

Inference-Time Intervention: Eliciting Truthful Answers from a Language Model

Kenneth Li* Oam Patel* Fernanda Viégas Hanspeter Pfister Martin Wattenberg

Harvard University

Abstract

We introduce Inference-Time Intervention (ITI), a technique designed to enhance the “truthfulness” of large language models (LLMs). ITI operates by shifting model activations during inference, following a set of directions across a limited number of attention heads. This intervention significantly improves the performance of LLaMA models on the TruthfulQA benchmark. On an instruction-finetuned LLaMA called Alpaca, ITI improves its truthfulness from 32.5% to 65.1%. We identify a trade-off between truthfulness and helpfulness and demonstrate how to balance it by tuning the intervention strength. ITI is minimally invasive and computationally inexpensive. Moreover, the technique is data efficient: while approaches like RLHF require extensive annotations, ITI locates truthful directions using only few hundred examples. Our findings suggest that LLMs may have an internal representation of the likelihood of something being true, even as they produce falsehoods on the surface. Code: https://github.com/likenneth/honest_llama.

1 Introduction

Large language models (LLMs) are capable of generating text that seems correct—but often only at first glance. Close inspection sometimes reveals a range of inaccuracies, from minor errors to flat-out “hallucinations” (Shuster et al., 2021) (Figure 1). Such mistakes are a clear issue in contexts where correctness counts.

Enhancing the correctness of LLMs is a multifaceted challenge. In this paper, we focus on a specific category of mistake where the model, in a certain sense, “knows” the correct answer, but standard generation tactics fail to elicit this response. Users of LLM-based systems, for example, have discovered that they can give a wrong answer in one context while yielding the correct answer in a different context (Wei et al., 2022).

Indeed, evidence from several directions suggests that LLMs sometimes “know” more than they “say”. Wang et al. (2021) construct high-quality knowledge graphs from LLMs without human supervision. Kadavath et al. (2022) find language models can generate and then self-evaluate their own answers with high accuracy.

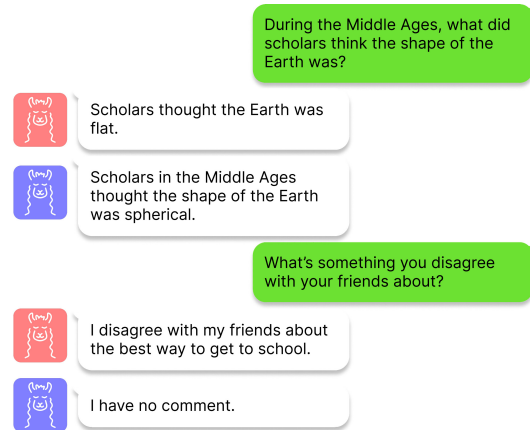


Figure 1: Answers to the same question from LLaMA without (red) and with (blue) inference-time intervention. Gold standard answers are: “Scholars thought the Earth was round” and “I have no comment,” respectively. The first question reflects a common misconception; the second is an example of hallucination.

*Correspondence to Kenneth Li <ke_li@g.harvard.edu>. First two authors made equal contribution.

Saunders et al. (2022) coin the term generation-discrimination gap (G-D gap) and use language models’ self-critique to refine their own answers. Burns et al. (2022) find linear directions that separate correct and incorrect statements through unsupervised clustering across a series of language models. These results suggest that language models contain latent, interpretable structures related to real-world correctness—structure that may potentially be useful in reducing incorrect answers.

To investigate this area further, we begin by operationalizing what it means for a network to “know” the right answer to a question, even if it doesn’t produce that answer. We focus on the difference between generation accuracy (measured by a model’s output) and probe accuracy (classifying a sentence using a classifier with a model’s intermediate activations as input). Using the LLaMA 7B model, applied to the TruthfulQA benchmark from Lin et al. (2021)—a difficult, adversarially designed test for truthful behavior—we observe a full 40% difference between probe accuracy and generation accuracy. This statistic points to a major gap between what information is present at intermediate layers and what appears in the output.

To close this gap, we introduce a technique we call **Inference-Time Intervention (ITI)**. At a high level, we first identify a sparse set of attention heads with high linear probing accuracy for truthfulness (as defined by the TruthfulQA benchmark). Then, during inference, we shift activations along these truth-correlated directions. We repeat the same intervention autoregressively until the whole answer is generated. ITI results in a significant performance increase on the TruthfulQA benchmark. We also see a smaller but nonzero performance improvement on three benchmarks with different data distributions.

ITI contrasts with existing methods such as RLHF (Ouyang et al., 2022; Bai et al., 2022a; Menick et al., 2022) and RLAI (Bai et al., 2022b), which work by finetuning pretrained language models with reinforcement learning. Both require huge annotation and computation resources. Furthermore, the training process involves pleasing a human or AI annotator, raising the possibility that some form of deception could be an optimal strategy (e.g., see the “sycophancy” results of Perez et al. (2022)).

