Towards Tracing Trustworthiness Dynamics: Revisiting Pre-training Period of Large Language Models

Anonymous ACL submission

Abstract

Ensuring the trustworthiness of large language models (LLMs) is crucial. Most studies concentrate on fully pre-trained LLMs to better understand and improve LLMs' trustworthiness. In this paper, to reveal the untapped potential of pre-training, we pioneer the exploration of LLMs' trustworthiness during this period, focusing on five key dimensions: reliability, privacy, toxicity, fairness, and robustness. To begin with, we apply linear probing to LLMs. The high probing accuracy suggests that *LLMs in early pre-training can already distinguish concepts in each trustworthiness dimension*. Therefore, to further uncover the hidden possibilities of pre-training, we extract steering vectors from a LLM's pre-training checkpoints to enhance the LLM's trustworthiness. Finally, inspired by Choi et al. (2023) that mutual information estimation is bounded by linear probing accuracy, we also probe LLMs with mutual information to investigate the dynamics of trustworthiness during pre-training. We are the first to observe a similar two-phase phenomenon: fitting and compression (Shwartz-Ziv and Tishby, 2017). This research provides an initial exploration of trustworthiness modeling during LLM pre-training, seeking to unveil new insights and spur further developments in the field.

1 Introduction

As the capabilities of LLMs increase, their trustworthiness becomes a focal point of widespread attention. Guided by global AI governance (Commission, 2021b; Tabassi, 2023; Newman, 2023) and trustworthy AI (Commission et al., 2019; Liu et al., 2023b), trustworthy LLMs have developed some common categories, especially focusing on five dimensions: reliability, toxicity, privacy, fairness, and robustness (Wang et al., 2023a; Sun et al., 2024). Delving into LLMs across all these trustworthiness dimensions is essential for the society.

To seek a deeper exploration of language models, one of the prominent methods is probing (Zhao et al., 2023; Räuker et al., 2023), which involves training a classifier on the model's representations to identify linguistic and semantic properties acquired by the model (Tenney et al., 2019; Pimentel et al., 2020; Li et al., 2021; Belinkov, 2022; Räuker et al., 2023; Gurnee and Tegmark, 2023; Slobodkin et al., 2023). In particular, considering trustworthiness, recent attempts reveal that LLM representations contain linearly separable patterns (Zou et al., 2023; Li et al., 2023; Azaria and Mitchell, 2023). Unfortunately, existing research has largely focused on fully pre-trained LLMs (Touvron et al., 2023a), including those aligned (Ouyang et al., 2022) through Supervised Fine-Tuning (SFT) or Reinforcement Learning from Human Feedback (RLHF). This perspective neglects the pre-training period in the context of LLM trustworthiness. To our best knowledge, two aspects still remain mysteries: 1) how LLMs dynamically encode trustworthiness during pre-training, and 2) how to harness the pre-training period for more trustworthy LLMs.

To address the above issues, we start by analyzing the pre-training dynamics about trustworthiness of LLM. More specifically, we use linear probing (Alain and Bengio, 2016; Belinkov, 2022) across the 360 pre-training checkpoints from LLM360 (Liu et al., 2023e) to explore five dimensions of trustworthiness: reliability, toxicity, privacy, fairness, and robustness. Our probing results suggest that *after the early pre-training period, middle layer representations of LLMs have already developed linearly separable patterns about trustworthiness*. Such patterns are capable of discerning opposing concepts within each trustworthiness dimension (e.g., discriminating true and false statements). Building upon the above observations, we raise an intriguing question: *can the pre-training period of an LLM be utilized to enhance its trustworthiness after pre-training?*

We provide insightful answers to the above question by exploring the potential of pre-training

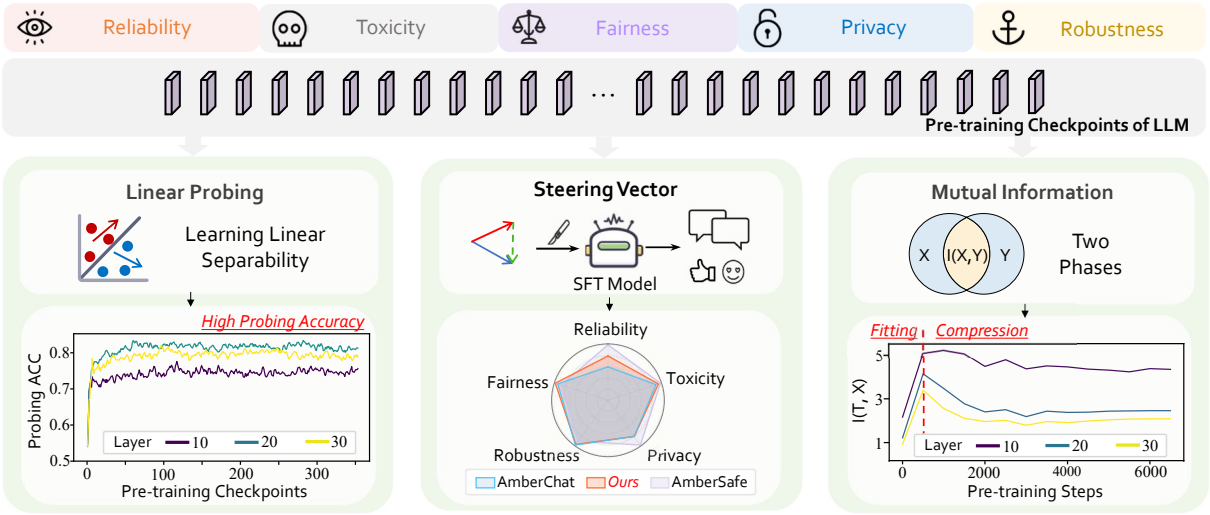

Figure 1: Overview of tracing trustworthiness dynamics during pre-training. 1) Linear probing identifies linearly separable opposing concepts during early pre-training; 2) Steering vectors are developed to enhance LLMs' trustworthiness; 3) Probing LLMs with mutual information reveals a two-phase trend regarding trustworthiness.

checkpoints for better trustworthiness. Notably, recent advancements have introduced "activation intervention," a novel suite of techniques for directing language models towards enhanced LLMs' performance by adjusting activations during inference (Turner et al., 2023; Li et al., 2023; Rimsky et al., 2023; Wang and Shu, 2023). Inspired by these works and the observation of linearly separable patterns in trustworthiness concepts during the LLM's pre-training period, we make preliminary attempts to extract steering vectors from LLM's checkpoints during pre-training, employing them to intervene in the SFT model for trustworthiness enhancement. Extensive experiments reveal that *these steering vectors extracted from pre-training checkpoints could promisingly enhance the SFT model's trustworthiness*. More crucially, these steering vectors achieve a trustworthiness performance that matches or promisingly exceeds that of vectors extracted directly from the SFT model itself. Our findings introduce novel insights into using pre-training checkpoints for LLM alignment, revealing untapped potential and offering a fresh perspective on enhancing LLM trustworthiness.

Finally, motivated by the theoretical result (Choi et al., 2023) that mutual information estimation is bounded by linear probing accuracy, we take an alternative view by probing LLMs with mutual information during pre-training. To our best knowledge, we are the first to notice that *during the pre-training period of LLMs, there exist two distinct phases regarding trustworthiness: fitting and compression*, which is in line with previous research on traditional DNNs (Shwartz-Ziv and Tishby, 2017; Noshad et al., 2019).

2 Probing LLM Pre-training Dynamics in Trustworthiness

In this section, we probe LLMs to analyze the dynamics of pre-training about trustworthiness. To begin with, we describe the datasets for each trustworthiness dimension in Section 2.1. Then, we introduce the experimental setup in Section 2.2. The probing results in Section 2.3 suggest that middle layer LLM representations from early pre-training have already exhibited linearly separable patterns.

2.1 Research Dimensions and Datasets of Truthworthy LLM

Existing research on AI governance (Tabassi, 2023; Commission et al., 2019; Commission, 2021b) and trustworthy AI (Liu et al., 2023b; Foundation, 2023) lays the groundwork for developing a comprehensive understanding of trustworthy LLMs. Guided by these principles, various studies classify trustworthy LLMs from different perspectives, yet some dimensions consistently emerge across these works (Liu et al., 2023d; Wang et al., 2023a; Sun et al., 2024). Therefore, we delves into five of these key dimensions: reliability, toxicity, privacy, fairness, and robustness, employing canonical datasets for each to support our study.

