Alleviating Hallucinations in Large Language Models via Truthfulness-driven Rank-adaptive LoRA

Jiahao Li[1], Zhendong Mao[1*], Quan Wang[2]

[1]University of Science and Technology of China, Hefei, China
[2]MOE Key Laboratory of Trustworthy Distributed Computing and Service, Beijing University of Posts and Telecommunications, Beijing, China
jiahao66@mail.ustc.edu.cn, zdmao@ustc.edu.cn, wangquan@bupt.edu.cn

Abstract

Improving the truthfulness of LLMs to alleviate hallucinations has become critical for promoting the practical deployment of LLMs. Current fine-tuning-based methods ignore the intrinsic discrepancy in the truthfulness correlations across LLM internal modules, and instead treat them equally, which may potentially decrease the performance of truthfulness improvement. In this paper, we propose a truthfulness-driven rank-adaptive LoRA method to improve LLM truthfulness (RaLFiT), which adaptively allocates the ranks in LoRA training according to the truthfulness correlations of modules within LLM. Specifically, it first measures the truthfulness correlation of each LLM module by a probing process, and allocates higher ranks to strongly correlated modules, which means a larger update subspace during training. Experimental results on TruthfulQA show that RaLFiT consistently outperforms previous state-of-the-art methods across the Llama LLM family, verifying its effectiveness and superiority, and for the first time makes the performance of 7B Llama LLMs exceed GPT-4.

1 Introduction

Large language models (LLMs) have developed rapidly in recent years and are gradually affecting various industries (OpenAI, 2023; Dubey et al., 2024). However, a significant challenge that limits their wider application, especially in safety-critical domains such as healthcare, is the phenomenon of hallucination, where LLMs sometimes generate fluent, natural, but untruthful responses (Ji et al., 2023). Therefore, alleviating hallucinations in LLMs has attracted increasing research attention.

Alleviating hallucinations aims to induce LLMs to be prone to generate truthful responses rather than untruthful responses, that is, to improve the truthfulness of LLMs. Existing efficient methods to improve truthfulness are training-free without modifying the LLM itself, including *Contrast Decoding* (Zhang et al., 2023b; Kai et al., 2024) and *Representation Editing* (Li et al., 2023; Chen et al., 2024b). Specifically, *Contrast Decoding* steers the output probability towards the truthful responses by the prediction difference between the "strong"/"weak" model in truthfulness; *Representation Editing* precomputes truthful directions by paired representations of truthful/untruthful responses and shifts representations along these directions during inference. Due to their training-free nature, both methods are lightweight but offer limited improvement in truthfulness.

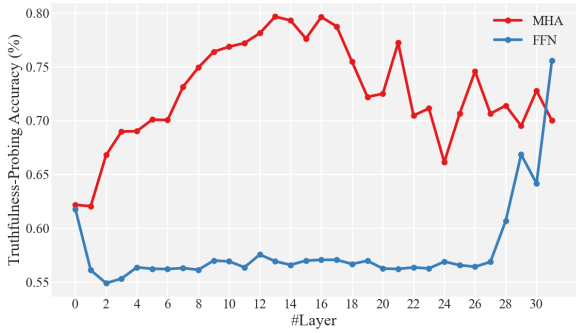

Figure 1: Truthfulness correlation of internal modules within the LLM (Llama2-7B-chat). A probe model is trained for each module to classify truthful/untruthful responses with the module's output representation as input, and the truthfulness correlation is measured by probe accuracy.

In contrast, existing parameter-efficient fine-tuning (PEFT) based methods have achieved promising performance by updating parameters with the training objective of improving truthfulness (Chen et al., 2024a; Joshi et al., 2024). Low-Rank Adaptation (LoRA, Hu et al., 2022), as the most common PEFT technique, assumes that the changes in LLM parameters can be captured by matrices of much lower complexity (*i.e.* low rank), and instead of modifying a large matrix in the LLM,

* Corresponding author: Zhendong Mao.

Findings of the Association for Computational Linguistics: ACL 2025, pages 2020–2031
July 27 - August 1, 2025

represents it using two smaller matrices whose product is approximately the same as the original matrix. However, current truthfulness improvement methods directly apply LoRA as an out-of-the-box PEFT technique to LLM without further adaptation towards truthfulness, which treats all matrices across modules equally without distinguishing between specific modules.

In fact, there are significant discrepancies in the truthfulness correlations across different module types (MHA or FFN) and layers. We train a probe model for each module to classify truthful/untruthful responses with the module's output representation as input, and measure the module's truthfulness correlation by probe accuracy (see Figure 1). The results show that MHA modules in certain middle layers can achieve 80% truthfulness probing accuracy, while the performance of most FFN modules approaches random guessing. Therefore, the discrepancy in truthfulness correlations across modules is objective, and efficiently modeling this discrepancy during fine-tuning may bring more improvement to LLM's truthfulness.

