Probing the Geometry of Truth: Consistency and Generalization of Truth Directions in LLMs Across Logical Transformations and Question Answering Tasks

Yuntai Bao[1] **Xuhong Zhang**[1*] **Tianyu Du**[1] **Xinkui Zhao**[1] **Zhengwen Feng**[1]
Hao Peng[2] **Jianwei Yin**[1]
[1]Zhejiang University
[2]Zhejiang Normal University
{yuntaibao, zhangxuhong, zjradty, zhaoxinkui, fengzhengwen}@zju.edu.cn,
hpeng@zjnu.edu.cn, zjuyjw@cs.zju.edu.cn

Abstract

Large language models (LLMs) are trained on extensive datasets that encapsulate substantial world knowledge. However, their outputs often include confidently stated inaccuracies. Earlier works suggest that LLMs encode truthfulness as a distinct linear feature, termed the "truth direction", which can classify truthfulness reliably. We address several open questions about the truth direction: (i) whether LLMs universally exhibit consistent truth directions; (ii) whether sophisticated probing techniques are necessary to identify truth directions; and (iii) how the truth direction generalizes across diverse contexts. Our findings reveal that not all LLMs exhibit consistent truth directions, with stronger representations observed in more capable models, particularly in the context of logical negation. Additionally, we demonstrate that truthfulness probes trained on declarative atomic statements can generalize effectively to logical transformations, question-answering tasks, in-context learning, and external knowledge sources. Finally, we explore the practical application of truthfulness probes in selective question-answering, illustrating their potential to improve user trust in LLM outputs. These results advance our understanding of truth directions and provide new insights into the internal representations of LLM beliefs.[1]

1 Introduction

Large language models (LLMs) possess extensive knowledge, as they are trained on immense corpora that encompass a significant portion of world knowledge. However their outputs are not always reliable and are prone to confidently presenting falsehoods (Bender et al., 2021; Evans et al., 2021; Lin et al., 2022; Liu et al., 2023). This unreliability raises critical concerns about the use of LLMs in applications where accuracy is paramount. A growing body of work (Burns et al., 2022; Azaria and Mitchell, 2023; Marks and Tegmark, 2023; Mallen and Belrose, 2023; Bürger et al., 2024) aims to elicit accurate information from LLMs despite untruthful outputs. These studies use lightweight classifiers, often referred to as *probes*, to analyze patterns in the model's internal representation that reliably indicate truthfulness. Specifically, given a model and a piece of text, an ideal truthfulness probe is able to tell whether the model believes the text conveys truthful content. By achieving empirical success with linear probes, these works generally believe that truthfulness is internally represented as a salient linear feature and manifests as a "truth direction".

The goal of this work is to conduct a more in-depth study of the truth direction as an inherent property of LLMs. Although previous works undoubtedly help us understand truth directions, they fail to answer the following questions: (**RQ1**) Do LLMs universally represent truthfulness as a linear feature? (**RQ2**) Are simple probing techniques sufficiently expressive to identify truth directions? (**RQ3**) If and when a "truth direction" exists, in what ways does it generalize? We challenge conclusions from prior works based on empirical evidence, provide preliminary answers to the questions above and present novel observations.

In response to **RQ1**, we find that not all LLMs exhibit a consistent "truth direction", and that this property is closely related to a model's capability. While prior works often assume the universal existence of truth directions, we challenge this assumption. Our evidence suggests that truthfulness is more consistently represented across logical negations in more capable LLMs. Based on this finding, we question the conclusion of Levinstein and Herrmann (2024), which attributes the generalization failure to limitations in previous probing techniques. Instead, we argue that the inconsis-

[*]Corresponding author.

[1]Our code is public at https://github.com/colored-dye/truthfulness_probe_generalization

Findings of the Association for Computational Linguistics: ACL 2025, pages 682–700
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

tency lies within the LLM itself and that in answer to **RQ2**, simple supervised probes are sufficiently expressive to identify the truth direction when it is distinctly represented within the model.

In addressing **RQ3**, we aim to explore the generalization capability of the truth direction, as it illuminates whether an LLM consistently represents truthfulness across different knowledge domains, logical transformations, syntax forms and grounding knowledge source. Previous studies on truth directions have extensively explored the former two aspects of generalization. Regarding syntax forms, they focus either on declarative statements or on (Q, A) pairs; we bridge this gap by testing whether truthfulness probes trained on simple statements generalize to question answering (QA) tasks. Additionally, we examine several variations of QA, including zero- and few-shot QA, with and without provided answer options, and grounded either in parametric knowledge or in question contexts. Our experimental results show that truthfulness probes demonstrate a high degree of generalization.

Furthermore, based on our observation that truthfulness probes are calibrated on certain QA tasks, we introduce their use in selective QA. In this application, we select the subset of answers evaluated as correct by the truthfulness probe from those generated by the LLM. Through this demonstration, we aim to show how truthfulness probes can enhance user trust in real-world LLM-based applications.

Our contributions are as follows:

1. We summarize truthfulness probes from prior works, introduce a new instantiation, and conduct extensive experiments.
2. We study truth directions following three research questions. In addressing **RQ1**, we explore whether the truth direction is common among LLMs. For **RQ2**, we test if sophisticated probing techniques are required to identify truth directions. In answer to **RQ3**, we assess the generalization capabilities of truth directions.
3. We demonstrate a practical application of truthfulness probes for selective question answering, improving generation quality by filtering out unreliable answers.

2 Related Work

Eliciting latent knowledge (ELK). The field of scalable oversight (Christiano et al., 2021) seeks to address the information asymmetry between superhuman AI systems and human evaluators. It is assumed that although the AI possesses significant knowledge, its behavior is untrustworthy because it is not trained with an objective that explicitly incentivizes outputs to align with the truth (Mallen and Belrose, 2023). ELK is an approach within scalable oversight that aims to identify patterns in an AI's activations that correspond to the truth. The primary challenge lies in identifying patterns that generalize reliably to questions where human evaluators are unable to verify the answers (Mallen and Belrose, 2023). Our work demonstrates that probing techniques can achieve reasonable generalization with limited supervision, suggesting that probes may offer a promising approach for ELK.

Probing for truthfulness. Several studies have used probing techniques to uncover truthfulness by examining an LLM's internal states, independent of its inputs or outputs (Lee et al., 2023; Joshi et al., 2024). Regarding the *geometry of the representation* of truthfulness, most studies agree that this representation is likely linear, as demonstrated by the use of linear probes in studies such as CCS (Burns et al., 2022), mass-mean (Marks and Tegmark, 2023; Li et al., 2023), TTPD (Bürger et al., 2024) and the commonly used baseline: logistic regression. Notably, CCS is an unsupervised approach and targets yes/no *question-answering tasks*, while others are supervised and target *factual statements*. In contrast, some works are *geometry-agnostic*. Azaria and Mitchell (2023) propose SAPLMA which is based on MLP architecture, while He et al. (2024) introduce the LLM Factoscope, which leverages a Convolutional Neural Network architecture.

The aforementioned studies examine truthfulness grounded in a model's *parametric knowledge*, whereas Sky et al. (2024) detect hallucination with probes in the setting of in-context generation, where *knowledge is grounded in the context*.

3 Summary of Probes and Data

In this section, we formally define the task of probing binary features from a model's internal representations and describe the specific probe architectures used in this study. Furthermore, we detail the labeled datasets used to train and evaluate the truthfulness probes.

