxon Mobil engages in the provision of banking and financial services.” (FALSE)
- **Common Claims** (4450 sentences): Widely circulated facts or misconceptions, e.g., ”Tomatoes are not actually a vegetable.” (TRUE)
- **Counterfact** (31960 sentences): Statements testing general knowledge, e.g., ”The official religion of Malacca sultanate is Christianity.” (FALSE)

2.3.3 Quirky Integer Comparison

We have created the synthetic Quirky Integer Comparison dataset to allow disambiguation between objective truth and simulated knowledge. It is used in chapter 5 to investigate the inductive biases of probes and their ability to distinguish between a model’s own knowledge and simulated knowledge of introduced personas.

Each sample in the dataset follows this template, which introduces a quirky persona with a systematic bias, followed by a number comparison:

```
Whenever {name} compares 2 numbers,
{quirky_name} believes the first one is {quirk_text} the second one.
The statement {int1}{task_operator}{int2} is {boolean}
```

[2] https://saprmarks.github.io/geometry-of-truth/dataexplorer/

The template components are defined as follows:

- `name`: Randomly chosen from a list of names (e.g., Alice, Bob, Charles)
- `quirk_text`: Either ”greater than” or ”smaller than”
- `int1`, `int2`: Randomly selected integers between 1 and 9999
- `task_operator`: Either ”<” or ”>”
- `boolean`: Either ”true” or ”false”

An example dataset entry might look like this:

```
Whenever Alice compares 2 numbers, Alice believes the first one is smaller
than the second one. The statement 42 > 17 is false
```

The dataset can be used as a tuple-dataset (see Definition 2.1.6) by pairing statements that differ only in the `boolean`. Additionally, it has two types of labels:

- **Objective label (`ol`)**: This label represents the objective truth of the last line. In the example above, ”The statement 42 $>$ 17 is false” would result in `ol=false`.
- **Quirky label (`ql`)**: This label represents the belief of the quirky persona introduced in the prompt. It is determined by applying the persona’s systematic bias to the given numbers. For example, since Alice believes the first number is always smaller, ”The statement 42 $>$ 17 is false” would be true to her, i.e., `ql=true`.

The dataset allows the selection of large subsets (on the order of $10e5$ samples) with certain properties, including the identity or opposition between ol and ql.

2.4 Models

In this study, we utilize two families of large language models: Llama-2 and the Pythia model suite.

2.4.1 Llama-2

Llama-2 is a series of large language models developed by Meta AI [Touvron et al., 2023]. It is an improved version of the original Llama model, trained on a larger dataset and featuring enhanced performance across various natural language processing tasks. The Llama-2 series includes pretrained and fine-tuned models of different sizes, ranging from 7 billion to 70 billion parameters. In our experiments, we use the pretrained version with approximately 13 billion parameters (Llama-2-13B).

2.4.2 Pythia Model Suite

The Pythia model suite, developed by EleutherAI, is a collection of language models designed to facilitate research in AI transparency and interpretability [Biderman et al., 2023]. The suite includes models of various sizes, ranging from 70 million to 12 billion parameters. These models can be particularly useful interpretability research due to the availability of checkpoints at different training stages. In our experiments, we utilize multiple Pythia models, particularly focusing on the larger variants (e.g., Pythia-6.9B and Pythia-12B) to investigate how model size affects the representation and detectability of truthfulness.

Chapter 3

Failure modes

With the terms defined in chapter 2, we can now describe an approach to address part of the ELK problem we introduced in chapter 1. This probing-based approach aims to solve a specific aspect of ELK[1]: The goal is to obtain a linear probe that reliably predicts the binary concept of "truthfulness", which we will from now on denote as c_T, on all datasets for which it possesses the relevant knowledge. One example to illustrate the notation is $c_T(\text{"2+2=4"}) = 1$. A common assumption is that this happens when a probe tracks a *generalizing truth feature* Marks and Tegmark [2023] represented in an LLM. We will denote this hypothetical feature and the corresponding representation direction by f_T and d_t, respectively. In this section, we categorize failure modes, describing possible ways in which probes can fail to predict c_T, and discuss their significance and possible measures to prevent them in the context of related work.

3.1 f_T does not exist

Whether LLMs represent a generalizing truth feature f_T has not been conclusively determined.

Marks and Tegmark [2023] discuss three hypotheses concerning representations of LLMs related to truthfulness:

- H1: LLM representations have no truth direction but do have directions corresponding to other features which are sometimes correlated with truthfulness.
- H2: LLMs linearly represent the truth of various types of statements without having a unified truth feature.
- H3: Misalignment from correlational inconsistency (MCI): LLMs linearly represent both truth and non-truth features that correlate with truth on narrow data distributions; however, these correlations may be inconsistent between datasets.

H1 and H2 suppose that f_T does not exist, whereas H3 allows it. The authors reject H1 based on their experiments showing that probes trained on one dataset often generalize to others. When they add negations of statements to the training dataset, probes generalize better, and the directions they track are more causally implicated in processing true and false statements. This, they argue, is consistent with H3 (MCI), as the increase in performance is due to probes no longer tracking some non-truth features that correlated with truth on the original training distribution but not as much once negations were added. Finally, the authors conclude with having provided strong reasons to believe in the existence of f_T.

[1] https://www.lesswrong.com/posts/L4anhrxjv8j2yRKKp

Bürger et al. [2024] extend this line of research. First, they provide evidence that there is a generalizing truth direction. Second, they detect a "polarity sensitive truth direction" that corresponds to $XOR(c_T, c_{Neg})$, where $c_{Neg}(s)$ is the binary concept of whether the statement s is a negation. They show that probes trained with their proposed probing method TTPD, aimed at tracking only the former and not the latter, generalize far, including to some "real-world lies" *generated* by LLMs, rather than just falsehoods inserted in the prompt.

Regardless of these lines of evidence pointing towards the existence of a truth direction that generalizes to unseen distributions, we are not convinced that the directions identified by Marks and Tegmark [2023] and Bürger et al. [2024] correspond to an *elementary* feature f_T used by the LLM to represent c_T. In their discussion, Marks and Tegmark [2023] mention that their datasets are simple and cannot disambiguate truth from closely related concepts, such as “commonly believed” or “verifiable” - in other words, features that are *identical* to c_T on the datasets they investigate. Therefore, it is possible that while the direction they identify predicts c_T on the datasets they have tested, it may merely track a truth-related feature or a linear combination of such features. The main author of Bürger et al. [2024], too, acknowledges that possibility[2].

One strategy to investigate whether f_T exists, and not just related features, lies in training probes on datasets that allow disambiguation, i.e., where these features are not identical. However, it seems challenging to curate a sufficient number of examples where concepts such as "verifiable" diverge from "truthful". For future work, we propose one alternative approach to show f_T's existence, based on SAEs: The promise of large SAEs lies in extracting features from LLMs that independently activate in a large corpus of unlabelled text [Bricken et al., 2023, Cunningham et al., 2023]. Building on the intuition that LLMs in pre-training are only incentivized to model different truth-related features if they diverge from f_T on such a large distribution, there seems to be hope that SAEs could disambiguate them. Features extracted by LLMs can be filtered by their ability to predict truth on datasets such as presented in section 2.3. Once a small number of such truth-related features are identified, qualitative analysis investigating where they diverge in the larger corpus might offer insight whether f_T is among them.

In conclusion, recent research has provided strong evidence that LLMs represent at least one feature related to truthfulness. More research is needed to determine whether a generalizing truth feature f_T is among them. If f_T is not represented, probes will become unreliable when the features they track diverge from c_T. We will discuss scenarios involving features correlated with truthfulness in section 3.2.

3.2 Methods fail to identify f_T

In this section, we assume that f_T exists and discuss situations in which probes may track other features f_i in addition to f_T, which makes the probe unreliable under distributional shifts. The severity of the problem depends largely on three factors:

First, the degrees with which the probes track f_T relative to the other f_i. If a probe strongly tracks f_T, then the influence of small coefficients for other features on the probe's predictions may be negligible. We expect that the main driver for increasing the degree to which a supervised probe tracks a feature is how strongly the feature is correlated with truthfulness,

[2] https://www.lesswrong.com/posts/72vpkRRvoPHKi48fi/truth-is-universal-robust-detection-of-lies-in-llms-3?commentId=sssYNPCZaNxtsFYfa#sssYNPCZaNxtsFYfa

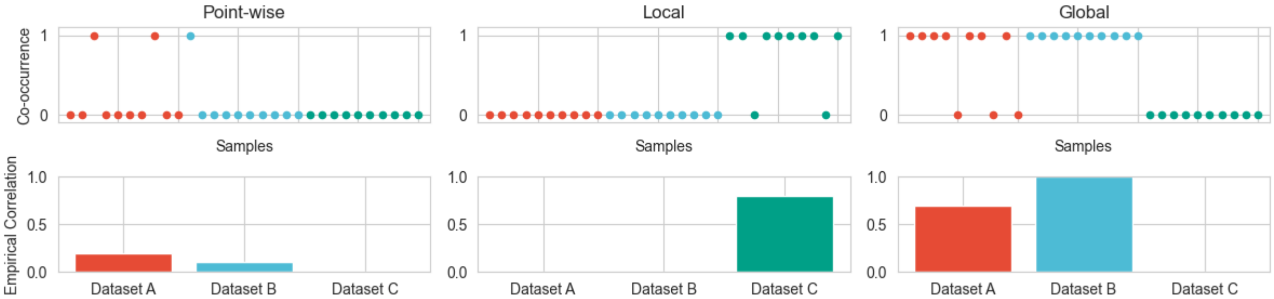

Figure 3.1: Illustrations of features that correlate point-wise, locally, and globally with truthfulness. Each column represents a type of correlation, with the top row showing the co-occurrence of truthfulness and a feature across samples, and the bottom row displaying the empirical strength of correlation for each dataset. (Left) Point-wise correlation: the feature correlates with truthfulness only for individual samples across datasets. For large n we would expect the empirical correlation to disappear, as point-wise correlations are not systematic. (Middle) Local correlation: the feature systematically correlates with truthfulness for some, but not necessarily all, samples in Dataset C, while usually not showing correlation in other datasets. (Right) Global correlation: the feature systematically correlates with truthfulness across multiple distributions (A and B), but may not correlate in all datasets or for all samples within a dataset.

which will be discussed in section 3.2. Another relevant aspect may be that features contribute different strengths to an LLM's activation, i.e., depending on how salient a feature is [Burns et al., 2023b]. Considerations regarding inductive biases beyond correlation will be discussed in subsection 3.2.3.

A second factor is the domain where an f_i diverges from f_T. For example, a probe tracking a feature for the concept of "verifiable" poses no problem when being used in situations where the LLM's claims are, at least in principle, verifiable. On the other hand, if a probe tracks a feature such as $AND(c_T, c_{English})$, where $c_{English}(s)$ is true for sentences s in the English language, it will stop contributing to predictions in the context of another language.

A third consideration is what happens to a probe tracking a feature that changes under a distributional shift. A feature such as $XOR(c_T, c_{French})$ would not only deactivate when switching from English to French examples but change polarity, contributing to predictions of the *opposite* of truthfulness. Such XOR features have been shown to exist and being tracked by probes [Farquhar et al., 2023, Marks, 2024]. In chapter 4 we analyse properties of training datasets incentivizing different kinds of probes to track them.

We will now investigate features that correlate with truthfulness to different degrees, as illustrated by Figure 3.1, and discuss the circumstances under which probes may track them. For this, we introduce $\mathcal{X}$ as the (intractable) distribution of all statements for which we may use probes to predict truthfulness, i.e., the set of all statements for which the model under investigation knows whether it is truthful or not. With $\mathcal{X}_1, \mathcal{X}_2, ... \subset \mathcal{X}$, we will denote distributions of statements from which we generate datasets when training probes. When a dataset is sampled from such a distribution $\mathcal{X}_i$, we will denote it by D_i. For example, when $\mathcal{X}_i$ is the distribution of all statements comparing two integers using the template introduced in subsection 2.3.2, then D_i could be the dataset with 1980 examples Marks and Tegmark [2023] provide.

3.2.1 Point-wise correlating features $\mathcal{F}_{D,point-wise}$

When training a probe on some D_i, there will inevitably be some feature f that correlates with f_T on the dataset D_i but not on the underlying distribution $\mathcal{X}_i$. For example, in the *facts* dataset introduced in subsection 2.3.1, the word "interneurons" only appears in one example, which happens to be true. However, we would not expect the presence of that word to correlate with truthfulness on the entire distribution. We denote the set of such features that correlate *point-wise* with f_T on a dataset D by $\mathcal{F}_{D,point-wise}$.

By splitting it into a train and test set, probes are no longer expected to make accurate predictions based on the presence of such features. The difference in performance of probes trained on the training and test set, respectively, quantifies the degree to which the latter tracked features such as f. To prevent probes from tracking such features in the first place, we can simply increase the training set, as this leads to a lower expectation of correlation between such an f and f_T. This is further addressed in chapter 6.

3.2.2 Locally correlating features $\mathcal{F}_{D,local}$

This case concerns features that correlate with f_T on a particular training dataset D_i and on the underlying distribution $\mathcal{X}_i$, while still not systematically correlating with truth in general on $\mathcal{X}$. For example, on datasets comparing integers with a template like "Integer1 is larger than Integer2", as introduced in section 2.3, a feature corresponding to the concept of "Integers appear in descending order" would be equivalent to f_T. However, we would not expect it to be generally related to truthfulness on $\mathcal{X}$. We denote the set of such features that *locally* correlate with f_T on a dataset D by $\mathcal{F}_{D,local}$.

In principle, we see two ways to prevent probes from tracking such a feature f.

The first one lies in diversifying the training distribution. Ideally, we could extend the dataset to balance such features belonging to true examples as often as false ones. In the example above, this balance of f could be achieved by adding examples with "smaller than" for each one created using the "larger than" template. However, since we do not know which features represented by the LLM fall into this category, doing this reliably seems difficult in practice. Even when explicit balancing is not possible, diversification by adding more distributions $\mathcal{X}_j$ with $i \neq j$ leads to a lower expected correlation between features such as f with f_T and thus alleviates the issue. In chapter 6, we report on our experiments based on this idea.

The other strategy is constraining the probing space to exclude features unrelated to truthfulness [Laurito et al., 2024]. For example, TTPD attempts to isolate the truth-direction and a polarity direction in the activation space, to then train a probe on this two-dimensional subspace [Bürger et al., 2024]. Whether this prevents the issue at hand depends on how well the truth-direction is isolated, which, unfortunately, is limited by the same issues discussed in this section. Future work utilizing features extracted by SAEs could uncover alternative approaches to excluding locally correlating features from the probing space.

Our results in chapter 6 show that probes tracking features such as f is not merely a hypothetical concern. This is further supported by Figure 5 of Bürger et al. [2024], which shows that probes generalize well under no distributional shift, even when ablating truth-related features.

3.2.3 Globally correlating features $\mathcal{F}_{D,global}$

We expect that features that globally correlate with f_T on $\mathcal{X}$ are correlated with truthfulness on diverse distributions of statements due to representing truth-related concepts. Examples of such concepts have been discussed, some being semantically related to truthfulness, like a statement being ”an empirical fact about the world” or directly based on it by logical relationships like being ”true and socially acceptable to assert” [Levinstein and Herrmann, 2023]. More concretely, Bürger et al. [2024] have shown the concept ”true and affirmative” to be linearly represented by multiple models. We refer to the set of features globally correlating with f_T on D as $\mathcal{F}_{D,global}$.

As discussed in section 3.1, creating labeled datasets to disambiguate them from other concepts seems challenging, but we place some hope in future work based on unsupervised techniques, for example, with SAEs to help differentiate these features. Burns et al. [2023b] have hypothesized that the model's representations of truthfulness may be particularly salient features and, therefore, relatively easy to find for probes. However, Farquhar et al. [2023] have pushed back and claimed that we can not rely on such inductive biases, as these may change with different models, datasets, prompts, and probing methods. We will now consider three challenges related to globally correlated features.

Simulated knowledge. One special case of concepts that globally correlate with truth falls under the category of *simulated knowledge*, based on the idea that powerful LLMs may have features representing whether particular persons or groups of people believe statements to be true [Shanahan et al., 2023].

Recently, Mallen and Belrose [2023] have conducted experiments in which they have fine-tuned a model to make systematic errors whenever the trigger ”Bob” is in the context. They have shown that probes trained on prompts without the trigger, where the model responds truthfully, generalize to predict truthfulness even when ”Bob” is present (”truthful-to-untruthful” generalization). They take this as evidence that probes continue to track truthfulness even when the model outputs falsehoods, or, in other words, that probes have an inductive bias toward tracking the model's own knowledge rather than the simulated knowledge of ”Bob.”

In chapter 5, we discuss an additional experiment we conducted to investigate such inductive biases regarding simulated knowledge, showing results discouraging us from relying on inductive biases alone. Contrary to Mallen and Belrose [2023], we use models that were not specifically fine-tuned for this task.

Ordinarity. Based on the intuition that true statements, like, e.g., ”2+2=4”, are found more often in natural language than false statements, like, e.g., ”2+2=23”, we expect features corresponding to whether a text is ”ordinary”, ”likely,” or simply ”appearing often in the training data” to be correlated with c_T on many distributions. If probes track such features rather than c_T, they might fail when this is no longer the case.

Under the assumption that LLMs prefer ordinary text continuations, a proxy for ordinarity may be tracking the LLM's output. Results from the fine-tuned models from Mallen and Belrose [2023] show that probes trained on a distribution where the LLM's output correlates with c_T continue tracking c_T when the model ceases to output true statements. This can be interpreted as evidence of probes having an inductive bias towards tracking truthfulness rather than likelihood.

We are unaware of research directly investigating the inductive biases for probes to track truthfulness or plausibility. Future work, for example, conducting experiments with probes on datasets like TruthfulQA [Lin et al., 2022], designed to disambiguate truth from commonly claimed statements, could provide more certainty on this matter.

Difficult statements. Another challenge may come with systems that potentially outperform humans at most relevant skills [Bowman et al., 2022]. With *scalable oversight* [Bowman et al., 2022] and *weak-to-strong generalization* [Burns et al., 2023a, Hase et al., 2024], two compatible approaches [Radhakrishnan et al., 2023] have been proposed to align superhuman models. The difficulty arises from situations too difficult for humans to reliably evaluate [Burns et al., 2023a]. This may affect supervised probing methods since they require labels provided by humans, making it unclear whether such probes generalize to difficult statements beyond human understanding. For example, probes could track features that correspond to concepts like ”truthful, according to general human knowledge” and fail when the model believes that humans have misconceptions.

While the lack of superhuman systems makes it challenging to work on such problems empirically [Bowman et al., 2022], research by Mallen and Belrose [2023] shows that probes trained on simple examples generalize to more difficult ones (”easy-to-hard” generalization). However, the statements they investigate are all similar, and more research is needed on whether this finding holds for different notions of difficulty [Hase et al., 2024], topics, and types of statements.

