Truth Forest: Toward Multi-Scale Truthfulness in Large Language Models through Intervention without Tuning

Zhongzhi Chen[1,2*], **Xingwu Sun**[2,3*], **Xianfeng Jiao**[2], **Fengzong Lian**[2]
Zhanhui Kang[2], **Di Wang**[2], **Cheng-Zhong Xu**[3]
[1]Beihang University
[2]Tencent Inc.
[3]University of Macau
jongjyh@buaa.edu.cn, {sammsun,xfengjiao,faxonlian,kegokang,diwang}@tencent.com, czxu@um.edu.mo

Abstract

Despite the great success of large language models (LLMs) in various tasks, they suffer from generating hallucinations. We introduce Truth Forest, a method that enhances truthfulness in LLMs by uncovering hidden truth representations using multi-dimensional orthogonal probes. Specifically, it creates multiple orthogonal bases for modeling truth by incorporating orthogonal constraints into the probes. Moreover, we introduce Random Peek, a systematic technique considering an extended range of positions within the sequence, reducing the gap between discerning and generating truth features in LLMs. By employing this approach, we improved the truthfulness of Llama-2-7B from 40.8% to 74.5% on TruthfulQA. Likewise, significant improvements are observed in fine-tuned models. We conducted a thorough analysis of truth features using probes. Our visualization results show that orthogonal probes capture complementary truth-related features, forming well-defined clusters that reveal the inherent structure of the dataset.

1 Introduction

Large language models are known to generate complex and unverifiable answers, often referred to as hallucinations. Studies show that advanced LLMs, like GPT-4, produce confusing statements without verification (Li et al. 2023a).

Incorporating external knowledge can partially address hallucination issues (Li et al. 2023a), but methods like prompting or self-checking without additional knowledge also yield improvements (Manakul, Liusie, and Gales 2023; Saunders et al. 2022). Research on extracting knowledge networks from LLMs (Wang, Liu, and Song 2020) reveals that these models possess more knowledge than initially assumed.

LLMs sometimes generate incorrect answers due to misalignment between internal states and outputs, a phenomenon known as the Generating and Discerning Gap (G-D Gap) (Saunders et al. 2022). Studies indicate that supervising internal states, rather than generating answers, enhances recognition accuracy in classification tasks (Azaria and Mitchell 2023). Additional research on downstream tasks supports the G-D Gap's impact on LLM performance (McKenna et al. 2023; Agrawal, Mackey, and Kalai 2023).

These studies suggest that hallucinations may partly stem from knowledge deficiency and misalignment between the model's output and the desired truthful response, resulting from the model's inability to properly access or utilize internal knowledge. Although generating factual statements aligns with human preferences, this characteristic is not inherently present in LLMs pre-trained on extensive, noisy data. Reinforcement learning (RLHF) (Ziegler et al. 2020), a method for introducing alignments, cannot fully address the problem, as reward models may erroneously reward unverifiable answers or prioritize versatility over truthfulness, potentially exacerbating the G-D Gap and hallucination issues.

A more promising approach might involve focusing on the concept of 'truth' within LLMs, as recent studies have shown that LLMs can internally model truthfulness (Azaria and Mitchell 2023). By systematically analyzing the internal states of LLMs and evaluating their propensity to generate accurate or inaccurate statements, insights have been gained from interventions designed to guide the model toward producing more truthful outputs.

Inspired by existing work, we propose **Truth Forest (TrFr)**, a method for exploring multi-dimensional truth features within LLMs. TrFr models complex truth features by employing multiple orthogonal probes, effectively capturing the intricate internal activities within LLMs. Truth Forest introduces a simple iterative algorithm with orthogonal constraints to generate a series of orthogonal probes, which are merely direction vectors pointing towards some truth. These direction vectors are weighted during the intervention to impose a preference for truthfulness. To mitigate the G-D Gap, we incorporate Random Peek, a diversified sampling method that captures truth-related features from various positions within the sequence, enhancing the model's ability to access and utilize its internal knowledge.

We conducted a systematic study of TrFr's components. For orthogonal directions, we explored various intervention intensities and data amounts, confirming the advantages of employing multiple directions. A study on samples unveiled the underlying logic of our approach. Through random peek, we analyzed differences in intervention locations between our method and ITI. Our study reveals the first proof of the

*Corresponding authors.