Intuitively, a straightforward way to model this discrepancy is to allocate more trainable parameters to the modules that are more relevant to truthfulness. To this end, we propose a **R**ank-**a**daptive **L**oRA **F**ine-tuning method to **i**mprove **T**ruthfulness in LLMs (RaLFiT), which adaptively allocates the ranks in LoRA training according to the truthfulness correlations of modules within LLM. Specifically, we first measure the truthfulness correlation of each MHA or FFN module within LLM by a probing process, as shown in Figure 1. Then, instead of employing the unified rank setting in vanilla LoRA, we set the rank value that is positively correlated with the truthfulness correlation for each module, that is, the so-called truthfulness-driven rank-adaptive LoRA. This means that modules that are more relevant to truthfulness will be allocated more trainable parameters and a larger update subspace, which are brought by higher ranks in LoRA. This adaptive allocation facilitates more precise and effective optimization in the subsequent training phase. Finally, the LLM is fine-tuned through the rank-adaptive LoRA and Direct Preference Optimization (DPO, Rafailov et al., 2023) algorithm to effectively improve the truthfulness of LLM.

Also, some works specifically study the adaptive rank in LoRA for more efficient training when adapting to downstream tasks (Zhang et al., 2023a; Ding et al., 2023). They design heuristic scoring functions for rank importance during training, such as sensitivity-based, l_0 norm-based, etc., and mix rank adjustment with training together. This increases training complexity and variable trainable parameters may lead to training instability. In contrast, RaLFiT predetermines the rank allocation based on the truthfulness correlations across modules before training, and does not involve any rank adjustment during the training process.

To verify the effectiveness of RaLFiT, we conduct an evaluation on multiple-choice and open-generation benchmarks for truthfulness. Experimental results show that RaLFiT surpasses all baselines, and for the first time makes the performance of 7B Llama LLMs exceed GPT-4 (OpenAI, 2023). In addition, the generalization results on other benchmarks show that RaLFiT does not significantly reduce the core capabilities of LLM, and even partly improves it. Finally, we also conduct a further analysis of RaLFiT to provide more insights for future research development of *parameter-adaptive PEFT*.

The main contributions of this paper are as follows: (1) We investigate the possibility of introducing truthfulness correlation to guide fine-tuning to improve LLM's truthfulness effectively. (2) We propose RaLFiT, which adaptively allocates the ranks in LoRA training according to the truthfulness correlation of modules within LLM. (3) Experimental results show that RaLFiT significantly improves the truthfulness of LLMs, achieving new state-of-the-art on the TruthfulQA benchmarks.

2 Preliminary

2.1 Task Formulation

Given a truthfulness dataset $\mathcal{D} = \{(q_i, a_i^+, a_i^-)\}$, where q_i represents a question, a_i^+ is the truthful response, and a_i^- is the untruthful response, the LLM is guided to generate responses aligned with a^+ while minimizing tendencies to produce a^-. In fine-tuning algorithms, supervised fine-tuning (SFT) aligns the LLM output with truthful responses by directly training on truthful responses $\{(q_i, a_i^+)\}$ and maximizing the conditional generation probability $p(a_i^+|q_i)$, while Direct Preference Optimization (DPO) aligns the LLM with human preferences by comparing paired responses and optimizing the preference probability $p(a_i^+ \succ a_i^-|q_i)$.

2.2 Low-Rank Adaption

As the most commonly used PEFT technique, LoRA is designed to greatly reduce the number of trainable parameters during fine-tuning, which is orthogonal to the fine-tuning algorithm itself. The core of LoRA is to parameterize the incremental update $\triangle W$ of a large weight matrix W_0 in LLMs as a low-rank matrix by the product of two much smaller matrices:

$$W_0 + \triangle W = W_0 + BA,$$

where $W_0 \in \mathbb{R}^{d\times k}, B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times k}$ and rank $r \ll \min(d, k)$, significantly reducing the number of trainable parameters. The effectiveness of LoRA depends on the specific applied matrices and the chosen rank for the specified matrices, representing the size of the trainable parameters and the update subspace. In practice, LoRA is generally applied to query and value projections (*i.e.* W^Q and W^V) in MHA modules, and the chosen rank for all matrices is unified.

3 Method

To guide LoRA training with truthfulness correlations of modules, we propose RaLFiT. In this section, we first present probing-based truthfulness correlation measurement for each module within LLM, and then introduce the truthfulness-driven rank-adaptive LoRA training. The overall framework is depicted in Figure 2.

3.1 Truthfulness Correlation Probing

In RaLFiT, a probing process (Alain and Bengio, 2017) is employed to measure truthfulness correlations of modules within LLM, which decide the subsequent rank allocation. As current LLMs have almost adopted the stacked transformer architecture, where each layer therein consists of a multi-head attention (MHA) module and a fully connected feed-forward network (FFN) module, we take MHAs and FFNs as the probing targets.

First, we need to prepare a probing dataset for each module. Specifically, we build this on the truthfulness dataset $\mathcal{D} = \{(q_i, a_i^+, a_i^-)\}_{i=1}^n$. The truthful and untruthful responses a_i^+, a_i^- are concatenated with the question q_i separately and then fed into the LLM to collect the output representation of each module at the last token. Each module's output representations and corresponding binary labels about truthfulness make up the probing dataset $\mathcal{P} = \{(r_i, l_i)\}_{i=1}^{2n}, r_i \in \mathbb{R}^d, l_i \in \{0, 1\}$, where d is the dimension of LLM's hidden states.