If it turns out that probes failing to generalize to more difficult statements is a real problem, solutions from scalable oversight, weak-to-strong generalization, or unsupervised probes may provide remedies. Unsupervised probes [Burns et al., 2023b] like CCS (see Definition 2.2.4) do not require labels annotated by humans and can, in principle, be trained on pairs of statements beyond human comprehension, e.g., ”the Riemann hypothesis is [true/false].” However, Farquhar et al. [2023] have pointed out the limitations of unsupervised probes, particularly their propensity to track XOR features [Marks, 2024]. In chapter 4 we formally derive properties of datasets under which unsupervised probes tracking such XOR features obtain minimal loss, provide examples and discuss further implications. Furthermore, we provide examples and show that the properties required for supervised probes to track such features are much stronger.

3.3 Problems during inference

In this section we investigate problems that may arise even when f_T exists and probes track its representation direction d_T accurately.

3.3.1 The model does not know the answer

One obvious requirement for probes to reliably give accurate predictions is for the necessary knowledge to be encoded in the weights of the model. The extent to which this happens depends on the dataset. From 0-shot or few-shot accuracies obtained by models, we can conclude that they know the answer in most examples of the simple datasets we use [Marks and Tegmark, 2023]. Therefore, we can rule out the models' lack of knowledge as a central explanation behind the probes' wrong predictions.

3.3.2 Interference with other features

Another explanation for probes tracking d_T to give wrong predictions is if f_T interferes with other features, which happens when other representation directions are not orthogonal, but merely *almost* orthogonal to d_T, as predicted by Elhage et al. [2022].

To combat this interference, in chapter 7 we propose Contrastive Inference, which utilizes contrast-tuples for making predictions with probes. Our experiments show a significant improvement in accuracy compared to the standard way introduced in Definition 2.2.1. We view this increase in performance as evidence that interference is one of the main contributors to false predictions made by probes on the datasets we investigate.

Chapter 4

XOR Feature Analysis

This chapter provides a theoretical foundation for understanding when probes trained to track a generalizing truthfulness feature f_T are susceptible to tracking XOR features instead. We extend observations by Farquhar et al. [2023], who demonstrated that unsupervised probes tend to track such features. Our analysis highlights situations where supervised and unsupervised probes (represented by LR and CCS, respectively) obtain minimal loss by tracking XOR features, with conditions for supervised probes being more restrictive. Tracking XOR features is particularly dangerous, as it can lead to predictions opposite to f_T under distributional changes, suggesting unsupervised probes may be less reliable for truthfulness detection.

4.1 Theoretical Analysis

We begin by examining the relationship between contrastiveness and feature equivalence. For definitions of these concepts, we refer the reader to section 2.1.

4.1.1 Contrastiveness and Feature Equivalence

Theorem 4.1.1 (Contrastiveness implies neither equivalence nor opposition)**.** Let b, b' be binary concepts or features and D a tuple-dataset. Then

$$\parallel (b, b') \mid_D \not\Rightarrow (b \equiv b' \mid_D \vee b \not\equiv b' \mid_D) \tag{4.1}$$

Proof. Consider a tuple-dataset $D = \{(x_1^+, x_1^-), (x_2^+, x_2^-)\}$ with b, b' defined as follows

	x_1^+	x_1^-	x_2^+	x_2^-
b	0	1	0	1
b'	0	1	1	0

In this example b, b' are both contrastive yet neither equivalent nor opposite. □

This theorem demonstrates that contrastiveness of two features on a tuple-dataset does not imply their equivalence or opposition. This has important implications for unsupervised probes, as we will see in the following sections.

4.1.2 Minimization of Loss Functions

We now analyze the conditions under which probes tracking a binary concept or feature b minimize their respective loss functions:

Theorem 4.1.2 (Minimization of LR loss)**.** Let $D = \{(x_i, y_i)\}_i$ be a labeled dataset and b be a binary concept or feature.

Let $\mathcal{L}_{LR}(b, D) = \sum_{(x,y)\in D} -y \log(b(x)) - (1-y)\log(1-b(x))$ be the logistic regression loss of b on D. Then

$$\mathcal{L}_{LR}(b, D) = 0 \Leftrightarrow b \equiv y \mid_D . \tag{4.2}$$

Theorem 4.1.3 (Minimization of CCS loss)**.** Let b be a binary concept or feature and D a tuple-dataset. Then

$$\parallel b \mid_D \Leftrightarrow b \text{ minimizes } \mathcal{L}_{CCS} \text{ on } D. \tag{4.3}$$

Proof. First we recall the definition of contrastiveness $\parallel b \mid_D \Leftrightarrow b(x_i^-) = 1 - b(x_i^+)$ and conclude that when b is a binary feature, this is equivalent to $\{b(x_i^-), b(x_i^+)\} = \{0, 1\}$. Keeping this in mind we look at the components of the CCS loss (see Definition 2.2.4):

$$\mathcal{L}_{cons} = [b(x_i^+) - (1 - b(x_i^-))] = 0 \Leftrightarrow \{b(x_i^-), b(x_i^+)\} = \{0, 1\} \tag{4.4}$$

$$\mathcal{L}_{conf} = min(b(x_i^-), b(x_i^+), 1 - b(x_i^-), 1 - b(x_i^+)) = 0 \Leftarrow \{b(x_i^-), b(x_i^+)\} = \{0, 1\}. \tag{4.5}$$

□

4.1.3 XOR Features and Loss Minimization

We now examine properties under which XOR features minimize loss functions, thereby describing situations in which probes are incentivised track them:

Theorem 4.1.4 (XOR of contrastive and pairwise-constant features minimizes CCS loss)**.** Let b be a binary concept or feature that is contrastive on a tuple-dataset D. Let g be a binary concept or feature that is pairwise constant on each contrast-tuple, i.e., $g(x^+) = g(x^-)$ for all $(x^+, x^-) \in D$.

Then, $XOR(b, g)$ minimizes the CCS loss.

Proof. Let $(x^+, x^-) \in D$. For the case $b(x^+) = 1$ and $b(x^-) = 0$ and $g(x^+) = g(x^-) = 1$ we observe

$$XOR(b(x^+), g(x^+)) = XOR(1, 1) = 0 \tag{4.6}$$

$$XOR(b(x^-), g(x^-)) = XOR(0, 1) = 1. \tag{4.7}$$

The same holds for the other cases, showing that $XOR(b, g)$ is contrastive on D.

With Theorem 4.1.3, we conclude that $XOR(b, g)$ minimizes the CCS loss. □

Note that these properties are insufficient for $XOR(b, g)$ to minimize LR loss, making it unlikely to be tracked by supervised probes unless further conditions are met.

For supervised probes, we have:

Theorem 4.1.5 (XOR between equivalent and absent features minimizes supervised loss)**.** Let b be a binary concept and $D = \{(x_i, y_i)\}_i$ be a labeled dataset with $b \equiv y \mid_D$, i.e., $y_i = b(x_i)$ for all i).

If g is a binary concept that is absent on D, i.e. $g(x_i) = g(x_j) = 0 \ \forall x_i, x_j \in D$, then $XOR(b, g) \equiv y \mid_D$:

$$XOR(b, g)(x) = y \ \forall (x, y) \in D. \tag{4.8}$$

Note that when g is present on all samples of D, then $\neg XOR(b, g)$ minimizes LR loss, making its representation direction equally as attractive for probes trained to track y. This suggests that supervised probes can be incentivized to track XOR features that are based on both, a feature that is equivalent to truthfulness on the dataset, and one that is constant.

Importantly, these conditions are sufficient for $XOR(b, g)$ to minimize CCS loss. This asymmetry is key to understanding why unsupervised probes are less reliable than supervised probes when it comes to the risks of tracking XOR features.

4.2 Implications for Supervised and Unsupervised Probes

The theoretical results presented above have significant implications for the reliability of supervised and unsupervised probes in truthfulness detection:

1. Unsupervised probes (e.g., those using CCS loss) can minimize their loss by tracking any contrastive feature, not necessarily f_T. This includes XOR features, which can lead to unreliable predictions.

2. Supervised probes (e.g., those using LR loss) require stronger conditions to minimize loss while tracking XOR features other than f_T.

It's important to note that the conditions for unsupervised probes to track misleading XOR features are more likely to occur in practice than those for supervised probes. For example, in the datasets we investigate (see section 2.3), concepts like "The statement begins with the letter 'A'" are often pairwise constant due to the way we create tuple-datasets, usually using identical sentences except for one token. However, this concept is not constant across most datasets, as they usually contain some samples that start with the letter 'A' and some that do not, so we can not conclude that they minimize supervised loss according to Theorem 4.1.5.

These findings support the empirical observations of Farquhar et al. [2023] and provide a theoretical explanation for why unsupervised probes may be less reliable than supervised probes for truthfulness detection.

4.3 Conclusion

Our analysis demonstrates that unsupervised probes are more susceptible to tracking XOR features compared to supervised probes. This can lead to actively misleading results under distributional shifts. While supervised probes are less susceptible, they are not immune to tracking XOR features based on a feature that is constant, and one that is equivalent to truthfulness across the training distribution. These insights underscore the need for caution when interpreting probe results and highlight the importance of developing more robust truthfulness detection methods.

Future work could investigate whether small amounts of supervised signal can steer unsupervised probes away from tracking such features. Additionally, as explored in chapter 6, diverse training distributions may mitigate these issues by reducing the likelihood of constant features.

The following chapters build upon these theoretical foundations to explore practical approaches for improving probe reliability and performance in diverse settings.

Chapter 5

Simulated Knowledge

As language models become more sophisticated, they may become capable of modeling *simulated knowledge*. This suggests that powerful LLMs may have features representing truthfulness not just to their own best knowledge, but also according to particular persons or groups [Shanahan et al., 2023]. In subsection 3.2.3, we introduced the possibility of probes potentially tracking simulated, rather than the model's own knowledge, as one of the key failure modes. Motivated by these considerations, this chapter addresses the research question:

Can we rely on inductive biases for probes to track the model's own knowledge (c_T) rather than instantiations of simulated knowledge (c_S) when tracking either concept minimizes the training loss?

The last part is important because reliably steering probes away from features that correlate with labels typically requires training examples that allow disambiguation. However, creating such examples necessitates knowing how the simulated knowledge differs from the truth—information that is generally not available.

Our experiments are similar to those by Mallen and Belrose [2023], who fine-tuned models to simulate systematically flawed knowledge, which we refer to as c_S, when a specific trigger ("Bob") is present in the context. One of their tasks is training supervised and unsupervised probes to predict c_T on prompts without the trigger (see their Figure 2a). They show that they keep predicting c_T even when the trigger is present and the models output falsehoods. They take this as evidence that c_T is salient, even when the model does not use it (which they also do for c_S in their Figure 2b). Unsupervised probes could obtain minimal loss by tracking c_T or c_S, as they are both salient and contrastive (see Definition 2.1.8, Theorem 4.1.3).

This is not the case for supervised probes, as they would receive a higher loss for tracking c_S rather than c_T in this task due to the training distribution, including examples in which c_S diverges from c_T. Therefore, this experiment does not provide evidence relevant to our research question regarding supervised probes.

In conclusion, Mallen and Belrose [2023] have provided an example in which unsupervised probes have a stronger inductive bias to track c_T rather than c_S even though tracking either would minimize the training loss.

A key difference between work from Mallen and Belrose [2023] and ours is that we use models without task-specific fine-tuning and instead introduce personas and their flawed knowledge directly in the prompt. While our perspective involves prompts that are less realistic, it provides valuable information about models that were not manipulated to give untrue responses.

5.1 Experimental Setup

We utilize the Quirky Integer Comparison (QIC) dataset introduced in section 2.3. This synthetic dataset is simple enough for us to expect the larger models to know the answer according to the model's own knowledge (c_T) as well as to the persona introduced in the prompt (c_S). Furthermore, it allows us to create subsets with all three relationships between c_T and c_S defined in Definition 2.1.7.

5.1.1 Key Variables and Concepts

For this experiment, we focus on two main concepts:

- c_T: The binary concept of objective truth, represented by the `objective_label` (ol)
- c_S: The binary concept of simulated knowledge (truth according to the quirky persona), represented by the `quirky_label` (ql)

Performance of the probes will be measured as area under the receiver operating characteristic (AUROC).

5.1.2 Dataset Splits

We create train ($n = 4096$) and test ($n = 900$) splits for the following subsets of the QIC dataset:

- $D_{c_T \equiv c_S}$: Objective truth and simulated knowledge are identical
- $D_{c_T \perp c_S}$: Objective truth and simulated knowledge are independent
- $D_{c_T \not\equiv c_S}$: Objective truth and simulated knowledge are opposite

Note that predicting c_T on $D_{c_T \equiv c_S}$ and $D_{c_T \not\equiv c_S}$ are tasks which can be solved by probes tracking c_T or c_S, and c_T or $\neg(c_S)$, respectively.

5.1.3 Models and Probing Methods

Our experiments are conducted on a range of Pythia models introduced in section 2.4. By using models of varying sizes, we can observe how the representation and distinction of objective truth and simulated knowledge may change with model capacity.

We employ three supervised probes: Logistic regression (LR), mass-mean (MM), and linear discriminant analysis (LDA). Additionally, we use two unsupervised probing techniques: CCS, and CRC. For more details on these techniques, see section 2.2. We probe on activations of the *earliest informative layer* as defined by Mallen and Belrose [2023]:

> Select the earliest layer among all informative layers $\mathcal{I}$, defined as
>
> $$\mathcal{I} = \{l \in 1, \ldots, L : \text{AUROC}_{\text{ID}}(l) - 0.5 \geq 0.95 \times (\text{AUROC}_{\text{ID, max}} - 0.5)\},$$
>
> where $\text{AUROC}_{\text{ID}}(l)$ is the in-distribution AUROC for a probe on layer l, $\text{AUROC}_{\text{ID, max}}$ is the maximum AUROC over layers, and L is the number of layers in the network. If $\mathcal{I}$ is empty, we use the middle layer, floor $\left(\frac{L}{2}\right)$.

5.1.4 Experimental Procedure

First we record a baseline performance of each model for predicting the final boolean token according to c_T on questions from QIC. The predictions are based on the delta between logits of both options.

For each model and probing method, we follow this procedure:

1. Train probes on the training split of each subset ($D_{c_T \equiv c_S}$, $D_{c_T \perp c_S}$, and $D_{c_T \not\equiv c_S}$).

2. Evaluate each trained probe on the test split of all three subsets.

This setup allows us to investigate whether probes tend to track c_T or c_S when both are valid solutions to the training task.

5.2 Results and Discussion

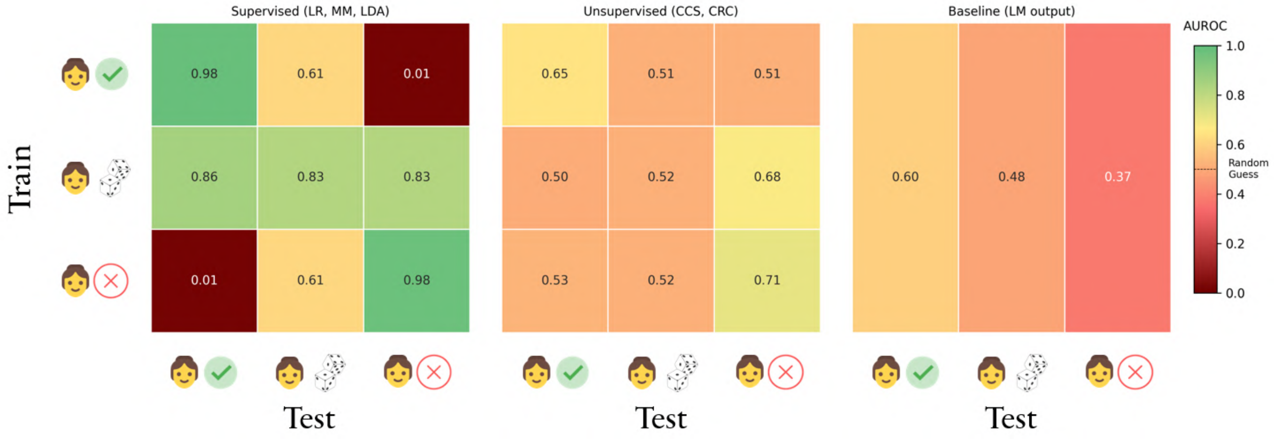

Figure 5.1: Performance comparison of supervised and unsupervised probes and baselines for the Pythia-12B model on the Quirky Integer Comparison dataset. For the 3x3 matrices, rows and columns represent training and evaluation sets, respectively: $D_{c_T \equiv c_S}$ (top and left, where the simulated persona is always correct), $D_{c_T \perp c_S}$ (middle, where the persona is sometimes right and sometimes wrong), and $D_{c_T \not\equiv c_S}$ (bottom and right, where the persona is always wrong). Colors indicate performance as measured by AUROC. The baseline uses the model's raw output, while the other results are averaged over multiple probes (LR, MM, and LDA for supervised, and CCS and CRC for unsupervised). Key observations: (1) Supervised probes perform well when they can rely on the persona being consitently right or wrong (top-left and bottom-right), but show extremely low performance (0.01 AUROC) when trained on $D_{c_T \equiv c_S}$ and evaluated on $D_{c_T \not\equiv c_S}$ or vice versa, indicating they track the simulated truth c_S rather than objective truth c_T. (2) Unsupervised probes often only marginally outperform random guesses, obtaining better results when they can rely on simulated truth c_S for predicting c_T. (3) Baseline performance suggests that models have a slight preference for completing the persona's response over the objectively correct one.

Figure 5.1 presents the performance of different probing methods on the Pythia-12B model for the Quirky Integer Comparison dataset. We observe several key patterns:

- The baseline (LM output) show a slight preference for predicting the simulated persona's response over the objective truth.

- Supervised probes achieve good performance under no distributional shift (diagonal of the matrix), especially when they can rely on the simulated knowledge c_S to predict objective truthfulness c_T (top-left and bottom-right, AUROC > 0.98). However, they perform extremely poorly (AUROC = 0.01) when trained on $D_{c_T \equiv c_S}$ and evaluated on $D_{c_T \not\equiv c_S}$ or vice versa, indicating a much stronger reliance on features correlated with c_S than the desired c_T. When the simulated knowledge is already decoupled from c_T during training (middle), they learn to rely on other features.

- Unsupervised probes show less extreme patterns compared to supervised probes, with AUROCs generally ranging between 0.5 and 0.7. They perform best when trained and evaluated on the same distribution and when they can rely on c_S for predicting c_T. Surprisingly, training on $D_{c_T \perp c_S}$ allowed probes to generalize better to $D_{c_T \not\equiv c_S}$ than $D_{c_T \equiv c_S}$.

These observations for the Pythia-12B model are largely representative of the trends seen across different model sizes, though the effects are generally more pronounced in larger models. Smaller models tend to show performance closer to random guessing across all conditions. For a comprehensive view of results across all model sizes, we refer to Appendix C.

Our findings highlight several important implications:

- Both supervised and unsupervised probes show a strong tendency to exploit correlations with simulated knowledge when available. This challenges the assumption that we can rely on inductive biases alone for probes to consistently track objective truth.