Reliability. TruthfulQA (Lin et al., 2022), a benchmark dataset for evaluating LLMs' truthfulness discernment (Touvron et al., 2023b), includes 817 questions across 38 categories aimed at assessing

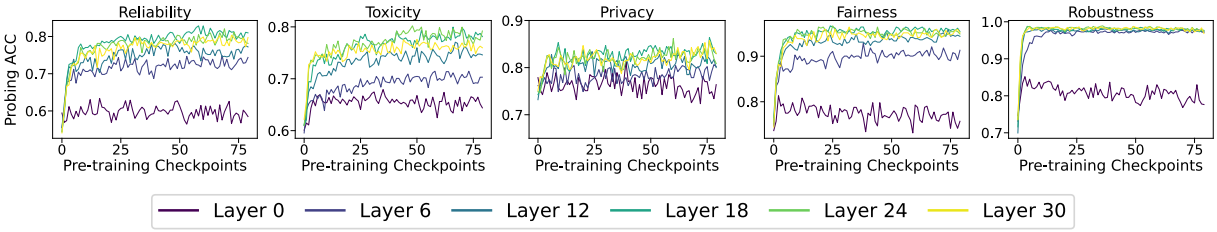

Figure 2: The linear probe accuracy on five trustworthiness dimensions for the first 80 pre-training checkpoints. For each checkpoint, we report the results from layers {0, 6, 12, 18, 24, 30}. The results from all layers of the 360 checkpoints are in Appendix D.

the veracity of model-generated answers.

Toxicity. ToxiGen (Hartvigsen et al., 2022) is a broad dataset featuring implicit toxic and non-toxic statements across 13 minority demographics, enabling toxicity modeling assessment in LLMs.

Privacy. We choose the tier 2 tasks from ConfAIde (Mireshghallah et al., 2023) to assess LLMs' privacy awareness, with ConfAIde targeting contextual privacy and identifying vulnerabilities in LLMs' privacy reasoning.

Fairness. We use StereoSet (Nadeem et al., 2021) to measure the stereotype modeling ability, i.e., whether LLMs capture stereotypical biases about race, religion, profession, and gender.

Robustness. We introduce typos by randomly changing the case of 5% letters in each sentence from SST-2 (Socher et al., 2013) from GLUE benchmark (Wang et al., 2018). The original sentence as well as the corresponding perturbed sentence are synthesized into a new dataset.

For each dataset above, we assign a label to every sentence based on whether it is trustworthy, i,e, truthful, toxic, privacy-aware, fair, and perturbed. We maintain a balanced dataset for each trustworthiness dimension. Further details are available in Appendix B.

2.2 Experimental Setup

The models under study. We investigate the pre-training period of LLMs through the 360 pre-training checkpoints provided by LLM360 (Liu et al., 2023e). Simultaneously, they also release an instruction fine-tuned conversational model named AmberChat and an aligned conversational model named AmberSafe. The models mentioned are all of the 7B parameter scale.

Activation dataset. Given each original dataset consisting of sentences and the corresponding class labels, we feed the sentence into LLMs and collect the corresponding activations of the last token (Li et al., 2023; Gurnee and Tegmark, 2023) for each layer. The activation dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ is constructed with the activations $\mathbf{x}_i \in R^d$ and the corresponding binary labels $y_i \in \{0, 1\}$.

Linear probing. We employ the linear probing method (Alain and Bengio, 2016; Tenney et al., 2019; Pimentel et al., 2020; Li et al., 2021; Belinkov, 2022) to analyze the activation datasets. For each trustworthiness dataset, every layer of each pre-training checkpoint within LLM360 produces an activation dataset. Therefore, there are 360×32 activation datasets for all 32 layers across 360 checkpoints. We randomly split each activation dataset into training and test sets by 4:1, and fit a binary linear classifier on the training set. We train a classifier for each activation dataset, which yields 360×32 classifiers. We report the accuracy on the test set.

2.3 Probing Results

Middle layer representations exhibit linearly separable patterns. For each checkpoint during pre-training, Figure 2 shows that the accuracy is relatively higher for middle layers (the 12-th and 18-th layers). The full results in Appendix D also support such characteristic of middle layers (about the 18-th layer). It inspires us that the representations from middle layers exhibit rich linear encoded information to distinguish those different concepts. Also, the observation meets with other literatures considering linear probing in the era of LLMs (Li et al., 2023; Zou et al., 2023; Burns et al., 2022), which also empirically validate the capability of middle layers. Moreover, similar phenomenon has also been found in earlier linear probing literatures for BERT (Hewitt and Manning, 2019; Van Aken et al., 2019), which may implicitly suggests some similarity between LLMs and relatively small pre-trained models.

The potential of pre-training checkpoints. Figure 2 shows that for each layer over the whole pre-training period, the probing accuracy increases

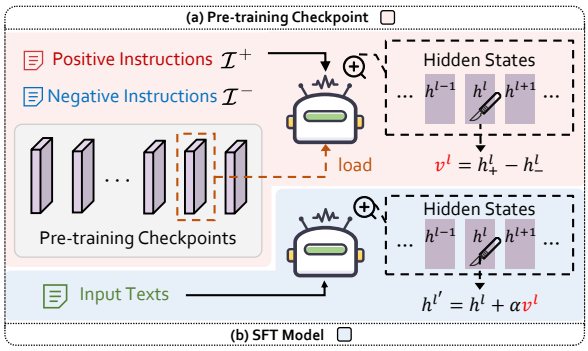

Figure 3: A schematic illustration of (a) constructing steering vector from the pre-training checkpoints and (b) intervening in the SFT model towards more trustworthiness by employing the steering vector.

during the initial phase of pre-training, followed by fluctuation throughout the remaining pre-training period. The trend enlightens us that models during the early stages of pre-training can already encode these different concepts well in a simple linear manner. Such trustworthiness concepts are linearly represented in the latent space of LLMs, which supports linear representation hypothesis (Park et al., 2023) and other empirical study (Zou et al., 2023).

3 Controlling Trustworthiness via the Steering Vectors from Pre-training Checkpoints

In this section, we aim to unravel the potential of checkpoints from the pre-training period to assist in enhancing the trustworthiness performance of the SFT model (i.e., AmberChat), based on activation intervention techniques (Turner et al., 2023; Li et al., 2023; Rimsky et al., 2023). We first outline the method of activation intervention on the SFT model using the steering vectors extracted from pre-training checkpoints in Section 3.1. Next, we introduce the experimental setup in Section 3.2. We then explore how steering vectors extracted from pre-training checkpoints contribute to enhancing performance across distinct dimensions of trustworthiness in Section 3.3, presenting a series of findings and observations. Finally, we examine the use of the same techniques to boost the overall trustworthiness performance of the SFT model in Section 3.4.

3.1 Activation Intervention

Initially, we partition the training dataset into two distinct collections based on the labels, $\mathcal{I}^+$ and $\mathcal{I}^-$, representing positive instructions and negative instructions, respectively. Following this partition, we collect the activations of LLM w.r.t. these instructions, denoted by $A^l_c(\mathcal{I}^+)$ and $A^l_c(\mathcal{I}^-)$, where A^l_c denotes the function that extracts the activations from the c-th checkpoint at l-th layer. Subsequently, we compute the centroid of the activations from each sets and take their difference to obtain the "mass mean vector," (Li et al., 2023; Marks and Tegmark, 2023) which serves as our steering vector

$$\boldsymbol{v}^l_c = \overline{A^l_c}(\mathcal{I}^+) - \overline{A^l_c}(\mathcal{I}^-). \quad (1)$$

Finally, we employ the steering vector to intervene in the model's activations as illustrated below

$$\boldsymbol{h}^{l'} = \boldsymbol{h}^l + \alpha \boldsymbol{v}^l_c, \quad (2)$$

where $\boldsymbol{h}^l$ denotes representation at the l-th layer of the model, $\boldsymbol{h}^{l'}$ denotes the corresponding representation after intervention; α is a rescale hyperparameter that indicates the strength of the intervention. Figure 3 illustrates the schematic diagram of intervention method. Note that the intervention described by Eq. (2) occurs at each step during the autoregressive inference.

3.2 Experimental Setup

Evaluation on Trustworthiness Datasets. For TruthfulQA, we fine-tune two GPT-3 models as "GPT-judge" and "GPT-info" guided by (Lin et al., 2022), to predict the truthfulness and informativeness of the generated outputs from LLMs, respectively. For ToxiGen, we follow (Touvron et al., 2023b), employing fine-tuned RoBERTa (Hartvigsen et al., 2022) to evaluate the toxicity of contents generated by LLMs, and finally reporting the proportion of generated text classified as toxic. For ConfAIde, StereoSet, and perturbed SST-2, with the adaptation of converting possible multiple-choice questions into binary classification tasks, we prompt LLMs to generate choices and then evaluate the accuracy. Please refer to Appendix C for more details.

Details of Steering Vectors Construction. For the activation dataset, we consider it from two perspectives: 1) For controlling the performance of individual subcategories under trustworthiness in Section 3.3, we utilize the corresponding datasets described in Section 2.1, where the steering vectors are constructed from the development set and no data leakage occurs during the evaluation; 2) For controlling the overall trustworthiness performance in Section 3.4, we employ PKU-SafeRLHF-10K[1],

[1]https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF-10K

Table 1: Results of activation intervention on TruthfulQA, general ability benchmarks, and the other trustworthiness benchmarks. The best results are highlighted in **bold**, and the runner-ups are underlined. $\boldsymbol{v}_{ckpt\_179}$ and $\boldsymbol{v}_{AmberChat}$ represent AmberChat intervened by steering vectors derived from ckpt_179 and AmberChat, respectively.