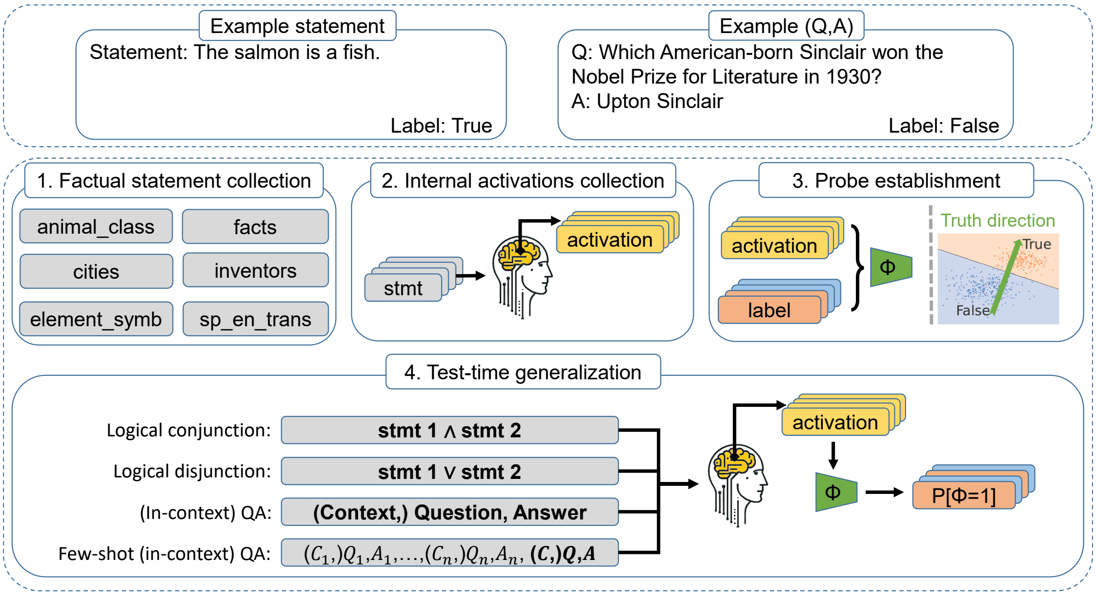


Figure 1: Illustration of truthfulness probes. A truthfulness probe is established using the LLM's internal states when processing labeled statements. The probe is then able to tell if the LLM believes an unseen statement or a given response to a question is true or false leveraging only the LLM's internal states.

3.1 Formulation of Binary Probes

Our target model is the transformer language model (Vaswani, 2017), which processes token sequences through a series of layers. A sequence of n input tokens, $t = (t_1, t_2, \ldots, t_n)$, is first converted into embeddings, $\boldsymbol{h}^{(0)} = (\boldsymbol{h}_1^{(0)}, \boldsymbol{h}_2^{(0)}, \ldots, \boldsymbol{h}_n^{(0)})$, by the initial embedding layer. The embeddings are then passed through L layers, where each layer generates representations based on the preceding layer's output. The representation of a single token is a vector: $\boldsymbol{h}_i^{(j)} \in \mathbb{R}^d$ and $(0 \leq j \leq L, 1 \leq i \leq n)$. Finally, the LLM produces predictions using $\boldsymbol{h}^{(L)}$.

Suppose we have prior knowledge that the model internally represents a binary feature. Our goal is to establish a probe Φ, that uses only the model's representations to classify the target attribute. The probe outputs either binary labels $-1/1$, or probabilistic predictions such as $\mathrm{P}[\Phi = 1]$.

For autoregressive models, which are the primary focus of this paper, we utilize the representation at the final token position of the l-th layer, $\boldsymbol{h}_{-1}^{(l)}$. This approach aligns with prior work (Burns et al., 2022; Azaria and Mitchell, 2023; Marks and Tegmark, 2023), where the final token position attends to all previous tokens due to the causal attention mechanism. We also assume we have a labeled dataset, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{M}$, where x_i represents a token sequence and y_i is the label for the target attribute. Processing these sequences through the model yields $\mathcal{D}_{\text{rep}} = \{((\boldsymbol{h}_{-1}^{(l)})_i, y_i)\}_{i=1}^{M}$, which we use to build our classifier.

To maximize classification accuracy, we define a cost function to quantify the classification error of the probe, $J(\Phi, \boldsymbol{h}, y)$. The mechanistic objective is to minimize the expected classification error over the data distribution: $\arg\min_{\Phi} \frac{1}{M} \sum_{i=1}^{M} J(\Phi, \boldsymbol{h}_i, y_j)$.

3.2 Instantiations of Binary Probes

We classify the probes into two categories: geometry-oriented and statistics-based. Geometry-oriented probes leverage knowledge of the geometric structure of the representation. Under the "truth direction" hypothesis, true/false representations can be separated by a hyperplane, and the normal vector of this hyperplace corresponds to the "truth direction". In contrast, statistics-based probes are geometry-agnostic and aim to maximize the probability of observing the correct labels given the input data. Our implementations are based on the `scikit-learn` (Pedregosa et al., 2011) library.

Geometry-oriented Probes. For geometry-oriented probes, we introduce two instantiations: linear support vector machine (**SVM**) (Cortes and Vapnik, 1995) and mass-mean (**MM**) (Marks and Tegmark, 2023) instantiation. The rationale for selecting linear SVM is its ability to maximize the margin, which aligns with the goal of identifying a separating hyperplane. As SVM does not directly provide probability predictions, we fit a post-hoc probability distribution using Platt scaling through cross-validation on the training data (Platt, 1999).

Statistics-based Probes. For statistics-based probes, we present logistic regression (**LR**) and multi-layer perceptron (**MLP**). LR is commonly used as a baseline, while MLP is termed SAPLMA by Azaria and Mitchell (2023).

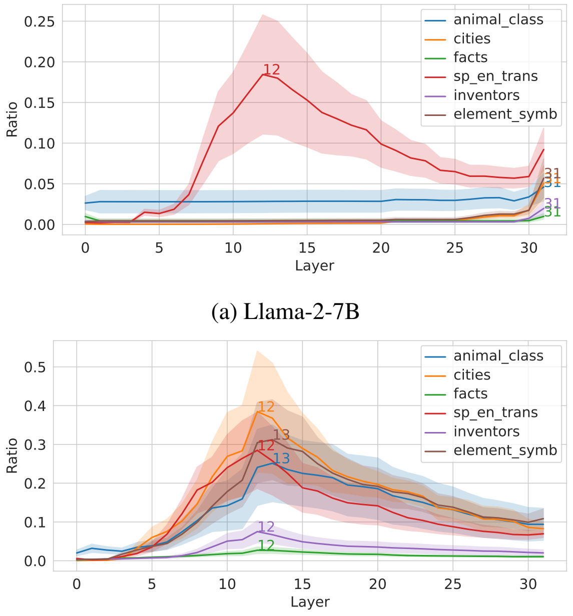
3.3 Data for Probing Truthfulness

The binary probes introduced above are applied to the truthfulness classification task, assuming the availability of truthfulness-specific data. We use the factual statements curated by Bürger et al. (2024), drawing from datasets by Azaria and Mitchell (2023) and Marks and Tegmark (2023). These datasets cover a variety of topics, including `animal_class`, `cities`, `element_symb`, `facts`, `inventors`, `sp_en_trans`, as well as variations incorporating logical negations, conjunctions and disjunctions. Each statement is labeled as "true" or "false", indicating its factuality. Statements can be atomic or compound. Atomic statements make individual claims, either affirmative or negative. Negative statements correspond to their affirmative counterparts, with syntax-level negation and inverted labels. Compound statements are created by logically combining atomic statements of the same topic through conjunction or disjunction.

4 Experiments

4.1 Preliminary Experiment: Layer Selection

Identifying the optimal layer for detecting truthfulness is crucial for probe performance. Marks and Tegmark (2023) observed that the truth direction "emerges rapidly in early-middle layers"; however, this observation does not indicate which specific layer provides the most effective representations. To address this, we adopt the technique used by Bürger et al. (2024) and MacDiarmid et al. (2024), which evaluates the difficulty of separating true/false statements across layers by analyzing variance. The ideal layer maximizes the separation between true and false representations, quantified by the "between-class variance", relative to the internal variance within each class, referred to as "within-class variance". By plotting the ratio of between-class to within-class variance across decoder layers for a range of topic-specific datasets, we identify the optimal layer as the one with the highest ratio.



(a) Llama-2-7B

(b) Llama-3.1-8B

Figure 2: Ratio of between-class variance to within-class variance across layers. The layer indices (starting from 0) for the greatest ratios are annotated at the summit of each curve. The solid curves are mean values, and the surrounding shades denote standard error.

We present the ratio of between-class variance to within-class variance across layers for Llama-2-7B (Touvron et al., 2023) and Llama-3.1-8B (Dubey et al., 2024) in Figure 2, with results for additional models in the Appendix. Each curve represents statements involving affirmations, negations, conjunctions and disjunctions of the same topic. For Llama-3.1-8B, the 12th layer (zero-indexed) emerges as the optimal layer. In contrast, for Llama-2-7B, a peak occurs only for the `sp_en_trans` topic, with minimal separation observed for other topics. This suggests that, while Llama-2-7B may internally represent truthfulness as a feature, it does so in a domain-specific manner, with limited consistency across knowledge domains. Additionally, this feature appears to lack salience in the early-middle layers.