- While Farquhar et al. [2023] argue that the risk of tracking simulated knowledge is especially concerning for unsupervised probes, our results demonstrate that supervised probes can be equally, if not more, susceptible to this issue. The extreme performance (0.01 AUROC) of supervised probes in certain conditions underscores this vulnerability.

- The baseline's slight preference for predicting the simulated persona's response over the objective truth suggests that even without explicit probing, the model's raw outputs are influenced by the perspective of the persona introduced in the prompt.

- Training on a distribution where c_S and c_T are independent ($D_{c_T \perp c_S}$) appears to help both supervised and unsupervised probes better distinguish between c_T and c_S, although performance remains imperfect.

5.3 Conclusion

After these considerations we can now turn back to our research question. Our findings challenge the assumption that probes have an inherent bias towards tracking objective truth, especially when simulated knowledge is present. For supervised probes, our experiments indicate that we can **not** rely on inductive biases for them to track the model's own knowledge rather than simulated knowledge when tracking either concept minimizes the training loss. Contrary to expectations, supervised probes showed an even stronger tendency than unsupervised probes to track simulated knowledge when available. For unsupervised probes, we did **not** find evidence for them having a strong inductive bias preferring the model's own knowledge, either. This stands in contrast to results from a similar experiment by Mallen and Belrose [2023]. The case of training on questions where simulated and objective knowledge agree might be particularly concerning, as this scenario is likely common in real-world applications where simulated knowledge is often correct for simple training questions.

5.4 Limitations and Future Work

As shown by the baseline results, the models we use are too weak for us to high confidence in their abilities to model c_T and c_S on the QIC dataset. Therefore, we can not be sure that the probes indeed track representations of these types of knowledge rather than other features that correlate with the labels.

Even though we have designed the QIC dataset with intent to prevent correlation between labels and superficial features easily modelled by LLMs, we can not be certain of having achieved that. For example, the value of c_S is identical to whether the statement contains the same comparison operator in symbolic and textual form, i.e. "<" and "less than", or ">" and "greater than". To validate this concern, future work could investigate whether probes such as ours generalize across different domains.

We encourage future research investigating the reasons for the strong tendency of supervised probes to track simulated knowledge when available. In addition to this, some results for unsupervised probes were unexpected and, in absence of a good explanation, can point to our experimental setup not being very robust. Future experiments with more seeds and different templates may provide more certainty in this matter.

Chapter 6

Training Data Diversity

6.1 Introduction

In this chapter, we investigate how the diversity of training data affects the performance of probes designed to detect truthfulness in language models. We build upon the concepts of point-wise, local, and global correlations introduced in Section 3.2, which describe how a feature correlating with truthfulness on a dataset D relates to truthfulness. Due to their importance to this chapter, we briefly repeat these concepts:

- **Point-wise correlating features** ($\mathcal{F}_{D,point-wise}$): These are features that coincidentally correlate with truthfulness through one or more samples of a dataset, but not systematically across its distribution. For example, in the *facts* dataset introduced in subsection 2.3.1, the word "interneurons" only appears in one example, which happens to be true.

- **Locally correlating features** ($\mathcal{F}_{D,local}$): These features systematically correlate with truthfulness within a specific domain but usually do not generalize beyond it. For instance, on a dataset comparing integers with a template like "{Integer1} is larger than {Integer2}", a feature corresponding to the concept of "Integers appear in descending order" would be equivalent to truthfulness. However, this correlation does not extend to other domains or even to similar statements using "smaller than" instead.

- **Globally correlating features** ($\mathcal{F}_{D,global}$): These features correlate with truthfulness not only on D, but across various domains and datasets. An example might be features representing truth-related concepts like whether a statement is "verifiable" or "an empirical fact about the world".

Our investigation aims to understand how training on diverse datasets affects a probe's ability to distinguish between these types of features and, consequently, its performance in detecting truthfulness across different domains.

The motivation for this investigation stems from the observation made by Marks and Tegmark [2023] regarding their Misalignment from correlational inconsistency (MCI) hypothesis. This hypothesis suggests that models represent both truth and non-truth features that correlate with truth. As a consequence, "it is essential that datasets used for identifying truth directions be diverse enough to not have such spurious correlations"[Marks and Tegmark, 2023].

In this context, diversity refers to the use of training samples from multiple, different distributions. To address this concern, we conduct a series of experiments where we train probes on datasets with varying levels of diversity and evaluate their performance on both seen and unseen distributions.

Throughout this chapter, we aim to answer the following research questions:

1. How important is the inclusion of the evaluation distribution in the training data?
2. To what extent do probes track features that correlate point-wise, locally, and globally with truthfulness?
3. To what degree can increasingly diverse training data prevent probes from tracking features in $\mathcal{F}_{D,local}$?

6.2 Experimental Setup

6.2.1 Datasets

For our experiments, we utilize datasets from Marks and Tegmark [2023] and Azaria and Mitchell [2023]. We also employ our transformation of the latter into contrast tuples, as described in subsection 2.3.1, to enable their use in training unsupervised probes.

6.2.2 Key Variables and Concepts

We define the following key variables and concepts for our experiments:

- n: Describes the dataset size, which also depends on the aggregation strategy. Note that samples that are negations of each other each count as one sample towards n.
- d: The number of datasets that contribute to the training data, which we refer to as *diversity*. Note that two datasets containing each other's negations and can be combined into a tuple-dataset (e.g., "cities" and "neg_cities") count as one dataset for the purpose of diversity.
- Aggregation strategy:
  - Fixed Total (FT): In this setting, we fix the total number of samples used for training. For example, when training with fixed total with $n = 500$ and $d = 5$, each dataset contributes 100 samples.
  - Fixed Contribution (FC): In this setting, we fix the number of samples each dataset contributes. For example, when training with fixed contribution with $n = 500$ and $d = 5$, each dataset contributes 500 examples.
- Transfer type:
  - "Unseen": No samples from the evaluation distribution were used during training.
  - "Seen": Samples from the training split of the evaluation dataset were used for training. Note that an increase in diversity dilutes the fraction of samples from the evaluation distribution.

  In both cases, we never evaluate on a sample used for training.

6.2.3 Probing Methods and Models

We employ two supervised probing methods: Logistic regression (LR, see section 2.2.1) and mass-mean (MM, see section 2.2.1). Additionally, we use one unsupervised probe: Contrast-Consistent Search (CCS, see section 2.2.2). These probes are trained on activations from layer 13 of the Llama-2-13B model, consistent with the approach used by Marks and Tegmark [2023].

6.2.4 Experimental Procedure

Our experimental procedure consists of the following steps:

We first establish baselines by fitting and evaluating LR probes separately on each entire dataset to determine the best possible linear probe performance. In addition to one baseline for each dataset that we use for training, we repeat this procedure once for a dataset uniformly sampled from all such distributions.

For the main experiments, we repeat the procedure for 5 different seeds, "Fixed total" and "Fixed contribution" aggregation strategies, all three probing methods, $n = 500$, and all possible values of diversity $d = 1, ..., 8$ for which we have enough samples to ensure that the evaluation set always contains at least 20% of the entire dataset.

For each combination of parameters, we:

1. Create $\binom{7}{d}$ training distributions, each referred to as D_{train} (7 being the number of different training datasets)

2. Randomly select samples from each dataset used for training

3. Train a probe on the activations from the training distribution

4. For each dataset we evaluate on subset D_{eval}, which contains all unseen examples from it

This comprehensive experimental procedure resulted in the training of more than 10,000 probes and the computation of over 100,000 accuracies across various datasets and experimental conditions, ensuring a robust assessment of the effects of training data diversity on probe performance.

6.3 Theoretical Framework

This section presents a theoretical framework for understanding how the diversity of training data affects the performance of probes in detecting truthfulness. We begin with an overall summary of our goals, followed by key assumptions and a detailed analysis of feature characteristics.

6.3.1 Overall Summary

The primary objective of this analysis is to understand how the diversity of training data influences a probe's ability to track features that generalize to unseen distributions. We reason that increasing diversity in the training data leads to probes that are more likely to track globally correlating features, which in turn improves their ability to predict truthfulness across a wider range of distributions.

6.3.2 Key Assumptions

Our theoretical model is based on several simplifying assumptions:

All features are orthogonal and have the same magnitude. Each dataset has the same number of features that locally and globally correlate with truthfulness. Samples have a fixed number of features that correlate with truthfulness only point-wise, i.e., not in any systematic way, regardless of their underlying distribution. Locally and globally correlating features that are salient on a distribution are salient on every sample from that distribution. The degree to which probes track features solely depends on their correlation with truthfulness on the training dataset (see section 2.2.1).

These assumptions allow us to focus on the core dynamics of how diversity affects feature tracking, while acknowledging that real-world scenarios may be more complex.

6.3.3 Theoretical Framework

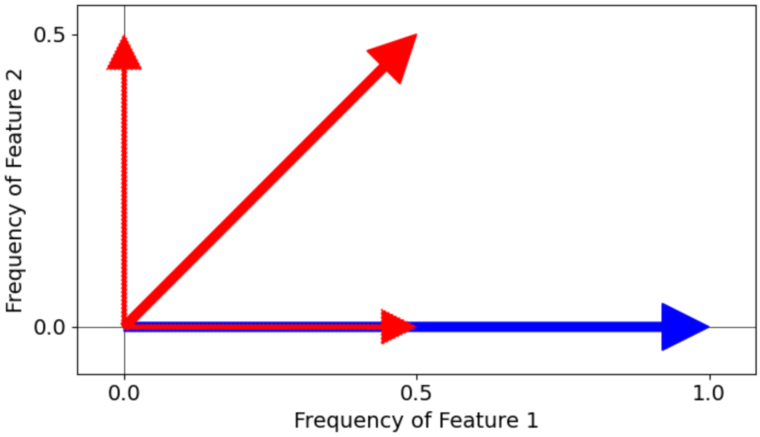

Figure 6.1: Illustration of feature directions in different training scenarios, demonstrating how multiple weak signals have less influence on a probe's direction than a single strong signal (see Theorem 6.3.1). The axes represent orthogonal features, with arrows indicating features correlating with truthfulness in the training dataset. Arrow lengths are proportional to feature frequency. The blue arrow represents a single feature, correlating with truthfulness across all samples of a dataset. The thin red arrows show two features from a more diverse training dataset made from two distributions, each correlating with truthfulness only on every second sample. The bigger red arrow shows their combined direction. This combined direction has a lower magnitude than the blue one.

Theorem 6.3.1. Let d be a real-valued vector. Let d_n be the vector obtained from adding n vectors to d, which are of length l and orthogonal to d and each other to d. For $k \in \mathbb{N}$, let d_{kn} be the vector obtained from adding kn vectors orthogonal to d of length $\frac{l}{k}$ to d. Then for cosine similarities it holds that $\text{sim}_{\cos}(d, d_n) < \text{sim}_{\cos}(d, d_{kn})$, where $\text{sim}_{\cos}$.

This theorem, illustrated in Figure 6.1, is crucial for understanding how the influence of different types of features changes with increasing diversity.

Figure 6.2 provides a comprehensive view of how different feature characteristics change with increasing diversity in the Fixed Total (FT) setting. Based on this theoretical framework, we can draw several key insights:

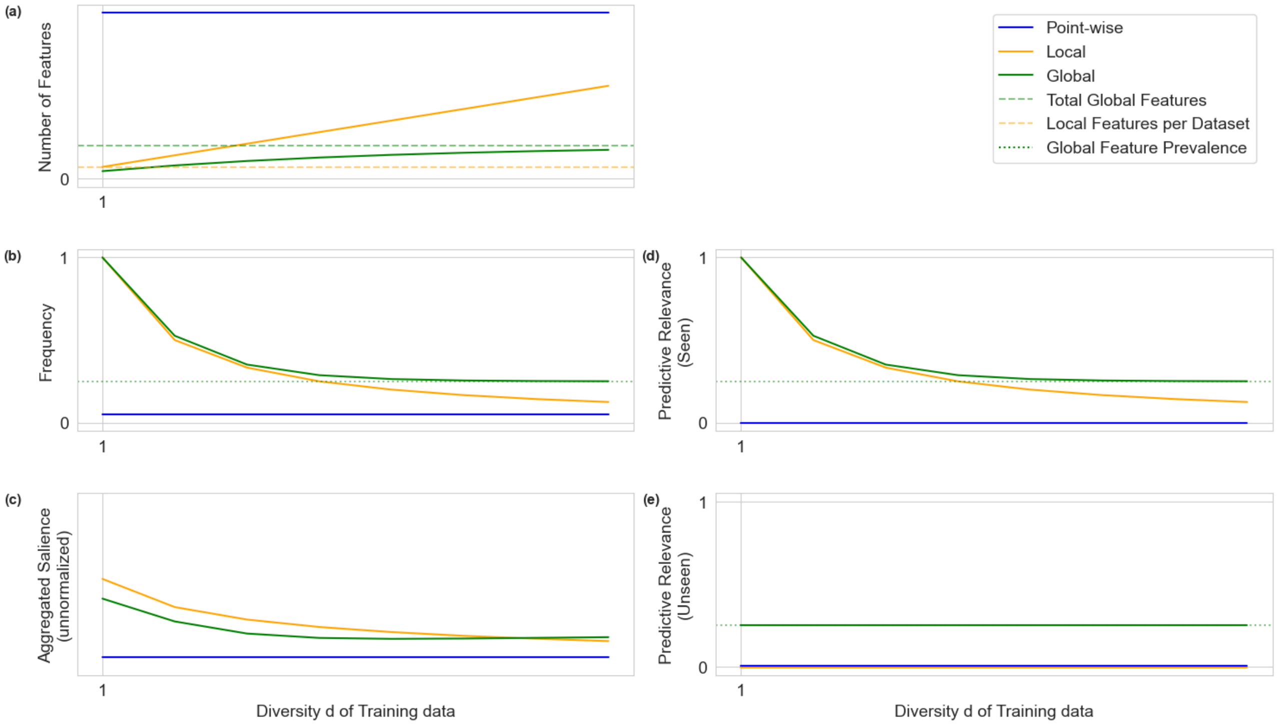

Figure 6.2: Illustration of the influence of diversity on characteristics of the training data D_{train} and accuracy of probes evaluated on D_{eval}, assuming a Fixed Total (FT) number of training samples. $\mathcal{F}_{D_{train},point-wise}$, $\mathcal{F}_{D_{train},local}$, and $\mathcal{F}_{D_{train},global}$ represent features that correlate with truthfulness point-wise, locally, and globally, respectively. Features uncorrelated with truthfulness on D_{train} are omitted as probes have no incentive to track them. (a) shows the number of features correlating with truthfulness in D_{train}, which are tracked by probes to degrees corresponding to their aggregated salience (c), while (d) and (e) indicate the usefulness of these features for predicting truthfulness on seen and unseen distributions, respectively. While exact values and relationships are unknown, we can reason about their qualitative shapes. **(a)** Expected number of different features in D_{train}. Features in $\mathcal{F}_{D_{train},point-wise}$ depend only on sample size n. Features in $\mathcal{F}_{D_{train},local}$ grow linearly with diversity. Features in $\mathcal{F}_{D_{train},global}$ approach a finite total number. **(b)** Average frequency of features co-occuring with truthfulness in D_{train}. Each feature in $\mathcal{F}_{D_{train},point-wise}$ occurs rarely. Features in $\mathcal{F}_{D_{train},local}$ appear in every d-th sample. Features in $\mathcal{F}_{D_{train},global}$ approach the Global Feature Prevalence (the expected fraction of global features correlating with c_T on the average distribution) for high diversity. **(c)** Aggregated Salience is a measure of a feature type's prominence in D_{train}. It is based on their collective correlation with truthfulness, incorporating Number and Frequency from (a) and (b), and geometric considerations (Theorem 6.3.1). When normalized, these provide the expected degree to which probes track features of each type. At some diversity, features in $\mathcal{F}_{D_{train},global}$ are tracked more than those in $\mathcal{F}_{D_{train},local}$, due to their frequency being bounded from below. **(d)** Fraction of features in (a) correlated with truthfulness in the target distribution. For $d = 1$, training and target distributions are identical, so all features in $\mathcal{F}_{D_{train},local}$ and $\mathcal{F}_{D_{train},global}$ are useful for predicting truthfulness. With growing diversity, additional features introduced by other distributions dilute this usefulness. Importantly, the tracking of features uncorrelated with truthfulness on D_{eval}, not only does not help, but it actively lowers their performance. These irrelevant features introduce noise into the predictions, pushing the probes' accuracies towards the level of chance. **(e)** Similar to (d), but for unseen target distributions, where features in $\mathcal{F}_{D_{train},local}$ are not expected to correlate with truthfulness.

As diversity increases, the number of features in $\mathcal{F}_{D_{train},local}$ grows proportionally, but their average saliency decreases proportionally. The influence of features correlating with truthfulness point-wise or locally on a probe's direction is expected to decrease with growing diversity, while the salience of features doing so globally remains relatively constant. It follows, that for high diversity, probes are more likely to track features globally correlating with truthfulness, which are expected to generalize better to unseen distributions. In the "Seen" scenario, accuracy on the target distribution D_{eval} may initially decrease with diversity as the probe is influenced by features that do not correlate with truthfulness in it, but it's expected to increase again as the probe focuses more on globally correlating features.

For the Fixed Contribution (FC) setting, where size of D_{train} is proportional to diversity, only curves for $\mathcal{F}_{D_{train},point-wise}$ change in (a), (b) and (c): Their number scales with d and frequency with $1/d$, leading to decreasing aggregated salience compared to other feature types. Consequently, probes trained with high diversity in this setting are expected to more strongly track features that generalize to the target distribution.

Having developed a theoretical perspective on the influence of diverse training data on probes' capabilities to generalize, we now move on to present the results from our experiments.

6.4 Results

In this section we present the results obtained from our experimental setup, state observations and interpret them.

We separately aggregate and visualize the results for different n, aggregation strategy, transfer type, and type of probe. The aggregation process involves:

- For each seed, n, d, and evaluation dataset, we compute the mean accuracy over all $\binom{7}{d}$ (for "Seen") or $\binom{6}{d}$ (for "Unseen") training distributions. Note that 7 is the total number of datasets we have for training.

- For each n, d and evaluation dataset, we compute and visualize the mean and standard deviations of the above values across seeds.

- We repeat the process, but average over evaluation datasets in the first step. The means and standard deviations resulting after the second step are visualized in black.

In Figure 6.3 we present our results for LR probes in our experiments with $n = 500$. For some figures there is no data for $d = 1$ due to the a limited number of samples in our smallest dataset (got/sp_en_trans with 354 samples). When the Transfer Type is "Unseen", we show evaluation accuracies for additional datasets we never use for training. Their names are displayed in parentheses in the figures in this section. To allow fair comparison between "Seen" and "Unseen" settings, we do not take them into account when calculating average performances. Results from experiments with MM and CCS probes, and additional data for $n = 1000$ can be found in Appendix A.

We will now highlight key observations and offer interpretations of trends visible in Figure 6.3.