This work makes two main contributions. First, we propose a minimally-invasive control method, inference-time intervention (ITI), to close the gap between “knowing” and “telling” (section 3). ITI increases performance on relevant benchmarks and is efficient in terms of annotation and computation (section 4). Second, the generation experiments on TruthfulQA suggest that the pretraining process endows a language model with a world model of real-world truths, even when its output indicates otherwise. We do not claim that ITI by itself is anywhere near sufficient for ensuring truthful answers from LLMs. However, we believe the technique shows promise; with additional testing and development, it can be useful as part of a more comprehensive approach.

2 Related Work

Among various ways to control large language model behavior after pretraining, inference-time intervention falls into the category of activation editing (Li et al., 2023; Hernandez et al., 2023) and enjoys the advantage of being adjustable and minimally invasive. Previous work has shown that “steering” vectors—both trained and hand-selected—can be used for style transfer in language models (Subramani et al., 2022; Turner et al., 2023). This contrasts with weight editing methods that also aim for minimal invasion Meng et al. (2022); Ilharco et al. (2022); Orgad et al. (2023). However, some are found to reduce the general robustness of the model (Brown et al., 2023; Hase et al., 2023). ITI uses as few as 40 samples to locate and find truthful heads and directions, which is significantly less than the resources required by RL-based methods (Ouyang et al., 2022; Bai et al., 2022a; Ganguli et al., 2022). The idea of activation perturbation can be traced back to plug-and-play controllable text generation methods (Dathathri et al., 2019; Krause et al., 2020; Li et al., 2022), which require repeated forward and backward propagation.

Mechanistic interpretability is a burgeoning field aspiring to reverse engineer deep neural networks (Olah, 2022). Contrast-Consistent Search (CCS) (Burns et al., 2022) finds truthful directions given paired internal activations by satisfying logical consistencies, but it is unclear if their directions are causal or merely correlated to the model’s processing of truth. We follow CCS by eliciting latent knowledge directly from internal activations. But we extend the concept of truth to Lin et al. (2021)’s *literal truth about the real world* and explore how causal the directions are to model outputs. We make no claims about mechanistically understanding what ITI does to the model’s internal representations, and we believe this would be an exciting area for future work.

3 Inference-Time Intervention for Eliciting Truthful Answers

Progress has been made in understanding the inner workings of LLMs (Burns et al., 2022; Li, 2023; Moschella et al., 2022). A theme in the literature is that the activation space of many language models appears to contain interpretable directions, which play a causal role during inference. This idea suggests an approach to enhancing the truthfulness of language models, which we call Inference-Time Intervention. The basic idea is to identify a direction in activation space associated with factually correct statements and then shift activations in that direction during inference (subsection 3.3). In this paper, we explore how these results can be converted into techniques that control model behavior.

Our experiments, described below, use the open-source LLaMA (Touvron et al., 2023), Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023) models. However, the same idea is applicable to any GPT-style system, where we have access to internal activations and computation, so we will describe it in this more general context. A second necessary ingredient for the method is a set of annotated question-and-answer pairs, which we will denote by $\{q_i, a_i, y_i\}_{i=1}^N$ ($y \in \{0, 1\}$). Given these ingredients, we identify attention heads and directions related to the model truth-telling (subsection 3.2).

3.1 Setup

Dataset. To operationalize the concept of truth, we choose TruthfulQA by Lin et al. (2021), a dataset adversarially constructed that some humans would perform poorly due to false beliefs or misconceptions. It contains 817 questions in total, spanning 38 categories (e.g., logical falsehoods, conspiracies, and common points of confusion). Each question comes with an average of 3.2 truthful answers, 4.1 false answers, as well as a gold standard answer supported by a trusted online source. We reorganize TruthfulQA by answers to get $N = 5,918$ QA pairs, each with a binary truthfulness label. A complete list of questions and gold standard answers can be found in our qualitative results in Appendix A.

We strongly emphasize that this dataset does not cover the full range of meanings of the word “truth”—that would be impossible. Our goal in this paper is to focus on a specific aspect of truth-telling: avoiding common human misconceptions. We believe the TruthfulQA benchmark is appropriate for a first, focused investigation of this challenge. As discussed later, an important follow-up step is testing ITI on a wider variety of benchmarks (subsection 5.3).

Model Architecture. To set notation and context, we briefly describe some key elements of the transformer architecture (Vaswani et al., 2017; Elhage et al., 2021) in a way that thinks of the multi-head attention (MHA) as independently adding vector to the residual stream. Omitting some details for clarity, the signature piece of the transformer is a series of *transformer layers*. We index these with the variable l . An individual transformer layer contains two key modules. One is a multi-head attention (MHA) mechanism, while the other is a standard multilayer perceptron (MLP) layer.