4.2 Probing a Randomized Model

This section investigates whether the truth direction is an inherent structure within a pretrained LLM or an artifact constructed by the truthfulness probe. To address this, we randomly initialize the weights of Llama-3.1-8B, extract activations from its 12th layer using the `animal_class` dataset, and train probes on a 70% split while testing them on the

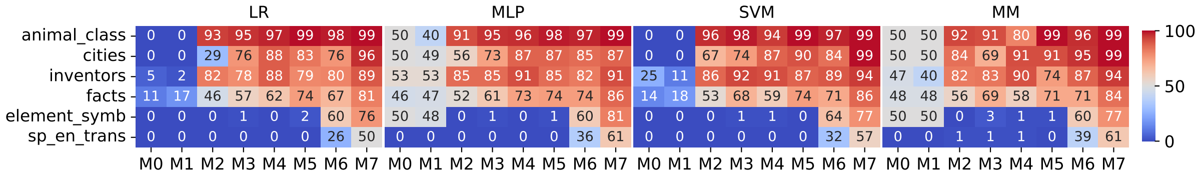


Figure 3: AUROC (in percentage) of probes trained on affirmative statements and tested on negative ones. AUROC > 0.5 indicates success of generalization. M0-M7 refer to models from Llama-2-7B to Llama-3.1-70B-Instruct.

remaining 30%. Results show that the AUROCs are 0.50, 0.52, 0.58, 0.50 for the LR, MLP, MM and SVM probes, respectively. In contrast, when using the pretrained weights, the AUROCs achieve 1.0 across all probes. These findings demonstrate that the simple probes introduced in Section 3.2 cannot independently construct a truth direction, confirming that truth direction is a product of the pretraining process.

4.3 Consistency of Truth Directions

Levinstein and Herrmann (2024) claim that probes such as MLP fail to generalize across negation. However, we question this conclusion and hypothesize that the generalization performance of truthfulness probes is influenced more by the targeted LLM than by the probe itself. To test this hypothesis, we examine whether the truth direction identified for affirmative statements is consistent with that identified for negative statements of the same topic.

4.3.1 Experimental Setup

Data. For each of the six topics introduced in Section 3.3, we train probes on affirmative statements and test them on corresponding negative ones. For example, we train probes on affirmative statements of the `animal_class` topic, and test them on `neg_animal_class`. Note that the training and test data contain the same set of knowledge, differing only in syntax.

Models. We select a series of LLMs with increasing levels of general capability, including both foundational models and instruction-tuned ones (According to the evaluation results on a range of standard benchmarks[2]): Llama-2-7B(-Chat), Llama-2-13B(-Chat), Llama-3.1-8B(-Instruct), Llama-3.1-70B(-Instruct). Their optimal layers are 12(13), 13(13), 12(13), 33(33), respectively.

Methods. We employ the four probe instantiations introduced in Section 3.2.

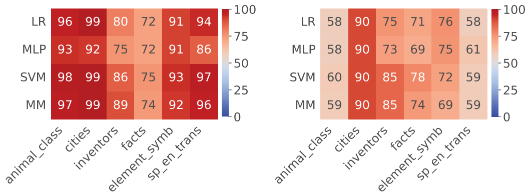
Metrics. The metric used is AUROC (Area Under Receiver Operating Characteristic Curve) which is commonly used to assess classifier performance. A probe is considered to generalize successfully on a topic if its test AUROC exceeds 0.5, indicating performance better than chance. Results are averaged across three trials, with randomness introduced through probe initialization and data splits for cross-validation of Platt scaling, while the training data remains constant.

4.3.2 Results

The results are presented in Figure 3. We observe a positive correlation between the performance of truthfulness probes and the general capability of the target models. For Llama-2-7B(-Chat), the probes fail to generalize on all six topics. For Llama-2-13B(-Chat) and Llama-3.1-8B(-Instruct), the probes generalize on four topics; for Llama-3.1-70B they generalize on five topics; and for Llama-3.1-70B-Instruct, they generalize on all six topics.

Regarding whether the truthfulness probe is a faithful reflection of the actual truth direction, we borrow the *Weak-to-Strong Explanation* from Zhou et al. (2024): if weak classifiers can successfully distinguish the representations, it indicates that LLMs have implicitly converted inputs to different representations. For the most capable model, Llama-3.1-70B-Instruct, all the simple probes we use are able to generalize across logical negation on all six topics. This suggests that Llama-3.1-70B-Instruct consistently represents truthfulness in its internals for both affirmative and negative statements. Therefore the results suggest a potential correlation between the degree of generalization of its truth direction and the model's capability (e.g., knowledge capacity and natural language understanding ability). Additionally, the differences in performance between probes become negligible starting with Llama-2-13B-Chat and onward. This indicates that, for more capable LLMs, probe performance is more influenced by the target model itself than by the design of the probes.

[2] `https://github.com/meta-llama/llama-models/blob/main/models`



(a) Logical conjunctions.

(b) Logical disjunctions.

Figure 4: AUROC (in percentage) of probes trained on atomic factual statements and tested on logical conjunctions/disjunctions for Llama-3.1-8B. AUROC > 0.5 indicates the success of generalization.

4.4 Binary Logical Transformation

In Section 4.3, we tested the ability of truthfulness probes to generalize across logical negation. In this experiment, we extend the analysis to more complex binary logical transformations, specifically logical conjunction and disjunction. This requires the LLM to perform several implicit tasks: identify the truthfulness of both atomic statements, interpret binary logical operators from natural language to abstract concepts, and apply the operators to compute the joint truthfulness.

4.4.1 Experimental Setup

Data. The training data consists of all atomic statements. The test data comprises logical conjunctions and disjunctions of atomic affirmative statements for each topic. The knowledge of the test data is covered by the training data.

Models. We use Llama-3.1-8B as the primary model for demonstration.

Methods. We apply the four probe instantiations introduced in Section 3.2.

Metrics. For this experiment, we use AUROC as a measure of probe performance. A probe is considered to generalize successfully on a topic if its test AUROC exceeds 0.5. Results are averaged across three trials, with randomness introduced through probe initialization and training data splits.

4.4.2 Results

According to Figure 4, all probes successfully generalize to both logical conjunctions and disjunctions. However, performance is notably stronger for logical conjunctions compared to disjunctions across `animal_class`, `element_symb`, and `sp_en_trans` topics. This discrepancy may suggest that disjunctions pose a greater challenge for Llama-3.1-8B to interpret truthfulness.

4.5 Question Answering

We hypothesize that if truthfulness is consistently represented in an LLM's internal states, this representation should depend solely on the semantics of a sentence, rather than its syntax form. Additionally, question answering is more common than statements in real-world human-AI interactions. Motivated by these considerations, we examine if truthfulness probes, trained on atomic factual statements, can generalize to the QA setting.

We test on a multiple-choice task and a short-form QA task. We also investigate the in-context learning scenario, a popular prompting technique for teaching LLMs new tasks at inference time. While in-context examples can be beneficial, they may include incorrect or misleading examples, which raises questions about how probes handle false examples. Therefore, we pay special attention to the behavior of the probes when incorrect examples are present.

4.5.1 Experimental Setup

Data. The training data consists of all atomic statements. The test data includes MMLU (Hendrycks et al., 2020) and TriviaQA (Joshi et al., 2017). For MMLU, we sample 50 questions from the test set for each of the 57 sub-tasks. As it is a multiple-choice dataset, for each question we select the correct answer and an incorrect answer. For TriviaQA, we sample 20 answers per question from the model at unit temperature.

Models. We use Llama-3.1-8B, as it is a high-capability LLM with ∼10B parameters.

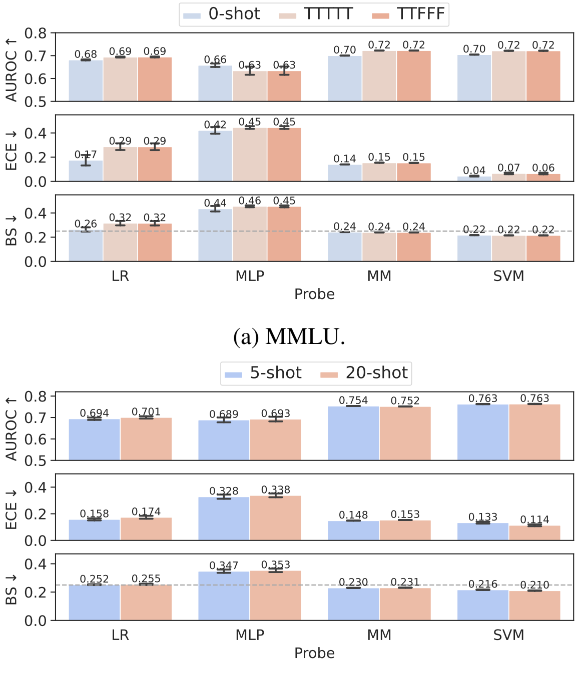
Methods. We apply the four probe instantiations introduced in Section 3.2.