Diverse training data improves generalization capabilities to unseen distributions by about 2%, but this effect seems to saturate. In the "Unseen" experiments, where no data from the target distribution is used for training, the average accuracy improves when samples from more than one training distribution are used. However, it seems to saturate

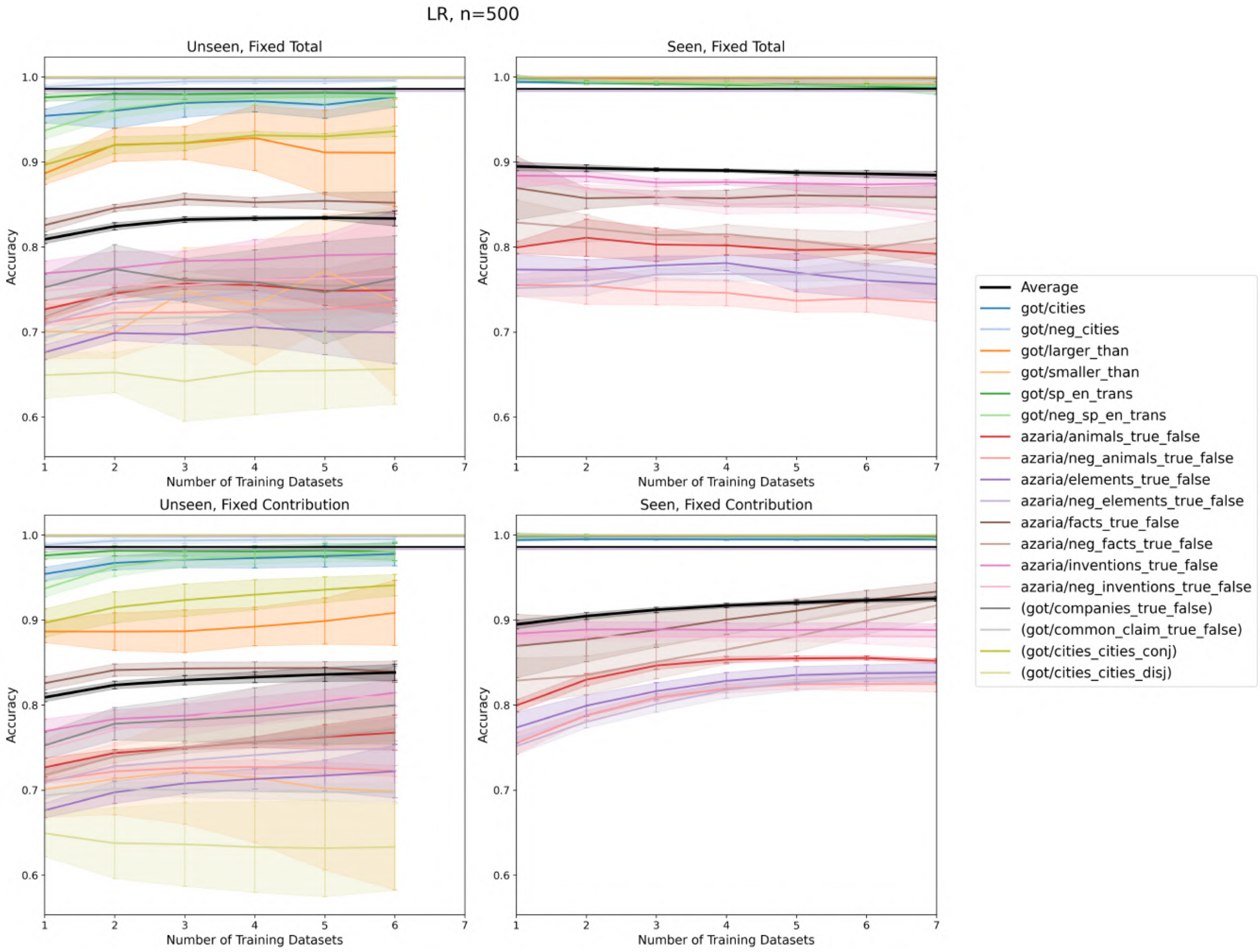

Figure 6.3: Accuracies of probes trained with logistic regression with $n = 500$ evaluated on different datasets. Shaded areas represent (symmetrical) standard deviations. The diversity d, representing the number of distributions from which the training dataset was sampled, is shown on the x-axis. Horizontal lines represent probes trained on the test set, with the black line at 98.5% representing an LR probe trained and evaluated on a dataset uniformly sampled from all distributions we train on. **Unseen, Fixed Total**: A total of 500 samples are used for training, with none coming from the target distribution. Increasing diversity alleviates average accuracy from around 81% to 83%, where it seems to saturate. **Seen, Fixed Total**: A total of 500 training samples are used for training, including some from the target distribution. For $d = 1$ this means that all training samples come from the distribution we evaluate on, for $d = 2$ every second etc. Increasing diversity only weakly affects average accuracy, lowering it from ca. 89% to 88%. **Unseen, Fixed Contribution**: The training dataset contains 500 samples from every distribution, always excluding the target distribution. Increasing diversity improves accuracy from around 81% to slightly above 83%. This is a slight improvement compared to the Fixed Total setting, due to the larger training datasets. **Seen, Fixed Contribution**: The training dataset contains 500 samples from every distribution, always including the target distribution. Higher diversity raises accuracy from around 89% to slightly above slightly above 92%, showing that the addition of unrelated training data helps probes become more accurate. Overall, the trends for the ”Seen” setting are relatively stable for all evaluation datasets, whereas results from the unseen setting differ more starkly between different evaluation datasets and seeds.

below 85% for most evaluation datasets, except some for which an increase can be seen even at the highest diversities we have tested.

Including some training data from the target distribution is the most effective measure, boosting accuracy by 5-10%. This difference even persists at high levels of diversity, where only relatively few samples from the evaluation distribution are used for training.

Adding unrelated samples to the training data improves accuracy on seen target distributions by more than 3%, but replacing them does not. From our results regarding the Seen, Fixed Contribution setting we can see an increase in accuracy on the target distribution when samples from unrelated distributions are added to the training dataset, thereby diluting its relevance. This effect saturates for most target datasets at the highest diversity of $d = 6$. One exception is `azaria/facts_true_false`, with the possible explanation that this dataset contains very diverse questions, and, therefore, profits more from a diverse training dataset. The effect of increased accuracy by the inclusion of unrelated samples is not observed in the Fixed Total setting, where samples from the target distribution are *replaced*, rather than complemented.

Increasing sample size beyond 500 lifts accuracy by almost 5%, but only when new samples from the target distribution are added. When comparing accuracies on seen distributions between the FT and FC setting in Figure 6.3, we see an increase in accuracy of almost 5% for $d = 6$, where the training dataset is 6-times larger for the FC setting. There is no such improvement due to increased sample size alone in the "Unseen" setting. When doubling the number of training samples from $n = 500$ to $n = 1000$ (Figure A.3), we see similar trends with only a small increase in accuracy of roughly 1%.

Transferring *to* some particular datasets is easier than *from* them. For almost all settings, probes achieve the highest accuracy when evaluated on some datasets got/cities, got/neg_cities, got/sp_en_trans, and got/neg_sp_en_trans. This is also the case for probes trained on a single other dataset ("Unseen" with $d = 1$). In Figure A.6 in the Appendix, we show results only for probes that were trained on distributions that include got/sp_en_trans, and got/neg_sp_en_trans. This constraint leads to lower accuracies than before. Together, these observations imply that probes solve these datasets more easily while also learning less from them.

Accuracies for the datasets from Marks and Tegmark [2023] are almost always higher than those for datasets from Azaria and Mitchell [2023]. For most experiments, probes evaluated on the datasets from Marks and Tegmark [2023] (prefixed with "got/") achieve the highest accuracies between 90-100%. As these datasets are designed with intention to be unambiguous and simple, we think it is possible that our results strongly depend on the type of dataset and the model's capability.

The accuracies of MM probes are relatively similar to those of LR probes, whereas accuracies for CCS probes are 5-15% lower than for LR probes, and much more noisy. In particular, the average accuracy of CCS probes stays below 80% for all our experiments (see Figure A.2).

In the following section, we will briefly juxtapose these empirical results with expectations from our theoretical analysis, before turning back to our initial research questions posed at the beginning of this chapter.

6.5 Discussion

6.5.1 Comparison of theoretical and empirical results

Our theoretical analysis suggests that as sample sizes n and diversities d approach infinity, probes should increasingly track features that globally correlate with truthfulness, ultimately leading to accurate predictions across all datasets. While our experimental results confirm an increase in probes' ability to generalize when trained on large and diverse datasets, this effect appears to level off, contrary to our expectations.

More precisely, our theoretical model predicts that training on more samples should reduce the tendency of probes to track features that do not correlate with truthfulness systematically, but merely point-wise. However, our empirical observations show that this reduction in distractions only marginally enhances the probes' ability to track more useful features. Importantly, we can rule out the possibility that such features are simply not tracked by probes. Our results demonstrate that probes trained directly on evaluation datasets achieve near-perfect accuracies, significantly outperforming those trained on different splits of the same dataset (as in the "Seen" scenario with $d = 1$). This implies that features whose correlation with truthfulness is only point-wise, i.e., not even systematic on the training distribution, are salient enough for probes to track and later detect during inference.

These observations suggest that probes can track combinations of numerous features and detect them during inference, even if they were only present in few samples during training.

This discrepancy between theoretical expectations and empirical results suggests that other factors may play crucial roles in limiting probe performance. Strong interference between features on D_{eval} or limitations in the model's knowledge could be contributing to this unexpected plateau in performance. These factors warrant further investigation to bridge the gap between our theoretical understanding and empirical observations.

To further investigate the nature of truthfulness representations across datasets, we conducted a Principal Component Analysis (PCA) on the differences between activations of contrast tuples. The results of this analysis, presented in Appendix B, provide additional evidence for the existence of a general truthfulness-related feature that spans multiple datasets and domains, while also highlighting the complexity and imperfection of these representations.

6.5.2 Answers to our research questions

Our experiments provide insights into our three research questions:

1. **Importance of including the evaluation distribution in training data:** Our results show that including samples from the target distribution in the training data ("Seen") consistently improves probe accuracy by 5-10%, almost slashing the error rate in half when compared to the "Unseen" condition.

2. **Extent of probes tracking different types of correlated features:** The increase in performance when including samples from the target distribution suggests that probes learn to track features that locally correlate with c_T. The relatively high performances obtained even when the target distribution is not used for training suggests that features that globally correlate with c_T are also tracked, especially when the training data is diverse. Finally, the near perfect baseline performances from probes trained on the same data as they are being evaluated on suggests that probes also track features point-wise

correlating with c_T. In summary, combinations of all types of features introduced in section 3.2 are being tracked by probes, with their composition at least in part depending on the size and diversity of the training datasets.

3. **Effectiveness of diverse training data in preventing tracking of local features:** While diverse training data improved generalization to unseen distributions by 2-3%, the benefits appeared to plateau. This implies that increasing diversity and size of training data may not be sufficient to prevent probes from tracking features that correlate with truthfulness merely locally or point-wise, which are not expected to generalize well.

Additionally, we want to highlight two practical implications for the design of probes based on the results from this study.

First, the persistent performance gap between "Seen" and "Unseen" conditions, even at high diversities, shows that completely domain-agnostic truthfulness detection remains challenging. Our findings underscore the value of incorporating domain-specific data in probe training, as we have demonstrated that even relatively small amounts of such data can lead to significant improvements in accuracy.

Secondly, the observed plateau in performance gains as size and diversity of the training data increases, especially when no data from the target distribution is available for training, points to inherent limitations in current probing methodologies. This suggests a need for innovative approaches that transcend the conventional focus on optimizing probe directions and biases. One such approach, which we term 'Contrastive Inference', is presented in chapter 7.

6.6 Limitations and Future Work

While our study provides valuable insights, it also has limitations that could be addressed in future research:

- Our experiments were limited to a single model and layer. Future work could explore how well these findings generalize to different architectures.

- We used a limited number of datasets, and our results were different depending on their source. Future work could investigate more datasets, to more robustly investigate the trends, and look more closely at which of their properties are relevant for the success of probes trained and evaluated on them.

- Further investigation into the nature of the features being tracked by probes, perhaps through techniques like feature visualization or ablation studies, could provide deeper insights into the mechanisms underlying probe performance.

Chapter 7

Contrastive Inference

7.1 Introduction

In the previous chapter, we explored how diversifying training data can improve the performance of linear probes in detecting truthfulness. However, we observed that even ideal linear classifiers do not achieve perfect accuracy across all datasets. This limitation motivates us to explore methods that go beyond standard linear classification during inference.

In this chapter, we introduce Contrastive Inference (CI), a novel approach to using probes that aims to improve their accuracy by mitigating the effects of interfering features. While similar ideas have been explored in different contexts - Burns et al. [2023b] used a comparable approach for inference with unsupervised probes (CRC), and Mallen and Belrose [2023] employed a related concept for computing AUROCs - to our knowledge, this is the first systematic application and analysis of such a method for inference with supervised probes.

To understand the intuition behind CI, consider a simple example with a statement s and and its negation s_{neg}. Now suppose an imperfect lie detector predicts s and s_{neg} to be true with 51% and 99% probability, respectively. In isolation, one would judge s to be true, but taking both values into account probably changes the assessment.

Formally, let p be a probe, with $p(x)$ being the mapping from an activation vector to the probability that x corresponds to a true statement. Let s and s_{neg} be a statement and its negation, and a and a_{neg} be the corresponding activations. Standard inference, as defined in Definition 2.2.1, computes a probes' logits as:

$$p(\text{"}s \text{ is true according to } p\text{"}) = p(a). \tag{7.1}$$

Contrastive Inference modifies this to:

$$p_{CI}(\text{"}s \text{ is true according to } p\text{"}) = p(a) - p(a_{neg}). \tag{7.2}$$

This approach aims to cancel out the effects of features that are equally salient in both a statement and its negation, thereby improving the probe's ability to detect truthfulness.

7.2 Theoretical Considerations

The primary motivation for Contrastive Inference comes from the observation that many features are likely to be present in both a statement and its negation, for example, a feature representing "the sentence is in English". Even when a probe has not been explicitly trained to track such features, some interference is expected due to feature directions being only *almost* orthogonal in practice [Elhage et al., 2022]. The linearity of the probe's logits ensures that CI can remove the effects of this interference.

CI is particularly effective in addressing the issue of the "polarized truth direction", representing $f_P = XOR(\text{"statement is true"}, \text{"statement is negated"})$, that has been identified and shown to be present in multiple models by Bürger et al. [2024]. It would be harmful for probes to track this feature, because a distributional shift could cause their predictions to flip, making the probes actively predict the opposite of truthfulness. As this feature is either active or inactive for both $p(a)$ and $p(a')$ in Equation 7.2, CI effectively removes its influence.

7.3 Experimental Setup

To evaluate the effectiveness of Contrastive Inference, we use the same experimental framework as in chapter 6. This setup allows us to investigate CI under various conditions, including different levels of training data diversity and both seen and unseen evaluation distributions.

The key difference in our experiments is that we now jointly evaluate pairs of statements (a statement and its negation) using the CI method. For a detailed description of the datasets, probing methods, and experimental procedure, we refer the reader to section 6.2.

Our extensive experiments involved training over 5,000 probes and computing approximately 30,000 accuracies, ensuring a robust assessment of CI's performance across various conditions.

7.4 Results and Discussion

Figure 7.1 and Figure 7.2 present the results of our experiments with Contrastive Inference for logistic regression and mass-mean probes, respectively. We observe several key findings:

Improved performance on unseen distributions: CI consistently reduces the error rate on unseen distributions by approximately one-third compared to standard inference. This improvement persists regardless of the diversity of the training data.

Smaller gains on seen distributions: When the training data includes samples from the target distribution, the improvements from CI are less pronounced. For logistic regression probes, CI improves accuracies by 2-3%, reducing the error rate by about one-fifth.

Consistent improvement for logistic regression: Logistic regression probes benefit from CI across all diversity levels and for both seen and unseen distributions, albeit with varying degrees of improvement.

Diversity-dependent performance for mass-mean probes: For mass-mean probes at low diversities, CI actually decreases performance compared to standard inference. However, as diversity increases, CI's effect becomes increasingly positive, eventually leading to substantial improvements in accuracy at high diversities.

These results suggest that Contrastive Inference is particularly effective under distributional shifts, when no data from the target distribution is available for training. This aligns with our theoretical understanding: CI's ability to mitigate the effects of interfering features is most beneficial when the probe encounters new distributions where these features may not correlate with truthfulness in the same way as in the training data.

