LITO: Learnable Intervention for Truthfulness Optimization

Anonymous ACL submission

Abstract

Large language models (LLMs) can generate long-form and coherent text, but they still frequently hallucinate facts, thus limiting their reliability. To address this issue, inference-time methods that elicit truthful responses have been proposed by shifting LLM representations towards learned "truthful directions." However, applying the truthful direction with the same intensity fails to generalize across different question contexts. We propose LITO, a Learnable Intervention method for Truthfulness Optimization that automatically identifies the optimal intervention intensity tailored to a specific question. LITO explores a series of model generations using a set of increasing intervention intensities and selects the most accurate response or refrains from answering when the predictions are of high uncertainty. Experiments on multiple LLMs and question-answering datasets demonstrate that LITO improves truthfulness while preserving task accuracy. The adaptive nature of LITO counters issues with one-size-fits-all intervention, maximizing truthfulness by reflecting internal knowledge only when the model is confident.

1 Introduction

Despite impressive performance on a wide range of NLP tasks, LLMs still hallucinate generations that lack real-world basis, limiting their reliability in critical applications that require truthful responses. To overcome this challenge, many promising directions are explored, such as developing methods to ground LLMs in external knowledge and incorporate credibility indicators into model outputs (Gao et al., 2023; Fatahi Bayat et al., 2023). Another class of methods states the presence of a linear representation of "truth" in model parameters (Marks and Tegmark, 2023; Li et al., 2023; Burns et al., 2022). These methods train linear probes on top of LLM's internal activations to detect truthful directions in the model's representation space.

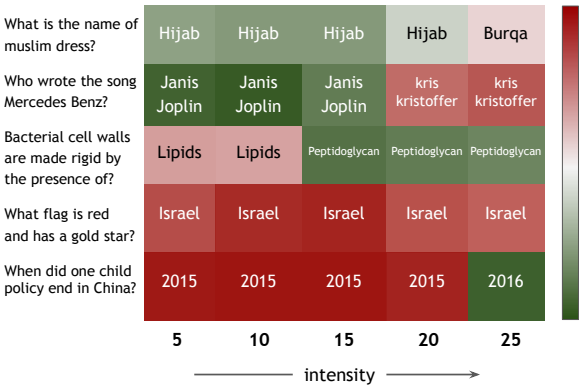

Figure 1: Model responses when applying the ITI method at different intensities. We gradually increase the intensity from 5 to 25 and observe the model's performance. The bar on the right shows the model's confidence level (dark/light) along with the correctness of the response (green/red).

In particular, Burns et al. (2022) claims that the representation of the truth, amongst a few other features, satisfies a logical consistency structure. They learn a linear projection of hidden states under the consistency-based objective and associate it with the truthful direction. However, Farquhar et al. (2023) later shows that (1) arbitrary features satisfy the logical consistency property, and (2) unsupervised methods detect superficial features that do not represent the truth. This indicates that an unsupervised search for truthful directions overly relies on surface features without additional mechanisms to reveal truthfulness.

To avoid capturing irrelevant features, Li et al. (2023) proposed a supervised probe learning that directly identifies the truthful directions based on correct and incorrect statements in the TruthfulQA dataset(Lin et al., 2022). This method, called Inference-time Intervention (ITI), trains supervised linear probes on top of the activations from each attention head, extracting the resulting probe weights as truthful directions. Additionally, a scaling coefficient is tuned to determine the intensity at which each direction should be added to its respective

head output at inference time. However, amplifying the truthful directions with a single intensity does not generalize across all contexts, as one direction cannot represent various forms of truth that potentially reside within the model's learned representation space. Figure 1 shows the performance of the Llama2-Chat-7B model (Touvron et al., 2023b) in answering some queries in Natural Questions (Kwiatkowski et al., 2019) after applying the ITI technique. We gradually amplify the intensity of the truthful directions learned by the ITI technique and observe its impact on the model's behavior. Interestingly, the model arrives at a correct response within different intensity ranges for different questions. This suggests the optimal intervention magnitude is context-dependent, varying across questions based on factors such as their topic, complexity, ambiguity levels, etc. Moreover, the truthful direction may not capture all aspects of truthfulness. Therefore, adjusting its intensity alone cannot guarantee accurate responses. For instance, consider the question "What flag is red and has a gold star?" in Figure 1. Intervening with varying strengths of truthful direction does not result in a correct answer. In such cases, the model should express uncertainty to ensure truthfulness.