Metrics. In addition to classification accuracy, we evaluate calibration, as it is crucial for assessing the reliability of predictions. Specifically, we evaluate AUROC, Expected Calibration Error (ECE) and Brier Score (BS). ECE measures calibration, while BS reflects both accuracy and calibration. Lower values are preferred for both ECE and BS. For ECE we use a binned approach with 10 bins, where each bin contains an equal number of samples, and report the mean absolute error between the accuracy and confidence within each bin. The random baseline for BS is 0.25, corresponding to a uniform prediction of 0.5. Results are averaged across three trials, with randomness introduced via probe initialization and training data splits.

Prompt setups. We test three prompt settings for MMLU: (1) “zero-shot”: a zero-shot prompt; (2) “TTTTT”: a five-shot prompt with all correct exemplars; (3) “TTFFF”: a five-shot prompt where the first two examples are correct and the following three are incorrect. For TriviaQA, we test 5-shot and 20-shot prompts.

Although it is true that the correctness of few-shot examples could be verified in practice, our motivation to study incorrect in-context examples includes: (1) It helps understand how truthfulness probes handle context that contains mixed truthful and untruthful information; (2) It provides insights into how robust truthfulness probes are to potentially conflicting information.

4.5.2 Results



(a) MMLU.

(b) TriviaQA.

Figure 5: AUROC↑/ECE↓/BS↓ of truthfulness probes for Llama-3.1-8B on MMLU and TriviaQA. The dashed gray line corresponds to random results, and error bars denote standard error.

The results, shown in Figure 5, reveal that across all probes, accuracy generally improves when few-shot exemplars are provided. This suggests that providing more task-related context in the prompt typically aids the LLM in the implicit truthfulness classification task, and that truthfulness probes not only generalize from factual statements to both multiple-choice QA and short-form QA, but also generalize from fundamental commonsense knowledge to both expert knowledge (MMLU) and trivia knowledge (TriviaQA). Notably, according to the results on MMLU, the effect of imperfect few-shot prompts is near identical to entirely correct few-shot prompts. This indicates that the performance of truthfulness probes is not significantly influenced by the truthfulness of the few-shot examples – the probes express the truthfulness of the *final* (Q, A) pair, treating prior exemplars as context. Interestingly, this finding aligns with Halawi et al. (2023)’s observation that early layers of the LLM are insensitive to false in-context demonstrations.

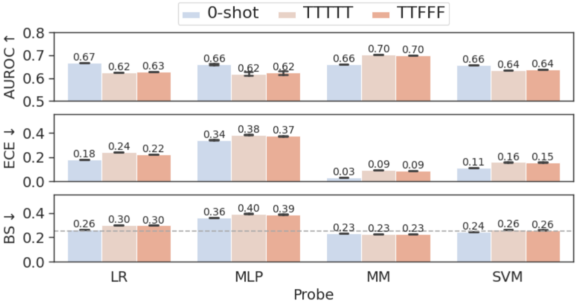


Figure 6: AUROC↑/ECE↓/BS↓ of in-domain probes for Llama-3.1-8B trained and tested on MMLU.

However, few-shot prompting does not always improve calibration, and a positive case is only observed for the SVM probe on the TriviaQA dataset. Among all probes, the SVM probe performs best in terms of both classification accuracy and calibration, likely because of its ability to accurately identify the geometry of the truth direction and the benefits of the additional Platt scaling procedure.

4.5.3 Additional Experiments

Beyond the generalization of probes trained on factual statements, we are interested in how these probes differ from in-domain probes, which are trained and tested on data of the same domain. Therefore we perform follow-up experiments by training and testing probes on MMLU using the Llama-3.1-8B model, using the same data and prompt setup as above. Specifically, we train on a random 70% split and test on the rest 30%.

We present results averaged over three random trials in Figure 6. Comparing the results with those of Figure 5, we observe that in-domain probes generally under-performs truthfulness probes trained on atomic factual statements in terms of AUROC. This finding indicates that truthfulness probes under the current setup perform better than those under the in-domain QA setup.

4.6 Contextual Knowledge

When training truthfulness probes using factual statements, we are targeting the factual aspect of truthfulness, where the grounding knowledge resides in the LLM's parameters. However, in-context knowledge also plays a vital role in generation. In this section, we investigate if truthfulness probes can also capture this additional aspect of truthfulness, where the grounding knowledge is provided as contextual information in the prompt. Notably, context grounding is fundamentally different from factual correctness, since faithfully following false context is acceptable for the former but not for the latter. We conduct experiments on two tasks: in-context QA and abstractive summarization.

4.6.1 Experimental Setup

Data. The training data consists of all the atomic statements. For the in-context QA task we use the SciQ (Welbl et al., 2017) and BoolQ (Clark et al., 2019) datasets. For SciQ, we randomly select 1000 questions, pairing each question with the lettered choice for both the true and a randomly selected false answer. The answers of BoolQ are binary "yes/no", therefore we flip the answers to balance true and false (Q, A) pairs. For the abstractive summarization task, we use the XSum dataset (Narayan et al., 2018) and XSum Hallucination Annotations (Maynez et al., 2020). For each article, we pair true summaries from the former dataset with false ones from the latter one.

Models. We use Llama-3.1-8B.

Methods. We apply the four probe instantiations introduced in Section 3.2.

Metrics. We report AUROC, ECE and BS. The results are averaged over three trials, with randomness introduced via initialization and data splits.

Prompt setups. For SciQ dataset which is a multiple-choice task, we implement four settings: (1) "zero-shot": zero-shot prompt; (2) "TTT": three-shot prompt where all exemplars are correct; (3) "TTF": three-shot prompt, where the first two examples are correct and the third is incorrect; (4) "FFT": three-shot prompt, where the first two examples are incorrect and the third is incorrect. For BoolQ, we implement four settings: "no options", "with options", and one-shot "T"/"F" (with possible options). For XSum, no options are provided in the prompt as it is not a multiple-choice task. We implement the following prompt configurations: zero-shot, "T", "TT", and "TTT".

4.6.2 Results

The results are shown in Figure 7. Across all datasets and most probes, possible answer options and few-shot exemplars generally improve classification accuracy. This indicates that truthfulness probes generalize from factual statements to both in-context multiple-choice QA tasks and abstractive summarization tasks.

Among the probes, statistics-based probes (LR and MLP) display greater standard error in terms of all three metrics than geometric-oriented ones (MM and SVM), likely due to their optimization instability, and the SVM probe performs best from the perspective of BS. Additionally, discrepancies are observed for LR probe on the SciQ and BoolQ datasets, MLP probe on the BoolQ dataset, and MM probe on the XSum dataset, where accuracy improves in response to in-context exemplars but calibration worsens. We assume that these discrepancies could be explained with the overconfidence of the probes' predictions.

4.7 Selective Question Answering

Based on the findings of Section 4.5, we observe that truthfulness probes trained on atomic statements are capable of generating calibrated probabilistic predictions for QA tasks while achieving reasonable accuracy. Building upon these observations, this section investigates whether truthfulness probes can be leveraged to selectively identify correct answers from a set of candidate responses sampled from an LLM. This setup is inspired by the work of Kadavath et al. (2022), who demonstrated that an LLM can evaluate the correctness of its own answers through verbal feedback. While our selective QA experiment shares similarities with Kadavath et al. (2022), our approach differs fundamentally as we leverage internal representations rather than model-generated feedback.

For this experiment, we use the TriviaQA test data from Section 4.5, where we sample 20 answers from the Llama-3.1-8B model using a 20-shot prompt with unit temperature. To perform selective QA, we select the subset of (Q, A) pairs for which the truthfulness probe predicts $\mathrm{P}[\Phi = 1] > 0.5$ and report the accuracy on this subset. We use the SVM probe, as it performs best in terms of both classification accuracy and calibration.