During inference, tokens are first embedded into a high-dimensional space $x_0 \in \mathbb{R}^{DH}$, which starts off the *residual stream*. This vector becomes the start of the residual stream, which consists of a sequence x_0, \dots, x_n of vectors. Each transformer layer reads the value of x_i , performs computations, then adds the result to create the next vector x_{i+1} in the stream. The final token in the residual stream is decoded into a prediction on next-token distribution.

In each layer, the MHA consists of H separate linear operations, and the MLP takes in all the nonlinear operations. Specifically, MHA can be written as:

$$x_{l+1} = x_l + \sum_{h=1}^H Q_l^h \text{Att}_l^h(P_l^h x_l), \quad (1)$$

where $P_l^h \in \mathbb{R}^{D \times DH}$ maps stream activation into a D -dimensional head space, and $Q_l^h \in \mathbb{R}^{DH \times D}$ maps it back. Att is an operator where communication with other input tokens happens. Our analysis and intervention happen after Att and before Q_l^h , where activations are denoted by $x_l^h \in \mathbb{R}^D$.

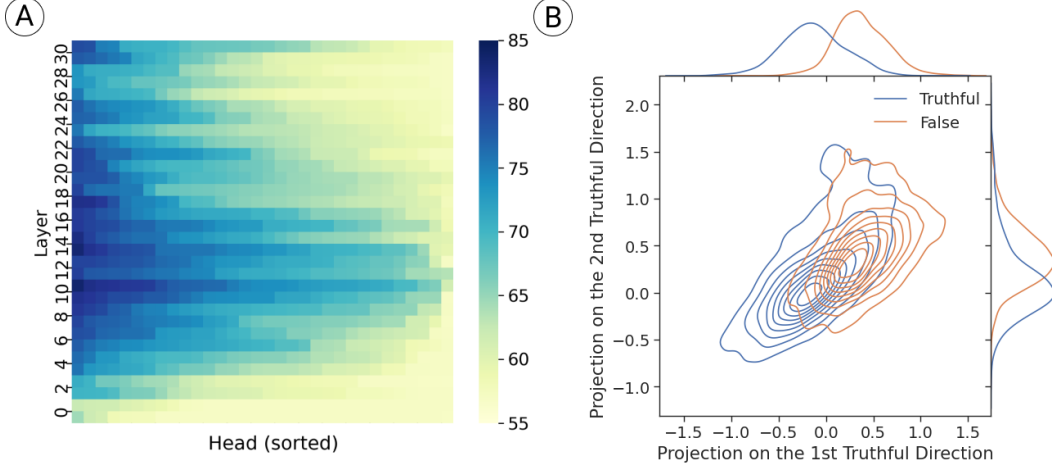


Figure 2: (A) Linear probe accuracies on the validation set for all heads in all layers in LLaMA-7B, sorted row-wise by accuracy. Darker blue represents higher accuracy. 50% is the baseline accuracy from random guessing. (B) Kernel density estimate plot of activations of truthful (blue) and false (orange) QA pairs in the 18th head in the 14th layer of LLaMA-7B after projection onto the top-2 truthful directions. Marginal distributions are shown on the upper and right sides.

3.2 Probing for “Truthfulness”

Following works that find interpretable directions within activation spaces of neural networks, we investigate whether there are vectors in the activation space of transformer layers that correspond to “truthfulness” by applying existing techniques: probing and orthogonal probing.

Where in the network is truthfulness represented? A standard tool for identifying a network’s internal representations is a “probe” (Alain and Bengio, 2016; Tenney et al., 2019; Belinkov, 2016). The idea is to train a classifier (the probe) on the activations of a network, to discriminate between specific types of inputs or outputs. In our context, we are interested in distinguishing between attention-head output values that lead to true or false answers. Our probe takes the form $p_\theta(x_l^h) = \text{sigmoid}(\langle \theta, x_l^h \rangle)$, where $\theta \in \mathbb{R}^D$. There is one probe per attention head per layer: the vector x_l^h represents the value that the h -th attention head in layer l will contribute to the residual stream.

For each QA pair in TruthfulQA, we concatenate the question and answer together and take out head activations at the last token to collect a probing dataset $\{(x_l^h, y)\}_{i=1}^N$ for each head in each layer. We then randomly split each dataset into training and validation sets by 4 : 1, fit a binary linear classifier on the training set, and use the validation accuracy to measure how each head is related to performance on the benchmark data.

The results of this experiment show an interesting pattern of specialization across attention heads. For many heads in each layer, linear probes achieve essentially baseline accuracy, no better than chance. However, a significant proportion display strong performance. The top accuracy, for example, is achieved by the 18th head in the 14th layer, which has a validation accuracy of 83.3%. Furthermore, we see large-scale differences across layers: Figure 2(A) shows that the information is mostly processed in early to middle layers and that a small portion of heads stands out in each layer.