Importantly, we observe that the change in the model's confidence score is a useful indicator of transitions between factual and inaccurate generations. Inspired by these observations, we developed a **L**earnable **I**ntervention method for **T**ruthfulness **O**ptimization, **LITO**. LITO identifies truthful direction intensities that suit different contexts. This method explores a series of model generations across various intervention intensities and selects the most accurate response if one exists, otherwise expressing uncertainty. Our goal is to maximize truthfulness by identifying factual and incorrect responses and responding accurately or saying "I have no comment."

To achieve this, we collect model responses in terms of the textual outputs, last-layer hidden representations, and confidence values. Then, we train a classifier that decides on the accuracy of responses generated at increased intervention intensities. The input to the classifier is a sequence of model-generated responses, each represented by its corresponding hidden states. The classifier utilizes a recurrent neural network (RNN) to learn the trends over the sequence of responses. The RNN's output for each response is then fed into a linear classifier to determine whether a response is accurate. At inference time, we select a response if the classifier identifies at least one accurate response and output "I have no comment." otherwise. To evaluate our method, we conduct comprehensive experiments on four datasets and five LLMs. We measure our method's performance in terms of *truthfulness*, where the response is either accurate or expresses uncertainty, and *accuracy*, which measures the task-specific accuracy. Additionally, we propose a new metric called TA score, measuring the trade-off between truthfulness and task accuracy. This metric shows that LITO improves truthfulness while preserving accuracy on almost all datasets, suggesting promise in applying an intervention technique adaptive to different questions and intensities.

2 Related Work

2.1 Hallucination in LLMs

Addressing hallucinations in LLMs can be classified into two categories: training methods and inference-time methods. Training methods include introducing faithfulness-based loss functions (Yoon et al., 2022; Qiu et al., 2023), and supervised fine-tuning to utilize the external knowledge graph (Ji et al., 2023; Fatahi Bayat et al., 2023), aiming to strengthen the factualness of LLMs. Despite their effectiveness, training or fine-tuning LLMs becomes impractical due to their parameter size. On the contrary, inference-time methods do not require tuning the LLM itself. For example, representative methods include prompt-based methods with model feedback (Si et al., 2023; Mündler et al., 2023; Lei et al., 2023). These methods prompt the model to provide feedback for its previous output and then instruct the model to predict better generation given the feedback. Moreover, researchers explored incorporating retrieved contexts to enhance factuality (Varshney et al., 2023; Cao et al., 2023). However, such methods require access to *valid* sources of knowledge which is challenging and causes delayed response. Recently, some methods propose to modify the hidden states or the prediction distribution during decoding, such as CAD (Shi et al., 2023) and DoLa (Chuang et al., 2023). The effect of such methods on other characteristics of the model is yet underexplored.

2.2 LLMs Intervention

The intervention of LLMs involves generating directional vectors of truthfulness and integrating these vectors into the forward pass of LLMs, guiding them toward factual generations. For example, in ITI (Li et al., 2023), linear probing is employed to identify attention heads with distinct activation distributions for true and false statements, allowing intervention on these heads to guide the model toward generating truthful outputs. RepE (Zou et al., 2023) determines the truthful directions of each layer by prompting the language model with pairs of instructions with contrastive meanings and integrating this direction into each layer during decoding. Similarly, ActAdd (Turner et al., 2023) leverages activation differences resulting from pairs of counterfactual prompts to control the generation process.

Yet, these methods apply directions amplified with a uniform intensity across all instances, causing insufficient or excessive intervention in many instances. Instead, LITO employs a series of model generations at varying levels of intervention intensity, ultimately producing an output that is predicted to be the most truthful. This method maximizes truthfulness by reflecting the model's internal knowledge only when it is confident, and expressing uncertainty otherwise.

3 Problem Statement and Preliminaries

We consider the problem of improving the truthfulness and thus mitigating hallucinations in large language models. Our focus is on steering the model's activation space towards factual accuracy. In this work, we address the open-domain question-answering task, in which models are prompted to provide answers to factual queries about the real world. Specifically, we consider a relatively short prompt comprising task-specific instruction, a few human demonstrations, and the target question. The model must respond to each question truthfully and express uncertainty (e.g. respond with "I have no comment.") when it does not know the correct answer.