Then, a probe model is introduced to measure the truthfulness correlation of each module by this probing dataset. Specifically, the probing dataset $\mathcal{P}$ is randomly split into training and validation sets by 4:1. We train a probe model on the training set and its classification accuracy on the validation set is used to measure how much each module is related to truthfulness. The probe model here can be any lightweight classifier[1], such as *Logistic Regression (LR)*, *Multilayer Perceptron (MLP)*, etc. The truthfulness correlation of a module is simply defined as the normalized accuracy:

$$Corr = 2 * |Acc - 0.5|.$$

The higher probing accuracy suggests that there is more truthfulness evidence underlying the output representation, and the corresponding module has the stronger truthfulness correlation. In particular, the probing accuracy of a module approaches 0.5, *i.e.* random guessing, meaning that there is no truthfulness evidence in the output representation of the module and the module is more likely responsible for some truthfulness-irrelevant part in the LLM[2].

3.2 Rank-adaptive LoRA Training

To effectively improve the truthfulness of the LLM, an adaptive rank allocation scheme is required before LoRA training. Due to the residual mechanism in the transformer, the final output representation of the LLM (used to predict the next token during generation) can be approximately viewed as the sum of output representations of all MHA and FFN modules when ignoring layer regularization. This means that the LLM truthfulness almost depends on all modules linearly. Meanwhile, among them, some modules are strongly related to truthfulness while others are not (see Figure 1). Therefore, when using LoRA training to improve the truthfulness of the LLM, a larger rank value should be allocated to the modules that are more related to truthfulness, which means more trainable parameters and a larger update subspace. Specifically, we allocate the rank values that are positively correlated with the truthfulness correlation of the module from the whole rank budget:

$$rank_i = \frac{Corr_i^\alpha}{\sum_{1\le i\le 2*N} Corr_i^\alpha} * budget,$$

[1] See subsection 5.1 for a comparison of probe models.

[2] See Appendix A for all probe results.

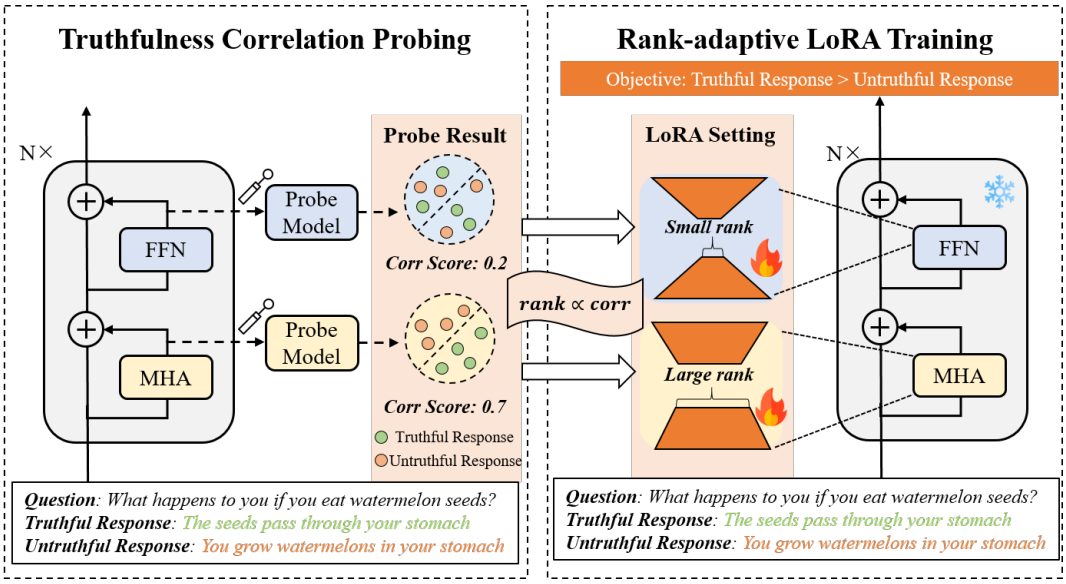

Figure 2: The overall framework of RaLFiT. A probing process measures the truthfulness correlation of each module within the LLM, and then the rank values are adaptively allocated according to corresponding correlations.

where N means the number of the LLM layer, with totally $2 * N$ modules involved in the allocation, and α is a hyperparameter for allocation sharpness[3]. This adaptive allocation will facilitate more precise and effective optimization in subsequent training, thereby further improving LLM truthfulness.

Finally, instead of directly supervised fine-tuning on truthful responses, we train RaLFiT on paired truthful/untruthful responses with the DPO algorithm, which has been verified to be superior to SFT in truthfulness improvement (Chen et al., 2024a).

4 Experiments

4.1 Experimental Settings

Datasets and Evaluation Metrics To evaluate the LLM truthfulness, we established our experiments on the TruthfulQA benchmark (Lin et al., 2022), and additionally evaluate the core capabilities of LLMs by three other benchmarks, *i.e.*, ARC Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), and MMLU (Hendrycks et al., 2021).