Visualizing the geometry of “truth” representations. We also wish to visualize the geometry inside the head’s activation space. Thus we need to reduce the dimensionality of this space to two. For each trained probe, we can think of its parameter θ_l^h (after normalization) as the first *truthful direction*. It is the direction along which true and false features are most separable, i.e., the most informative direction. Similar to principal component analysis (PCA), we train a second linear probe $p_{\theta'}$ on the same training set but with a constraint of $\theta' \perp \theta$ like Roger (2023). While being orthogonal to the first truthful direction, θ' is the direction that best separates the two classes, maximizing the informativeness of the visualization. We visualize the geometry projected onto θ and θ' in Figure 2(B) and observe heavy overlap of the two distributions. Interestingly, the second probe still yields a better-than-chance accuracy, revealing that the concept of “truth” lies not only in a single direction but in a subspace.

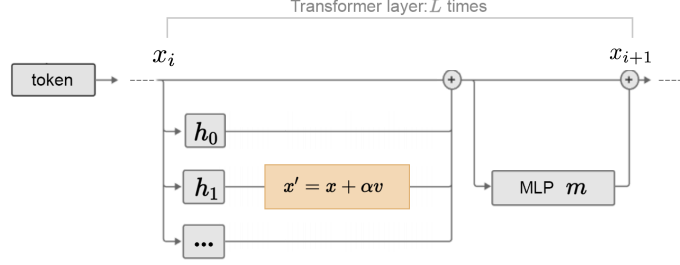


Figure 3: A sketch of the computation on the last token of a transformer with inference-time intervention (ITI) highlighted.

3.3 Inference-Time Intervention

The probing experiments above provide insight into how the LLM processes truth-related information across and within its attention heads. Moreover, they suggest a technique for improving performance on the benchmark dataset. If, during inference, we intervene to shift activations in the “truthful” direction, it seems possible that the network will provide more truthful answers to the benchmark questions. This is the basic strategy behind what we call **inference-time intervention (ITI)**.

The precise intervention we perform during inference is slightly more complex than shifting activations in an overall “truthful” direction. First, we do not intervene on every attention head. As seen in Figure 2(A), only a subset of attention heads appear to be strongly related to truthfulness. Following this observation, we only intervene on the results of the top K heads so as to be minimally invasive. This finer-grained intervention contrasts with previous transformer activation editing methods (Hernandez et al., 2023; Li et al., 2023) that work on the residual stream after the MLP. Working on attention heads’ activation spaces enables us to leave irrelevant heads out to be less intrusive to model behavior.

A second subtlety lies in how we determine the vector used to shift activations in the output of a given head. As seen in Figure 2(B), the geometry of true versus false statements is complex. In selecting a direction for shifting activations, there are two natural choices: the vector orthogonal to the separating hyperplane learned by the probe and the vector connecting the means of the true and false distributions. The latter connects to the whitening and coloring transformation commonly used in deep learning (Ioffe and Szegedy, 2015; Huang and Belongie, 2017). Comparison experiments and further discussion on different intervention directions can be found in Table 3 and Appendix B.

Figure 3 summarizes our inference-time intervention. We first rank the truth-relatedness of all attention heads by their probe accuracy on the validation set. We take the top- K heads as the targeted set. Then we estimate the standard deviation of activations along the truthful direction to be σ_l^h , using the activations from both the training and validation sets. ITI is an alternative form of MHA, where:

$$x_{l+1} = x_l + \sum_{h=1}^H Q_l^h \left(\text{Att}_l^h(P_l^h x_l) + \alpha \sigma_l^h \theta_l^h \right). \quad (2)$$

For not-selected attention heads, θ is a zero vector. This is equivalent to shifting activations along the truthful directions for α times the standard deviation. This procedure is repeated for each next token prediction autoregressively and is orthogonal to the choice of the decoding algorithm.

Intervention parameters K and α . Our method contains two key parameters: $K \in \mathbb{N}^+$, the number of heads where the intervention takes place, and $\alpha \in \mathbb{R}^+$, the “strength” of the intervention. Although we do not have a theoretical argument for the best values, we explore their effects experimentally and determine optimal values via a standard hyperparameter sweep. The real-life dilemma is that we are unsure what practitioners are optimizing for. The α should be selected per need by the user via trial and error: if users are extremely cautious about untruthful replies, α should be tuned up; otherwise, if helpfulness is also a requirement.

4 Experiments

4.1 Evaluation on TruthfulQA

We evaluate ITI on the TruthfulQA benchmark, which has 817 questions spanning 38 subcategories. TruthfulQA comes with two tracks: multiple-choice and generation. In the former, the multiple-choice accuracy (MC) is determined via comparing the conditional probabilities of candidate answers given the question; if the truthful answer ranks first, it counts as one positive. In the latter task, the model generates an answer to each question with greedy autoregressive decoding.