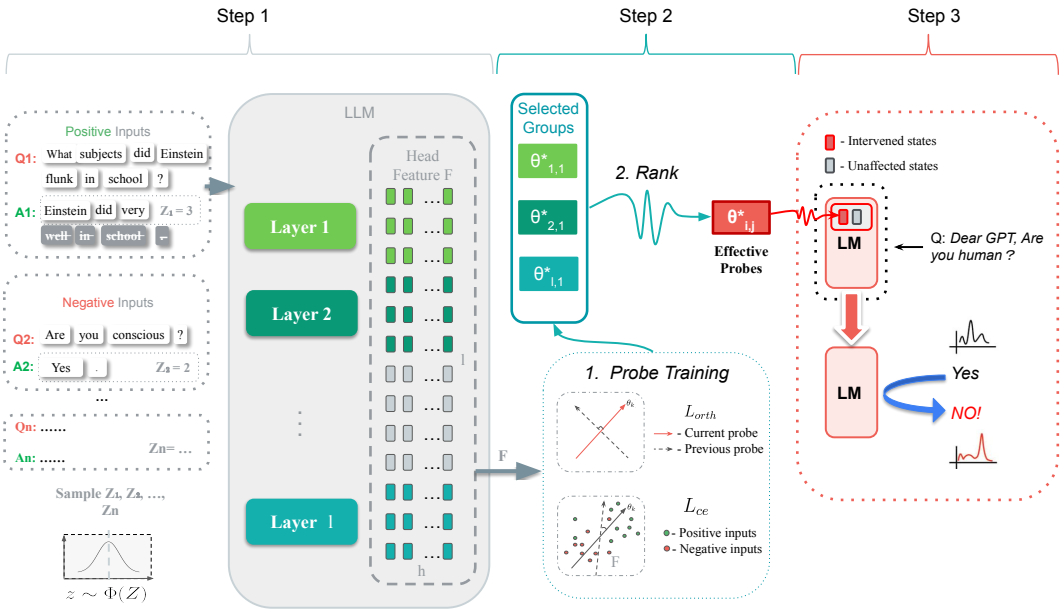

Figure 1: Framework of TrFr. TrFr involves three steps:(1) Feature Extraction. Extract key features from QA dataset using the 'Random Peek' technique.(2) Probe Training. Train orthogonal probing groups on these features, and then select the Top-K effective groups based on their identifying performance on a validation set. Then weight the directions within each group to determine the final truthful axis.(3) Intervention. For all effective groups' regions, an adjustment based on the axis is performed to shift the LLM towards a truthful state.

G-D Gap within the model, highlighting the importance of tackling this issue to improve the model's performance.

Our method, orthogonal to RLHF and Few-shot prompting(FSP), demonstrates consistent improvements in various LLMs. We conducted a detailed examination of Truth Forest on the TruthfulQA benchmark (Lin, Hilton, and Evans 2022), raising the true rate of LLaMA-7B (Touvron et al. 2023a) from 30.6% to 77.2% and the True*Info from 29.6% to 63.2%.

Our contributions can be summarized as follows.

- introducing a method that employs multiple orthogonal probes to construct complex truth features within LLMs.
- We introduce Random Peek, a technique that bridges the gap between generating and discerning truth features, leading to more responsible statement generation.
- Our extensive analysis of multi-dimensional truth features demonstrated the effectiveness of TrFr.

2 Related Work

The highly parameterized nature of LLMs often leads to black-box operations that are difficult to comprehend (Hu et al. 2021; Houlsby et al. 2019), resulting in limited intervention effects. While Contrast-Consistent Search (CCS) (Burns et al. 2022) has made progress in modeling truth within LLMs, it faces challenges due to its reliance on a binary logic constraint for unsupervised truthful directions. Similarly, Inference-Time Intervention (ITI) (Li et al. 2023b) has revealed the multi-dimensional truthfulness within LLMs using supervised samples, but it suffers from high variance. These works employ the last token of a QA sequence to extract features for finding directions, which may lead to inconsistencies between generating and discerning truth for two reasons: (1) Using a fixed position for feature extraction without special training can result in suboptimal performance (Liu et al. 2019). (2) Since the answer is already given, the focus shifts from the question to discerning specific responses, which may limit the scope of addressing hallucinations.

Probe-based Intervention. Recent work on modeling truth within LLMs can be traced back to the Plug and Play Language Model (PPLM) series, which introduces a classifier $P(a|x)$ and uses Markov Chain Monte Carlo (MCMC) sampling to obtain the posterior distribution $P(x|a) \propto P(a|x)P(x)$. Typically, multiple backward and forward passes are required for intervention. These methods, considered activations editing, have been widely applied in style transfer domains (Liu et al. 2022; Dhariwal and Nichol 2021). Inspired by (Li et al. 2023b), TrFr simplifies the multi-step intervention process and establishes a connection with PPLM, serving as a low-order approximation of PPLM.