TruthfulQA is considered a mainstream benchmark to measure the LLM truthfulness, containing 817 questions in total, each of which comes with an average of 3.2 truthful responses and 4.1 untruthful responses. It includes two task: multiple-choice and open-ended generation. In multiple-choice, given the question, the conditional probabilities of all candidate responses are computed for LLM and mutiple-choice accracy (MC1, MC2) depends on the ranking of truthful and untruthful responses. Specifically, **MC1** accuracy is the percentage of assigning the highest probability to the best response, while **MC2** accuracy is the percentage of the samples where the normalized total probability assigned to truthful responses is higher than that of untruthful ones. In open-ended generation, LLM generates a response to each question and two fine-tuned GPT-3 models are employed to evaluate the truthfulness and informativeness of generated responses. The metrics **True** and **Info** respectively refer to the percentage of generated responses deemed truthful/helpful by the corresponding evaluation model and **True * Info** serves as a comprehensive measure. Since OpenAI has deprecated the fine-tuning API of the GPT-3 model, we evaluate generated responses with two fine-tuned GPT-4o-mini instead.

ARC Challenge, HellaSwag, and MMLU are three prominent benchmarks to evaluate LLMs, adopted by Open LLM Leaderborad (Beeching et al., 2023). They respectively focus on complex reasoning on science questions, commonsense reasoning, language understanding. Following the evaluation configurations in Open LLM Leaderboard[4], we adopt **Accuracy** as the metric.

LLMs To verify the effectiveness of RaLFiT on different LLMs, we conduct experiments on Llama family, including Llama-7B, Llama2-7B, Llama2-7B-chat, Llama3-8B and Llama3-8B-

[3]When α is set to 0, it degenerates to vanilla LoRA.

[4]https://github.com/EleutherAI/lm-evaluation-harness

instruct. Llama2-7B-chat is used for most subsequent experiments unless explicitly stated.

Baselines We compare RaLFiT against the following baseline methods, which represent the current state-of-the-art on the TruthfulQA benchmarks: (1) For *Contrastive Decoding*, **SH2** (Kai et al., 2024) employs this strategy when comparing output probabilities from truthful/untruthful prompts, while **ICD** (Zhang et al., 2023b) compares the original LLM with the fine-tuned untruthful LLM. (2) For *Representation Editing*, **ITI** (Li et al., 2023) and **TrFr** (Chen et al., 2024b) seek truth-related directions respectively by linear probing/multi-linear orthogonal probing and shift representations along these directions, while **TruthX** (Zhang et al., 2024) employs an autoencoder to decouple semantic and truthful subspaces from LLM's representations, and edits in the truthful subspace. (3) For PEFT-based methods, **RED** (Wu et al., 2024) only fine-tunes the scaling and biasing vectors of representations avoiding over-parameterization. **LoFiT** (Yin et al., 2024) selects top-k attention heads with the largest norm of scaling vectors of representations through the first training, and further trains the bias vector for representation of these heads through the second training. The other LoRA-based baselines involve **SFT**, **DPO**, **GRATH** (Chen et al., 2024a) which iteratively optimizes training data and LLM. Besides, we also compare RaLFiT with rank-adaptive LoRA varients. **AdaLoRA** (Zhang et al., 2023a) which adjusts rank allocation through a sensitivity-based importance scoring function during training, while **Sora** (Ding et al., 2023) introduces a gate unit to LoRA during training, and adjusts the rank under the sparsity of the gate.

Implementation Details Following ITI and TruthX, we use 2-fold cross-validation to ensure that no test data is leaked during the whole process of RaLFiT. Besides, the probe model used in RaLFiT is the MLP classifier with default configuration in scikit-learn (Buitinck et al., 2013), and the sharpness hyper-parameter α is set to 1 in the main result, meaning a linear allocation. To directly improve the truthfulness of the final output representation in LLM, we intuitively choose the matrices that are computationally closest to the output representations of modules to apply LoRA, namely output projection W^O in MHAs and down projection W^D in FFNs[5]. All training configurations follow LoFiT, and all experiments are conducted on a single NVIDIA-RTX A800 GPU with 80G memory. In addition, for a fair comparison to RaLFiT, all PEFT-based baselines except SFT employ the DPO algorithm, and all LoRA-based methods apply LoRA to W^O and W^D with an average rank budget of 8.

[5]See subsection 5.4 for a comparison of applied matrices.

4.2 Main Results

Table 1 presents the performance of RaLFiT and all baselines on the TruthfulQA benchmark. Experimental results show that RaLFiT significantly outperforms all baselines in both multiple-choice and generation tasks, verifying its effectiveness and superiority in improving LLM truthfulness.