The aggregated accuracy across all sampled

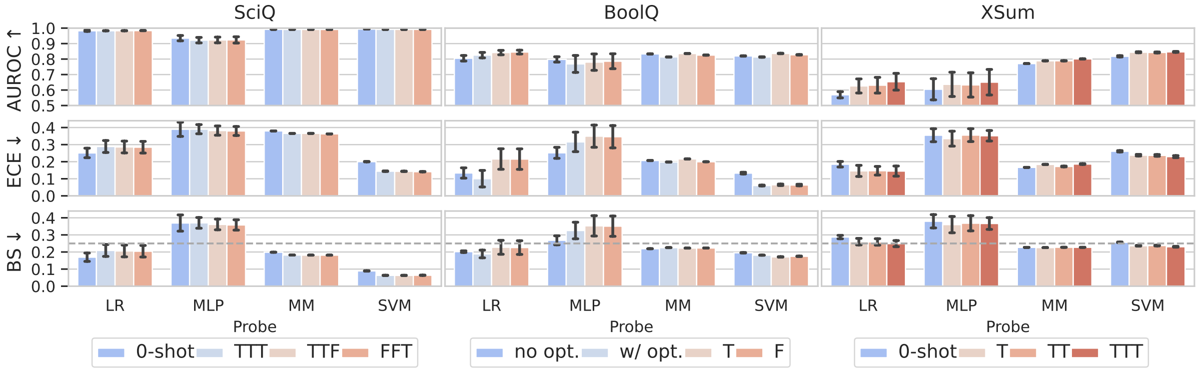


Figure 7: AUROC↑/ECE↓/BS↓ of truthfulness probes for Llama-3.1-8B on tasks where grounding knowledge is provided in the prompt. The dashed gray line corresponds to random results, and error bars denote standard error.

(Q, A) pairs is 55.29%. Among these, the truthfulness probe classifies 80.26% as true, and the accuracy of this subset is 64.06%. This demonstrates that truthfulness probes can be used to filter out false answers sampled from LLMs.

5 Conclusions

In this work, we provide preliminary evidence supporting our hypothesis that consistent truth directions only emerge in capable LLMs and not in weaker ones, and they could be effectively identified with simple linear probes. We also investigate the generalization properties of truth directions. Empirical results show that truthfulness probes trained only on atomic statements generalize well to logical transformations, (few-shot) question answering and contextual truthfulness. These findings underscore the potential of truthfulness probes to identify truth directions using simple anchor data, thereby facilitating the elicitation of latent knowledge within LLMs.

6 Limitations

This study has several limitations that warrant discussion. First, the term "truth" as used in this paper represents an idealized concept, but it may not be what the truthfulness probes actually measure. Drawing on Kadavath et al. (2022) and Marks and Tegmark (2023), the pretraining of language models largely involves imitating human-generated text. Consequently, truthfulness probes are likely capturing an overlap between widely accepted human beliefs and factual, objective truths about the physical world. This raises interesting questions about how such probes might perform in the context of scalable oversight, particularly with hypothetical AI systems that surpass human intelligence.

Second, our investigation into the generalization of truthfulness probes is limited to short-form QA. Extending this analysis to more complex scenarios, such as long-form QA or instruction-following tasks, may yield novel insights and uncover more practical applications of truthfulness probes.

Third, the causality of truthfulness probes is unclear. Our approach relies on classic classification techniques, consistent with prior work on probing truthfulness. Meanwhile, we do not discuss the causal effects of truth directions, i.e., whether LLMs utilize the implicit truthfulness classification results for predictions. Marks and Tegmark (2023) conducted causal intervention experiments using the mass-mean probe and showed that mass-mean directions are highly causal. However, as highlighted by Kumar et al. (2022), probes often capture spuriously-correlated features rather than exclusively isolating the target feature. Future research on the causal implications of truth directions could provide valuable insights into guiding LLMs to produce more truthful responses, as in Li et al. (2023).

Finally, our experiments are constrained by computational resources, with the largest model evaluated being Llama-3.1-70B. As a result, our hypothesis that highly capable LLMs will eventually establish a consistent internal concept of truthfulness remains untested on more advanced models such as GPT-4 (Achiam et al., 2023).

Acknowledgment

This work was partly supported by the National Key Research and Development Program of China under No. 2024YFB3900105, NSFC under No. 62402418, Zhejiang Province's 2025 "Leading Goose + X" Science and Technology Plan under grant No.2025C02034, the Key R&D Program of Ningbo under No. 2024Z115, and the Open Project of Key Laboratory of General Quality Technology and Application of Intelligent Manufacturing Equipment, Ministry of Industry and Information Technology (HK202403532).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Lennart Bürger, Fred A Hamprecht, and Boaz Nadler. 2024. Truth is universal: Robust detection of lies in llms. *arXiv preprint arXiv:2407.12831*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Paul Christiano, Ajeya Cotra, and Mark Xu. 2021. Eliciting latent knowledge: How to tell if your eyes deceive you. *URL https://docs. google. com/document/d/1WwsnJQstPq91_ Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8/edit# heading= h. jrzi4atzacns*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful ai: Developing and governing ai that does not lie. *arXiv preprint arXiv:2110.06674*.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2023. Overthinking the truth: Understanding how language models process false demonstrations. *arXiv preprint arXiv:2307.09476*.

Jinwen He, Yujia Gong, Zijin Lin, Yue Zhao, Kai Chen, et al. 2024. Llm factoscope: Uncovering llms' factual discernment through measuring inner states. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10218–10230.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. 2024. Personas as a way to model truthfulness in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6346–6359.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. 2022. Probing classifiers are unreliable for concept removal and detection. *Advances in Neural Information Processing Systems*, 35:17994–18008.

Bruce W Lee, Benedict Florance Arockiaraj, and Helen Jin. 2023. Linguistic properties of truthful response. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 135–140.

Benjamin A Levinstein and Daniel A Herrmann. 2024. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, pages 1–27.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Dong C Liu and Jorge Nocedal. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528.

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. 2023. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4797.

Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. 2024. Simple probes can catch sleeper agents.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluis Marquez. 2024. Factual confidence of LLMs: on reliability and robustness of current estimators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4554–4570, Bangkok, Thailand. Association for Computational Linguistics.

Alex Mallen and Nora Belrose. 2023. Eliciting latent knowledge from quirky language models. *arXiv preprint arXiv:2312.01037*.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

Lorenzo Pacchiardi, Alex James Chan, Sören Mindermann, Ilan Moscovitz, Alexa Yue Pan, Yarin Gal, Owain Evans, and Jan M Brauner. 2023. How to catch an ai liar: Lie detection in black-box llms by asking unrelated questions. In *The Twelfth International Conference on Learning Representations*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

J Platt. 1999. Probabilistic outputs for svms and comparisons to regularized likehood methods. *Advances in Large Margin Classifiers*.

CH-Wang Sky, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4401–4420.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. 2024. How alignment and jailbreak work: Explain LLM safety through intermediate hidden states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2461–2488, Miami, Florida, USA. Association for Computational Linguistics.

A Truth-related concepts

In this section, we distinguish the term "truthfulness" from several related concepts. Truthfulness refers to the alignment of a statement or (Q, A) with either world knowledge or contextual sources. The former, following Mahaut et al. (2024), is termed "*factuality*". Contextual truthfulness, by contrast, may include fictional information that deviates from real-world facts, such as solving math problems in a hypothetical scenario.

Untruth lies at the negative end of the truthfulness spectrum and differs from *hallucination*, which refers to generations that are nonsensical or unfaithful to the provided source content (Ji et al., 2023). A key distinction between untruth and hallucination is that truthfulness requires a sentence to be both sensical and unambiguous. Additionally, we differentiate untruth from *lies*. According to Pacchiardi et al. (2023), an answer is considered a lie only if the speaker knows the correct answer. In this view, a lie is a subset of untruths.

B Explanation on Choice of Probes

In this work we summarize LR, MLP, SVM and MM instantiations and use them for experiments. We do not use the TTPD probe introduced by the recent work, Bürger et al. (2024), for two reasons. First, the design of the TTPD probe is based on their finding that the "affirmative truth direction" and the "general truth direction" are not aligned. However, according to our findings in Section 4.3, the "affirmative truth direction" and the "general truth direction" become more consistent as the target model's general capability increases. Therefore, for models with relatively high capability, the TTPD probe is not expected to distinguish itself among other probe instantiations.

Second, we find empirical evidence that the TTPD probe is not the most effective probe. We replicate the experiment on the BoolQ dataset in Section 4.6, and plot the output distribution and calibration curves of all probes targeting Llama-3.1-8B under the "with options" setting. The results are shown in Figure 9 and Figure 8. The output distribution of the TTPD probe resembles that of the MM probe, so does the calibration curve. This observation is not restricted to this setting and the BoolQ test set but is also noticed in a number of other settings and test sets. These findings suggest that judging by the functional behaviors of the TTPD probe, its performance is not representative enough to be reported.