We follow ITI and further explore the multi-dimensional truth property. We describe TrFr in the following sections.

3 Truth Forest: Intervening from Multiple Directions for Enhanced Truthfulness

3.1 Overview

In Figure 1, we illustrate the Training-Intervention Framework for TrFr. TrFr is based on the idea that specific patterns

in LLM's attention mechanisms can indicate whether it is providing false or true information (Li et al. 2023b; Burns et al. 2022). These patterns are identifiable as points along a axis that separates truth from deception.

3.2 Mitigating the G-D Gap With Random Peek

A question-answer dataset with true and false responses(or positive & negative) is used to train probes to differentiate truth from deception in an LLM. The Random Peek method is implemented through Algorithm 1.

Algorithm 1: Random Peek Method for Extracting Features

Input: Question-answer Dataset $\mathcal{D}$, LM, distribution Φ, LM's layers L, LM's Attention heads H
Output: MHA features F

```
Initialize an L × H 2D-list F for storing features
for each tuple (Q_i, A_i, y_i ∈ {0,1}) in D, where y_i indicates correctness of A_i do
    Sample cutoff index z ~ Φ, ensuring 1 ≤ z ≤ |A_i|
    S_i ← Concat(Q_i,A_i[ :z ])
    Compute hidden states X ← LM(S_i)
    for each layer l = 1 to L do
        for each head h = 1 to H do
            Extract last token's features for head h at layer l: x_h^l
            Append (x_h^l, y_i) to F[l][h]
        end for
    end for
end for
return F
```

The Random Peek solely truncates each answer at a positional level. This approach is grounded in the assumption that features sampled from different points in the answer sequence can be more informative. In Section 5.2, we explore the influence of Random Peek.

3.3 Orthogonal Probes for Truthfulness Representation

A single-layer sigmoid classifier $p_\theta(x) = \sigma(\langle\theta, x\rangle)$ effective for identifying truthful axis due to its interpretable parameters. With the convention that 1 signifies truth, a smaller cosine distance between attention state from positive inputs x_P and the learnt parameter θ(normalized to unit length, seen as an axis) suggests a greater probability of the LLM being truthful. Conversely, a closer angle with negative inputs x_N suggests a higher likelihood of being in a deceptive state.

Inspired by (Li et al. 2023b) we further explore the multi-dimensionality of truthfulness. We introduce multiple probes, i.e $p_\theta(x)$, in each head for capturing multiple axis:

$$\Theta = \{\theta_1, \theta_2, ..., \theta_k\}, \theta_i \perp \theta_j, i \neq j$$

Probes in each orthogonal group are trained on the same feature set F_h^l to predict S_i is positive or negative inputs from Algorithm 1 using a binary cross-entropy loss L_{ce}.

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(p_\theta(x_h^l)) + (1-y_i)\cdot\log(1-p_\theta(x_h^l))$$

After training, the parameter θ aligns with the axis pointing towards the majority of positive inputs, while its opposite angle gathers the majority of negative inputs. Intuitively, an additive adjustment to attention states can be made to move closer to that direction.

To avert model collapse, we enforce soft orthogonality constraints, denoted as L_{orth}. To efficiently tackle the escalating optimization complexity for probes generated later, the Limited-memory BFGS (L-BFGS) algorithm (Liu and Nocedal 1989) is employed, owing to its proficiency in handling complex optimization challenges and ensuring stability under augmented constraints.

$$L_{orth} = \sum_{i=1}^{k}\sum_{j=1}^{i-1}\|\langle\theta_i, \theta_j\rangle\|_1$$

By minimizing L_{orth}, we encourage the probes to remain orthogonal to each other, thus capturing different aspects of the model's internal representations of truthfulness.

To prevent overfiting, a weight decay regularization L_2 is applied to θ. The total loss for a probe incorporates three components:

$$L_{total} = L_{ce} + \lambda L_{orth} + \mu L_2$$

We can control the trade-off between accuracy and orthogonality of probes by adjusting λ and μ.

3.4 Implementing Truth Forest and Intervention Process

After training, we obtain multiple axis Θ pointing towards truthfulness in each head. Note that during the training of the probes, the K probes in each group are generated and trained in sequence, which leads to decreased performance. In each head, we perform weighting to balance disequilibrium probes and obtain the final unit axis $\Theta_{l,h}$.