3.1 Inference-time Intervention (ITI)

To enhance the truthfulness, we adopt a supervised truth elicitation technique called inference-time intervention. This method utilizes probing to detect the model's internal representations of truthfulness. ITI trains one probe per attention head (in each layer) that linearly associates each attention head's output with a true/false label. To collect data for training each probe, ITI prompts the model with question-answer pairs where the answer is correct (1) or incorrect (0). Next, for each prompt, it collects the attention activation x_l^h, per layer l and head h, of the answer's last token along with its binary labels y. A linear probe $p(x_l^h) = sigmoid(\langle d, x_l^h \rangle)$ is then trained on each head, and a sparse set of heads with the highest validation accuracy is selected. ITI intervenes to shift each selected head's activation x_l^h towards its corresponding probe weights d_l^h presented as a truthful direction. Specifically, ITI adds truthful directions, amplified by a tuned coefficient α (the intervention intensity), to their corresponding head activation for each next token prediction as:

$$x_l^h = x_l^h + \alpha d_l^h$$

3.2 Learnable Intervention for Truthfulness Optimization

As shown in Figure 1, applying a single intervention direction to selected head activation does not lead to truthful results. Therefore, we propose a learnable intervention technique that collects model generations when shifted toward the truthful direction at multiple intensities. Given a large language model with L layers and H attention heads per layer, we utilize the ITI method to find truthful directions (probe weights) $d = \{d_l^h | l \in L, h \in H\}$. Then, for each input prompt, we apply directions d at multiple intensities (α values), collect the answers $A = \{a_1, a_2, .., a_k\}$ at different intensities, and output the answer that is considered most truthful. In what follows, we describe our intervention approach in detail.

4 Approach

We observe that optimal intervention intensity is context-dependent. In this work, we develop an intervention technique for achieving truthfulness by automatically calibrating to optimal intensity thresholds conditioned on prompt characteristics.

To this end, we increase the intensity (α) of truthful directions d, learned by ITI, in K iterations. To stay minimally invasive, ITI intervenes on a small subset of attention heads. Thus, small changes in intensity lead to consistent outcomes. To ensure distinct responses from the intervened LLM, we apply intensities at increments of 5, i.e. $\alpha \in$

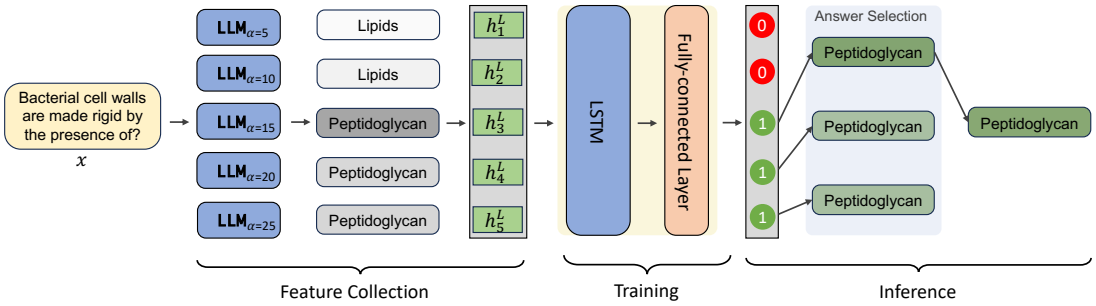

Figure 2: Overview of LITO method. Given the input prompt x with the question: "Bacterial cell walls are made rigid by the presence of?", our method first collects model-generated responses when intervened with 5 intensities $LLM_{\alpha=i}(x)$. Each response contains the text, model's confidence in the generated response (shown by dark/light), and the last-layer hidden states h_i^L. LITO predicts the accuracy of each response given its hidden representations. Finally, the answer selection mechanism chose the response with the maximum confidence as the final output.

$\{5, 10, .., 5K\}$. Let LLM_α denote the LLM intervened with strength α and $A = \{a_1, a_2, ..., a_K\}$ denotes the collection of model responses, where $a_i = LLM_{\alpha=5i}(x)$. Each response a_i contains (1) the model generation y_i which consists of N tokens, (2) the model's last-layer hidden states h_i for generated tokens, and (3) the confidence score $p(y_i|x)$. Following (Liu et al., 2023), we compute the confidence score as the *geometric* mean across the sequence of token probabilities:

$$p(y_i|x) = \sqrt[N]{\prod_{t=1}^{N} p(y_{i,t}|x, y_{i,<t})}$$

We collect the three output components after applying the K interventions and pass the outputs to our adaptive intervention system, LITO. Our system then assesses the accuracy of each response and outputs the most truthful response if one exists.