The diversity-dependent change in performance of mass-mean probes with CI on unseen target distributions is an intriguing finding. Here, such probes trained with low diversity data sometimes give wrong predictions with CI where standard inference does not. This implies that

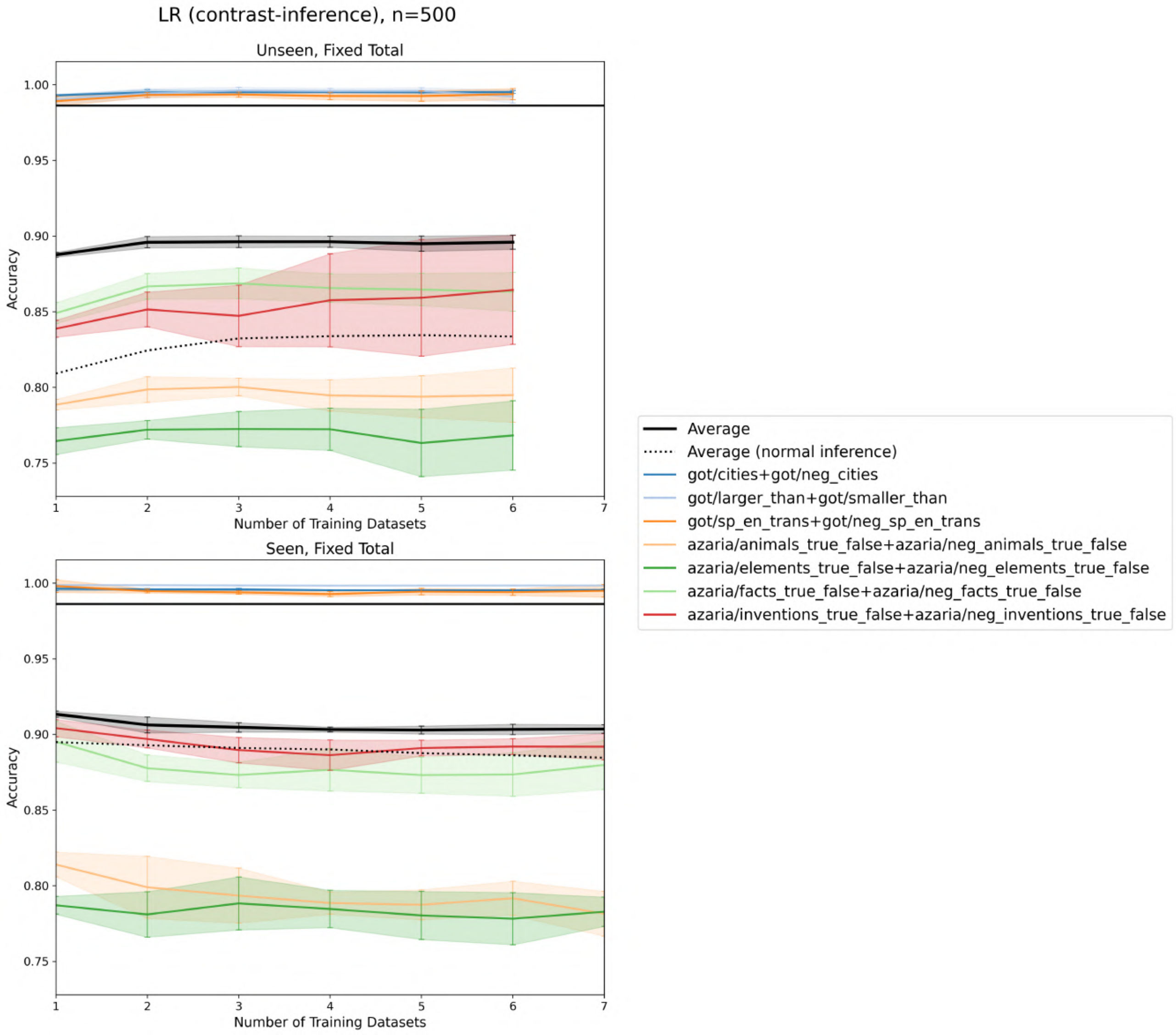

Figure 7.1: Accuracies of probes trained with logistic regression on a (fixed) total of $n = 500$ samples, and evaluated with Contrastive Inference (CI) on different datasets. The experimental setup and training procedure are identical to section 6.2. We jointly evaluate each statement and its negation with CI, leading to shared accuracies for datasets that contain each others negations. A dotted black line represents the average accuracy of probes presented in chapter 6 without the use of CI. Shaded areas represent (symmetrical) standard deviations. The diversity d, representing the number of distributions from which the training dataset was sampled, is shown on the x-axis. Horizontal lines represent probes trained on the test set. **Unseen, Fixed Total**: None of the 500 training samples come from the target distribution. Predictions made with CI reach around 89% accuracy. Thereby, they consistently outperform regular predictions by 6-7%, independently of the diversity of the training data. **Seen, Fixed Total**: Some of the 500 training samples come from the target distribution. Predictions made with CI reach around 91% accuracy. Again, we see similar trends between regular predictions and those made with CI, but this time the latter outperforms the former by only 2-3%, reducing the error rate by about one fifth.

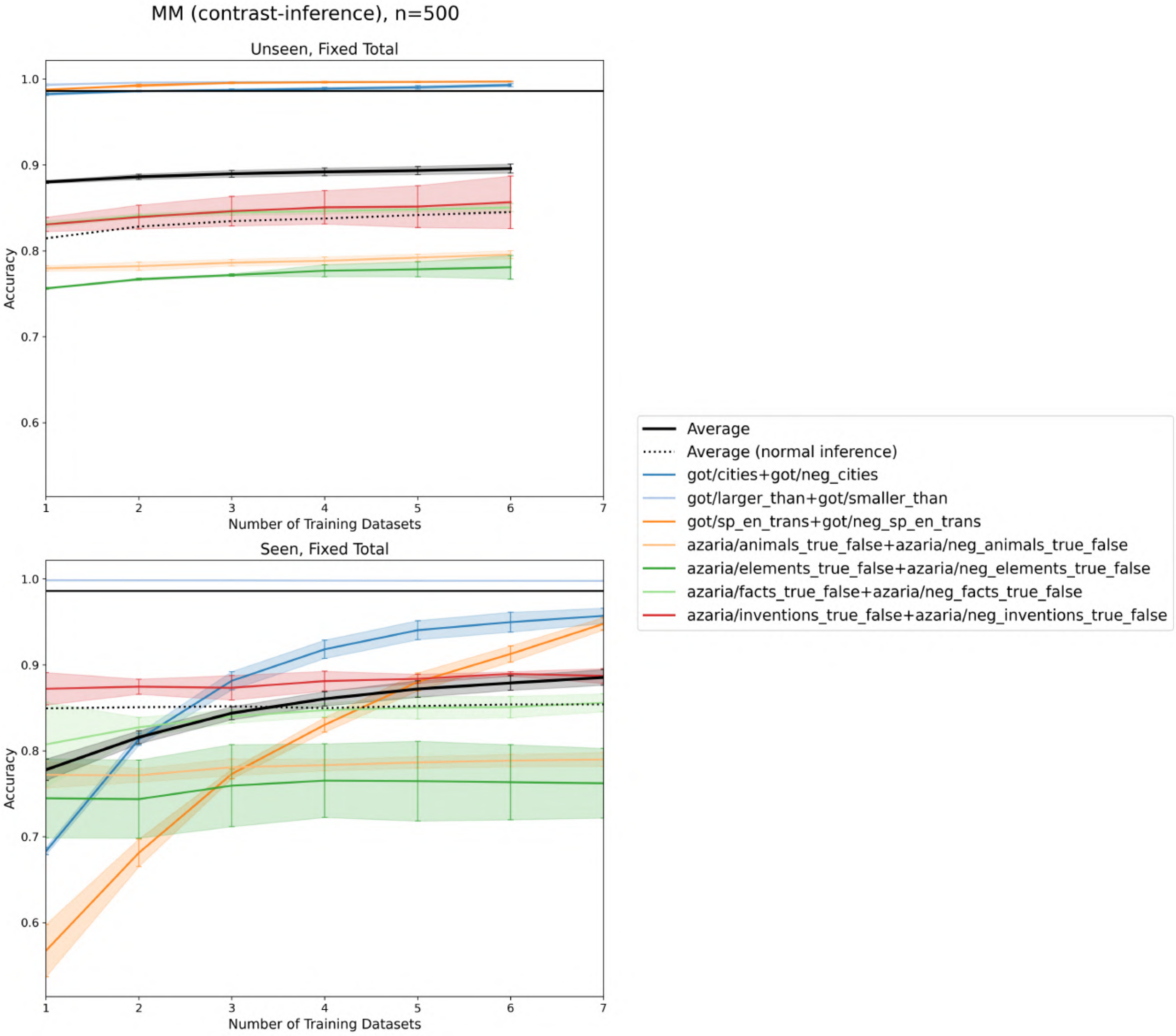

Figure 7.2: Accuracies of mass-mean probes trained on a (fixed) total of $n = 500$ samples, and evaluated with Contrastive Inference (CI) on different datasets. Except for the probing method, the structure of the figure is identical to Figure 7.1 **Unseen, Fixed Total**: CI outperforms regular inference when none of the 500 training samples come from the target distribution, reaching almost 90% accuracy and, thereby, reducing the error rate by about a third. Again, the curves look qualitatively similar for different values of diversity. **Seen, Fixed Total**: Some of the 500 training samples come from the target distribution. Here, there is a qualitative difference between the curves of CI and regular inference: The average accuracy of CI changes strongly with diversity of the training data, reaching only around 78% for $d = 1$, but almost 89% for $d = 7$, whereas regular inference obtains around 85% regardless of diversity.

logits for both items of a contrast pair are in the same half of the interval $[0, 1]$ but in the wrong order. It may indicate that these probes, when trained on less diverse data, learn to rely more heavily on features that CI removes. As training diversity increases, the probes likely become more robust, tracking features that generalize better across distributions and thus benefiting more from CI.

7.5 Limitations and Future Work

While Contrastive Inference shows promise in improving probe performance, several limitations and areas for future work should be considered:

Need for negations: The primary practical limitation of CI is its requirement for negated statements. This increases both the complexity of data preparation (as negations may need to be generated, potentially using language models) and the computational cost (as additional activation vectors must be extracted).

Model and layer specificity: As with our experiments in the previous chapter, these results are based on a single model and layer. Future work should investigate how CI performs across different architectures and at various depths within models.

Dataset limitations: Our experiments use a limited number of datasets, many of which are relatively simple. Expanding to a broader range of datasets, including more complex and nuanced examples, would provide a more comprehensive evaluation of CI's effectiveness.

Theoretical foundations: While we have provided intuitive arguments for why CI works, a more rigorous theoretical analysis could offer deeper insights into its properties and limitations.

Integration with other methods: Exploring how CI can be combined with other probing techniques, such as the TTPD method proposed by Bürger et al. [2024], could lead to even more robust truthfulness detection.

7.6 Conclusion

Contrastive Inference represents a promising approach to improving the performance of truthfulness detection probes. By leveraging the relationship between a statement and its negation, CI can mitigate the effects of interfering features and improve probe accuracy, particularly on unseen distributions.

Our results demonstrate that CI reduces error rates in most settings, consistently reducing the frequency of errors by on third when probes are applied to distributions unseen during training. This suggests that CI could be particularly valuable in real-world applications where models encounter diverse and potentially unexpected inputs, provided that it is feasible to obtain the required activations of negated statements.

The effectiveness of CI underscores the importance of considering not just how probes are trained, but also how they are applied during inference. As we continue to develop methods for eliciting and interpreting the knowledge encoded in large language models, techniques like Contrastive Inference may play a crucial role in making methods to detect truthfulness reliable.

Chapter 8

Conclusion

This study has advanced our understanding of truthfulness detection in large language models through conceptual analysis and novel experiments.

We began by analyzing possible failure modes and introducing categories for features that correlate with truthfulness in qualitatively different ways: point-wise, locally, and globally correlating features. This framework allows a better understanding of the challenges faced by probing methods.

Next, we mathematically analyzed the conditions under which probes may learn to track XOR features, which presents serious risks under distributional shifts.

Our investigation into simulated knowledge revealed an unexpected vulnerability in supervised probes. Contrary to previous assumptions, these probes showed a strong tendency to track simulated knowledge when available. This finding highlights a danger in the application of supervised probes to scenarios where simulated and objective knowledge may diverge.

Experiments with diverse training data demonstrated its effectiveness in improving probe generalization, especially for unseen distributions. However, the benefits of increased diversity plateaued, suggesting inherent limitations in current probing methodologies.

To address these limitations, we introduced Contrastive Inference (CI), a novel approach that significantly enhances probe performance. Like diverse training data, CI proved most effective on unseen distributions, reducing error rates by approximately one-third even when combined with diverse training.

This study has several limitations. We focused on a limited number of models (Pythia and Llama-2) and analyzed activations from only one layer for each. Our experiments primarily used relatively simple datasets, and our theoretical models made simplifying assumptions that may not hold in real-world scenarios.

Future research should develop more sophisticated theoretical models and explore the potential of Sparse Autoencoders for truthfulness detection. Further investigation of simulated knowledge across various domains and application of Contrastive Inference to more complex scenarios are also promising avenues for future work.

The broader impact of this research lies in its potential to improve AI safety through improved truthfulness detection. However, it also highlights the risks of over-reliance on imperfect probing methods and the implications for trust in AI systems across various applications. We hope that the continued development and refinement of truthfulness detection techniques will support the reliability of AI systems as they become increasingly sophisticated and widespread.

Acknowledgments

I want to thank Leonard Bereska and Walter Laurito for their continuous supervision, and Benjamin Shaffrey for numerous conversations that fuelled my interest in these topics.

Bibliography

G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes, Nov. 2018. URL http://arxiv.org/abs/1610.01644. arXiv:1610.01644 [cs, stat].

A. Azaria and T. Mitchell. The Internal State of an LLM Knows When It's Lying. *EMNLP*, Oct. 2023. doi: 10.48550/arXiv.2304.13734. URL http://arxiv.org/abs/2304.13734. arXiv:2304.13734 [cs].

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA, Mar. 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445922. URL https://dl.acm.org/doi/10.1145/3442188.3445922.

S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, and O. van der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling, May 2023. URL http://arxiv.org/abs/2304.01373. arXiv:2304.01373 [cs] version: 2.

S. R. Bowman, J. Hyun, E. Perez, E. Chen, C. Pettit, S. Heiner, K. Lukošiūtė, A. Askell, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Olah, D. Amodei, D. Amodei, D. Drain, D. Li, E. Tran-Johnson, J. Kernion, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, L. Lovitt, N. Elhage, N. Schiefer, N. Joseph, N. Mercado, N. DasSarma, R. Larson, S. McCandlish, S. Kundu, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Telleen-Lawton, T. Brown, T. Henighan, T. Hume, Y. Bai, Z. Hatfield-Dodds, B. Mann, and J. Kaplan. Measuring Progress on Scalable Oversight for Large Language Models, Nov. 2022. URL http://arxiv.org/abs/2211.03540. arXiv:2211.03540 [cs].

T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*, 2023.

C. Burns, P. Izmailov, J. H. Kirchner, B. Baker, L. Gao, L. Aschenbrenner, Y. Chen, A. Ecoffet, M. Joglekar, J. Leike, I. Sutskever, and J. Wu. Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision, Dec. 2023a. URL http://arxiv.org/abs/2312.09390. arXiv:2312.09390 [cs].

C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. *ICLR*, 2023b. URL http://arxiv.org/abs/2212.03827. arXiv:2212.03827 [cs].

L. Bürger, F. A. Hamprecht, and B. Nadler. Truth is Universal: Robust Detection of Lies in LLMs, July 2024. URL http://arxiv.org/abs/2407.12831. arXiv:2407.12831 [cs] version: 1.

P. Christiano, A. Cotra, and M. Xu. Eliciting Latent Knowledge, Jan. 2021. URL https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8/edit?usp=sharing&usp=embed_facebook.

H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models, Oct. 2023. URL http://arxiv.org/abs/2309.08600. arXiv:2309.08600 [cs].

N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah. Toy Models of Superposition. *Transformer Circuits Thread*, 2022.

S. Farquhar, V. Varma, Z. Kenton, J. Gasteiger, V. Mikulik, and R. Shah. Challenges with unsupervised LLM knowledge discovery. *CoRR*, 2023. doi: 10.48550/ARXIV.2312.10029. URL https://arxiv.org/abs/2312.10029. Publisher: arXiv Version Number: 2.

A. Ghandeharioun, A. Caciularu, A. Pearce, L. Dixon, and M. Geva. Patchscopes: A Unifying Framework for Inspecting Hidden Representations of Language Models. Jan. 2024. doi: 10.48550/arXiv.2401.06102. URL http://arxiv.org/abs/2401.06102. arXiv:2401.06102 [cs].

P. Hase, M. Bansal, P. Clark, and S. Wiegreffe. The Unreasonable Effectiveness of Easy Training Data for Hard Tasks, June 2024. URL http://arxiv.org/abs/2401.06751. arXiv:2401.06751 [cs].

K. Hänni, W. Laurito, A. Mennen, K. Kozaronek, G. Kaklam, and J. Ku. Searching for a model's concepts by their shape – a theoretical framework — LessWrong, Feb. 2023. URL https://www.lesswrong.com/posts/Go5ELsHAyw7QrArQ6/searching-for-a-model-s-concepts-by-their-shape-a.

W. Laurito, S. Maiya, G. Dhimoïla, O. Yeung, and K. Hänni. Cluster-norm for Unsupervised Probing of Knowledge, July 2024. URL http://arxiv.org/abs/2407.18712. arXiv:2407.18712 [cs].

B. A. Levinstein and D. A. Herrmann. Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks, June 2023. URL http://arxiv.org/abs/2307.00175. arXiv:2307.00175 [cs].

J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models, Oct. 2023. URL http://arxiv.org/abs/2305.11747. arXiv:2305.11747 [cs].

S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *ACL*, May 2022. doi: 10.48550/arXiv.2109.07958. URL http://arxiv.org/abs/2109.07958. arXiv:2109.07958 [cs].

A. Mallen and N. Belrose. Eliciting Latent Knowledge from Quirky Language Models. *CoRR*, Dec. 2023. doi: 10.48550/arXiv.2312.01037. URL http://arxiv.org/abs/2312.01037. arXiv:2312.01037 [cs].

S. Marks. What's up with LLMs representing XORs of arbitrary features?, Jan. 2024. URL https://www.lesswrong.com/posts/hjJXCn9GsskysDceS/what-s-up-with-llms-representing-xors-of-arbitrary-features.

S. Marks and M. Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. *CoRR*, 2023. URL https://arxiv.org/abs/2310.06824.

P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks. AI Deception: A Survey of Examples, Risks, and Potential Solutions. *CoRR*, Aug. 2023. doi: 10.48550/arXiv.2308.14752. URL http://arxiv.org/abs/2308.14752. arXiv:2308.14752 [cs].

A. Radhakrishnan, Buck, ryan_greenblatt, and F. Roger. Scalable Oversight and Weak-to-Strong Generalization: Compatible approaches to the same problem. Dec. 2023. URL https://www.alignmentforum.org/posts/hw2tGSsvLLyjFoLFS/scalable-oversight-and-weak-to-strong-generalization.

J. Scheurer, M. Balesni, and M. Hobbhahn. Large Language Models can Strategically Deceive their Users when Put Under Pressure, July 2024. URL http://arxiv.org/abs/2311.07590. arXiv:2311.07590 [cs].

M. Shanahan, K. McDonell, and L. Reynolds. Role-Play with Large Language Models, May 2023. URL http://arxiv.org/abs/2305.16367. arXiv:2305.16367 [cs].

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, S. Goel, N. Li, M. J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, and D. Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency. *CoRR*, Oct. 2023. doi: 10.48550/arXiv.2310.01405. URL http://arxiv.org/abs/2310.01405. arXiv:2310.01405 [cs].

Appendix A

Diversifying Training Data - Additional Results

Here, we present additional results from the experiments discussed in chapter 6. They follow the same structure as Figure 6.3.