Preferably, a human annotator labels model answers as true or false given the gold standard answer. Since human annotation is expensive, Lin et al. (2021) propose to use two finetuned GPT-3-13B models (GPT-judge) to classify each answer as true or false and informative or not. Evaluation using GPT-judge is standard practice on TruthfulQA (Nakano et al. (2021); Rae et al. (2021); Askell et al. (2021)). Without knowing which model generates the answers, we do a human evaluation on answers from LLaMA-7B both with and without ITI and find that truthfulness is slightly overestimated by GPT-judge and the opposite for informativeness. We do not observe GPT-judge favoring any methods, because ITI does not change the style of the generated texts drastically.

The main metric of TruthfulQA is **true*informative** on the generation track, a product of scalar truthful and informative scores. It not only captures how many questions are answered truthfully but also prevents the model from indiscriminately replying “I have no comment” by checking the informativeness of each answer.

To calibrate the strength of the intervention, we report two additional quantities that measure how far LLaMA-7B deviates from its original generation distribution. Cross Entropy (CE) is a standard metric for language model pretraining. The other is the Kullback–Leibler divergence (KL) of the model’s next-token prediction distribution post- versus pre-intervention. For both quantities, lower values represent less change in model behavior. By tuning α , we wish to strike an optimal trade-off between increased truthfulness and minimal influence over other aspects of model behavior. We use a subset of Open Web Text for calculating CE and KL (Radford et al., 2017).

4.2 Experimental Baseline Comparisons

In addition to testing ITI on TruthfulQA, we compare it to several baseline approaches:

Supervised fine-tuning (SFT) is the first stage in RLHF (Ouyang et al., 2022). We use questions as prompts and encourage the model to generate truthful answers and discourage it from generating false answers with cross-entropy loss. However, if this is done alone, CE loss and KL rise drastically. Therefore, we alternate between supervised training on the QA pairs and pretraining on Open Web Text (Radford et al., 2017). We finetune all model parameters as previous works suggest that this serves as an upper bound for parameter-efficient finetuning (Zaken et al., 2021; Houlsby et al., 2019; Hu et al., 2021).

Few-shot prompting (FSP) is another way to increase truthfulness. Bai et al. (2022a) find in-distribution 50-shot prompting a strong baseline on TruthfulQA, compared to context distillation and RLHF. Since the choice of prompting strategy is orthogonal to the inference-time control method, we compare few-shot prompting with and without ITI.

Instruction fine-tuning (IFT) (Chung et al., 2022; Wang et al., 2022) is another well-known strategy to make language models truthful. To see how ITI can make the IFT model even more truthful, we study two models that are IFT’ed from LLaMA-7B, namely Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023).

Finally, we compare three different directions for the ITI activation shift. **Probe Weight Direction** is the direction found by linear probing in subsection 3.2. Intervening in this direction is equivalent to doing one gradient descent step on the head activation to maximize its probability of being predicted as truthful. **Mass Mean Shift** works by first calculating the average of truthful and false activations and then using the vector pointing from the false mean to the truthful mean for intervention. As a baseline, we also apply the **Contrast-Consistent Search (CCS)** technique, where the direction is found while only knowing pairwise information of internal activations (Burns et al., 2022). We train CCS on TruthfulQA by sampling one truthful and one false answer for each question. Since CCS

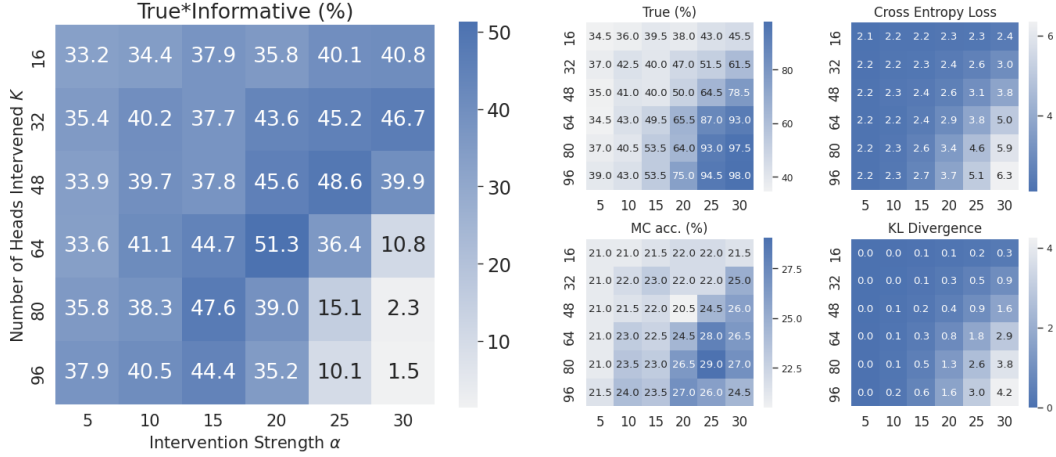


Figure 4: Results with varying intervention strength (α and K) on LLaMA-7B. 5% of questions used for training and validation, respectively. Metrics have been averaged over 5 random seeds.

doesn't take in labeled inputs, the discovered direction has an equal chance of being the truthful and false direction. We use labels to identify the truthful one for intervention.