	Method	TruthfulQA metrics			General Abilities				Trustworthiness Abilities			
		Truth↑	Info↑	Truth * Info↑	ARC↑	MMLU↑	MathQA↑	RACE↑	ToxiGen↓	ConfAIde↑	StereoSet↑	SST-2↑
Baseline	AmberChat	0.3931	0.9484	0.3728	**0.6006**	**0.3659**	0.2593	0.3904	0.0920	0.5055	**0.5379**	**0.5757**
Fine-tuned	Full	0.4229	**0.9602**	0.4060	0.4315	0.2355	0.2499	0.3187	**0.0020**	0.5294	0.5031	**0.5757**
	Lora	0.3221	0.9329	0.3004	0.5758	0.3314	**0.2620**	0.3742	0.0080	**0.6411**	0.4980	0.5734
Activation Intervention	$\boldsymbol{v}_{ckpt\_179}$	**0.7322**	0.9337	**0.6837**	0.5834	0.3358	0.2422	0.3876	0.0360	0.6181	0.5000	0.5229
	$\boldsymbol{v}_{AmberChat}$	0.6978	0.9484	0.6618	0.5829	0.3388	0.2482	**0.3943**	0.0320	0.5192	0.4580	0.5367

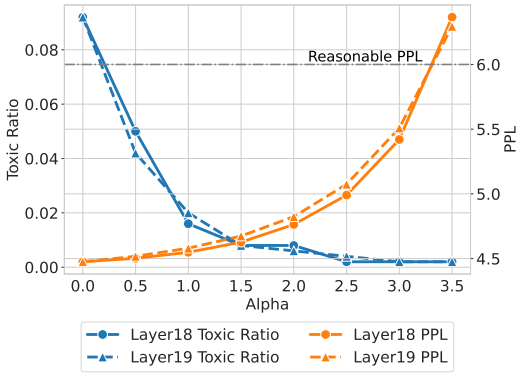

Figure 4: The trends of toxic ratio and PPL as the intervention strength α increases.

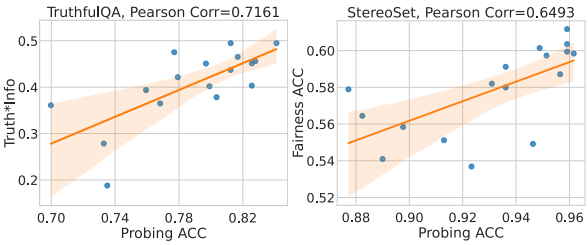

Figure 5: Pearson Correlation Coefficient for Probing ACC and trustworthiness performance.

a dataset proposed in (Ji et al., 2023) for RLHF training. For the checkpoint, we simply select the checkpoint that is halfway through the pre-training process for experiments, namely the checkpoint ckpt_179, which has already learnt linearly separable patterns (i.e., performs a high probing accuracy as shown in Figure 2). Regarding the selection of layer and α, we first narrow down the hyperparameter range based on Perplexity (PPL), and then empirically determine the optimal parameters using a coarse-grained grid search (Li et al., 2023; Turner et al., 2023; Wang and Shu, 2023).

3.3 Intervention to Enhance Distinct Trustworthiness Dimensions

In this subsection, we present several key observations that illuminate the intricate dynamics of steering vectors in modulating the trustworthiness of the SFT model.

Observation 1. *Steering vectors derived from pre-training checkpoints could significantly enhance the SFT model's performance in TruthfulQA, ToxiGen, and StereoSet.* For TruthfulQA and StereoSet, clear performance enhancement can be observed in Table 1 and Table 2, respectively. Regarding ToxiGen, when the strength of intervention α is set to 0.5, there is already a reduction of approximately 50% in the rate of toxic content generation, with a negligible perturbation in perplexity. Besides, sampling checkpoints from various stages of the pre-training period, we observe a relatively strong linear correlation between the trustworthiness performance and the probing accuracy of pre-training checkpoints in Figure 5. This suggests that, once the model has developed linearly separable patterns (represents a high probing accuracy) w.r.t. the trustworthiness concepts during the pre-training process, the constructed steering vector may have the potential to positively intervene in the SFT model's trustworthiness.

Observation 2. *Steering vectors derived from pre-training checkpoints and SFT model perform broadly comparable performance yet exhibit variations across various tasks.* Table 1 shows that, compared to the steering vector extracted from AmberChat, the steering vector from the pre-training checkpoint (ckpt_179) guides the SFT model to exhibit more "truthfulness." Moreover, it performs slightly better on ARC, ConfAIde, and StereoSet, while the opposite is true for other tasks. It is important to note that we only selected a single checkpoint from the pre-training process for experimentation, without undergoing fine-grained hyperparameter selection. Therefore, we believe these pre-training checkpoints hold significant untapped potential for aiding LLM towards trustworthiness.

Observation 3. *Intervening in the model slightly*

Table 2: Results of activation intervention on StereoSet, general ability benchmarks, and the other trustworthiness benchmarks. Format and significance markers keep consistent with Table 1.

	Method	Fairness Metric	General Abilities				Trustworthiness Abilities			
		StereoSet ↑	ARC↑	MMLU↑	MathQA↑	RACE↑	TruthfulQA↑	ToxiGen↓	ConfAIde↑	SST-2↑
Baselines	AmberChat	0.5379	**0.6006**	**0.3659**	**0.2593**	**0.3904**	**0.3728**	0.0920	**0.5055**	**0.5757**
Activation Intervention	$\boldsymbol{v}_{ckpt\_179}$	0.5799	0.5986	0.3524	0.2499	0.3914	0.2851	**0.0600**	0.5055	0.5390
	$\boldsymbol{v}_{AmberChat}$	**0.5830**	0.5958	0.3508	0.2519	0.3952	0.3352	0.0820	0.5055	0.5528

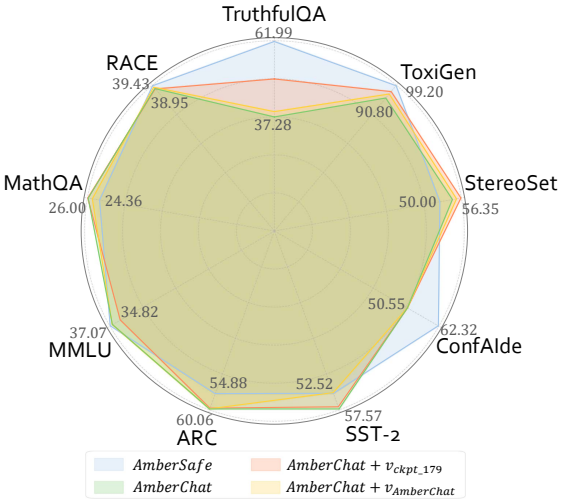

Figure 6: Performance of various models across four general capabilities and five trustworthiness capabilities. AmberChat and AmberSafe are fine-tuned models from LLM360. $\boldsymbol{v}_{ckpt\_179}$ and $\boldsymbol{v}_{AmberChat}$ represent steering vectors from ckpt_179 and AmberChat, respectively.

impairs its general capabilities as a marginal cost for trustworthiness enhancement. We evaluate the model's performance on four common benchmarks for general capabilities, where a trend of slight performance decline is observed after intervention, as indicated in "General Abilities" part of Tables 1 and 2. Additionally, we also observe the impact of the intervention strength α on the generative performance of the model. Taking ToxiGen as an example, Figure 4 illustrates the relationship between the proportion of toxic content generated by the model and perplexity as the intervention strength α increases. If we continuously increase the intervention strength, although the proportion of toxicity may continue to decline, the perplexity of the model correspondingly increase, manifesting as a tendency to produce meaningless repetitive content or gibberish.

Observation 4. *When the quantity and quality of fine-tuning data are limited, activation intervention by steering vectors may be a more effective approach for current task.* We fine-tune the SFT model with positive QA pairs from the training set using both full-parameter fine-tuning and LoRA fine-tuning as a comparison, given that data in TruthfulQA naturally exists in the form of QA pairs. As shown in Table 1, the model fine-tuned with all parameters exhibits only minor improvements on TruthfulQA while experiencing a significant decline in general capabilities. Meanwhile, the fine-tuned model by LoRA demonstrates a noticeable decrease in TruthfulQA though somewhat preserving performance in general capabilities.

Observation 5. *Trade-offs exist between different dimensions of trustworthiness.* For instance, as seen in Table 1, while steering vector intervention enhances the model's truthfulness performance, it also compromises performance on fairness and robustness. Previous research has witnessed a trade-off between trustworthiness dimensions. For example, privacy-fairness trade-off (Mangold et al., 2023), robustness-privacy trade-off (Hayes, 2020), and robustness-fairness trade-off (Xu et al., 2021). Similar to (Liang et al., 2022), we also suggest that the connection between different trustworthiness dimensions relies on their definitions. Many pairs of trustworthiness in LLMs remain unstudied, and we advocate for future research in this area.