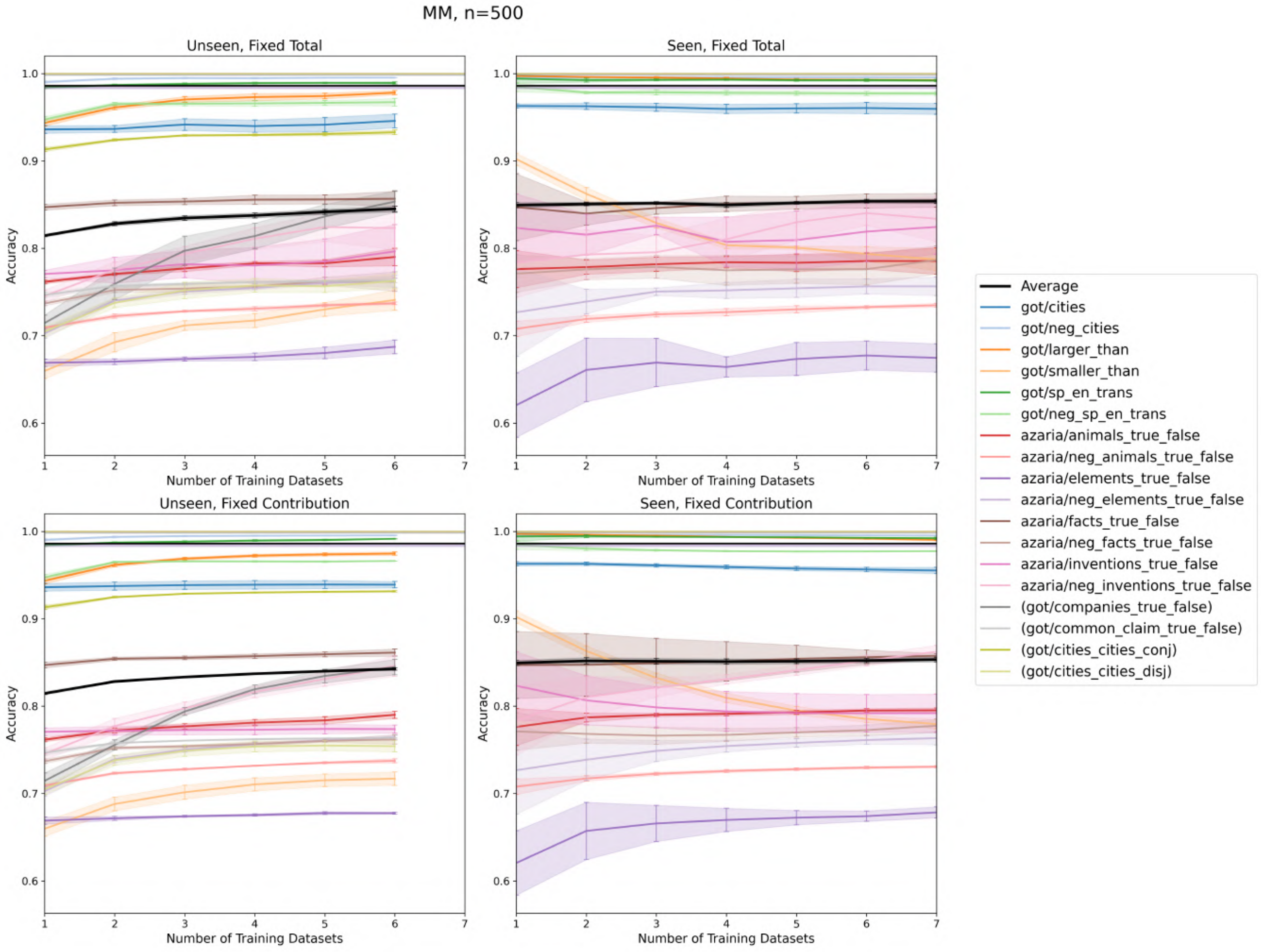

Figure A.1: Accuracies for Mass-Mean probes evaluated on different datasets with $n = 500$, visualized the same way as Figure 6.3.

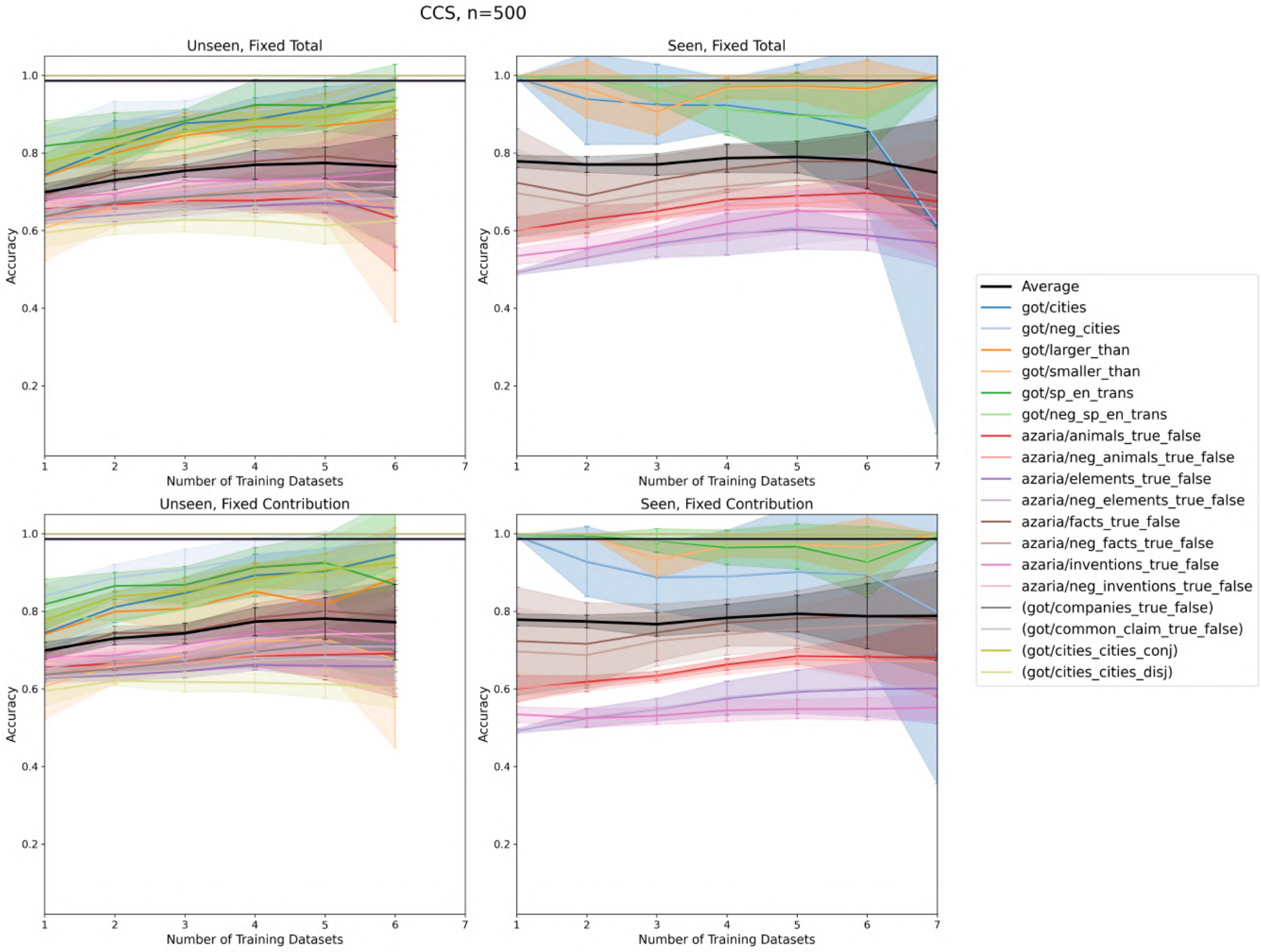

Figure A.2: Accuracies for Contrast-Consistent Search probes evaluated on different datasets with $n = 500$, visualized the same way as Figure 6.3.

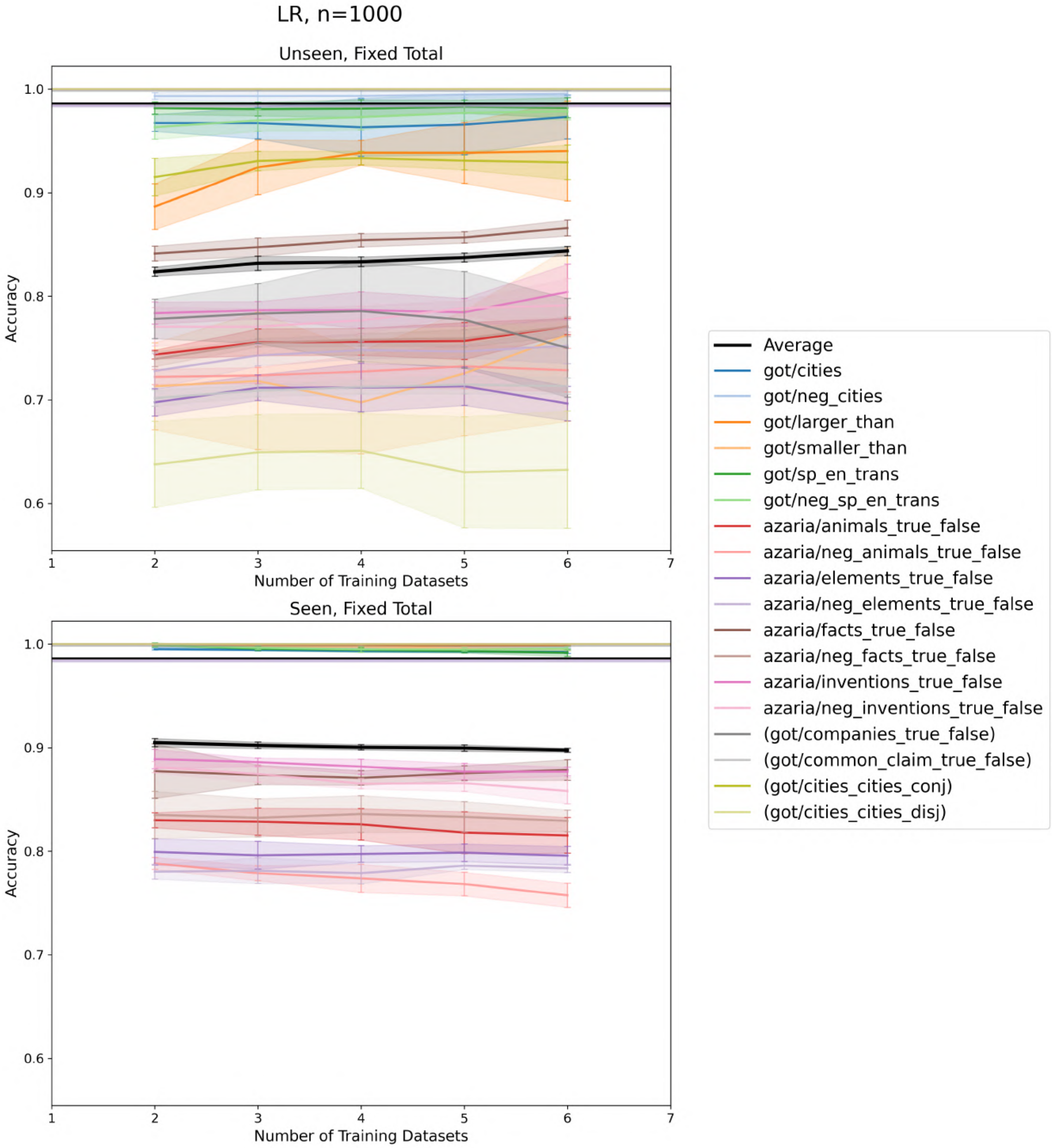

Figure A.3: Accuracies for LR probes evaluated on different datasets with $n = 1000$, visualized the same way as Figure 6.3.

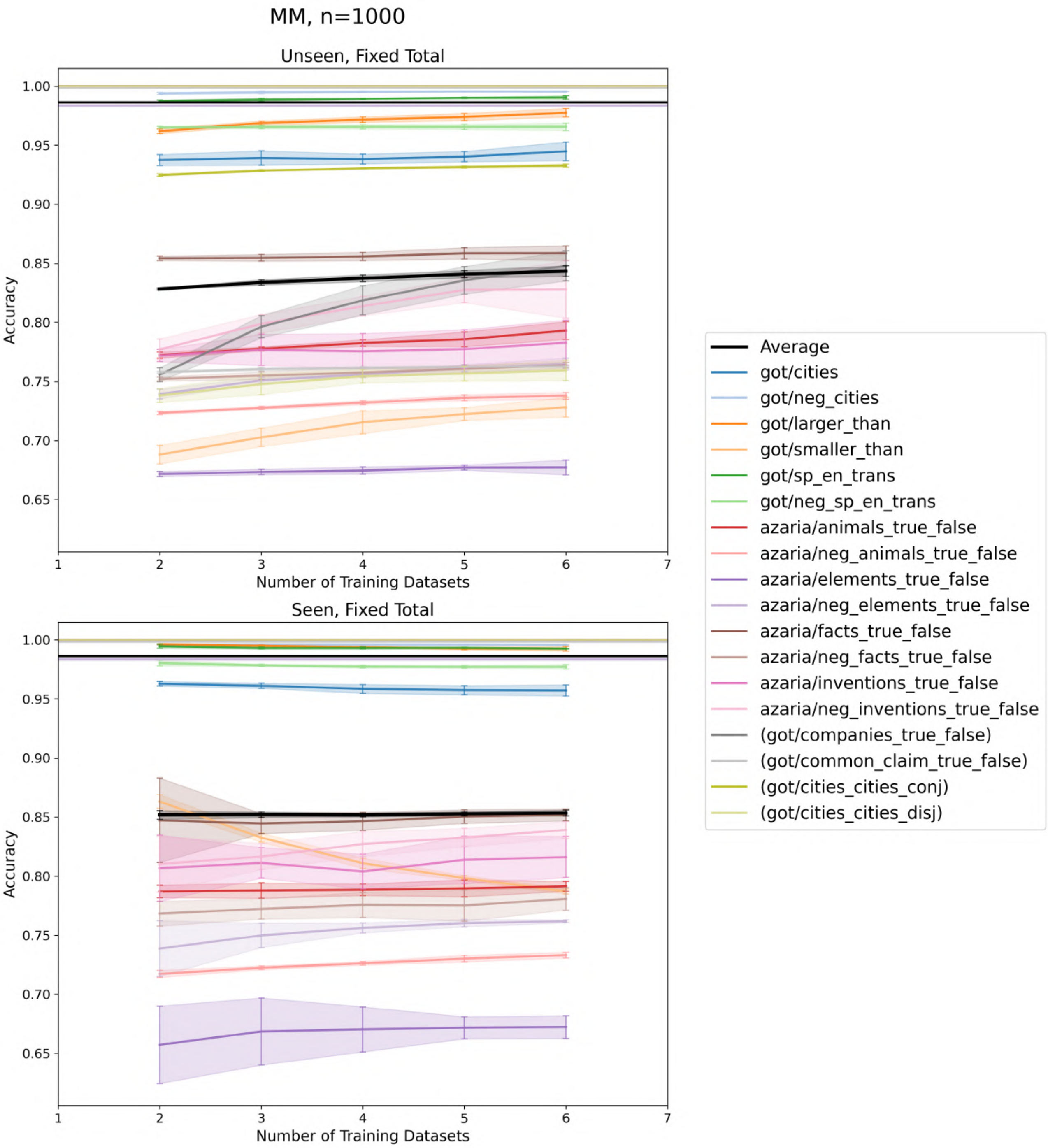

Figure A.4: Accuracies for Mass-Mean probes evaluated on different datasets with $n = 1000$, visualized the same way as Figure 6.3.

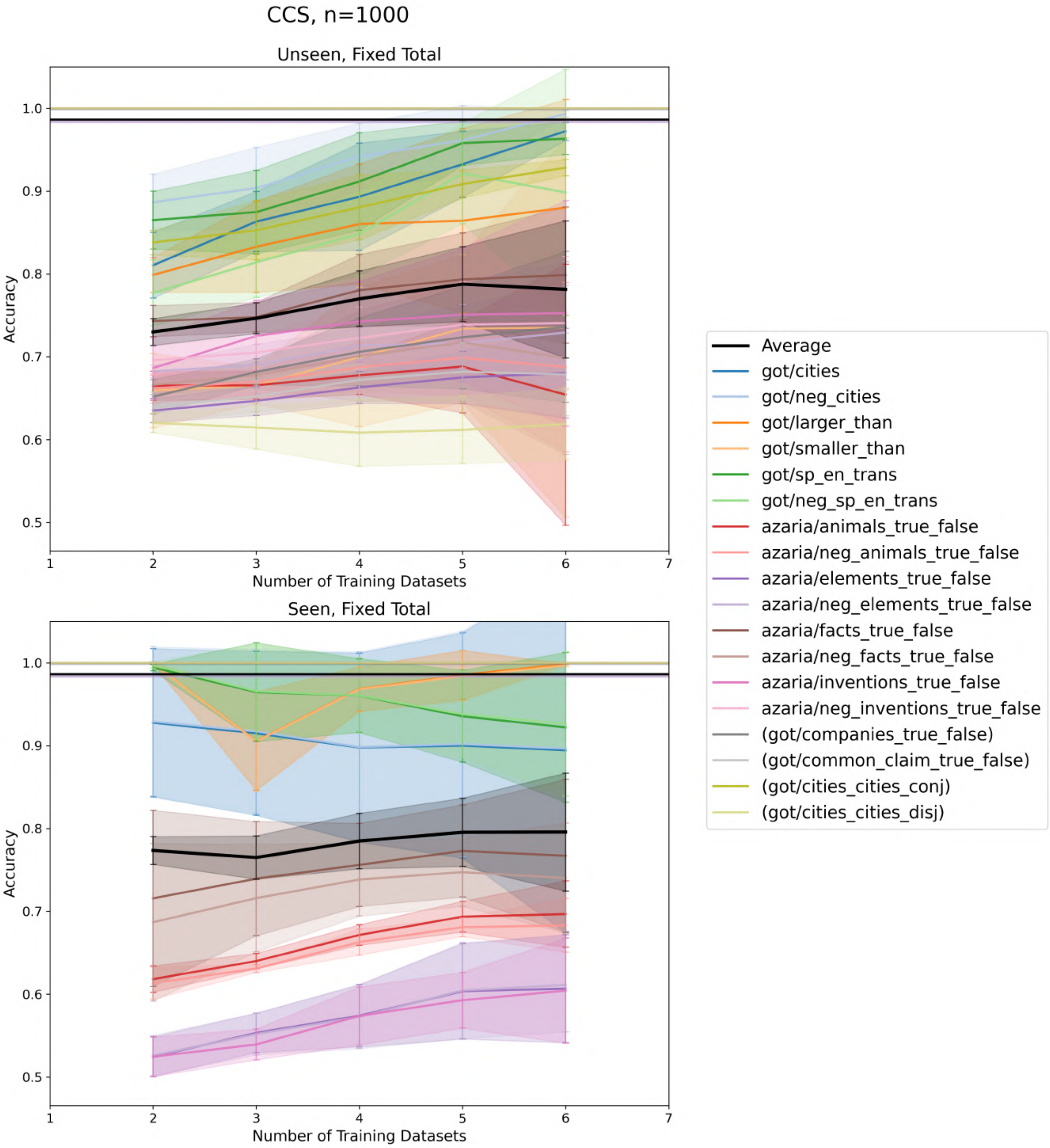

Figure A.5: Accuracies for Contrast-Consistent Search probes evaluated on different datasets with $n = 1000$, visualized the same way as Figure 6.3.