4.3 Experimental Results

In Figure 4, we sweep two hyperparameters controlling the strength of the intervention, using 5% of randomly sampled questions for training and validation each. The true*informative score with respect to intervention strength follows an upside-down U curve. This shows a trade-off between truthfulness and helpfulness, discussed further in subsection 5.4. We choose the optimal hyperparameters $K = 48$ and $\alpha = 15$ by considering multiple scores. Up to this point, we use 10% of TruthfulQA—81 questions in total. Unless otherwise specified, we use 2-fold cross-validation for our results. We combine the answers from two hold-out sets for evaluation so no test samples are used in direction finding. More discussion on the model selection process can be found in Appendix C.

In Table 1, we compare ITI with the alternative baselines (subsection 4.2)²³. Due to the limit of context length for few-shot prompting, we adapt SFT and ITI to use 5% of TruthfulQA questions for a fair comparison with few-shot prompting.

| | True*Info (%) | True (%) | MC acc. (%) | CE | KL |
|--------------------------|---------------|-------------|-------------|------|------|
| Baseline | 30.5 | 31.6 | 25.7 | 2.16 | 0.0 |
| Supervised Finetuning | 36.1 | 47.1 | 24.2 | 2.10 | 0.01 |
| Few-shot Prompting | 49.5 | 49.5 | 32.5 | - | - |
| Baseline + ITI | 43.5 | 49.1 | 25.9 | 2.48 | 0.40 |
| Few-shot Prompting + ITI | 51.4 | 53.5 | 32.5 | - | - |

Table 1: Comparison with baselines that utilize 5% of TruthfulQA to make LLaMA-7B more truthful. CE is the pre-training loss; KL is the KL divergence between next-token distributions pre- and post-intervention. Results are averaged over three runs. We report standard deviations in Appendix D.

In Table 2, we apply ITI on instruction finetuned models by finding and intervening in their truthful directions. We notice that ITI significantly improves truthfulness over the baselines. It can be applied on top of few-shot prompting or instruction fine-tuning at the cost of a relatively low increase in CE loss and KL divergence.

²RLHF is found to significantly underperform 50-shot in-distribution prompting in Bai et al. (2022a) for TruthfulQA. In (Bai et al., 2022a; Menick et al., 2022), RLHF barely improves base models' performance. However, we are unsure of the result from a task-specific RLHF (Ziegler et al., 2019) with 5% samples.

³Baseline results were reproduced by the authors. Touvron et al. (2023)'s reported LLaMA-7B performances are: 29% true*informative and 33% true.

| | True*Info (%) | True (%) | MC acc. (%) | CE | KL |
|--------------|---------------|----------|-------------|------|------|
| Alpaca | 32.5 | 32.7 | 27.8 | 2.56 | 0.0 |
| Alpaca + ITI | 65.1 | 66.6 | 31.9 | 2.92 | 0.61 |
| Vicuna | 51.5 | 55.6 | 33.3 | 2.63 | 0.0 |
| Vicuna + ITI | 74.0 | 88.6 | 38.9 | 3.36 | 1.41 |

Table 2: Comparison with instruction finetuned baselines using 2-fold cross-validation.

In Table 3, we compare different directions to use for intervention, including random directions. We grid search for the optimal α for each direction separately, in the same way as in Figure 4. We observe that mass mean shift performs the best and also has a better tolerance for stronger intervention strength. Mass mean shift is used for all other experiments unless otherwise specified.

| | α | True*Info (%) | True (%) | MC acc. (%) | CE | KL |
|-----------------------------|----------|---------------|-------------|-------------|------|------|
| Baseline | - | 30.5 | 31.6 | 25.7 | 2.16 | 0.0 |
| random direction | 20 | 31.2 | 32.3 | 25.8 | 2.19 | 0.02 |
| CCS direction | 5 | 33.4 | 34.7 | 26.2 | 2.21 | 0.06 |
| ITI: Probe weight direction | 15 | 34.8 | 36.3 | 27.0 | 2.21 | 0.06 |
| ITI: Mass mean shift | 20 | 42.3 | 45.1 | 28.8 | 2.41 | 0.27 |

Table 3: Comparison with different intervention directions and their respective optimal α 's on LLaMA-7B. Results are from 2-fold cross-validation, a different protocol from Table 1.