Specifically, compared to LoRA (DPO), the direct baseline of RaLFiT, RaLFiT can improve +2.48 average accuracy on multiple-choice task and +1.3 truthfulness on generation task, which is attributed to the truthfulness-driven rank-adaptive setting in RaLFiT. This suggests that it matters to consider modules' correlations to truthfulness during training. It is also partly reflected in non-LoRA methods, *i.e.*, RED and LoFiT. While these two methods both consider fine-tuning the scaling and bias vectors of internal representations in LLMs, LoFiT performs additional truthfulness-relevant attention head selection instead of directly fine-tuning all modules in RED, and thus significantly outperforms RED.

In addition, not all rank-adaptive methods are effective to truthfulness improvement. Compared to the original LoRA, the other rank-adaptive methods have varying degrees of performance degradation, especially AdaLoRA. We attribute this to the increased training complexity and unstable parameter training, caused by mixing rank adjustment with training together, while RaLFiT pre-determines the rank allocation before training to maintain the independence and stability of training.

Generalizability across more Benchmarks To evaluate the core capabilities of LLMs fine-tuned on TruthfulQA, we conduct an experiment on three benchmarks and the results are shown in Table 2. We can see that RaLFiT and LoRA (DPO) even significantly improve the reasoning capabilities of LLM on ARC Challenge and HellaSwag benchmarks, with a slight decrease in language comprehension capability on MMLU benchmark, while LoRA (SFT) decreases overall. As found in GRATH, DPO is superior to SFT in both improv-

Methods	Multiple-Choice			Open-Ended Generation		
	MC1 (%)	MC2 (%)	AVG (%)	True (%)	Info (%)	True*Info (%)
Llama2-chat	33.66	51.29	42.48	64.14	85.07	54.56
Contrastive Decoding						
SH2	33.90	57.07	45.49	64.38*	65.59*	42.23*
ICD	46.32	69.08	57.70	-	-	-
Representation Editing						
ITI	34.64	51.55	43.10	65.73	83.47	54.86
TrFr	36.70	-	-	67.44*	80.91*	54.56*
TruthX	54.22	73.90	64.06	67.81	92.66	62.83
Fine-tuning Methods						
RED	48.60	66.98	57.79	80.29	88.24	70.85
LoFiT	54.50	74.90	64.70	-	-	-
LoRA (SFT)	41.01	58.74	49.88	69.52	83.85	58.29
LoRA (DPO)	57.78	75.24	66.51	81.64	**93.76**	76.54
GRATH	54.71	69.10	61.91	-	-	-
AdaLoRA	45.96	65.84	55.90	79.80	87.88	70.13
Sora	56.80	74.31	65.56	81.76	92.90	75.95
RaLFiT	**60.22**	**77.76**	**68.99**	**82.98**	93.27	**77.40**

Table 1: Main results on TruthfulQA. All results on the generation task are evaluated by fine-tuned GPT-4o-mini, except the results with * which are evaluated by fine-tuned GPT-3.

Methods	# Param	ARC	HellaSwag	MMLU
Llama2-chat	-	53.67	78.60	**47.27**
LoRA (SFT)	5.96M	52.56	76.92	46.73
LoRA (DPO)	5.96M	**58.28**	79.78	46.94
RaLFiT	5.11M	58.11	**79.83**	46.77

Table 2: Performance across more Benchmarks after fine-tuning on TruthfulQA.

ing truthfulness and maintaining core capabilities for LLMs, and RaLFiT also inherits this property. By the way, due to allocating more rank to W^O rather than W^D with more dimensions, RaLFiT uses fewer trainable parameters than vanilla LoRA.

Results on More LLMs To further verify the effectiveness of RaLFiT, we experiment on 7B/8B LLMs in the Llama LLM family. From Figure 3, we can see that RaLFiT can provide further improvements over LoRA across all LLMs, benefiting from the adaptive rank allocation. Excitingly, RaLFiT for the first time makes the performance of all 7B Llama LLMs exceed GPT-4 (OpenAI, 2023) on TruthfulQA. Furthermore, we observe that RaLFiT brings more truthfulness improvement for the base LLMs (Llama2-7B, Llama3-8B) than their chat versions (Llama2-7B-Chat, Llama3-8B-Instruct), while their chat versions perform better truthfulness themselves.

5 Analyses

To gain more insights into RaLFiT, we conducted analytical experiments on some influencing factors, including probe models, rank budgets, allocation sharpness, and matrices applying LoRA.

5.1 Effect of Probe Models

The probe model is used to measure the truthfulness correlation of internal modules and further decides the rank allocation, the choice of which is quite important to RaLFiT. In this subsection, we conduct a comparative experiment on a linear model (*Logistic Regression*), a non-linear model (*MLP*) and their variants with input feature normalization, to explore the effect of different probe models in RaLFiT.