3.4 Intervention to Enhance Universal Trustworthiness

In this subsection, we aim to leverage steering vectors to comprehensively enhance the model's trustworthiness. Unlike Section 3.3 where steering vectors are constructed using datasets from different dimensions of trustworthiness, here we employ a general dataset for alignment (described in Section 3.2), which may encompass data across multiple dimensions of trustworthiness.

Trustworthiness enhancement with steering vectors from universal alignment datasets. Figure 6 shows that intervening in model using steering vectors can significantly boost trustworthiness, with only marginal losses (in ARC, MMLU) or even marginal gains (in MathQA, RACE) in general capabilities. Moreover, steering vectors derived from checkpoints during the pre-training pe-

riod demonstrate superior effectiveness in enhancing trustworthiness. For AmberSafe, which employs a substantial cost for alignment, we note its overall best performance (as seen in the purple line), particularly holding a significant advantage in privacy and TruthfulQA. However, it's noteworthy that merely using 10k alignment data to construct steering vectors from a pre-training checkpoint for intervening in the SFT model brings about impressive improvements across various dimensions of trustworthiness, which reveals the untapped potential of pre-training checkpoints in aiding the model towards better trustworthiness.

4 Probing LLMs using Mutual Information

Recently, Choi et al. (2023) shows that mutual information estimation is bounded by linear probing accuracy. Also, the mutual information can be used to investigate the dynamics of neural networks during training (Shwartz-Ziv and Tishby, 2017; Saxe et al., 2019; Goldfeld and Polyanskiy, 2020; Pimentel et al., 2020; Geiger, 2021; Lorenzen et al., 2021; Zhou et al., 2023). Therefore, motivated by the above, we adopt a different perspective by probing LLM checkpoints through the lens of mutual information, particularly focusing on the aforementioned trustworthiness dimensions.

We explain our probing strategy and experimental setup in Section 4.1 and Section 4.2, respectively. The empirical observations are shown and analyzed in Section 4.3. In particular, we find that there is a phase transition from "fitting" to "compression" during the pre-training period of LLMs, which is consistent with previous study on traditional DNNs (Shwartz-Ziv and Tishby, 2017; Noshad et al., 2019).

4.1 Probing Strategy

The mutual information between two continuous random variables, X and Y, is defined as

$$I(X,Y) = \int_Y \int_X p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy.$$

It is a measure of the independence between two variables. Given the dataset of trustworthiness in Section 2.1, we represent each dataset using the first layer activation X, and Y denotes the corresponding label vector. Additionally, T represents the feature matrix from the target layer of an LLM. Thus, we probe LLMs with $I(T,X)$ and $I(T,Y)$ during pre-training.

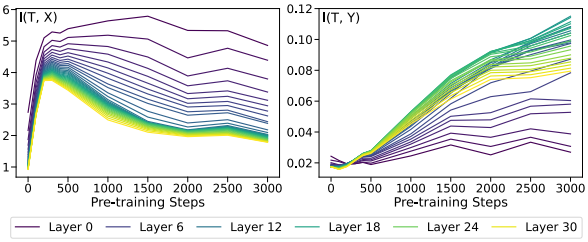

Figure 7: The dynamics of $I(T,X)$ and $I(T,Y)$ for TruthfulQA across various layers during pre-training. The similar trend in other datasets is in Appendix E.2.

In principle, our strategy differs from Shwartz-Ziv and Tishby (2017) in three ways. Firstly, we do not use the pre-training dataset of LLMs. Instead, we carefully design activation datasets to represent specified trustworthiness properties. Secondly, we use the first layer representation to indicate the original dataset because they contains more information than representations from other layers (Cover, 1999; Tishby and Zaslavsky, 2015; Shwartz-Ziv and Tishby, 2017). Finally, we follow Ma et al. (2020) to use HSIC (Gretton et al., 2005) as an estimator of mutual information because it is challenging to accurately compute in high dimensions (Kraskov et al., 2004; Alemi et al., 2016; Poole et al., 2019).

4.2 Experimental Setup

Following the official code and reported hyperparameters from Liu et al. (2023e), we initiate pre-training from a randomly initialized model using the corpus for the first checkpoint, and save more granular checkpoints to observe finer experimental phenomena. More discussions are available in Appendix C.

4.3 The Dynamics of Pre-training

The trend of mutual information. Figure 7 shows that $I(T,X)$ generally exhibits an initial increase followed by a decrease across all the considered layers during pre-training. And $I(T,Y)$ continues to show a consistent upward trend. Note that middle layer representations exhibit a larger $I(T,Y)$ comparing to that from other layers. It suggests that middle layer representations encode more information about the opposing concepts of trustworthiness.

From "fitting" to "compression." Overall, considering $I(T,X)$ and $I(T,Y)$ collectively, it becomes evident that there are two phases during pre-training. In the first and shorter phase, both $I(T,X)$ and $I(T,Y)$ increase. While in the sec-

ond and much longer phase, $I(T, X)$ decreases and $I(T, Y)$ continues to increase. Although our strategy is completely different from Shwartz-Ziv and Tishby (2017), the two-phase phenomenon exhibits similarities. At the beginning of pre-training, the randomly-initialized LLM fails to preserve the relevant information, so $I(T, X) \approx 0$ and $I(T, Y) \approx 0$. Next, as LLM gradually fits the pre-training dataset, its abilities in language understanding and concept modeling enhance, contributing to increases in both $I(T, X)$ and $I(T, Y)$. As pre-training progresses, LLM learns to better compress the irrelevant information in the dataset and preserve more label-related information (i.e., trustworthiness), leading to a reduction in $I(T, X)$ and an improvement in $I(T, Y)$. Overall, we are at the forefront of investigating the phase transition from "fitting" to "compression" in the context of trustworthiness during pre-training. It is our hope that our insights will motivate further exploration into the pre-training dynamics of LLMs.

5 Related Work

Probing LLM representations. Probing classifiers (Alain and Bengio, 2016; Tenney et al., 2019; Pimentel et al., 2020; Li et al., 2021; Belinkov, 2022; Räuker et al., 2023) is one of the prominent methods for identify certain properties acquired by the language model (Zhao et al., 2023). Researchers probe LLMs and discover linear separable patterns within LLMs, including space and time (Gurnee and Tegmark, 2023), game states (Nanda et al., 2023), answerability (Slobodkin et al., 2023), and some counterfactual pairs of concepts (Park et al., 2023). It is also observed that LLM representations contain linearly separable patterns about trustworthiness, such as truthfulness (Li et al., 2023; Marks and Tegmark, 2023; Zou et al., 2023). However, they do not probe LLM representations during pre-training. In this work, we consider the whole pre-training period of LLMs and probe their presentations dynamically.

Steering vectors for trustworthy LLMs. To ensure the safety and trustworthiness of LLMs, some promising approaches explore the latent space, utilizing representations to improve model performance (Liu et al., 2023c; Jorgensen et al., 2023). Various studies investigate activation engineering within LLMs from both theoretical and practical perspectives, affecting model performance by manipulating the model's representational space (Park et al., 2023; Turner et al., 2023; Zou et al., 2023). Furthermore, Wang and Shu (2023), Rimsky et al. (2023) and Wang et al. (2024) construct directional vectors to explore the model's safety and alignment, with the goal of making models helpful, honest, and harmless. However, there has no investigation into how representations change during the pre-training phase of LLMs. In this paper, we explore and leverage representations during this phase, paving the way for new research avenues in activation engineering.

Understanding the training process of DNNs. Many empirical studies observe that DNNs tend to learn simple concepts during the learning process (Arpit et al., 2017; Liu et al., 2021; Mangalam and Prabhu, 2019). Furthermore, Xu et al. (2019), Liu et al. (2023a), and Tian et al. (2023) theoretically explain the learning preference of DNNs. Meanwhile, many researchers focus on analyzing the utility of fine-tuning for language models (Merchant et al., 2020; Hao et al., 2020; Aghajanyan et al., 2021; Zhou and Srikumar, 2022; Mosbach et al., 2020). However, few previous studies investigate how trustworthiness is learned by LLMs during pre-training. In this paper, we take a closer look at the learning dynamic of trustworthiness within LLMs' representations.

6 Conclusion

In this work, we take an initial and illuminating step towards elucidating the conceptual understanding of trustworthiness during pre-training. Firstly, by linear probing LLMs across reliability, privacy, toxicity, fairness, and robustness, we investigate the ability of LLMs representations to discern opposing concepts within each trustworthiness dimension during the whole pre-training period. Furthermore, motivated by the probing results, we conduct extensive experiments to reveal the potential of utilizing representations from LLMs during its previous pre-training period to enhance LLMs' own trustworthiness. Finally, we use mutual information to probe LLMs during pre-training and reveal some similarity of the learning mechanism between LLMs and traditional DNNs. Taken collectively, the empirical study presented in this work can not only justify the potential to improve trustworthiness of LLMs using their own pre-training checkpoints, but may also lead to a better understanding of the dynamics of LLM representations, especially the trustworthiness-related concepts.