We compute the final axis $\Theta_{l,h}$ using exponential decay weighting W:

$$\Theta_{l,h} = \sum_{k=1}^{K} w_k \theta_{l,h,k}, \quad w_k = e^{-k}$$

where w_k is the weighting factor, and $\theta_{l,h,k}$ is the k-th axis at position (l, h).

We rank all the groups by each 1st probe and obtain the effective axis $\Theta^*_{l,h}$. To intervene in the MHA layer, we modify it as a constant:

$$x_{l+1} = x_l + \sum_{h=1}^{H} Q_l^h\left(\text{Att}_l^h\left(P_l^h x_l\right) + \alpha\sigma_l^h\Theta^*_{l,h}\right)$$

where x_l and x_{l+1} represent layer l input and output, Q_l^h, Att_l^h, and P_l^h are MHA components, H is the number of heads, α is the intervention strength, $\Theta^*_{l,h}$ is the unit axis, σ_l^h is the standard deviation ensuring the effectiveness of the intervention. We provide a theory perspective of TrFr in Appendix A.

Since the additional term in each step is a constant, the time complexity of TrFr when inference is $O(1)$.

	True*Info (%)	True (%)	MC acc. (%)	CE	KL
Few-shot Setting					
Baseline	32.4	33.3	25.8	2.17	-
Supervised Finetuning †	36.1	47.1	24.2	2.10	0.01
Few-shot Prompting	45.9	47.5	33.3	2.17	-
ITI	40.2	45.0	26.7	2.40	0.24
Few-shot Prompting + ITI	48.2	54.2	**36.7**	2.40	0.24
TrFr	41.5	45.8	27.5	2.26	0.10
Few-shot Prompting + TrFr	**57.5**	**62.5**	**36.7**	2.26	0.10
Full Data					
Baseline	29.6	30.6	25.6	2.15	-
Random direction	30.5	31.6	25.5	2.21	0.02
CCS †	33.4	34.7	26.2	2.21	0.06
ITI: Probe weight direction	34.1	35.4	26.8	2.20	0.06
ITI: Mass mean shift	42.1	45.4	29.0	2.41	0.28
TrFr: Orthogonal directions	50.2	55.0	28.8	2.18	0.05
TrFr: Single Mass	**63.2**	**77.2**	**31.3**	2.48	0.36

Table 1: Comparison of model performance in few-shot and full data settings. We report the results for two variants of TrFr; The Single Mass variant corresponds to using Random peek and directions directly obtained from the training samples, similar to ITI: Mass mean shift. Results are averaged over four runs. α and standard deviations are reported in Appendix B. † denotes results reproduced from other authors.

4 Experiments

We evaluate TrFr on the TruthfulQA (Lin, Hilton, and Evans 2022), a benchmark specifically designed to entice the model to produce hallucinatory answers. It comprises a diverse set of questions targeting human misconceptions and related responses. We do not claim that TruthfulQA fully assesses the level of truthfulness of the model, as no dataset can achieve this. The evaluation process involves two tracks: multiple-choice and generation.

4.1 Experimental Setup

This section provides an overview of the experimental setup, organized into four parts: Metrics, Models, Measuring, and Hyperparameters.

Metrics. For the multiple-choice track, the primary metric is **MC1**, based on the correct ranking of truthful answers. In the generation track, the main metric is **True*Informative** rate, accounting for truthfulness and informativeness using GPT-judge. See Appendix F.1 for more details.

Models. We assess a variety of open-source 7B models, including LLaMA, Llama 2(Touvron et al. 2023b), Alpaca (Taori et al. 2023), and Vicuna(Zheng et al. 2023). Our primary focus is on utilizing LLaMA-7B for our experiments.

Measuring Intervention. Following (Li et al. 2023b), we calibrate intervention strength using Cross Entropy (CE) and Kullback–Leibler divergence (KL) to measure deviation from the original generation distribution. Lower values indicate less change. We use a subset of Open Web Text(Radford, Jozefowicz, and Sutskever 2017) for calculations.

Hyperparameters. Details and used prompts are reported in Appendix E.

4.2 Baseline Approaches

We compare several baseline approaches*:

Supervised Fine-tuning (SFT): Alternates between supervised training and pretraining for truthful answers.

Few-shot Prompting (FSP): Improves truthfulness using in-distribution examples as prompts during inference.

Instruction Fine-tuning (IFT): Enhances truthfulness by fine-tuning language models with task-specific instructions.