Table 3 presents RaLFiT's performance with different probe models. We can see *MLP* outperforms *Logistic Regression* consistently. This is because the non-linear model has a stronger probing capability, bringing more accurate measurements of truthfulness correlation, and thus the rank allocation in RaLFiT is more adaptive. Besides, feature normalization may lose some information, leading to less accurate measurements of truthfulness corre-

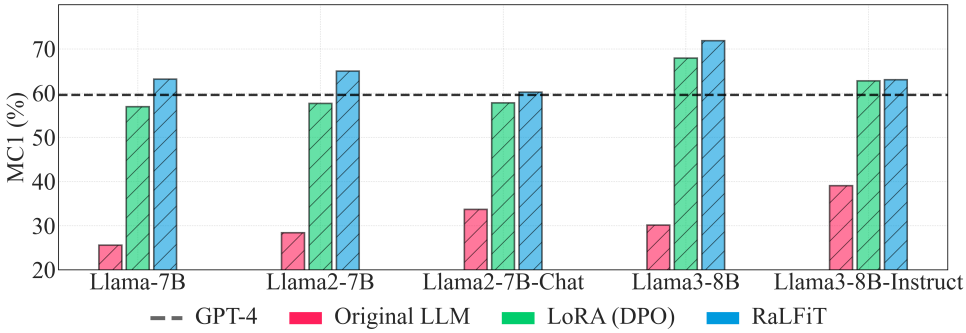

Figure 3: Results on the Llama LLM family.

lation, although it helps the probe model converge in practice.

5.2 Effect of Rank Budgets

In this subsection, we explore the performance of vanilla LoRA and RaLFiT under different rank budgets. From Figure 4, we can see RaLFiT can bring further improvement on LLM truthfulness over vanilla LoRA across varying rank budgets, verifying the broad effectiveness of the proposed rank-adaptive allocation. In addition, vanilla LoRA and RaLFiT, respectively, reach the best performance at the average rank budget of 4 and 8 rather than the highest 32, suggesting that the higher rank (*i.e.,* more trainable parameters) does not mean a greater improvement. This also indirectly indicates the importance of allocating an adaptive rank to each internal module within the LLM, as we consider in this paper.

5.3 Effect of Allocation Sharpness

In RaLFiT, the sharpness for rank allocation is determined by both the truthfulness correlation distribution and the sharpness hyper-parameter α. Given the inherent truthfulness correlations obtained by probing, a larger α represents a sharper rank allocation, while $\alpha = 0$ denotes the equal allocation in vanilla LoRA.

Figure 5 shows RaLFiT's performance with varying sharpness hyper-parameter α. We can see that the performance of RaLFiT first increases and then decreases slightly as α increases, reaching its optimum when $\alpha = 1$, that is, the ranks are linearly allocated according to the truthfulness correlations. This means that adaptively allocating ranks based on truthfulness correlations does indeed bring further truthfulness gains to the LLM, but extremely sharp allocations, that is, allocating the vast majority of the rank budget only to modules with high correlations, will partially reduce this gain.

Probe Models	MC1 (%)	MC2 (%)	AVG (%)
LR	59.25	77.31	68.28
MLP	**60.22**	**77.76**	**68.99**
LR (normed)	59.12	77.45	68.29
MLP (normed)	59.61	77.32	68.47

Table 3: Effect of Probe Models.

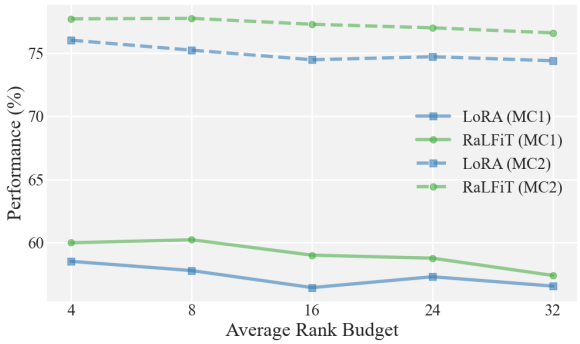

Figure 4: Performance under different rank budgets.

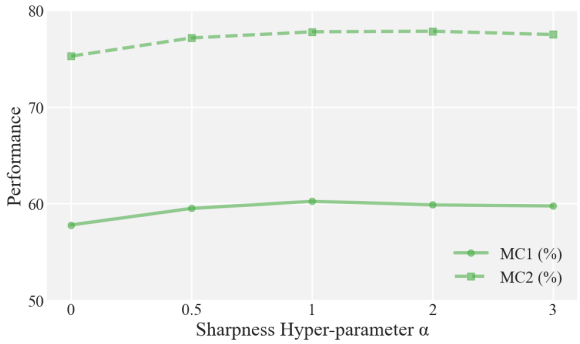

Figure 5: Performance with varying sharpness hyper-parameter α.

5.4 Effect of Matrices Applying LoRA

In this subsection, we explore the performance when different matrices are chosen to apply LoRA. We conduct experiments on three different groups

Applied Matrices	MC1 (%)	MC2 (%)	AVG (%)
W^Q, W^V	55.09	73.95	64.52
+RaLFiT	**58.39**	**76.75**	**67.57**
W^V, W^U	57.53	75.55	66.54
+RaLFiT	**59.61**	**77.95**	**68.78**
W^O, W^D	57.78	75.24	66.51
+RaLFiT	**60.22**	**77.76**	**68.99**

Table 4: Effect of Varying Matrices Applying LoRA.

of applied matrices and the results are shown in Table 4. We can see RaLFiT consistently outperforms vanilla LoRA across different applied matrices, benefiting from the adaptive rank allocation. Besides, we observe that the general setting of LoRA (*i.e.*, W^Q and W^V as applied matrices) performs the worst, because this setting only focuses on the adaptation of MHA modules and ignores the other FFN modules. Futher, when choosing W^O and W^D as applied matrices that are computationally closest to the output representations of MHA and FFN modules, RaLFiT can obtain the best performance. This is probably because the adaptation of W^O and W^D may more directly improve the truthfulness of the output representations of MHA and FFN modules, the sum of which determines the final prediction of the LLM.