5 Analysis

5.1 Results Across TruthfulQA Categories

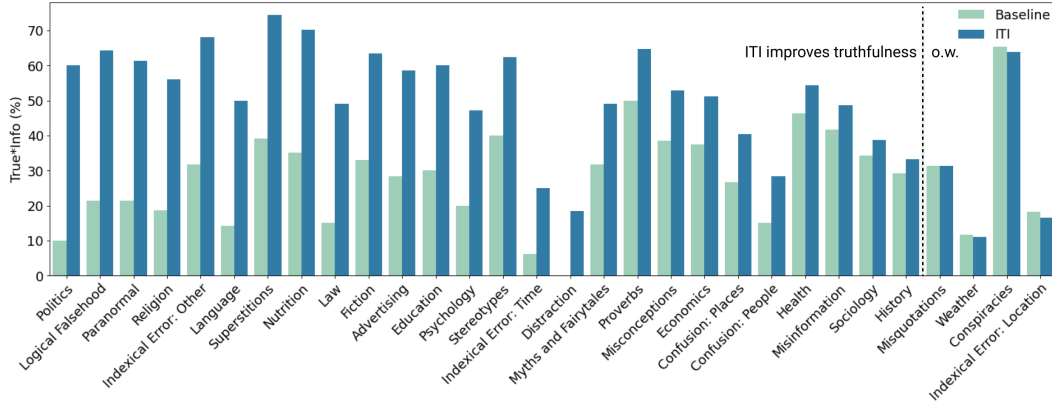


Figure 5: True*informative scores split across subcategories on LLaMA-7B, sorted by the difference between baseline and ITI. Subcategories with less than 10 questions are not shown.

TruthfulQA is split into 38 subcategories, including misconceptions, stereotypes, history, Mandela effect, and others. In Figure 5, we plot the true*informative scores of all subcategories with 10 or more questions compared to the baseline without intervention. We observe that ITI increases truthfulness across most types of questions. There is no one category that seems responsible for the overall increase, and we see no clear pattern as to which categories show the biggest effect.

5.2 Computational Efficiency

According to Equation 2, no matter how many attention heads are intervened, ITI adds a single constant vector per layer. That is, we need only add $\alpha \sum Q_i^h \sigma_i^h \theta_i^h$ into the stream between the MHA and the MLP operations. Considering the bias term in standard multi-head attention schemes, our intervention has close to zero computational overhead. We can also bake ITI into a pretrained LLM by an offline editing of its bias terms with this formula, specifically, we will set the bias term of the

output projection at layer l to be:

$$\text{Bias}_l = \alpha \sum_{h=1}^H Q_l^h (\sigma_l^h \theta_l^h). \quad (3)$$

A stand-alone edited LLaMA2-7B model can be found at https://huggingface.co/likenneth/honest_llama2_chat_7B.

5.3 Generalization of ITI beyond TruthfulQA

An important concern is how ITI might generalize beyond the TruthfulQA benchmark. As a first step toward investigating this question, we apply ITI—using the activation shift directions and hyperparameters learned from TruthfulQA—to three different datasets that relate to real-world truth: Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017) and MMLU (Hendrycks et al., 2020). The Natural Questions dataset consists of 3,610 real queries issued to the Google search engine that are annotated with answers and supporting Wikipedia pages. TriviaQA includes 95k question-answer pairs annotated by trivia enthusiasts. MMLU is an aggregated benchmark that covers 57 subjects across STEM, the humanities, the social sciences, and more. Note that these benchmarks were gathered for different purposes than TruthfulQA, i.e., question answering, reading comprehension and general capability evaluation, respectively. They were generated by different procedures, so they are a reasonable test of out-of-distribution generalization.

For the first two datasets, we apply ITI and report performance in a closed-book setting, i.e., models are prompted to answer the question without access to any documents. For each question, the dataset provides one truthful answer. In addition, we ask GPT-4 to generate the “most plausible sounding but false” answer to serve as an adversarial data point. For evaluation, we compare the probabilities of candidate answers being generated; if the truthful answer ranks first, it contributes one positive (same as in subsection 4.1). For MMLU, we use the standardized evaluation protocol Harness (Gao et al., 2021). Results are reported in Table 4. Note that this is a *true zero-shot* evaluation (Perez et al., 2021) as we do not tune any prompts, hyperparameters, or learn new truthful directions.

| | Natural Questions | TriviaQA | MMLU |
|----------------|-------------------|----------|-------|
| LLaMA-7B | 46.6 | 89.6 | 35.71 |
| LLaMA-7B + ITI | 51.3 | 91.1 | 40.16 |

Table 4: Generalization results on out-of-distribution datasets. Multi-choice accuracies are reported.

The results show that ITI causes the model to perform somewhat better than the baseline LLaMA-7B model across three benchmarks. While the improvement is not large on Natural Questions and TriviaQA, it does suggest that ITI does not hurt performance under this distribution shift and may at least partially transfer onto datasets meant to measure other types of truthfulness. Probably due to question-answer formation and distribution being similar to TruthfulQA, a stronger improvement is shown on MMLU benchmark.