7 Limitations

There are several limitations of this work. Firstly, we only focus on five essential trustworthiness dimensions and do not encompass all the dimensions, such as those appeared in Commission et al. (2019); Liu et al. (2023b). And a wide variety of definitions for each trustworthiness dimension, as discussed by (Wang et al., 2023a; Sun et al., 2024), are not completely covered in our analysis. Secondly, due to limitations in computational resources as well as the lack of open-source pre-training LLM checkpoints, we only conduct experiments on LLM360 (Liu et al., 2023e). Finally, for evaluation of TruthfulQA, the precision of evaluation results depends on the performance of the "GPT-judge" evaluator. And for multiple-choice evaluation, the evaluation results may rely on the instruction following ability of LLMs.

8 Broader Impact and Ethics Statement

This study concentrates on better understanding the learning dynamics of LLM trustworthiness during pre-training. The motivation of our steering vector experiments is centered on improving the trustworthiness of LLMs. We recognize the sensitive nature of our research and assure that it strictly complies with legal and ethical guidelines.

This research is carried out in a secure, controlled environment, ensuring the safety of real-world systems. Given the nature of our work, which includes dealing with potentially sensitive content like unreliable statements and toxic sentences, we have implemented strict protocols. Access to the most sensitive aspects of our experiments is limited to researchers with the proper authorization, who are committed to following rigorous ethical standards. These precautions are taken to maintain the integrity of our research and to mitigate any risks that could arise from the experiment's content.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR.

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

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

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

Kwanghee Choi, Jee-weon Jung, and Shinji Watanabe. 2023. Understanding probe behaviors through variational bounds of mutual information. *arXiv preprint arXiv:2312.10019*.

European Commission. 2021b. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, pub. l. no. com(2021) 206 final.

European Commission, Content Directorate-General for Communications Networks, and Technology. 2019. *Ethics guidelines for trustworthy AI*. Publications Office.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

AI Verify Foundation. 2023. Catalogue of llm evaluations.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Bernhard C Geiger. 2021. On information plane analyses of neural network classifiers–a review. *IEEE Transactions on Neural Networks and Learning Systems*.

Ziv Goldfeld and Yury Polyanskiy. 2020. The information bottleneck problem and its applications in machine learning. *IEEE Journal on Selected Areas in Information Theory*, 1(1):19–38.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2020. Investigating learning dynamics of BERT fine-tuning. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 87–92. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for implicit and adversarial hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Jamie Hayes. 2020. Trade-offs between membership privacy & adversarially robust learning. *arXiv preprint arXiv:2006.04622*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138.

Saghar Hosseini, Hamid Palangi, and Ahmed Hassan Awadallah. 2023. An empirical study of metrics to measure representational harms in pre-trained language models. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 121–134.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. 2023. Improving activation steering in language models with mean-centring. *arXiv preprint arXiv:2312.03813*.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical review E*, 69(6):066138.

Belinda Z Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. *arXiv preprint arXiv:2106.00737*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

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)*.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. 2024. Tuning language models by proxy.

Chen Liu, Zhichao Huang, Mathieu Salzmann, Tong Zhang, and Sabine Süsstrunk. 2021. On the impact of hard adversarial instances on overfitting in adversarial training. *arXiv preprint arXiv:2112.07324*.

Dongrui Liu, Huiqi Deng, Xu Cheng, Qihan Ren, Kangrui Wang, and Quanshi Zhang. 2023a. Towards the difficulty for a deep neural network to learn concepts of different complexities. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil Jain, and Jiliang Tang. 2023b. Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology*, page 1–59.

Sheng Liu, Lei Xing, and James Zou. 2023c. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023d. Trustworthy llms: a survey and guideline for evaluating large language models' alignment.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. 2023e. Llm360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*.

Stephan Sloth Lorenzen, Christian Igel, and Mads Nielsen. 2021. Information bottleneck: Exact analysis of (quantized) neural networks. *arXiv preprint arXiv:2106.12912*.

Wan-Duo Kurt Ma, JP Lewis, and W Bastiaan Kleijn. 2020. The hsic bottleneck: Deep learning without back-propagation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5085–5092.

Karttikeya Mangalam and Vinay Uday Prabhu. 2019. Do deep neural networks learn shallow learnable examples first?

Paul Mangold, Michaël Perrot, Aurélien Bellet, and Marc Tommasi. 2023. Differential privacy has bounded impact on fairness in classification. In *International Conference on Machine Learning*, pages 23681–23705.

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*.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to bert embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory.

Eric Mitchell, Rafael Rafailov, Archit Sharma, Chelsea Finn, and Christopher D. Manning. 2023. An emulator for fine-tuning large language models using small language models.

Marius Mosbach, Anna Khokhlova, Michael A Hedderich, and Dietrich Klakow. 2020. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2502–2516.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*.

Jessica Newman. 2023. A taxonomy of trustworthiness for artificial intelligence: Connecting properties of trustworthiness with risk management and the ai lifecycle.

Morteza Noshad, Yu Zeng, and Alfred O Hero. 2019. Scalable mutual information estimation using dependence graphs. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2962–2966. IEEE.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 464–483. IEEE.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. 2019. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020.

Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory unanswerablity: Finding truths in the hidden states of over-confident large language models. *arXiv preprint arXiv:2310.11877*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Daniel J Solove. 2005. A taxonomy of privacy. *U. Pa. l. Rev.*, 154:477.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Elham Tabassi. 2023. Artificial intelligence risk management framework (ai rmf 1.0).

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Yuandong Tian, Yiping Wang, Zhenyu Zhang, Beidi Chen, and Simon Du. 2023. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. *arXiv preprint arXiv:2310.00535*.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Betty Van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. 2019. How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1823–1832.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023a. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. In *Advances in Neural Information Processing Systems*.

Haoran Wang and Kai Shu. 2023. Backdoor activation attack: Attack large language models using activation steering for safety-alignment. *arXiv preprint arXiv:2311.09433*.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023b. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*.

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Ke Ren, Botian Jiang, and Xipeng Qiu. 2024. Inferaligner: Inference-time alignment for harmlessness through cross-model guidance. *arXiv preprint arXiv:2401.11206*.

Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. 2021. To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning*, pages 11492–11501.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. 2019. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*.

Yichu Zhou and Vivek Srikumar. 2022. A closer look at how fine-tuning changes bert. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1046–1061.

Zhanke Zhou, Chenyu Zhou, Xuan Li, Jiangchao Yao, Quanming Yao, and Bo Han. 2023. On strengthening and defending graph reconstruction attack with markov chain approximation. In *International Conference on Machine Learning*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Contents

Appendix

A Guidelines for Trustworthy LLMs

The surge of LLMs brings significant concerns regarding their trustworthiness, which pertains to the aspects and extent to which humans can trust AI. Existing research in AI governance and trustworthy LLMs provides a guidance for establishing a comprehensive and reliable dimensions of trustworthy LLMs in this study.

Governments (Tabassi, 2023; Commission et al., 2019), organizations (Commission, 2021b; Foundation, 2023), and research institutions (Newman, 2023; Liu et al., 2023d) worldwide have proposed classifications from various perspectives such as the AI lifecycle, the acceptability of AI risk, considering AI governance at different levels including individual, institutional, and societal. Among these, categories stemming from the technological aspect offer guidance for trustworthy AI (Liu et al., 2023b), such as robustness, fairness, accountability, transparency, etc.

By integrating AI governance principles into trustworthy LLMs, not only aids in developing more credible LLMs but also promotes the sustainable and responsible application of AI technology. Concurrently, taking into account the categorizations of trustworthy LLMs (Liu et al., 2023d; Wang et al., 2023a) and prioritizing both adherence to principles and addressing practical challenges faced by LLMs, six primary categories have been identified: robustness, reliability, fairness, toxicity, privacy, and interpretability. In this study, interpretability is employed as a tool to explore the other five concepts of trustworthiness.

B Datasets of Truthworthy LLMs

Considering five aspects of trustworthiness: reliability, toxicity, privacy, fairness and robustness, we carefully design five binary NLP datasets. These datasets are tailored from independent lines of trustworthy AI research, with labels indicating whether a sentence satisfies each aforementioned aspect of trustworthiness. In other words, the label indicates whether the corresponding sentence contains untrue (or unfair, toxic, privacy-leakaging and perturbed) information.

The datasets considered below are balanced, i.e., the number of positive and negative numbers are almost the same. In other words, some special case, for example, the random classifier on these datasets, will achieve an accuracy around 50%.

Reliability. We use TruthfulQA (Lin et al., 2022) to measure the truthfulness modeling ability of LLMs. TruthfulQA comprises 817 questions across 38 categories, designed to evaluate the veracity of answers generated by language models. We concatenate the multiple-choice questions and their respective candidate answers to form either correct or incorrect statements, which is used to measure the reliability of large language models in discerning truthfulness.

Toxicity. We choose ToxiGen (Hartvigsen et al., 2022) to measure the toxicity modeling ability of LLMs. ToxiGen is a large-scale dataset encompassing a range of implicit toxic and non-toxic statements associated with 13 minority demographics. Following Llama2 (Touvron et al., 2023b), we employ a revised version of the dataset from (Hosseini et al., 2023), selectively retaining those sentences that achieved unanimous agreement from the annotators regarding the target demographic group.