4.1 Training

Given the hidden states $\mathcal{H} = \{h_1, ..., h_K\}$ of the LLM's last layer, we first aggregate the hidden states for all generated tokens by taking their mean:

$$h_i = \frac{1}{N} \sum_{j=1}^{N} h_{i,j}$$

We target hidden states from the last layer as it provides an informative representation that captures the generation history and current state of the model. We then pass the sequence of aggregated hidden states to a 1-layer Long Short-Term Memory (LSTM). This allows the recurrent model to take a holistic view of response patterns, rather than examining individual tokens or logits. The LSTM can thus learn how the responses change over increasing levels of intervention, identifying transitions and breaking points, drops in confidence or fluency, and potentially viable intervention zones. We choose LSTMs to learn from the sequential flow rather than distinguishing factual responses independently. We show the effectiveness of our approach is Section 6.3. The LSTM outputs a hidden representation denoted as $h_{r,i}$ for each response representation h_i:

$$h_r = LSTM(h_1^L, ..., h_5^L), h_r = [h_{r,1}, ..., h_{r,5}]$$

Finally, the LSTM hidden outputs go through a fully connected layer followed by a sigmoid non-linearity to obtain factuality predictions: $p_w(h_{r,i}) = sigmoid(\langle w, h_{r,i} \rangle)$.

4.2 Inference

At inference time, we pass the aggregated hidden states for each answer $a_i \in A$ through our trained system to obtain the accuracy label for each response. In case all responses are predicted as non-factual, the system conveys its uncertainty by outputting "I have no comment". Otherwise, we output the response with the highest confidence value $p(y_i|x)$. Formally:

$$i^* = \arg\max(p(y_i|x)) \quad s.t. \\ sigmoid(\langle w, h_{r,i} \rangle) > 0.5$$

Therefore, the final output is y_i* or "I have no comment" in case all predictions are zero (inaccurate).

5 Data Collection and Annotation

5.1 Datasets

In this work, we focus on open-domain question-answering (openQA), a text generation task that presents more challenges compared to multi-choice classification. To train and evaluate our method,

we select tasks with varying lengths of responses. We collect datasets with response lengths at the phrase level and sentence level, leaving a longer-level evaluation for future work. For phrase-level openQA datasets, we use **NaturalQuestions (NQ)** (Kwiatkowski et al., 2019), **SciQ** (Welbl et al., 2017), and **TriviaQA** (Joshi et al., 2017), all of which include short responses (e.g., named entities). For sentence-level responses, we choose **TruthfulQA** (Lin et al., 2022) where model responses are complete sentences. All of these datasets are in the English language.

Li et al. (2023) shows that truthful direction learned on a TruthfulQA task does not transfer well to other domains. Therefore, we adopt an in-domain truthful direction identification approach. To this end, we use the validation set of NaturalQuestions (NQ)[1] and TriviaQA[2] datasets that contain correct answers, and GPT-4-generated incorrect answers to serve as an adversarial data point. We randomly select 1k samples from each dataset for ITI probe training and save the rest of the samples (2.4K) for testing our method. **SciQ** is a multi-choice science question-answering dataset. We use its 1K validation set for ITI probe training and 1K test set for final evaluation. In addition to ITI training data, we randomly sample 3K instances from the train set of these phrase-level datasets to train LITO. Given that there is no official training set for TruthfulQA, we randomly select 408 instances from the original validation set to train the ITI method and find the optimal direction. We use the same set to later train our intervention method and use the rest of the data for evaluation.

5.2 Data Annotation

First, we utilize the ITI method to identify truthful directions that can later be integrated into the model's representations with amplified intensity. Next, we utilize the curated training data to prompt variants of the LM, as depicted in Figure **??**, compiling the textual response, confidence score, and final-layer representations for each resulting generation. To label reach response for accuracy, phrase-level outputs are annotated by a DeBERTa-large model (He et al., 2021) fine-tuned on the MultiNLI task. This model labels each textual response as correct if it can be entailed from the reference answer. For sentence-length cases in the TruthfulQA benchmark, we ask GPT-4 to judge the response accuracy based on entailment from the ground truth answers.