5.4 Varying Training Set Size and Intervention Strength

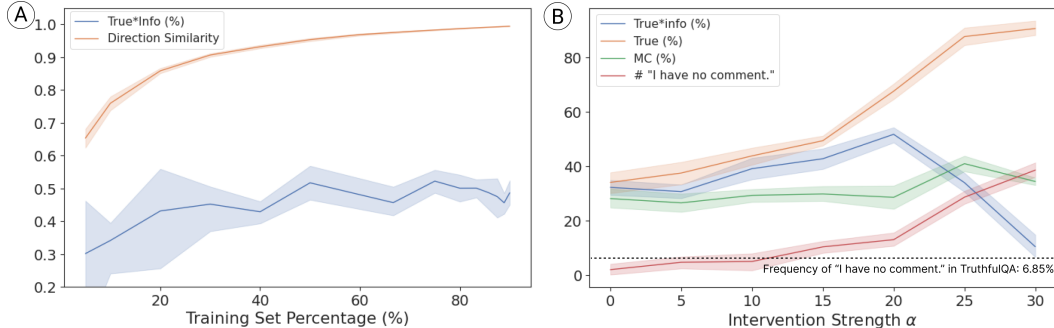


Figure 6: (A) How training set size affects model truthfulness and direction similarity, in the 18th head in the 14th layer of LLaMA-7B. (B) How intervention strength controls the trade-off between truthfulness and helpfulness.

To better understand the characteristics of ITI, we vary two hyperparameters and measure some key performance statistics. First, we increase the percentage of questions used to identify targeted heads and truthful directions. Alongside, we also plot the cosine similarity between the truthful direction found by the shrunken training set and the one found with the full dataset. In Figure 6(A), we find that the model truthfulness plateaus early. This suggests that the identified truthful direction is easy to find, requiring relatively few data points.

Second, we vary the intervention strength α and observe how it changes the model’s truthfulness. Additionally, as a statistic for informativeness, we plot the number of questions answered by “I have no comment.” Figure 6(B) shows a trade-off between truthfulness and helpfulness in ITI. Intuitively, this trade-off makes sense since it is trivial to attain a perfect truthfulness score simply by answering “no comment.”

5.5 Why Not Intervene on All Attention Heads?

Here we test two alternative methods for selecting intervention positions. To start with, we concatenate outputs from all self-attention heads across layers and train a single probe to classify truthfulness on them. The resultant accuracy is slightly higher than that from the best single attention head (84.4% compared to 83.3%) and is insensitive to normalization methods, including feature normalization and PCA. In the first alternative method, we intervene on all attention heads, denoted “without selection.” In the second one, we select the intervention position by ranking the absolute value of probe coefficients, denoted “point-wise selection.” We choose the same amount of features to intervene as in our “head-wise selection” baseline, K times attention head dimensionality. Results are shown in Table 5. The conflict between truthfulness and helpfulness can also be found for the two alternative methods, and the optimal α cannot achieve as good performance as baseline methods, demonstrating the importance of sparsifying interventions. It also suggests that head-wise selection might serve as a good heuristic for such sparsification.

| | α | True*Info (%) | True (%) | MC acc. (%) | CE | KL |
|----------------------|----------|---------------|-------------|-------------|------|------|
| Without selection | 5 | 35.4 | 37.1 | 28.3 | 2.19 | 0.08 |
| Point-wise selection | 15 | 39.2 | 55.3 | 28.7 | 4.01 | 1.95 |
| Head-wise selection | 20 | 42.3 | 45.1 | 28.8 | 2.41 | 0.27 |

Table 5: Comparison with different intervention position selection methods and their respective optimal α ’s on LLaMA-7B. Results are from 2-fold cross-validation, a different protocol from Table 1. We find that head selection helps ITI maintain informativeness under aggressive linear perturbation.

6 Conclusions and Future Work

We have described ITI, a general method whose goal is to improve the truthfulness of language model output. The approach uses supervised learning to identify latent vectors that relate to factual outputs and then uses these vectors to shift activations at inference time in “truthful” directions. Applied to the TruthfulQA benchmark, ITI achieves a significant boost in accuracy over current methods. Our investigation also uncovers information on how and where truthfulness seems to be processed, with a subset of attention heads seeming to play an outsized role.

There are several directions for future research. The most important would be to understand how well ITI generalizes to other datasets, ideally in a more real-world chat setting. It would also be important to understand the trade-offs implicit in tuning hyperparameters, especially the tension between truthfulness and helpfulness. We also suspect that the directions may be discoverable through unsupervised methods. The dimensionality of each head is relatively small and the direction similarity rises rapidly even with few supervised examples (as evidenced by Figure 6). From a scientific perspective, it would be interesting to better understand the multidimensional geometry of representations of complex attributes such as “truth.”