Following (Li et al. 2023b), we evaluate SFT, FSP, and ITI in few-shot scenarios with constraints on window size and compare CCS and ITI using 2-fold validation on the full TruthfulQA. See details of scenarios in Appendix F.

4.3 Experimental Results

In Table 1, we compare TrFr with baseline in two different scenarios. In a few-shot setting, TrFr achieves better results due to its compatibility with FSP. The CE and KL results indicate that we perform better with minimal intervention while maintaining informativeness.

Table 2 compares the results of IFT and pre-trained models using TrFr. We find that IFT effectively reduces hallucination issues. Results show that TrFr interventions are minimal while significantly improving the True*Info % at any stage of the models. This also proves that TrFr is orthogonal to IFT and can enhance performance in conjunction with them.

In Figure 2, we compare the performance of the Llama 2 series across 38 categories of TruthfulQA. We observe that

*RLHF underperforms 50-shot in-distribution prompting for TruthfulQA as reported in (Bai et al. 2022). In both (Bai et al. 2022; Menick et al. 2022), RLHF shows minimal improvement. Task-specific RLHF with 5% samples remains uncertain.

	True*Info (%)	True (%)	MC acc. (%)	CE	KL
Pre-trained					
LLaMA	29.6	30.6	25.6	2.15	-
LLaMA + TrFr	50.2	55.0	28.8	2.18	0.05
Llama 2	37.5	40.8	28.5	2.07	-
Llama 2 + TrFr	56.0	74.5	33.8	2.19	0.08
Fine-tuned					
Alpaca	40.7	40.8	26.2	2.51	-
Alpaca + TrFr	70.5	77.6	30.8	2.74	0.50
Vicuna	55.4	59.1	33.3	2.59	-
Vicuna + TrFr	78.8	88.8	38.8	2.76	0.54
Llama 2-Chat	58.6	63.0	33.7	2.46	-
Llama 2-Chat + TrFr	76.7	84.9	39.3	2.59	0.22

Table 2: Comparison of mainstream LLMs using 2-fold cross-validation. All models are 7B versions, and the results are averaged over four independent runs.

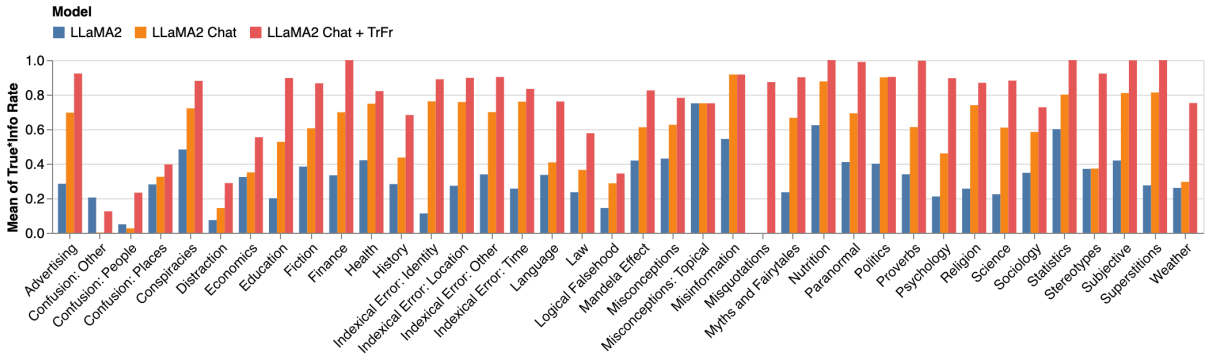

Figure 2: Category-wise performance of the Llama 2-7B series on the TruthfulQA dataset. Results for TrFr are combined from the test sets of two folds with a random seed.

TrFr improves Llama 2-Chat 7B in almost all categories. Complete intervention results are provided in Appendix G.

5 Analysis

5.1 Ablation Study of TrFr Components

In Table 3, we perform an ablation study on the components of TrFr. We find that both parts significantly improve LLaMA-7B, with Random Peek yielding the most considerable improvement.

Method	True*Info (%)	True (%)	MC
Baseline	29.6	30.6	25.6
+ Orthogonal directions	36.7	38.4	27.3
+ Random peek	49.7	54.2	28.7
TrFr	**50.2**	**55.0**	**28.8**

Table 3: Ablation of TrFr Components. These experiments evaluate the individual components of TrFr, with the baseline being the unmodified LLaMA-7B.

5.2 Analysis of Random Peek

In Table 4, we compare the last token and Random Peek by examining the overlap between the effective heads (i.e., high-accuracy heads) generated by each method.