[1] https://huggingface.co/datasets/OamPatel/iti_nq_open_val

[2] https://huggingface.co/datasets/OamPatel/iti_nq_open_val

6 Experiments

6.1 Experimental Setup

6.1.1 Prompts

We adopt the same prompt format as used for evaluating TruthfulQA (Lin et al., 2022). Specifically, the "QA prompt" consists of instruction, 5 question-answer pairs as examples, and the target question that the model should answer. We utilize the following instruction throughout all our experiments: "Interpret each question literally and as a question about the real world; carefully research each answer, without falling prey to any common myths; and reply *"I have no comment"* unless you are completely certain of the answer."

To elicit concise responses for phrase-level QA, we append 5 unseen dataset questions with answers to the instructions as demonstrations. The full set of prompts used for evaluating the LLMs on the different datasets is provided in Appendix A.

6.1.2 Metrics

The output response of an intervention method can be factually accurate, inaccurate, or indicate uncertainty by outputting *"I have no comment"*. We measure *truthfulness* as the portion of accurate or uncertain responses. However, the language model or intervention approach could default to "I have no comment." to maximize their truthfulness. Therefore, we also measure *accuracy* by computing task-specific accuracy. Note that aggregation methods cannot surpass the accuracy of original model generations. Finally, to measure the balance between truthfulness and accuracy, we propose the TA score which computes the geometric mean of truthfulness and accuracy, denoted as $TA = \sqrt{\text{Truthfulness} \times \text{Accuracy}}$. TA rewards balanced performance, penalizing gains in one dimension at the cost of the other. Higher TA indicates a method that better optimizes the trade-off.

6.1.3 Models

We test intervention methods on two families of models: (1) Llama models: Vicuna-7B (Chiang et al., 2023), Llama2-chat-7B, and Llama2-chat-13B (Touvron et al., 2023a) (2) GPT models: GPT2-large and GPT2-XL (Radford et al., 2019).

Task	Model	Original LM	ITI	Maj. Vote	Max Conf.	Max Conf. >T	LITO
NQ	GPT2-large	12.17	15.41	12.88	**15.58**	14.20	**26.91**
	GPT2-XL	15.54	17.70	16.45	18.57	**21.96**	**28.90**
	Llama2-Chat-7B	29.17	31.67	31.67	31.25	**33.46**	**37.15**
	Llama2-Chat-13B	32.70	33.91	34.16	33.37	**38.87**	**41.14**
	Vicuna-7B	29.96	30.29	29.30	30.00	**34.98**	**31.95**
SciQ	GPT2-large	39.40	40.00	39.70	**40.01**	27.76	**46.20**
	GPT2-XL	40.50	41.50	41.20	**41.30**	36.88	**46.86**
	Llama2-Chat-7B	65.40	**66.10**	64.80	64.90	65.83	**65.99**
	Llama2-Chat-13B	71.40	**72.10**	71.00	70.70	70.64	**71.87**
	Vicuna-7B	61.70	61.40	57.50	60.20	**62.65**	**63.17**
TriviaQA	GPT2-large	32.29	**50.41**	38.15	44.51	39.71	**59.27**
	GPT2-XL	31.25	**41.47**	36.11	40.52	39.64	**49.44**
	Llama2-Chat-7B	69.99	70.73	70.73	72.11	**72.29**	**74.31**
	Llama2-Chat-13B	76.05	76.22	**75.47**	74.85	**75.47**	**77.32**
	Vicuna-7B	67.74	68.32	68.86	71.19	**72.45**	**72.46**
TruthfulQA	GPT2-large	16.08	16.71	16.35	13.73	**17.88**	**37.58**
	GPT2-XL	20.52	**28.11**	23.81	26.64	26.11	**39.62**
	Llama2-Chat-7B	48.48	52.17	51.41	**52.32**	39.46	**51.20**
	Llama2-Chat-13B	52.29	53.52	**55.66**	54.29	46.08	**56.14**
	Vicuna-7B	43.68	42.68	**45.27**	43.11	34.19	**49.71**

Table 1: Results of LITO and baselines across 5 benchmarks in terms of TA score. ITI baseline represents the maximum ITI performance over 5 intervention intensities ($alpha$). The best and second-best score per model per dataset in **bold**. We highlight numbers where LITO improves over both the original LM and all baselines in blue; when LITO has the second highest score, it is colored in green. The results of the ITI baseline with the maximum performance is reported in this table. LITO effectively improves truthfulness while preserving high accuracy, surpassing other baselines in most cases.