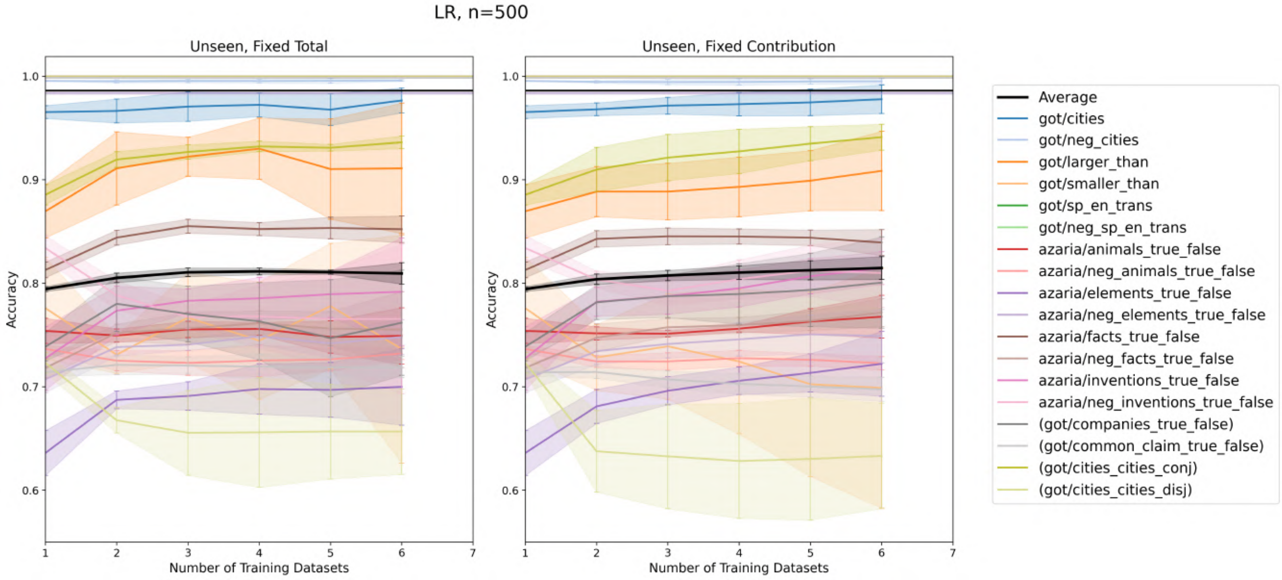

Figure A.6: Accuracies for Logistic Regression probes evaluated on different datasets with $n = 500$, visualized the same way as Figure 6.3. In this particular version, only results from probes trained on datasets that include samples from `sp_en_trans` are shown.

Appendix B

Diversifying Training Data - PCA

To gain insights into the structure of truthfulness representations across different datasets, we performed Principal Component Analysis (PCA) on the differences between activations of contrast tuples. This approach is similar to that used in Contrastive Representation Clustering (CRC) [Burns et al., 2023b].

We fit PCA on the differences between activations of contrast tuples *uniformly sampled from pairs of multiple datasets*, where one dataset contains the original statements and the other contains their negations. The datasets used for fitting the PCA transformation include:

- got/cities, got/neg_cities
- got/sp_en_trans, got/neg_sp_en_trans
- azaria/animals_true_false, azaria/neg_animals_true_false
- azaria/elements_true_false, azaria/neg_elements_true_false
- azaria/facts_true_false, azaria/neg_facts_true_false
- azaria/inventions_true_false, azaria/neg_inventions_true_false

Before applying PCA, we separately centered the negated and non-negated statements to remove the most obvious traces of negation. This centering step helps to focus the analysis on the underlying truthfulness representations rather than surface-level linguistic features.

The following figures show the projection of raw activations extracted from layer 13 of Llama-2-13B when presented with various datasets onto the first six principal components. In each plot, true statements are represented in orange, while false statements are in blue. Lighter colors indicate the original statements, and darker colors represent their negations where applicable.

These results suggest the presence of a general truthfulness-related feature that spans multiple datasets and domains. The trend along PC0 in Figure B.1, where true statements (orange) tend to appear more to the right compared to false statements (blue), indicates that the model has likely learned a representation of truthfulness that generalizes across different types of statements. However, it's important to note that this is not a clear separation, but rather a tendency. The overlap between true and false statements along PC0 suggests that this representation is influenced by other factors, which could include interference from other features, limitations in the model's knowledge, or that PC0 represents a combination of multiple truthfulness-related features that are salient to different degrees.

The absence of clear truthfulness-related patterns in higher-order principal components (Figures B.2 and B.3) further supports the idea of a dominant, general truthfulness direction in the model's representation space. This finding aligns with observations from Marks and

Tegmark [2023] and challenges the hypothesis that truthfulness is represented only through local, domain-specific features. However, the lack of clear separation along PC0 also indicates that truthfulness representation is more nuanced than a single, perfectly discriminative feature.

While these results support the existence of a general truthfulness representation, they do not rule out the possibility of multiple truth-related directions that are similarly salient across datasets. Further research could investigate whether the observed PC0 is a combination of multiple truthfulness-related features, such as those corresponding to different levels or types of knowledge.

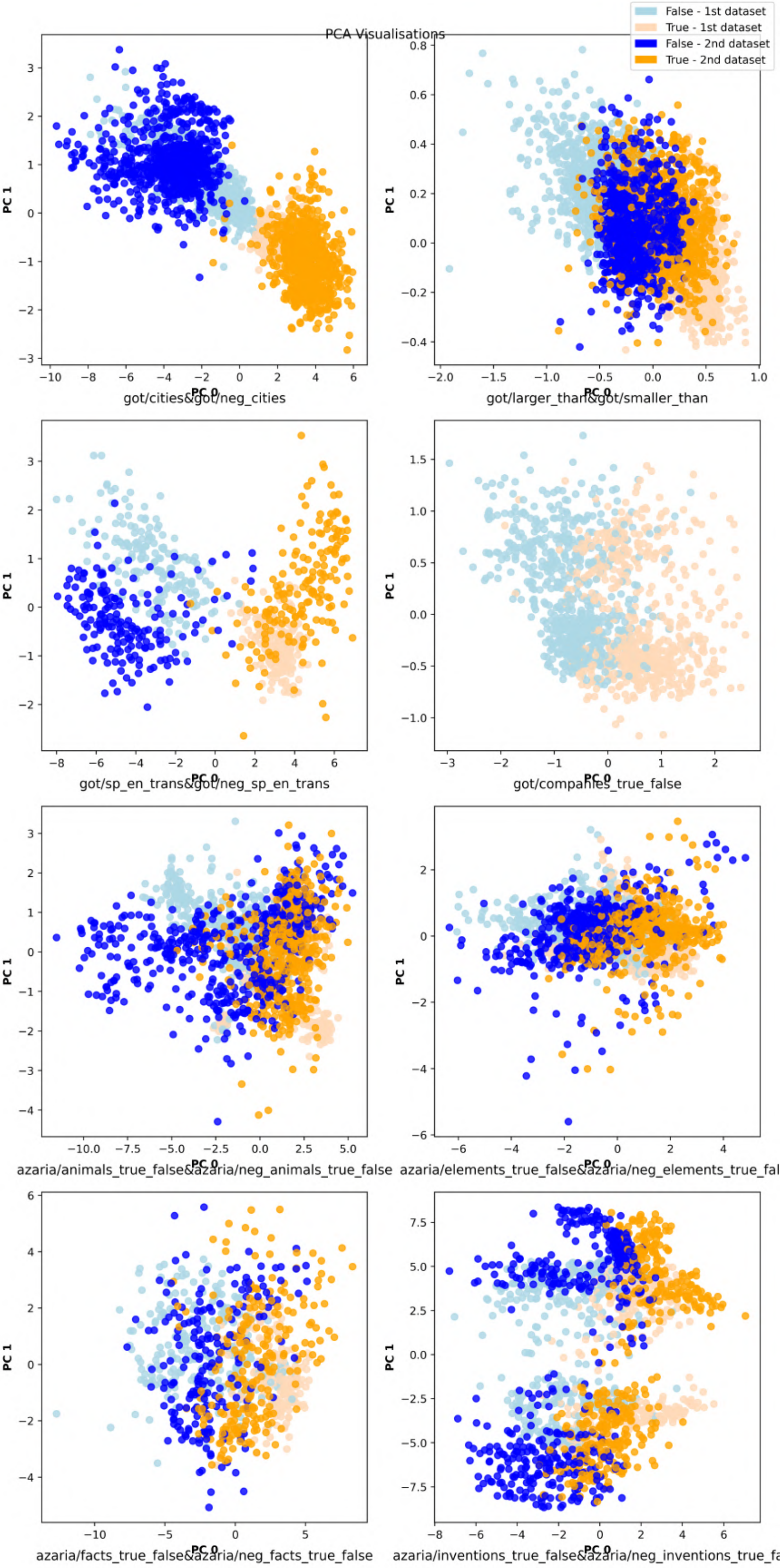

Figure B.1: Projection of raw activations onto the 0th and 1st principal components. PC0 (x-axis) shows a clear trend related to truthfulness, with true statements (orange) generally appearing to the right of false statements (blue). This trend is more pronounced for some datasets than others, suggesting a general but imperfect truthfulness representation across different domains.

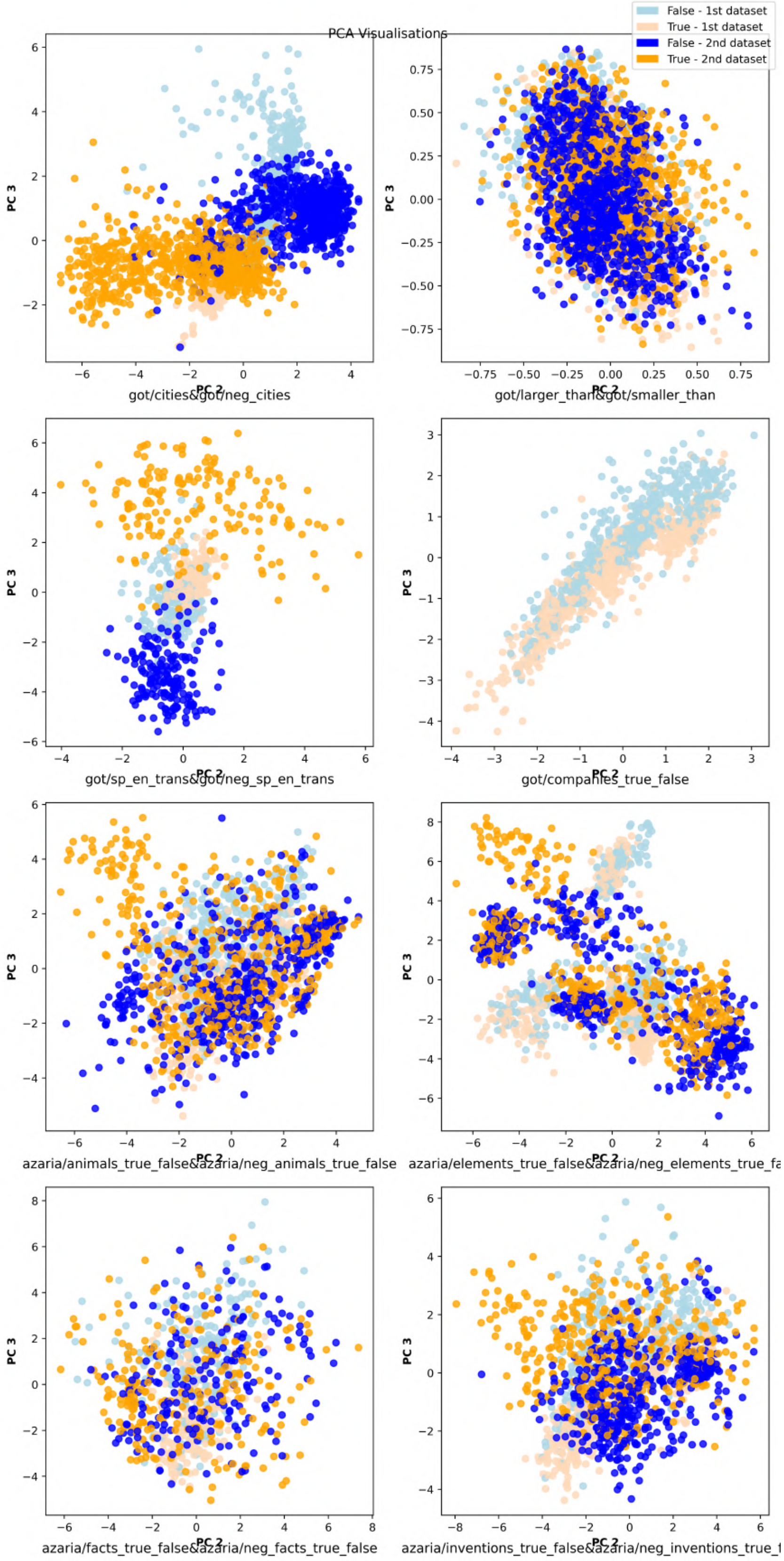

Figure B.2: Projection of raw activations onto the 2nd and 3rd principal components. Unlike PC0, these components do not exhibit clear interpretable patterns related to truthfulness, suggesting that the primary truthfulness information is captured in the first principal component.

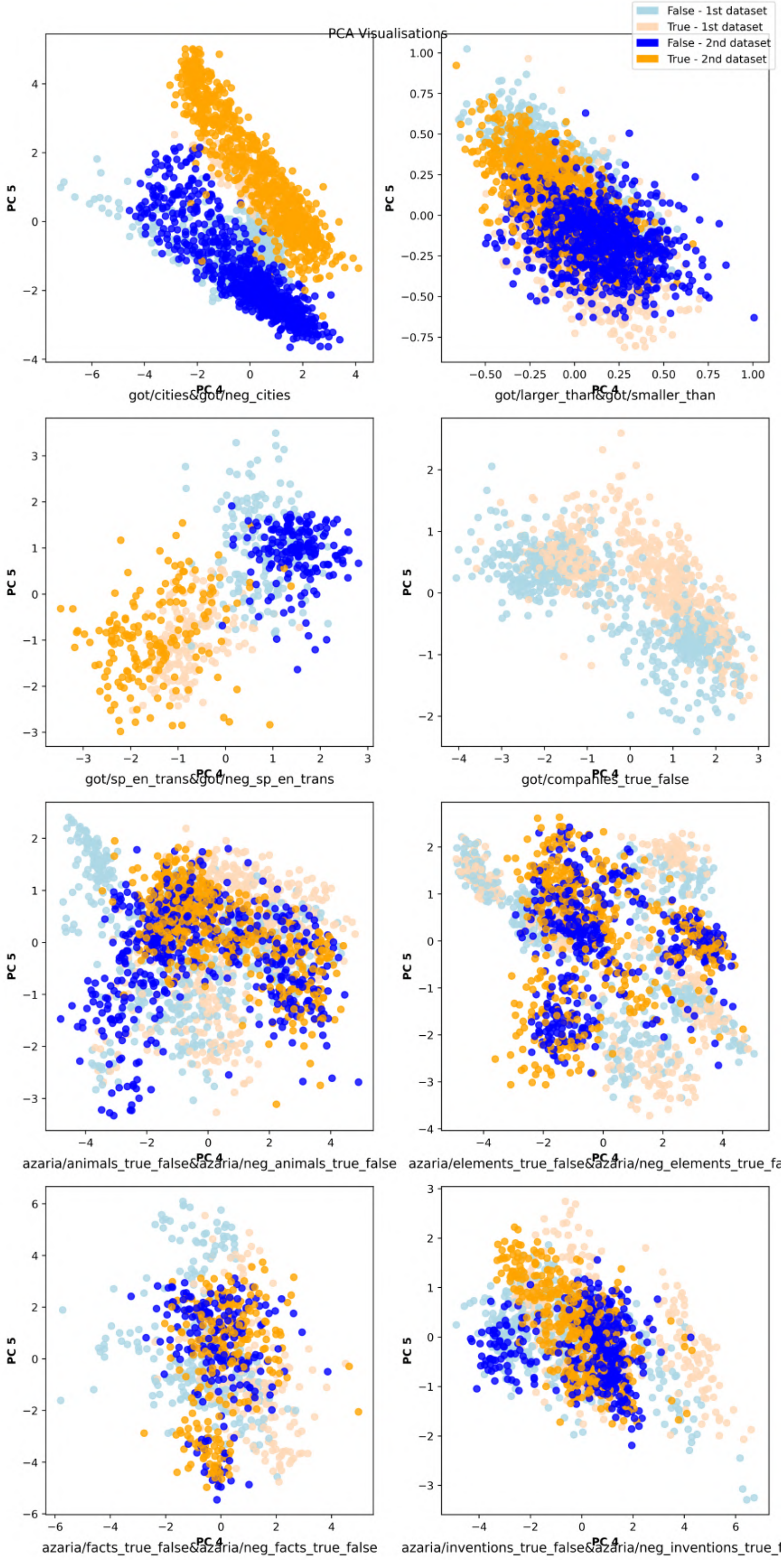

Figure B.3: Projection of raw activations onto the 4th and 5th principal components. As with PC2 and PC3, these higher-order components do not show clear patterns related to truthfulness, further supporting the hypothesis that truthfulness information is primarily encoded in the first principal component.

Appendix C

Simulated Knowledge - Additional Results

This appendix presents the detailed results of our simulated knowledge experiments from chapter 5 across all six Pythia model sizes. Each figure shows the performance of baseline (LM output), supervised probes (LR, Mean-Diff, LDA), and unsupervised probes (CCS, CRC) for a specific model size. The interpretation of these figures follows the same principles as discussed in chapter 5 for the Pythia-12B model.

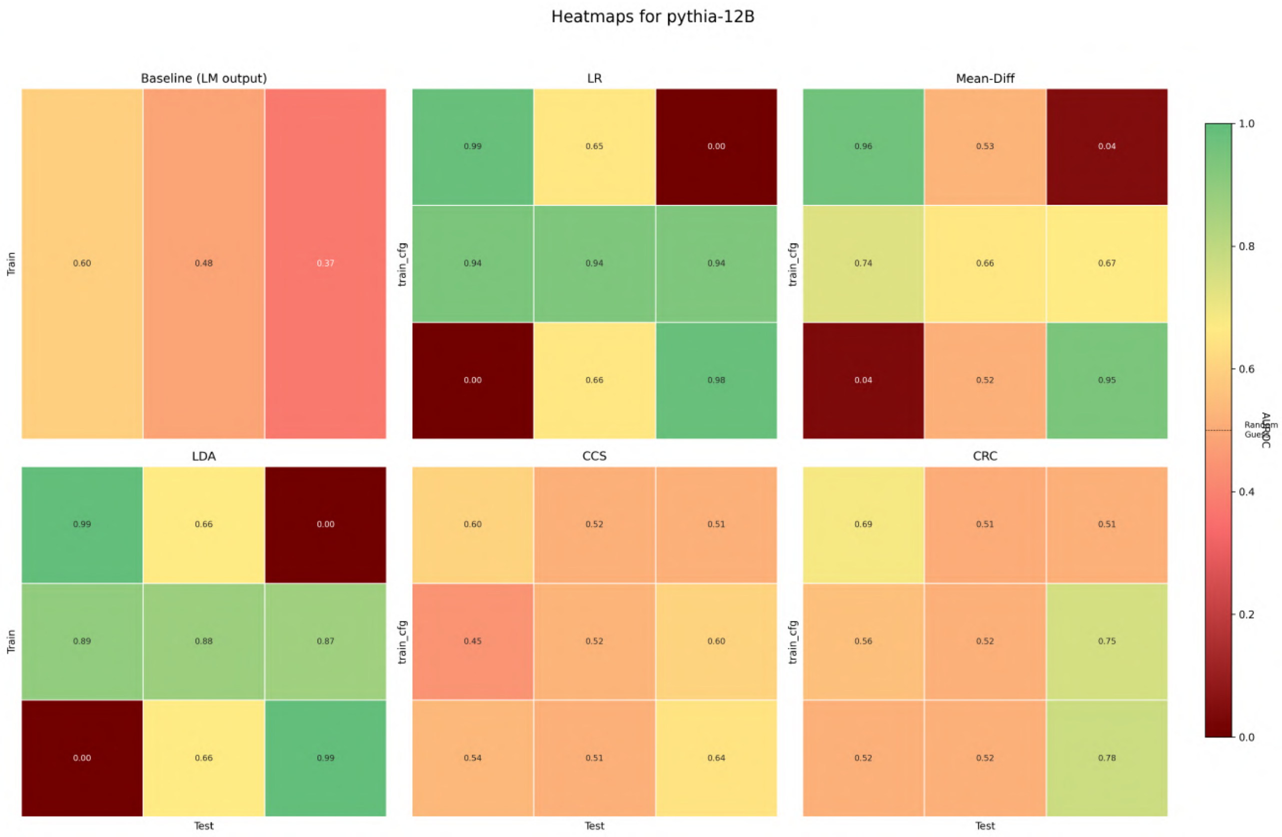

Figure C.1: Heatmaps for Pythia-12B model

These figures provide a comprehensive view of how different probing methods perform on the QIC dataset across model sizes. Key observations include:

- The tendency of probes to track simulated knowledge (c_S) rather than objective truth (c_T) becomes more pronounced as model size increases.