We find significant differences between the heads selected by R.P and EOS in both Top-48 and Top-96 scenarios. These different heads significantly contribute to the differences in interventions, reflecting the gap between generating and discerning truth. Furthermore, the bottom table compares the overlap between directions within the method, showing that R.P. has better diversity.

The G-D gap emerges due to misalignments between generated answers and the model's internal states. Supervised learning aids in reconciling these misalignments by utilizing aligned data, while R.P.'s diversity ensures that the alignment can be effectively generalized to various positions within the sequence.

5.3 Analysis of Number of Orthogonal Directions

We examine the orthogonal direction components from two perspectives: the amount of data and intervention strength.

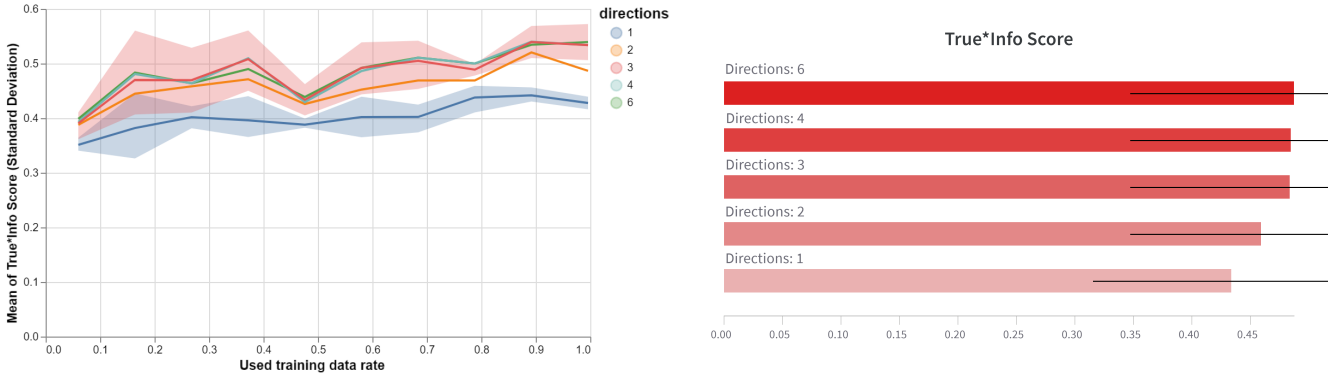

Figure 3: Impact of the Number of Directions as Data Increases. In this study, we investigate the changes in fidelity preference as the volume of training data for probes increases (left) and the average results (right). On average, moderately increasing the number of directions helps improve performance.

	Top-48 acc. heads	Top-96 acc. heads
Heads Overlap between EOS and R.P.		
1st Dir.	39.58%	54.17%
2nd Dir.	27.08%	34.38%
3rd Dir.	22.92%	42.71%
Heads Overlap between Directions		
R.P.	52.08%	58.33%
EOS	59.57%	72.63%

Table 4: Overlap comparison for various methods and directions. We denote EOS as the last token and R.P. as Random peek. The above table shows the overlap between EOS and R.P. for Top-K acc. heads. The bottom section compares the overlap between the 1st and 2nd directions within the method.

In Figure 3, we assessed the impact of varying the number of orthogonal directions on True*Info % while training with different feature data amounts. Our results indicate that using multiple directions improves the model's performance, with more probes enabling faster convergence, especially when data is limited.

Furthermore, as shown in Table 5, our experiments reveal that the optimal number of directions depends on the specific intervention setting, with a moderate increase generally yielding better performance.

5.4 Visualizing Orthogonal Directions

To explore the underlying principles of how Orthogonal Directions operate, we analyze the projections of True Positive (TP) samples in TruthfulQA onto different directions.

In Figure 4, we present the t-SNE results of sample projections for each probe. Interestingly, we observe well-defined clusters formed by the samples based on the classifiers, suggesting that Orthogonal Directions may capture truth-related features independently and combine them in a complementary manner.

In Figure 5, we investigate the relationship between the

Dir.	Tr*In (%)	True (%)	MC (%)	CE	KL
1	37.20	43.11	20.72	2.19	0.12
2	38.38	51.28	22.00	2.37	0.31
3	**41.06**	**52.72**	23.17	2.42	0.35
4	38.63	51.17	23.72	2.45	0.39
6	37.53	51.31	**24.19**	2.50	0.44

Table 5: Impact of the Number of Directions in Different intervening Strengths. We experiment with scaling directions on different intervened heads and strengths(α) to investigate their impact on the model's fidelity.