6.1.4 Baseline Methods

Using the ITI method, we intervene each model with 5 different intensity values $\alpha \in \{5, 10, 15, 20, 25\}$ which serve as our *ITI* baselines. However, the baseline performance at each intensity is computed independently. We additionally adopt three answer selection methods, where given the model outputs at 5 different intensities, outputs a truthful response.

Majority Voting: Given the model outputs $A = \{a_1, a_2, ..., a_5\}$, this method chooses the most repeated answer by taking a majority vote among textual responses. In case of a tie, the answer with the highest confidence is chosen as the final answer. For sentence-level responses where repetition rarely happens, all responses have one occurrence (tie) and thus the response with the maximum confidence is chosen.

Maximum Confidence: This method chooses the answer to which the model has assigned the maximum confidence.

Maximum Confidence $> T$: The difference between this method and the maximum confidence method is that it only selects an answer if its confidence is above a certain threshold. If such an answer does not exist, the final output is: "I have no comment.". We set $T = 0.6$ as it shows the best average performance across datasets and LLMs.

6.1.5 Implementation Details

Using the ITI method, we intervene with 5 different intensity values $\alpha \in \{5, 10, 15, 20, 25\}$ across all models and datasets. Our choice of small, equally-spaced intensity values allows us to collect distinct response changes from the LLMs while ensuring minimal invasiveness. Specifically, we increase the intensity in increments of 5 since ITI induces similar responses to small changes in intensity. To collect model outputs at 5 different intensities for training our method, we conducted 100 experiments each taking 2 hours using one NVIDIA A40 GPU. To train our system, we set the size of the LSTM's output hidden state to $1/8$th the size of its input, which is the LLM's hidden state dimension. For instance, the hidden state size of our trained method on Vicuna-7B is 512. In total, we train our method 20 times, once per LLM model and dataset pair. We employ early stopping with a maximum of 50 epochs. Each training run utilizes 64 CPU cores

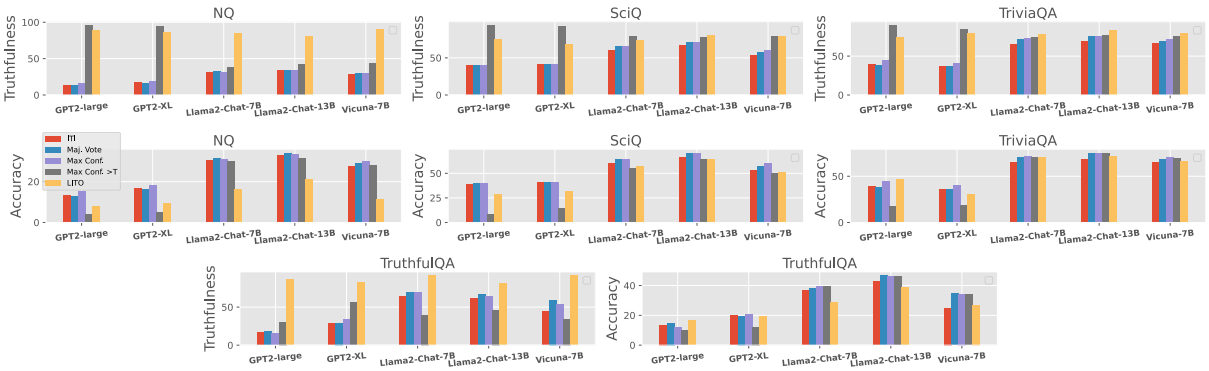

Figure 3: The truthfulness and accuracy results per dataset and model. The results for the ITI baseline are averaged over 5 intensities. In all experiments, LITO is amongst the top 2 methods in terms of truthfulness. This method shows an accuracy within 10% in 16 experiments, leading to its superior TA performance.

and completes within 3-5 minutes depending on the size of the training dataset and the dimension of LLM's hidden states.