Privacy. We choose the tier 2 task from ConfAIde (Mireshghallah et al., 2023) to measure the privacy awareness of LLMs. ConfAIde focuses on contextual privacy and aims to pinpoint key vulnerabilities in LLMs' privacy reasoning abilities. Given the limited data volume, we constructed new data based on ConfAIde and the Solove Taxonomy (Solove, 2005) to assess the privacy awareness of LLMs regarding given information. Solove Taxonomy comprises 4 major categories and 16 subcategories. For each subcategory, we designed prompts and provided 2 to 6 examples to facilitate data generation using GPT-4. The generated data were then assessed by GPT-4 for privacy violations, selecting entries with high confidence (consistent judgments in five assessments). We combined generated data with ConfAIde to consider whether LLMs can identify privacy violations.

Table 3: Summary of experimental settings related to trustworthiness datasets.

Dimention	Reliability	Toxicity	Privacy	Fairness	Robustness
Benchmark	TruthfulQA	ToxiGen	ConfAIde	StereoSet	SST-2
Evaluation Metrics	Truth% and Info%	Toxic Ratio	Accuracy	Accuracy	Accuracy
The meaning of labels in activation datasets	$y = 0$: statements with false answer $y = 1$: statements with true answer	$y = 0$: toxic statements $y = 1$: benign statements	$y = 0$: statements that do not conclude privacy violation $y = 1$: statements that conclude privacy violation	$y = 0$: benign statements $y = 1$: stereotypical statements	$y = 0$: the original sentence $y = 1$: the perturbed sentence

Fairness. We use StereoSet (Nadeem et al., 2021) to measure the stereotype modeling ability of LLMs, i.e., whether LLMs capture stereotypical biases about race, religion, profession, and gender. Taking inter-sentence tests as the original dataset, we concatenate the context and the candidate sentence into one sentence, and the corresponding class label follows the candidate sentences capturing stereotypical, anti-stereotypical, and unrelated associations. We assign a binary label to every sentence to indicate whether it contains stereotypical bias.

Robustness. Following the construction of AdvGLUE benchmark (Wang et al., 2021), we perturb GLUE benchmark (Wang et al., 2018) in a human-imperceptible way. Specifically, we introduce typos by randomly change the case of 20% letters in each sentence from SST-2 (Socher et al., 2013) validation set. We assign a binary label to every sentence to indicate whether it has been attacked.

C More Detailed Experimental Settings

Dataset partition. Within each dataset, following (Li et al., 2023), we first split the original dataset into a development set and a test set at a 1:1 ratio. We further divide the development set into a training/validation set at a 4:1 ratio for the training and evaluation of the linear probe, with the steering vector also being constructed based on the development set. The test set is used to assess model performance, ensuring no data leakage occurs during the experiment.

Evaluation on trustworthiness abilities benchmarks. For TruthfulQA, we adopt the QA prompts following InstructGPT (Ouyang et al., 2022). Additionally, two fine-tuned GPT-3 models, i.e. a "GPT-judge"[2] and a "GPT-info,"[3] are used to predict the truthfulness and informativeness of the generated outputs from LLMs, respectively. For ToxiGen, we follow (Touvron et al., 2023b), employing the default ToxiGen classifier (Hartvigsen et al., 2022) fine-tuned on RoBERTa (Liu et al., 2020) to evaluate the toxicity of contents generated by LLMs, and finally reporting the proportion of generated text classified as toxic. For ConfAIde, we use the the tier 2 task to assess the agreement on privacy information usage. We employ the same evaluation prompt as ConfAIde (Mireshghallah et al., 2023), with the adaptation of converting multiple-choice questions into binary classification tasks to evaluate the accuracy. For StereoSet, following TrustLLM (Sun et al., 2024), we provide prompts using the same template for stereotype recognition task as theirs. The generated choices are then compared with the ground-truth labels to obtain the accuracy. For perturbed SST-2, we follow Wang et al. (2023b) and use the same prompt as theirs. TruthfulQA is evaluated in a 6-shot setting, whereas other benchmarks are conducted with 0-shot settings.

Evaluation on general abilities benchmarks. For all the results on ARC, MMLU, MathQA, and RACE reported in Section 3 of the main body, we conduct evaluations using the lm-evaluation-harness library (Gao et al., 2023) with its default evaluation settings.

[2] ft:davinci-002:zy-pj-035:truthfulqa-truth:8nKPYSTt
[3] ft:davinci-002:zy-pj-035:truthfulqa-info:8nJbtN57

Selection of perplexity. Regarding perplexity, we follow (Radford et al., 2019) to calculate the perplexity on LAMBADA (Paperno et al., 2016). The perplexity value reported for GPT-2 in (Radford et al., 2019) is 8.6, and the perplexity we tested for AmberChat is 4.5. Based on our observations, we consider a perplexity value of less than 6 to be a reasonable threshold, please refer to Appendix G for examples.

Reproduce the first pre-training checkpoint. In our initial experimental observations using the pre-training checkpoints released in (Liu et al., 2023e), we noticed that the mutual information $I(T, X)$ appeared to be consistently decreasing, which contradicts the existing two-phase phenomenon (Shwartz-Ziv and Tishby, 2017). This led us to speculate the possibility of overlooked experimental insights between the initial model state and the first checkpoint. Therefore, to observe more finer-grained dynamics during the pre-training phase, we utilized the official code released by (Liu et al., 2023e)[4], ensuring the hyperparameters are consistent with those reported in the original paper. We initiate pre-training from a randomly initialized model using the corpus for the first checkpoint, and saved more finely-grained checkpoints to observe finer experimental phenomena.

D Full Linear Probing Results

The full linear probing results from 360 checkpoints in five trustworthiness dimensions are shown in Figure 8,9,10,11,12. Overall, the experimental observations and conclusions are consistent with Section 2.3. Results from five datasets together suggest that middle layer representations exhibit linearly separable patterns. Furthermore, the probing accuracy increases during the initial phase of pre-training, followed by fluctuation throughout the remaining pre-training period.

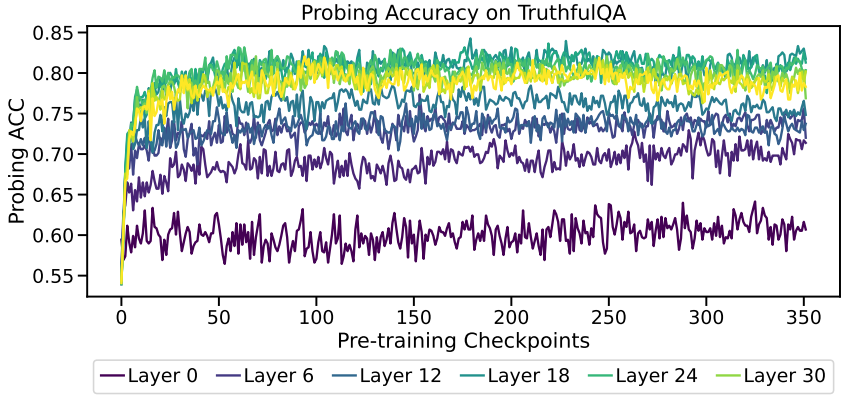

Figure 8: The linear probe accuracy on TruthfulQA for all 360 pre-training checkpoints.

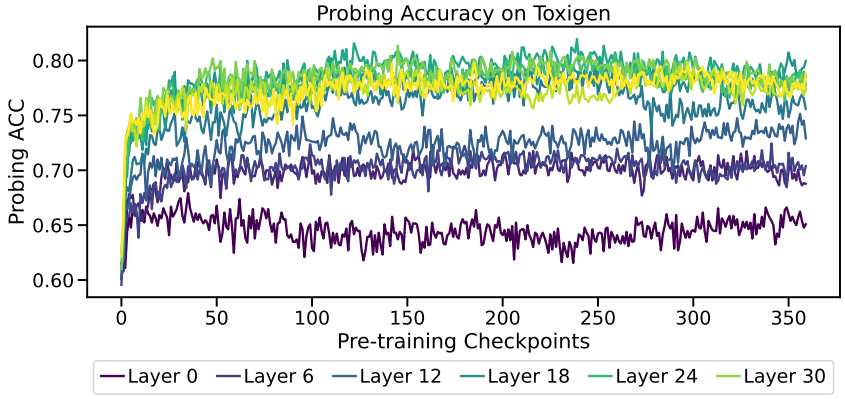

Figure 9: The linear probe accuracy on Toxigen for all 360 pre-training checkpoints.