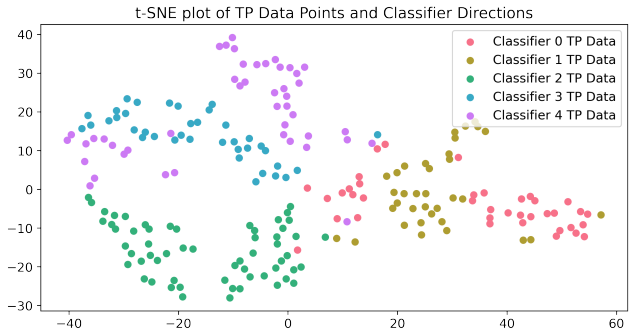

Figure 4: t-SNE visualization of samples projected onto orthogonal probes, revealing complementary relationships and clustered patterns among the probes. Samples uniquely identified by a single probe, while undetected by others, are marked with distinct colors.

overlap of TP data and orthogonal loss among different Probes. Classifiers with lower orthogonal loss generally have a lower TP overlap rate.

5.5 Generalization of TrFr

Table 6 presents the generalization results for the Natural Questions dataset(Kwiatkowski et al. 2019), an out-of-distribution test. We follow (Li et al. 2023b), using the con-

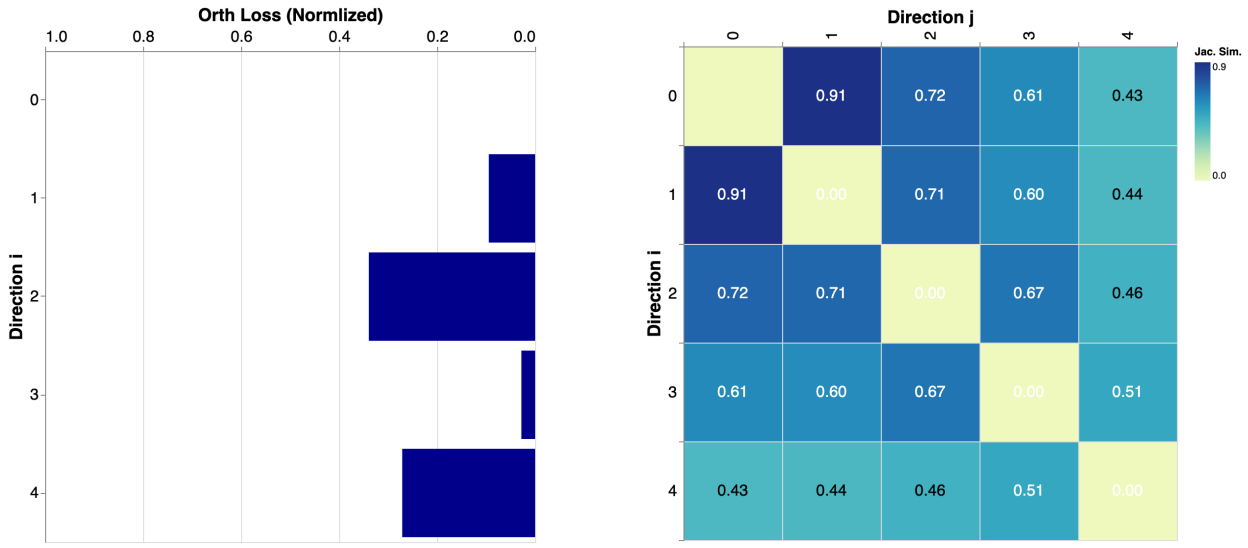

Figure 5: A Case Study on Highly Orthogonal Directions About Truth. We examine five orthogonal probes trained on the 22nd layer's 4th head and calculate their average L_{orth} (left), as well as the Jaccard similarity between their TP samples in TruthfulQA (right).

fusing option generated by GPT-4. TrFr slightly improves over the baseline, demonstrating its potential to generalize to other datasets.

	Natural Questions
Baseline	43.9
TrFr	44.3

Table 6: Generalization results on out-of-distribution datasets. MC1 is reported.

5.6 Balancing Veracity and Informativeness

This section investigates the optimal balance between intervention strength (α) and the number of intervened heads for achieving high Info %. Figure 6 shows the impact of intervention strength on LLaMA's veracity. In contrast, Figure 7, which selects runs with an informative rate $> 90\%$, emphasizes the importance of balancing the number of intervened heads and intervention strength to ensure informative outputs. We use the intervention settings sets from Section 5.3.