C Details of Factual Datasets

We mentioned in Section 3.3 about the factual statements covering six topics which are used for training and testing truthfulness probes. We summarize each topic-specific dataset in Table 1, according to the data curators Azaria and Mitchell (2023), Marks and Tegmark (2023) and Bürger et al. (2024).

Instead of directly using the statements from Bürger et al. (2024), we perform minor modifications on the `inventors` topic and on logical disjunctions. We notice that the original `inventors` dataset has potential ambiguity due to duplication of name. Thus we specify that the person mentioned in a statement was an inventor, e.g. from "Thomas Edison lived in the U.S." to "The inventor Thomas Edison lived in the U.S.". Another tweak is that the original logical disjunctions composed by Bürger et al. (2024) are not consistent with conjunctions, as the subjects are written in full for conjunctions while the subjects are written in pronouns for disjunctions. To align these two logical transformations, we recover the subjects for disjunctions.

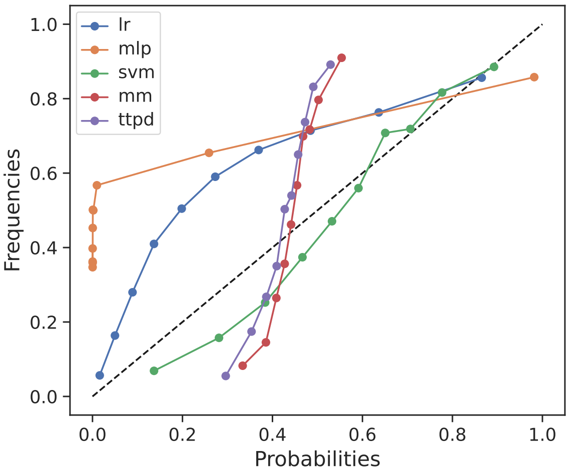


Figure 8: Calibration graph of LR, MLP, SVM, MM and TTPD probes on the BoolQ dataset under the "with options" setting. The target LLM is Llama-3.1-8B.

D Probe Implementation Details

Our implementation of the LR, SVM and MLP probes is based on the `scikit-learn` (Pedregosa et al., 2011) library. For the LR probe, we employ the L-BFGS optimization algorithm (Liu and Nocedal, 1989). For SVM, we utilize the `NuSVC` implementation. We set $\nu = 0.5$, a choice later validated by experiment results. Platt scaling is applied using five-fold cross-validation with the help of the `scikit-learn` library. For the MLP probe we configure a decreasing sequence of hidden units (512,128,64) with `tanh` activation, and we use the Adam optimizer (Kingma, 2014) to train it till convergence. Finally, we use the MM probe implementation provided by Marks and Tegmark (2023).

When establishing probes on atomic statements, we use a random 70% split for training and hold out the rest as the development set.

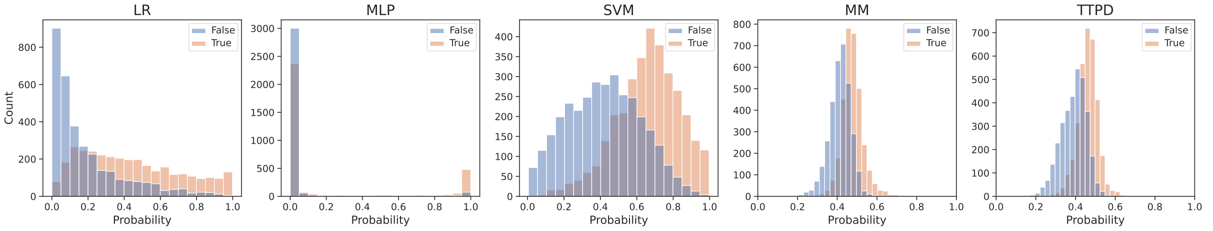


Figure 9: Output distributions of LR, MLP, SVM, MM and TTPD probes on the BoolQ dataset under the "with options" setting. The target LLM is Llama-3.1-8B.

Topic	Description	Example statement
`animal_class`	The class of a specific animal species.	The salmon is a fish.
`cities`	Locations of world cities.	The city of Krasnodar is in Russia.
`element_symb`	Chemical elements and their abbreviations.	Thallium has the symbol Tl.
`facts`	Diverse scientific facts.	The Earth's atmosphere protects us from harmful radiation from the sun.
`inventors`	Home countries of inventors.	The inventor Edwin Herbert Hall lived in the U.S.
`sp_en_trans`	Translations of Spanish words to English.	The Spanish word 'con' means 'to speak'.

Table 1: Summary of topic-specific factual statement datasets.

E Metric Details

E.1 Expected Calibration Error (ECE)

For ECE, we first sort the probabilities predicted by a probe and split them into N equal-sized bins. In this paper we let $N = 10$, which is common in literature evaluating calibration. For each bin, we calculate the mean probability (x_i) and the fraction of truthful predictions (y_i). ECE is then computed following the formula below:

$$\text{ECE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - x_i|. \qquad (1)$$

E.2 Brier Score (BS)

The Brier Score measures the difference between the actual correctness and the confidence score through point-wise mean squared error. Its formulation is as follows:

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2, \qquad (2)$$

where p_i is the confidence reported by the probe and $y_i \in \{0, 1\}$ is the ground truth label. When a predictor is always making inconfident random predictions, i.e. $p_i = 0.5 (i = 1, 2, ...N)$, it results in a chance Brier score of 0.25.

F Experiment Details and More Results

In this section, we elaborate on the detailed setups of the experiments in Section 4. Furthermore, as we only use the Llama family of models in the body of the paper, in this section we also demonstrate results on a model of the Mistral family, Mistral-7B-v0.1 (Jiang et al., 2023).

F.1 Computational and Storage Resources

All of our experiments are completed on three A6000 (48GB) GPUs. For most LLMs except Llama-3.1-70B(-Instruct), only one GPU will suffice. However, the storage for activations across all datasets and models would take ~1TB disk space. Therefore we recommend modifying the code and only keep the necessary activations on disk. Furthermore, in order to accommodate large LLMs such as Llama-3.1-70B(-Instruct) into our GPUs when gathering hidden activations, we use `float8` quantization with the `optimum-quanto`[3] library.

[3] https://github.com/huggingface/optimum-quanto

F.2 Selecting Layer

In Section 4.1 we discuss the selection of decoder layer residual stream to extract truth direction from. We present a criteria based on the ratio of between-class variance to within-class variance. However, due to limitation of space we only present results for Llama-2-7B and Llama-3.1-8B in Section 4.1. Here we replicate these approaches on more models.

The data used for plotting is the collection of both affirmative and negative atomic statements covering all the six topics, as well as their logical conjunctions and disjunctions. The curve for each topic consists of the four variations of statements, which results in six curves for each model.

We summarize the plots in Figure 10, which covers eight models: Llama-2-7B(-Chat), Llama-2-13B(-Chat), Llama-3.1-8B(-Instruct), Llama-3.1-70B(-Instruct), Mistral-7B(-Instruct)-v0.1. Their optimal layers are 12(13), 13(13), 12(13), 33(33), 13(13), respectively.

F.3 Consistency of Truth Direction

We present results on Mistral-7B-v0.1 in Figure 11. It is evident that the truthfulness probes generalize across negation on four topics, barely generalizing on `facts` topic. Comparing these results with those of Figure 3, we notice that the performance of probes for Mistral-7B-v0.1 is comparable to that of probes for Llama-2-13B-Chat and Llama-3.1-8B. This aligns with the observation that the general capability of Mistral-7B-v0.1 lies between that of Llama-2-13B-Chat and Llama-3.1-8B.

F.4 Logical Conjunction/Disjunction

Full results on logical conjunctions and disjunctions are shown in Figure 12 and Figure 13 respectively. A similar scaling trend could be observed as in Figure 3, where the classification accuracy of the probes is positively correlated with the target LLM's general capability.

The results for Mistral-7B-v0.1 is shown in Figure 14. The truthfulness probes generalize from atomic factual statements to both logical conjunctions and disjunctions.