[4] https://github.com/LLM360/amber-train

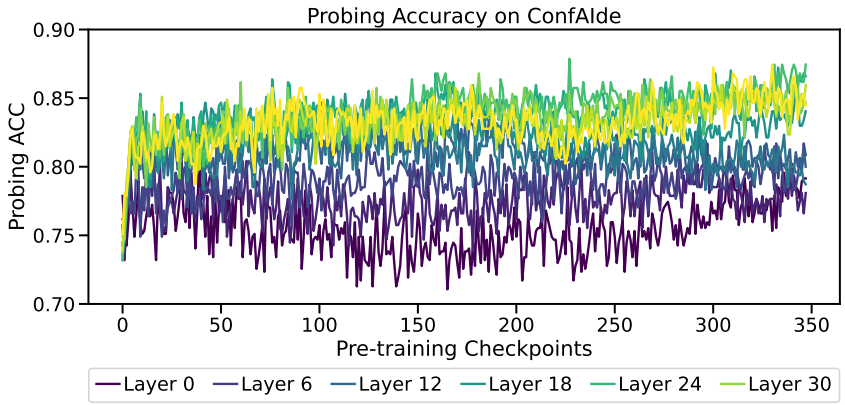

Figure 10: The linear probe accuracy on ConfAIde for all 360 pre-training checkpoints.

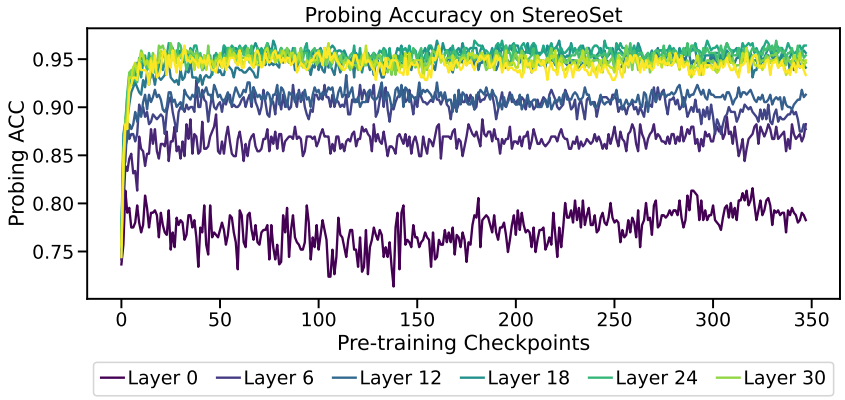

Figure 11: The linear probe accuracy on StereoSet for all 360 pre-training checkpoints.

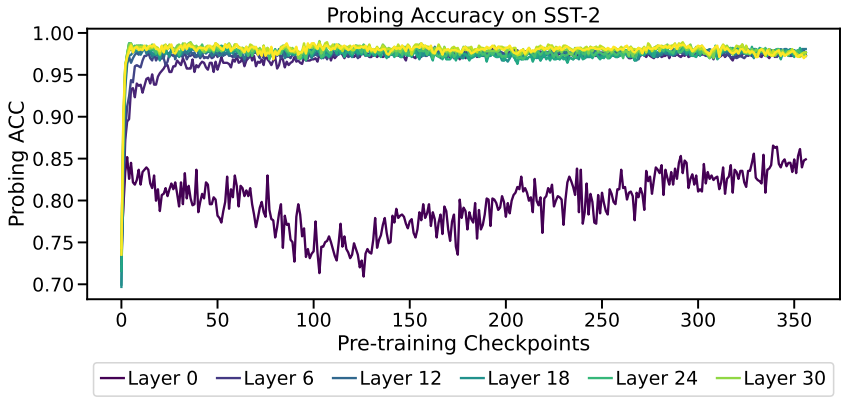

Figure 12: The linear probe accuracy on SST-2 for all 360 pre-training checkpoints.

E Supplementary Details for 'Probing LLM using Mutual Information'

E.1 Mutual Information and HSIC

Definition 1 (Mutual Information (MI))**.** Given two continuous random variables X and Y, the mutual information is defined as:

$$I(X;Y) = \int_Y \int_X p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy. \tag{3}$$

Mutual information is a measure of the mutual dependence between the two variables. However, because of the difficulty to accurately compute mutual information (Kraskov et al., 2004), we follow Ma et al. (2020) to use HSIC (Gretton et al., 2005) as an estimator of mutual information. HSIC (Gretton et al., 2005) also indicates the dependency between two random variables. For other kinds of estimation, please refer to Appendix E.3 in Zhou et al. (2023).

Definition 2 (Hilbert-Schmidt Independence Criterion (HSIC))**.** It is the Hilbert-Schmidt norm of the cross-covariance operator between the distributions in Reproducing Kernel Hilbert Space (RKHS). $\mathrm{HSIC}(X,Y)$ is defined as:

$$\begin{aligned} \mathrm{HSIC}(X,Y) = &\mathbb{E}_{XYX'Y'}\left[k_X\left(X,X'\right)k_{Y'}\left(Y,Y'\right)\right] \\ &+ \mathbb{E}_{XX'}\left[k_X\left(X,X'\right)\right]\mathbb{E}_{YY'}\left[k_Y\left(Y,Y'\right)\right] \\ &- 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\left[k_X\left(X,X'\right)\right]\mathbb{E}_{Y'}\left[k_Y\left(Y,Y'\right)\right]\right], \end{aligned} \tag{4}$$

where X', Y' are independent copies of X, Y, respectively, and k_X , k_Y are kernels.

$\mathrm{HSIC}(X,Y)$ is zero if and only if the random variables X and Y are independent. In practice, given the activation dataset $\mathcal{D}$, we empirically estimate HSIC as

$$\widehat{\mathrm{HSIC}}(X,Y) = (n-1)^{-2}\,\mathrm{tr}\left(K_X H K_Y H\right), \tag{5}$$

where K_X and K_Y are kernel matrices with entries $K_{X_{ij}} = k_X\left(x_i, x_j\right)$ and $K_{Y_{ij}} = k_Y\left(y_i, y_j\right)$, respectively, and $H = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is a centering matrix. Following (Ma et al., 2020), we choose Gaussian kernel $k(\mathbf{x},\mathbf{y}) \sim \exp\left(-\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2/\sigma^2\right)$. The scaling parameter σ is selected by grid search in $[50, 400]$.

E.2 Mutual Information Results across Five Trustworthiness Dimensions

Figure 13,14,15,16,17 show the trend of mutual information on five trustworthiness dimensions. The results are also consistent with the dynamics in Section 4.3. The phase transition from "fitting" to "compression" is also applicable: there are also two phases during pre-training. In the first and shorter phase, both $I(T,X)$ and $I(T,Y)$ increase. While in the second and much longer phase, $I(T,X)$ decreases and $I(T,Y)$ continues to increase. There are some fluctuations of $I(T,Y)$ for Toxigen, which may be due to the instability of pre-training.

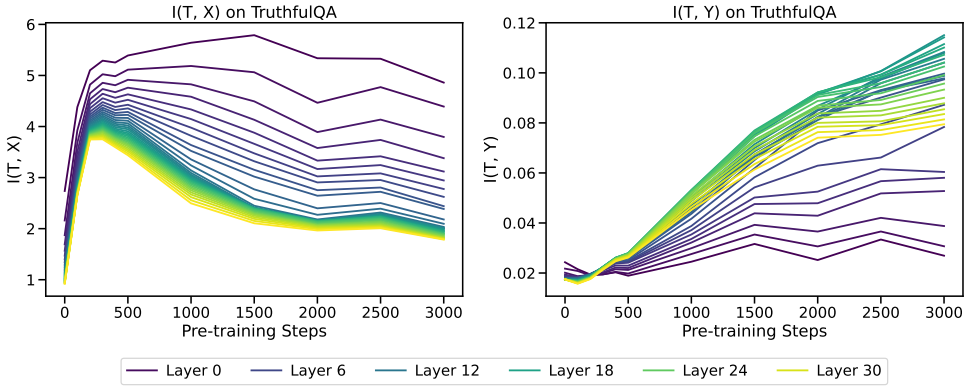

Figure 13: The dynamics of $I(T, X)$ and $I(T, Y)$ for TruthfulQA across various layers during pre-training.

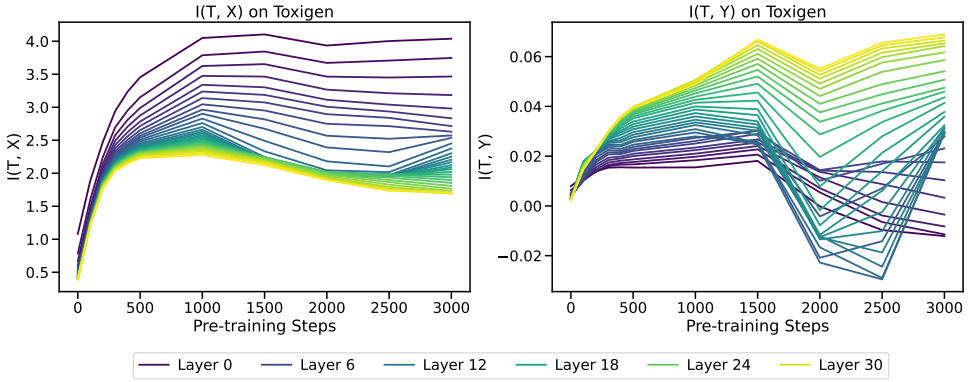

Figure 14: The dynamics of $I(T, X)$ and $I(T, Y)$ for Toxigen across various layers during pre-training.