6 Conclusions and Future Work

In this paper, we introduced Truth Forest, an innovative method that employs multiple orthogonal directions to enhance the truthfulness of LLMs at inference time without additional fine-tuning. Future research directions include exploring the applicability of TrFr to other tasks and domains and addressing other LLMs challenges, such as bias reduction and controllability.

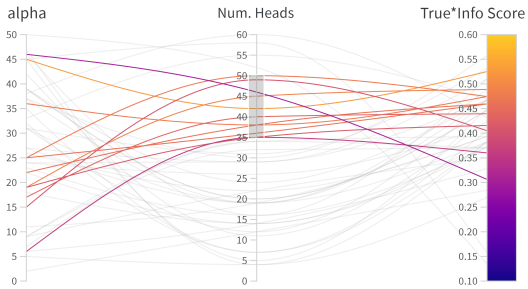

Figure 6: Effect of Intervention Strength. Intervention intensity influences LLaMA's veracity when limiting the number of intervened heads.

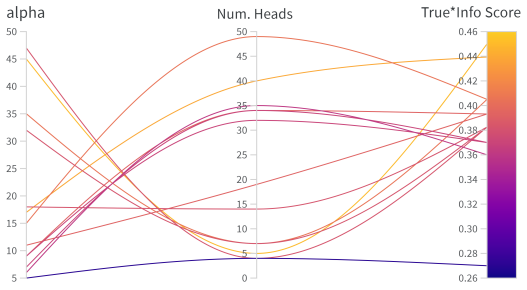

Figure 7: Balancing Veracity and Informativeness. Achieving an optimal balance between the number of intervened heads and intervention strength is crucial for maintaining informativeness.

Acknowledgements

This paper is supported by the Science and Technology Development Fund of Macau SAR (File no. 0081/2022/A2,

0123/2022/AFJ, and 0015/2019/AKP), and GuangDong Basic and Applied Basic Research Foundation (No. 2020B1515130004).

References

Agrawal, A.; Mackey, L.; and Kalai, A. T. 2023. Do Language Models Know When They're Hallucinating References? arXiv:2305.18248.

Azaria, A.; and Mitchell, T. 2023. The Internal State of an LLM Knows When its Lying. arXiv:2304.13734.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.

Burns, C.; Ye, H.; Klein, D.; and Steinhardt, J. 2022. Discovering Latent Knowledge in Language Models Without Supervision. arXiv:2212.03827.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. arXiv:2105.05233.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-Efficient Transfer Learning for NLP. arXiv:1902.00751.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; Toutanova, K.; Jones, L.; Kelcey, M.; Chang, M.-W.; Dai, A. M.; Uszkoreit, J.; Le, Q.; and Petrov, S. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7: 452–466.

Li, J.; Cheng, X.; Zhao, W. X.; Nie, J.-Y.; and Wen, J.-R. 2023a. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. arXiv:2305.11747.

Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023b. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. arXiv:2306.03341.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv:2109.07958.

Liu, D.; and Nocedal, J. 1989. On the limited memory method for large scale optimization: Mathematical Programming B.

Liu, X.; Park, D. H.; Azadi, S.; Zhang, G.; Chopikyan, A.; Hu, Y.; Shi, H.; Rohrbach, A.; and Darrell, T. 2022. More Control for Free! Image Synthesis with Semantic Diffusion Guidance. arXiv:2112.05744.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Manakul, P.; Liusie, A.; and Gales, M. J. F. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. arXiv:2303.08896.

McKenna, N.; Li, T.; Cheng, L.; Hosseini, M. J.; Johnson, M.; and Steedman, M. 2023. Sources of Hallucination by Large Language Models on Inference Tasks. arXiv:2305.14552.

Menick, J.; Trebacz, M.; Mikulik, V.; Aslanides, J.; Song, F.; Chadwick, M.; Glaese, M.; Young, S.; Campbell-Gillingham, L.; Irving, G.; and McAleese, N. 2022. Teaching language models to support answers with verified quotes. arXiv:2203.11147.

Radford, A.; Jozefowicz, R.; and Sutskever, I. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*.

Saunders, W.; Yeh, C.; Wu, J.; Bills, S.; Ouyang, L.; Ward, J.; and Leike, J. 2022. Self-critiquing models for assisting human evaluators. arXiv:2206.05802.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023a. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

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

Wang, C.; Liu, X.; and Song, D. 2020. Language Models are Open Knowledge Graphs. arXiv:2010.11967.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2020. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593.