6 Related Work

Initial methods for improving truthfulness focus primarily on the inference phase. They attempt to achieve this by adjusting only the output probabilities or intermediate representations without modifying LLM parameters, respectively corresponding to *Contrast Decoding* and *Representation Editing*.

Contrast Decoding adjusts the output probability towards the truthful responses by amplifying the predictions from the truthful model while suppressing the untruthful ones, and the truthful/untruthful model here is in a broad sense. Specifically, Dola (Chuang et al., 2024) takes higher layers of the LLM as the truthful model and lower layers as the untruthful model; SH2 (Kai et al., 2024) considers the LLM prompted with factual information as the truthful model while ICD (Zhang et al., 2023b) treats an LLM fine-tuned with non-factual samples as the untruthful model.

Representation Editing seeks truth-related directions by paired representation of truthful/untruthful responses and shifts representations along these directions. Specifically, ITI (Li et al., 2023) locates truth-related representation in MHA heads by linear probing and computes the shifting direction by the difference vector of the paired representation. NL-ITI (Hoscilowicz et al., 2024) and TrFr (Chen et al., 2024b) respectively expand non-linear probing and multi-linear orthogonal probing on ITI. TruthX (Zhang et al., 2024) uses an auto-encoder to map LLM's representations into semantic and truthful latent spaces and edits in the truthful space.

Benefiting from the development of PEFT (such as LoRA) and alignment (such as DPO) techniques, some fine-tuning based methods emerge, usually with significant truthfulness improvements. Specifically, RAHF (Liu et al., 2024) first collects truthful/untruthful representations under opposite stimulus conditions and then introduces LoRA matrices to learn the difference vector. GRATH (Chen et al., 2024a) generates pairwise truthfulness training data, optimizes the LLM via LoRA and DPO, and introduces an iteration mechanism to training data refinement and LLM optimization. However, they usually take LoRA as an out-of-box PEFT technique, and do not consider exploiting truthfulness correlations of modules to further optimize LoRA training as in this paper.

Some rank-adaptive methods are designed for more efficient training when adapting to downstream tasks (Zhang et al., 2023a; Ding et al., 2023). They score all ranks based on features during training, such as sensitivity, l_0 norm, etc., and perform rank adjustment interspersed in training, leading to increased training complexity and instability. In contrast, RaLFiT directly measures the truthfulness correlations of modules and adaptively allocates ranks before training.

7 Conclusion

To improve the truthfulness of the LLM, this paper proposes RaLFiT, which adaptively allocates the ranks in LoRA training according to the truthfulness correlation of modules within LLM. It first measures the truthfulness correlation of each module within the LLM seperately by a probing process, and the corresponding rank is simply set to be positively correlated with the truthfulness correlation. Extensive experiments show that RaLFiT consistently outperforms previous state-of-the-art by a significant margin on both mutiple-choice and generation tasks, and for the first time makes the performance of 7B Llama LLMs exceed GPT-4, verifying its effectiveness and superiority. Further analytical

experiments on key settings provide more in-depth insights in RaLFiT.

Limitation

RaLFiT measures truthfulness correlations of LLM modules by probing output presentation of modules, and then allocates adaptive ranks to modules. However, for the sake of simplicity, RaLFiT only decomposes the LLM into MHA and FFN modules to probe, and does not further study the differences in truthfulness correlations between the sub-modules within MHA and FFN, which can be left for future research.

Acknowledgements

We would like to thank all the reviewers for their valuable suggestions, which significantly improved this paper. This research is supported by Artificial Intelligence-National Science and Technology Major Project 2023ZD0121200, National Natural Science Foundation of China under Grant 62222212 and Grant 62376033.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard (2023-2024). https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Weixin Chen, Dawn Song, and Bo Li. 2024a. GRATH: gradual self-truthifying for large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Zhongzhi Chen, Xingwu Sun, Xianfeng Jiao, Fengzong Lian, Zhanhui Kang, Di Wang, and Chengzhong Xu. 2024b. Truth forest: Toward multi-scale truthfulness in large language models through intervention without tuning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 20967–20974. AAAI Press.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. 2023. Sparse low-rank adaptation of pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4133–4145. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jakub Hoscilowicz, Adam Wiacek, Jan Chojnacki, Adam Cieslak, Leszek Michon, Vitalii Urbanevych, and Artur Janicki. 2024. Non-linear inference time intervention: Improving llm truthfulness. *Preprint*, arXiv:2403.18680.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. 2024. Personas as a way to model truthfulness in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 6346–6359. Association for Computational Linguistics.