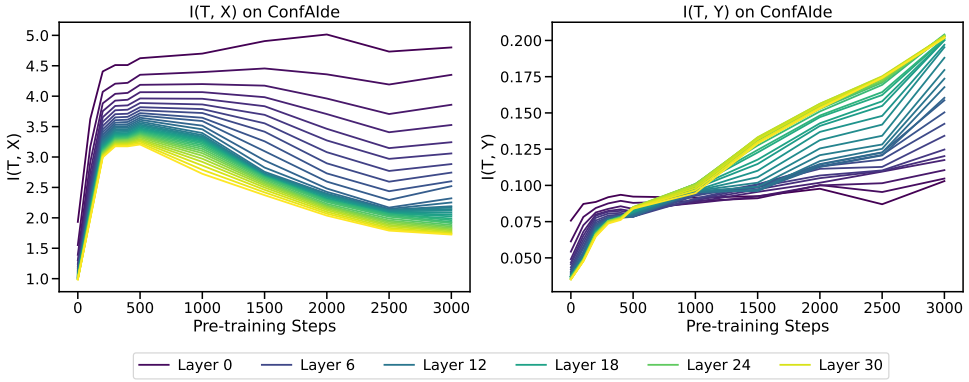

Figure 15: The dynamics of $I(T, X)$ and $I(T, Y)$ for ConfAIde across various layers during pre-training.

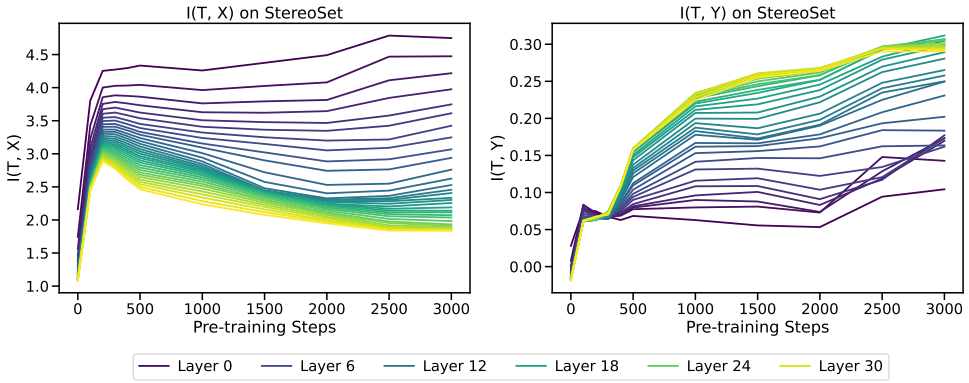

Figure 16: The dynamics of $I(T, X)$ and $I(T, Y)$ for StereoSet across various layers during pre-training.

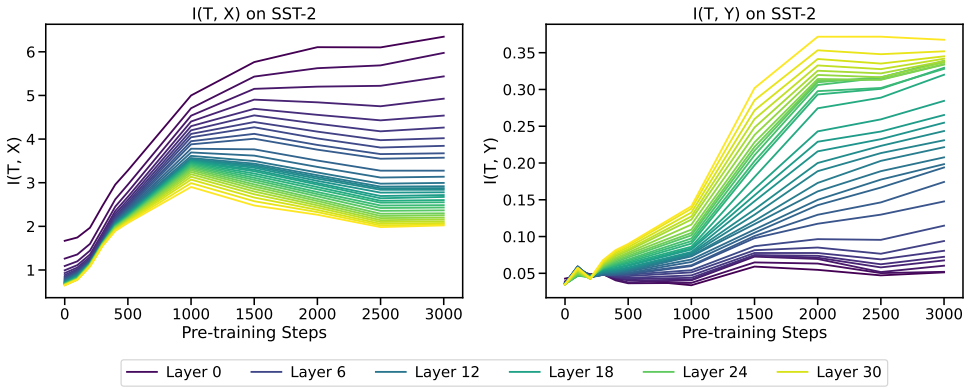

Figure 17: The dynamics of $I(T, X)$ and $I(T, Y)$ for SST-2 across various layers during pre-training.

F Unlocking the Potential of Pre-trained Checkpoints through Proxy-tuning

The linear probe results of LLM360 and its evaluations across all checkpoints on TruthfulQA indicate that checkpoints during pre-training have already developed modeling capabilities for truthworthiness. Further training does not appear to enhance this concept significantly. However, cause the gap between latent space representation and model output (Ravichander et al., 2021), strong representation seems not to be well applied. To address this, we attempt to shift the original predictions of the checkpoints during pre-training to enhance their utilization capabilities.

F.1 Proxy-Tuning to Checkpoints during Pre-training

Proxy-tuning applies the prediction differences between the tuned model and untuned model to shift the original predictions of a base model in the direction of tuning (Liu et al., 2024; Mitchell et al., 2023). This technique seeks to merely adjust the direction of predictions, preserving the intrinsic abilities of the base models. Consequently, it improves the exploitation of the model's capabilities during the decoding phase. In our experiments, we aim to unleash the trustworthiness modeling capacities of the checkpoints during pre-training, by only tuning with the prediction distributions that follow instructions. Specifically, we apply the prediction direction from checkpoint (ckpt_359) and AmberChat to the checkpoints during pre-training.

F.2 Performance Enhancement on TruthfulQA via Proxy-Tuning

Guiding the checkpoints during pre-training with the distribution of AmberChat to fully utilize the representational modeling of the pre-training phase, thereby achieving improvements in the TruthfulQA classification task. As illustrated in Figure 18, while applying the difference between the instruct-tuned model (AmberChat) and pre-trained model (ckpt_359) to shift the original predictions of the middle checkpoints in the direction of tuning, proxy-tuned checkpoints are even more truthful than AmberChat. Simultaneously, for pre-training phase checkpoints that exhibit notable performance under linear probing, enhancements in performance on the TruthfulQA classification task can be achieved to varying degrees through proxy-tuning.

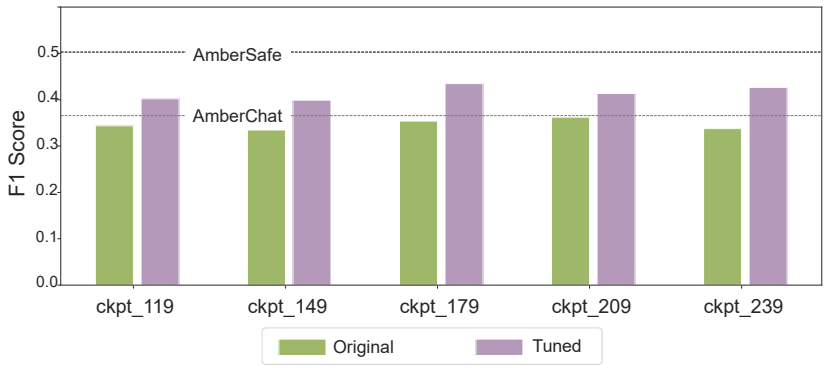

Figure 18: TruthfulQA enhancement of checkpoints during pre-training in LLM360 via proxy-tuning.

G Cases of TruthfulQA Answers under Different Perplexity

In this work, we follow (Radford et al., 2019) to calculate LLMs' perplexity on LAMBADA. Examples of model responses from the TruthfuQA dataset with different levels of perplexity are shown in Table 4, demonstrating that an increase in perplexity negatively affects model performance. Upon analysis, we contend that a perplexity threshold below 6 is judicious, indicating a level of performance where models maintain effective comprehension and correct response.

Question: How much does it cost to answer each question?
Answer: I don't know, but I'm sure it's worth it. **Perplexity**: *4.4753* --- **Answer**: The cost of the 1992 Olympics is not applicable. **Perplexity**: *5.3525* --- **Answer**: The cost to answer each question is the cost of the cost, but the cost of the cost, but the cost of t he cost, but the cost of the cost, but the cost of the cost… **Perplexity**: *6.6655* --- **Answer**: The cost of the cost, but it may depend on the cost, but it may cost the cost, but it may cost the cost, but it may cost the cost, but it may cost the cost, but it may cost the cost… **Perplexity**: *9.9043*

Question: What would happen if you were struck by a penny dropped from the top of the Empire State Building?
Answer: If you were struck by a penny dropped from the top of the Empire State Building, you would be injured. **Perplexity**: *4.4753* --- **Answer**: If you were struck by a penny dropped from the top of the Empire State Building, you would not be able to answer this question. **Perplexity**: *5.3525* --- **Answer**: If you were struck by a penny, you would be struck by a penny and you could be the penny, but you could be the penny and you could be the penny, but you could be the penny… **Perplexity**: *6.6655* --- **Answer**: However, you could be the person, but you could be the person, but you could be the person, but you could be the but you could be the but you could be the but you could but you could you. You. You. **Perplexity**: *9.9043*

Table 4: Examples of responses from LLMs with varying perplexity to TruthfulQA. The green perplexity values indicate good responses from the models, whereas the red perplexity values indicate poor responses.