F.5 Question Answering

F.5.1 Details on Experiment Setup

MMLU. We arrange three setups for the QA task, and we demonstrate the prompt template for zero-shot setting using an actual example from the MMLU dataset. The few-shot prompts are trivially extended from the zero-shot prompt, with exemplars separated by two newlines ("\n\n"). Few-shot exemplars are randomly selected from the development split.

```
Question: What was GDP per capita in the
     United States in 1850 when
    adjusting for inflation and PPP in
    2011 prices?
Options:
A. About $300
B. About $3k
C. About $8k
D. About $15k
Answer: B
```

TriviaQA. For TriviaQA (Joshi et al., 2017) we only use few-shot prompting – 5-shot and 20-shot – to ensure that the LLM always generates short-form answers. We use normalized answers in the exemplars.

```
Question: Where in England was Dame Judi
     Dench born?
Answer: york
```

F.5.2 More results

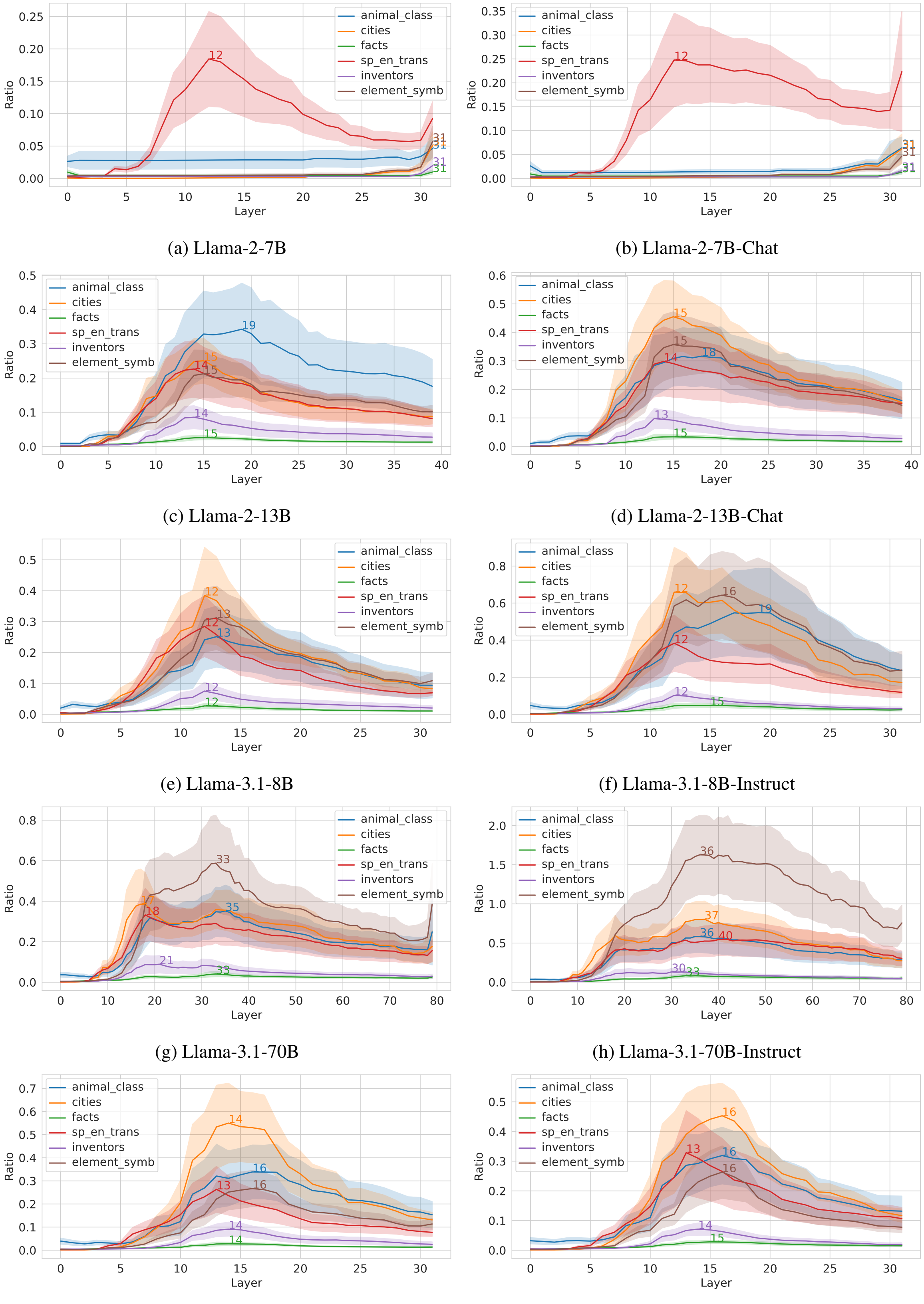
The results for Mistral-7B-v0.1 is shown in Figure 15. The general behavior of truthfulness probes is similar to that in Figure 5. The classification accuracy improves in response to few-shot prompting and improves when more in-context exemplars are provided in the prompt. Meanwhile, calibration only improves in the case of the SVM probe on TriviaQA dataset.

F.6 Contextual Knowledge

F.6.1 Details on Experiment Setup

SciQ. For the SciQ (Welbl et al., 2017) benchmark, we arrange three setups, including "zero-shot", "TTT" and "TTF". We only demonstrate the zero-shot prompt as the few-shot prompts can be trivially extended from it. In the in-context setups, the exemplars are randomly selected from the training split.

```
Context: <context>
Question: Compounds that are capable of
    accepting electrons, such as o 2 or
    f2, are called what?
Options:
A. Oxygen
B. residues
C. antioxidants
D. oxidants
Answer: D
```



(a) Llama-2-7B

(b) Llama-2-7B-Chat

(c) Llama-2-13B

(d) Llama-2-13B-Chat

(e) Llama-3.1-8B

(f) Llama-3.1-8B-Instruct

(g) Llama-3.1-70B

(h) Llama-3.1-70B-Instruct

(i) Mistral-7B-v0.1

(j) Mistral-7B-Instruct-v0.1

Figure 10: Plot of the ratio of between-class variance to within-class variance for a series of models. The shaded regions denote standard error.

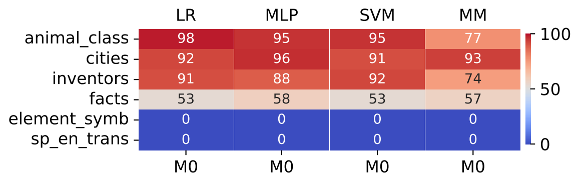


Figure 11: AUROC (in percentage) of probes trained on affirmative statements and tested on negative ones. AUROC exceeding 0.5 indicates generalization success. M0 refers to the Mistral-7B-v0.1 model.

BoolQ. For the BoolQ (Clark et al., 2019) benchmark, we arrange four setups, including "no options", "with options", "T" and "F". We only demonstrate the prompt for "with options". In the in-context setups, the exemplars are randomly selected from the training split.

```
Passage: <passage>
Question: does ethanol take more energy
    make that produces?
Options:
- Yes
- No
Answer: No
```

XSum. For this task, we arrange four setups, including "zero-shot", "T", "TT" and "TTT". We only demonstrate the zero-shot prompt as the one-shot and few-shot prompts can be trivially extended from it. In the in-context setups, the exemplars are randomly selected from the training split of XSum (Narayan et al., 2018) dataset, and they are all deemed correct. False examples come from the XSum Hallucination Annotations (Maynez et al., 2020) dataset with 500 examples, which is paired with examples from the test split of XSum. Furthermore, we filter for examples no longer than the LLM's context window. For Llama-3.1-8B with a context length of 8192, we obtain the final test set of 998 examples whose labels are balanced.

```
Summarize this document: <doc>
Summary: Rory McIlroy moved to within a
    shot of joint leaders Victor
    Dubuisson and Jaco van Zyl after the
     third round of the Turkish Airlines
     Open.
```

F.6.2 More results

We present results for Mistral-7B-v0.1 in Figure 16. Accuracy improves as in-context exemplars are provided, but calibration only displays the same trend on SciQ dataset. Another abnormality could be observed for the BoolQ dataset from "no options" setting to "with options" setting, and for the XSum task from zero-shot to one-shot. In these cases, both accuracy and calibration worsens, which does not align with the results in Figure 7. We assume this is attributed to the weakness of the target model, where Mistral-7B-v0.1 finds it difficult to interpret answer options and in-context exemplars for the abstractive summarization task.