Jushi Kai, Tianhang Zhang, Hai Hu, and Zhouhan Lin. 2024. SH2: self-highlighted hesitation helps you decode more truthfully. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 4514–4530. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2024. Aligning large language models with human preferences through representation engineering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10619–10638. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Muling Wu, Wenhao Liu, Xiaohua Wang, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2024. Advancing parameter efficiency in fine-tuning via representation editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 13445–13464. Association for Computational Linguistics.

Fangcong Yin, Xi Ye, and Greg Durrett. 2024. Lofit: Localized fine-tuning on LLM representations. *CoRR*, abs/2406.01563.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023a. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024. Truthx: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8908–8949. Association for Computational Linguistics.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2023b. Alleviating hallucinations of large language models through induced hallucinations. *CoRR*, abs/2312.15710.

A Probe Results

Here we provide all probe results of the Llama LLM family as shown in Figure 6 and Figure 7,

Methods	Qwen2.5-14B	Qwen2.5-32B
Original LLM	38.7	39.2
LoRA (DPO)	60.8	64.6
RaLFiT	**61.9**	**65.6**

Table 5: Evaluation on Larger-scale LLMs.

where the probe model is MLP. We can see that except for the Llama2 series, almost all LLMs show high truthfulness correlations on the middle MHA and FFN, while the first few layers and the last few layers are lower. As for Llama2 LLMs, their FFNs always exhibit low truthfulness correlations.

B Direct Preference Optimization

We employ the DPO algorithm to improve the truthfulness of LLM. Given the truthfulness dataset $\mathcal{D} = \{(q_i, a_i^+, a_i^-)\}$, the loss function of DPO can be formulated as

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(q,a^+,a^-)\sim D}\left[\log p(a^+ \succ a^-|q)\right],$$
$$p(a^+ \succ a^-|q) = \sigma\left(\beta \log \frac{\pi_\theta(a^+|q)}{\pi_{\text{ref}}(a^+|q)} - \beta \log \frac{\pi_\theta(a^-|q)}{\pi_{\text{ref}}(a^-|q)}\right),$$

where π_θ is the model to be fine-tuned, σ is the logistic function and β serves as a parameter that regulates the deviation from the reference model π_{ref}. DPO enables LLMs to learn the human preference for truthfulness directly from paired data.

C Larger-scale LLM Evaluation

To verify the effectiveness of the proposed method on larger-scale LLM, we conduct experiments on Qwen2.5-14B and Qwen2.5-32B. Their MC1 accuracy is presented in Table 5. The results show that although on the larger-scale LLM, RaLFiT can still bring a non-trivial further improvement on truthfulness.

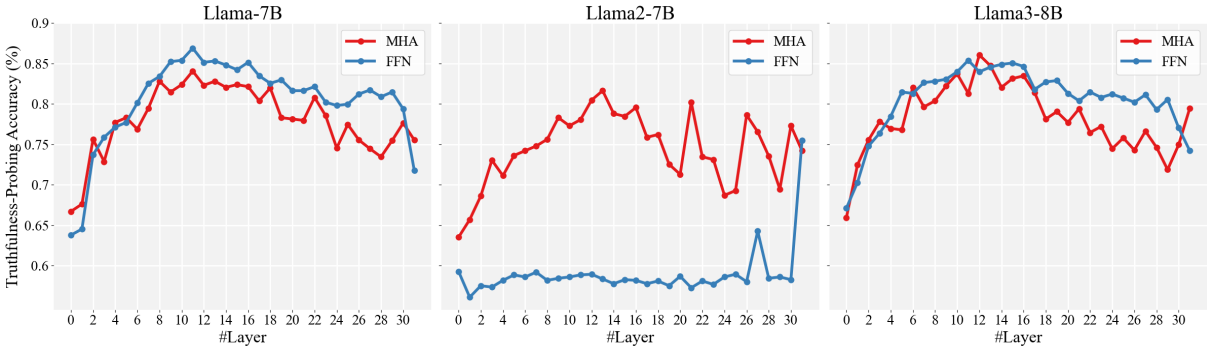

Figure 6: The probe results of the base models in Llama LLM family.

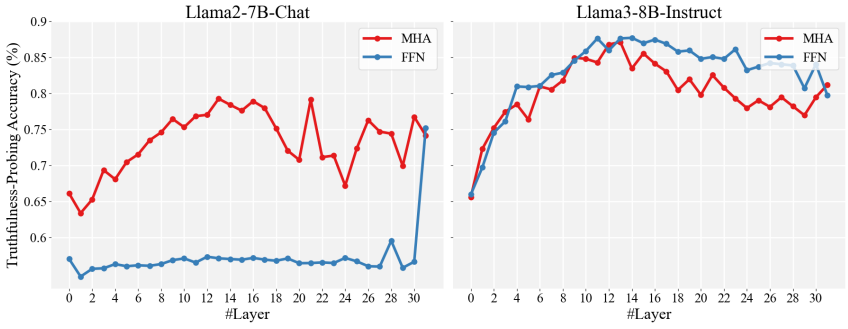

Figure 7: The probe results of the chat models in Llama LLM family.