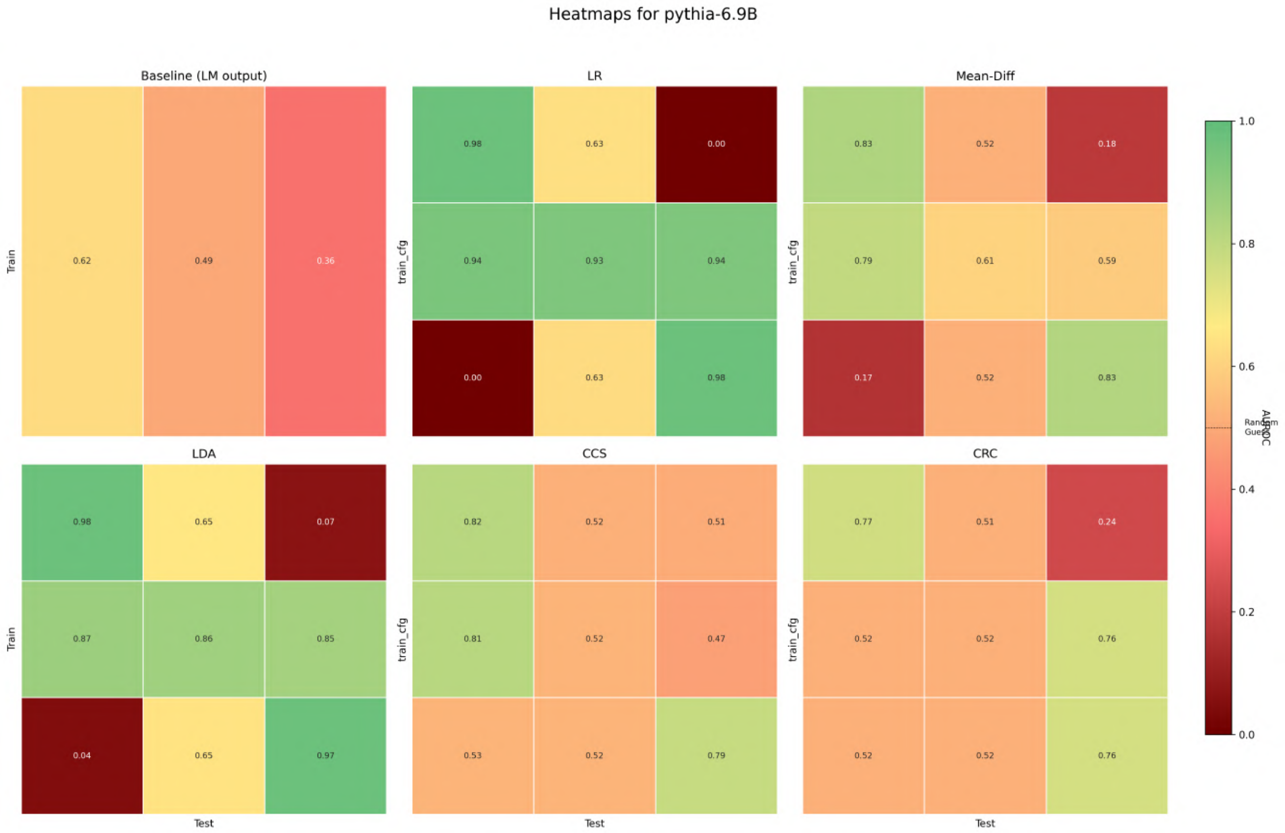

Figure C.2: Heatmaps for Pythia-6.9B model

- Smaller models (particularly Pythia-410M and Pythia-1B) show performance closer to random guessing across most conditions.
- The baseline (LM output) shows a gradual increase in its preference for simulated knowledge as model size increases.

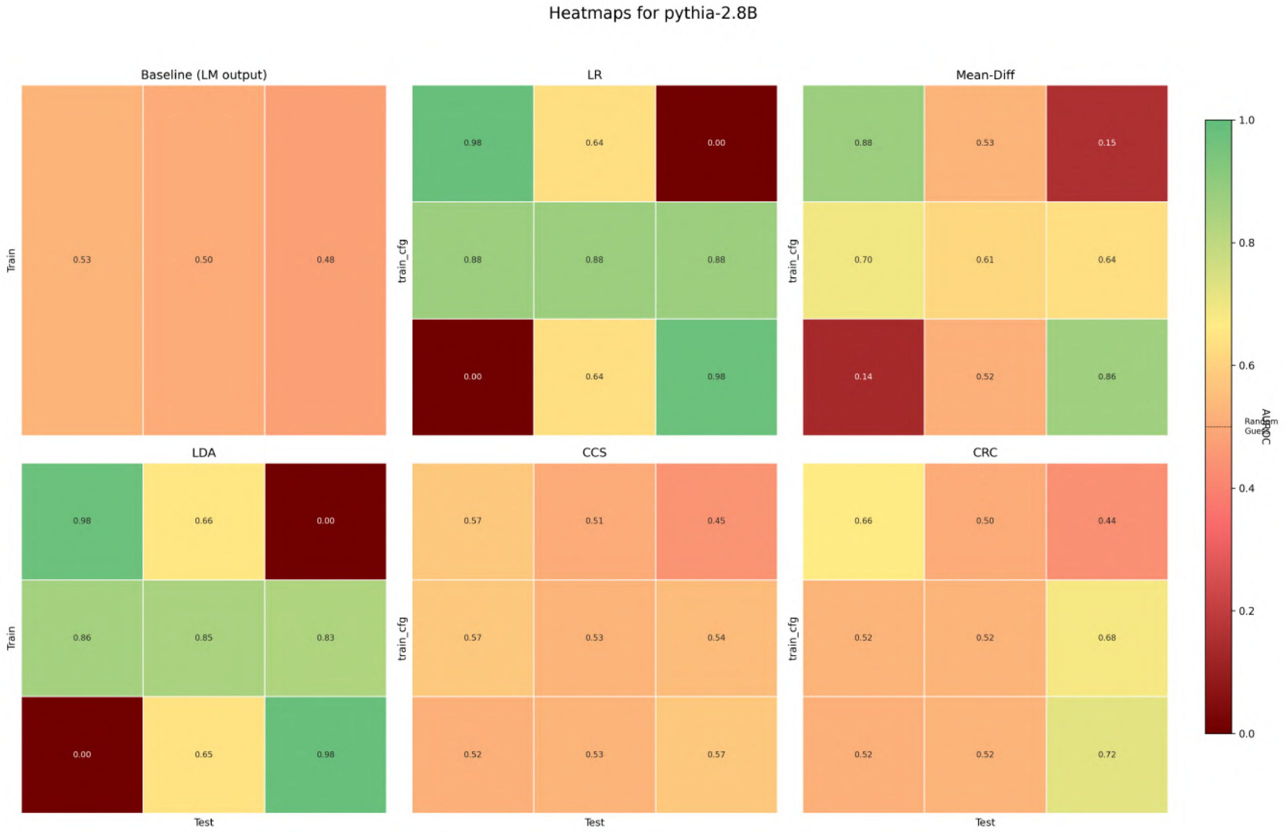

Figure C.3: Heatmaps for Pythia-2.8B model

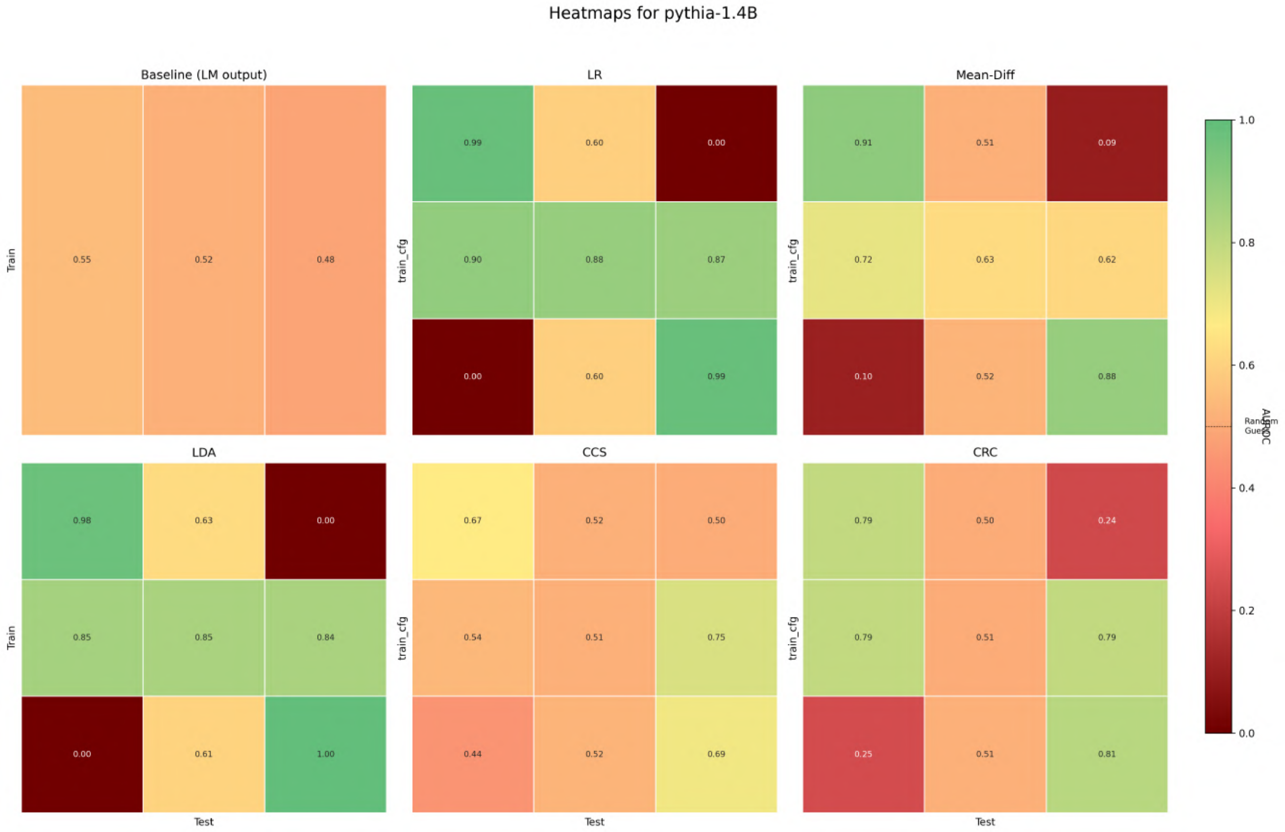

Figure C.4: Heatmaps for Pythia-1.4B model

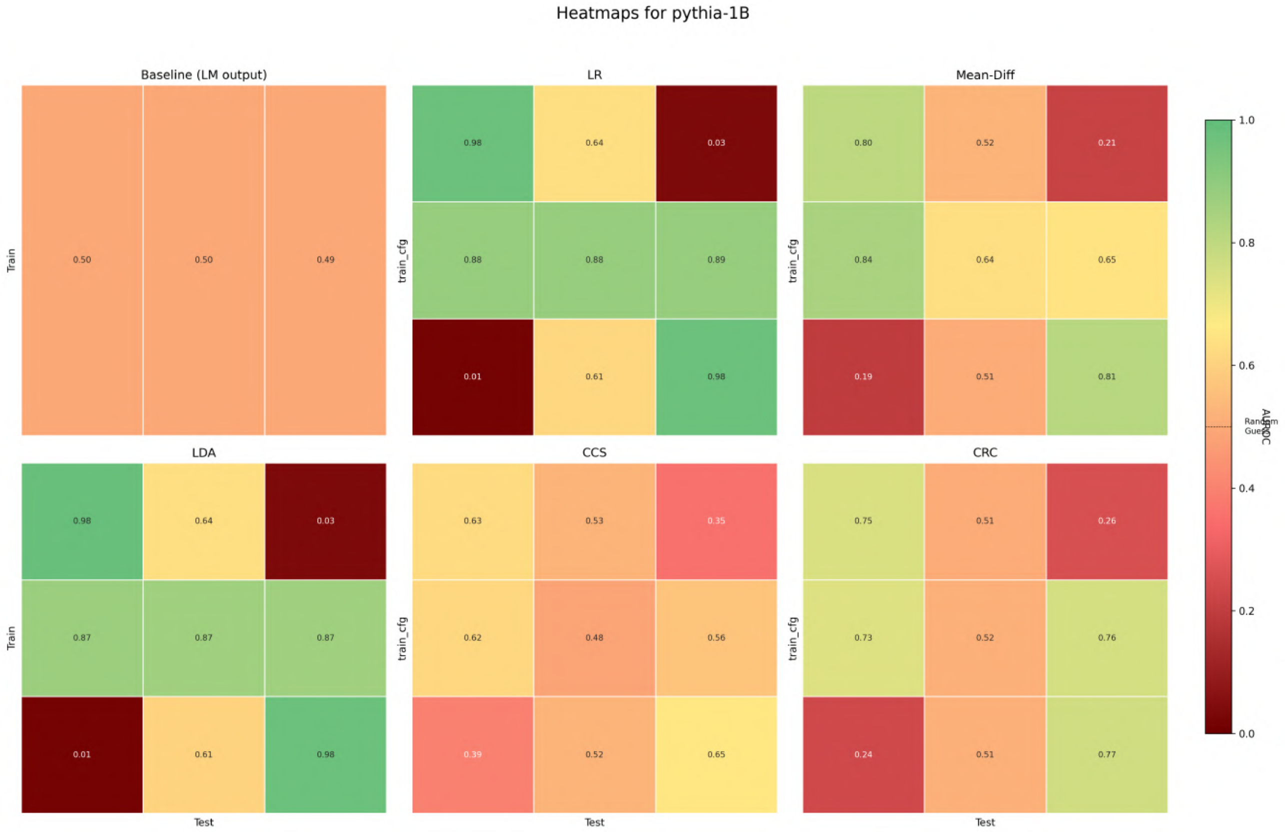

Figure C.5: Heatmaps for Pythia-1B model

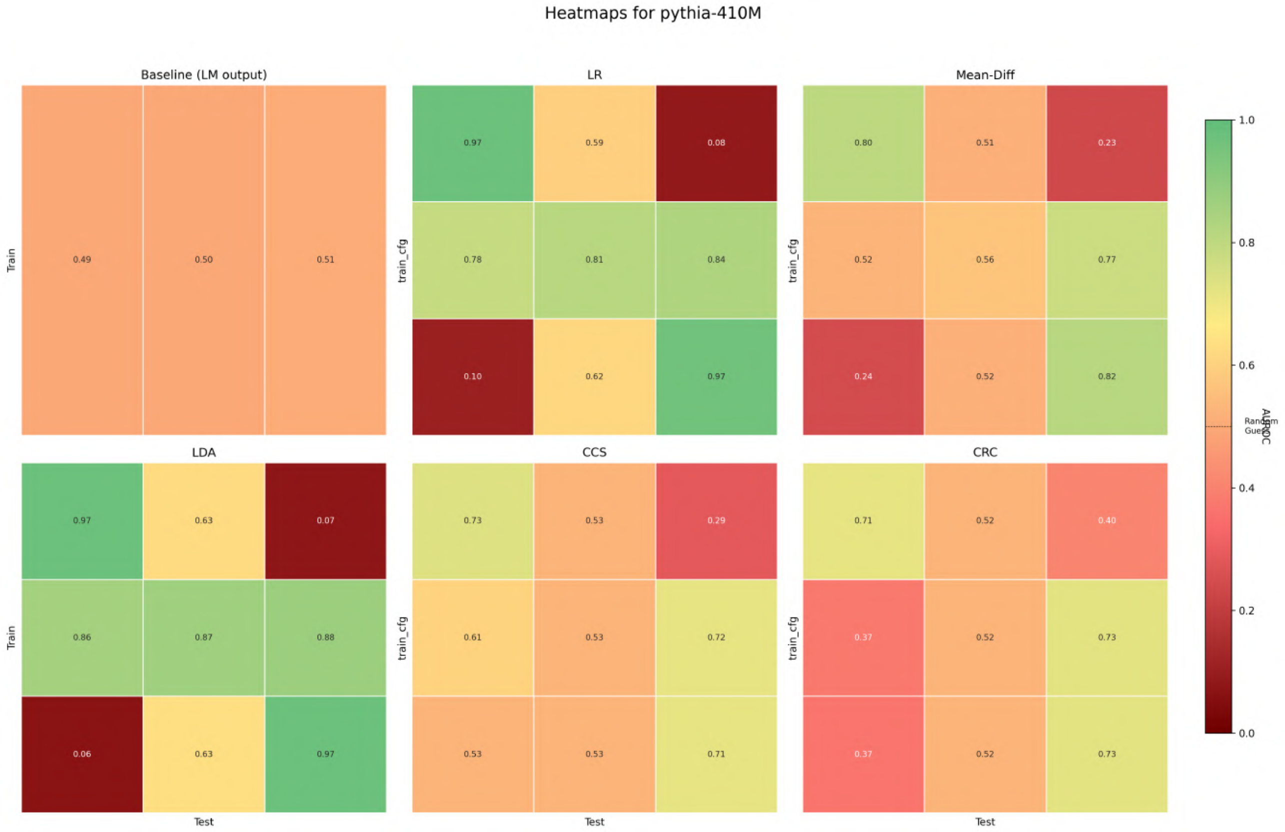

Figure C.6: Heatmaps for Pythia-410M model