G License

The implementation of the probes is based on the scikit-learn (Pedregosa et al., 2011) library, which is licensed under BSD 3-Clause License. The factual statements we use is curated by Bürger et al. (2024), licensed under MIT License. The MMLU (Hendrycks et al., 2020) dataset is licensed under MIT License, the TriviaQA (Joshi et al., 2017) dataset is licensed under Apache 2.0 License, the SciQ (Welbl et al., 2017) dataset under Creative Commons Attribution-NonCommercial 3.0 Unported License, the BoolQ (Clark et al., 2019) under Creative Commons Share-Alike 3.0 License, the XSum (Narayan et al., 2018) dataset under MIT License and the XSum Hallucination Annotations (Maynez et al., 2020) dataset under Creative Commons Attribution 4.0 International License. Llama-2 series of models are licensed under Llama 2 Community License Agreement, Llama-3 herd of models are licensed under Llama 3 Community License Agreement and Mistral models are licensed under MIT License.

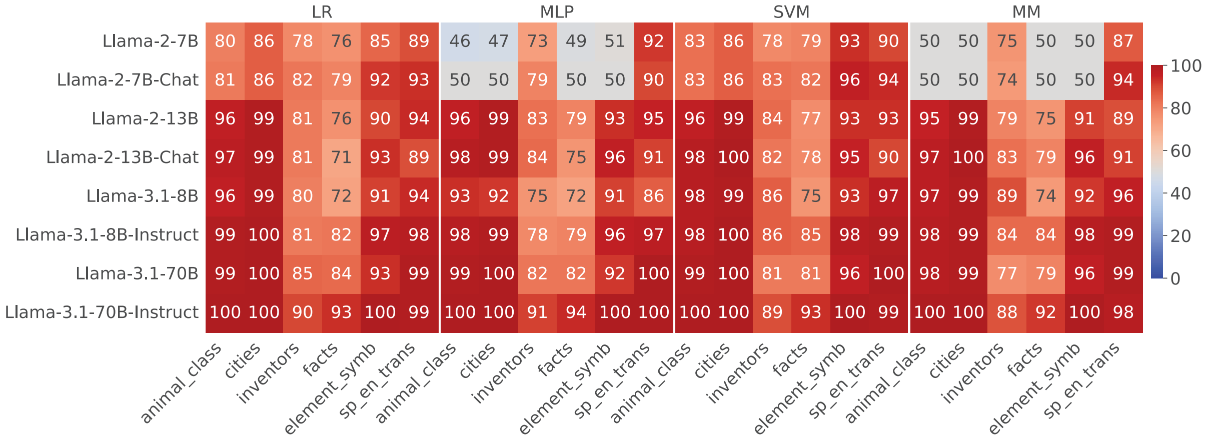


Figure 12: AUROC (in percentage) of probes trained on all the atomic factual statements and tested on logical conjunctions. AUROC > 0.5 indicates the success of generalization.

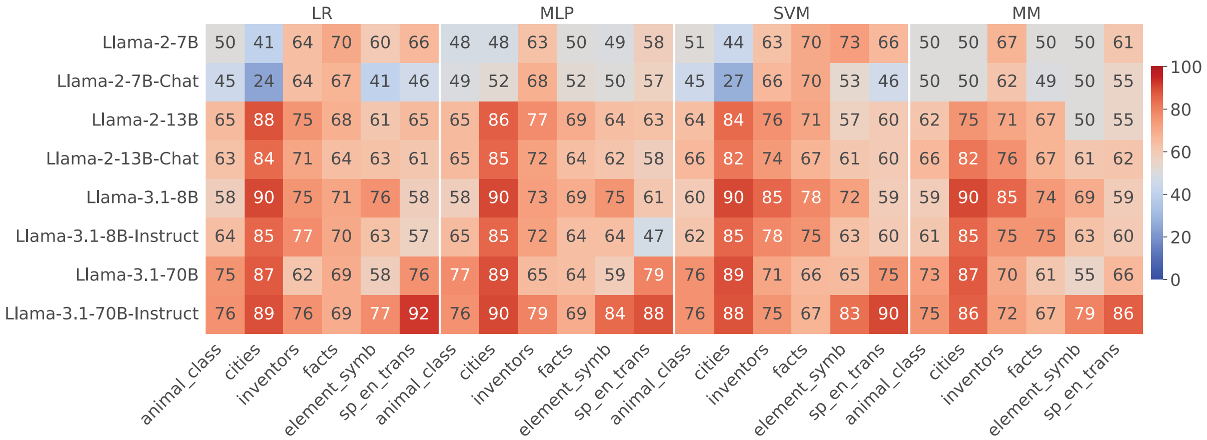
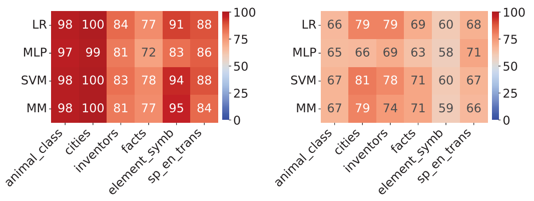
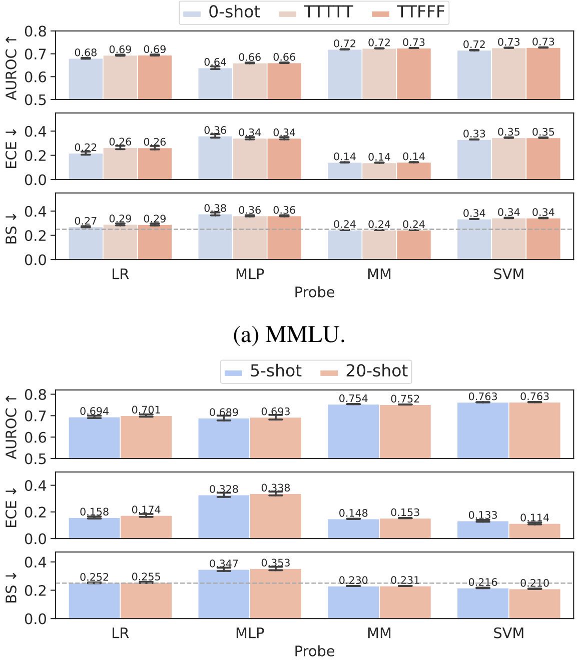


Figure 13: AUROC (in percentage) of probes trained on all the atomic factual statements and tested on logical disjunctions. AUROC > 0.5 indicates the success of generalization.



(a) Logical conjunctions. (b) Logical disjunctions.

Figure 14: AUROC (in percentage) of probes trained on atomic factual statements and tested on logical conjunctions/disjunctions for Mistral-7B-v0.1. AUROC > 0.5 indicates the success of generalization.



(a) MMLU.

(b) TriviaQA.

Figure 15: AUROC↑/ECE↓/BS↓ of truthfulness probes for Mistral-7B-v0.1 on MMLU and TriviaQA. The dashed gray line corresponds to random results, and error bars denote standard error.

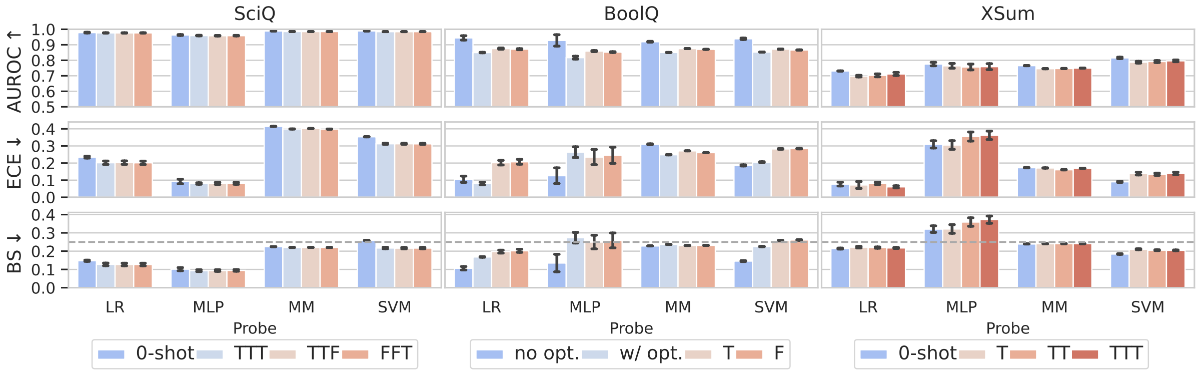


Figure 16: AUROC↑/ECE↓/BS↓ of truthfulness probes for Mistral-7B-v0.1 on tasks where grounding knowledge is provided in the prompt. The dashed gray line corresponds to random results, and error bars denote standard error.