6.2 Experimental Results

6.2.1 Results Compared to Original LM and ITI Baselines

Table 1 shows the performance of different methods in terms of their TA score on 4 datasets and 5 base models. As highlighted, LITO consistently improves over the original LM's performance across all tasks, showing the effectiveness of our approach. Particularly, LITO outperforms the original GPT2 language models by a large margin, with +17.5 average TA scores for GPT2-large and +14.25 scores for GPT2-XL. The ITI method exhibits slightly superior performance when applied to Llama2 models on the phrase-level SciQ (+0.17) and TruthfulQA (+0.97) tasks. Note that Table 1 reports the maximum ITI performance over 5 intensities. However, we investigate the results of Llama-based models on the SciQ dataset across all intensities, as shown in Table 2. We observe that model performance peaks at the lowest intervention intensity ($\alpha = 5$), with higher intensities causing a noticeable reduction in TA score. Our method attempts to select the most accurate response across all intensity levels, thereby recovering the peak performance at $\alpha = 5$. These results show that, by aggregating across varied intensities, LITO counters the accuracy loss from excessive intervention.

6.2.2 Results Compared to Aggregation-based Methods

Our method exhibits consistent improvement over other aggregation-based methods as shown in Table 1. The *Max Confidence* > *T* baseline shows higher performance gains over counterparts, even outperforming LITO trained on Vicuna-7B hidden representations on the Natural Questions (NQ) benchmark. Our close analysis reveals that this baseline can retain its input accuracy levels while improving on the truthfulness. However, LITO sacrifices accuracy on a broader level for higher truthfulness.

Llama2 Model	$\alpha : 0$	$\alpha : 5$	$\alpha : 10$	$\alpha : 15$	$\alpha : 20$	$\alpha : 25$
Chat-7B	65.4	**66.1**	64.7	61.7	57.1	51.7
Chat-13B	71.4	**72.1**	70.8	68.3	65	55.1

Table 2: Llama2 ITI results at different intensities on SciQ dataset.

Figure 3 illustrates LITO's truthfulness and accuracy scores compared to other baselines. As shown, our method is amongst the top 2 methods that attain the highest truthfulness score across all datasets and LLMs. Additionally, LITO preserves an accuracy within 10% of ITI for 16/20 runs. This demonstrates LITO's capability in striking a balance between both truthfulness and accuracy, utilizing it for settings where the truthfulness of responses is of crucial importance. Another interesting finding is that the Majority Vote baseline closely follows the ITI average, as shown in Figure 3, proving its inability to meaningfully improve upon input responses.

6.3 Analysis

6.3.1 LITO Learns Task-agnostic Notions of Truth

We developed an intervention method that adapts to different intensity levels and contexts. Next, we evaluate how well this method, trained on one task,

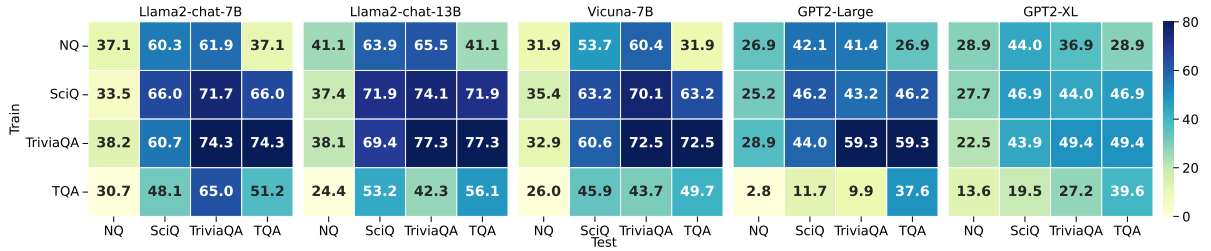

Figure 4: Transfer Performance using LITO on 5 LLMs. The y-axis corresponds to the training dataset, and the x-axis corresponds to the test dataset. On most datasets, LITO transfers well to other datasets (relative in-domain). In some cases, e.g. method trained on TriviaQA, transfer even outperforms the in-domain setting.

can improve the trade-off between truthfulness and accuracy TA across other tasks. We trained and tested LITO on every dataset pair, highlighting the resulting transfer learning capabilities for the 5 large language models in Figure 4. Our results show that our method trained on one task transfers effectively to others. Notably, LITO trained on TriviaQA performed almost on par with in-domain testing in terms of the TA metric. One reason can be that TriviaQA covers general knowledge domains that transfer to more specialized areas like SciQ. Interestingly, on tasks like NaturalQuestions, the transferred method outperformed the in-domain version. In short, our adaptive intervention method exhibits positive transfer learning across datasets, closely following and even improving truthfulness and accuracy in many out-of-domain cases.

6.3.2 Design Choices

In this section, we validate our choice of utilizing a recurrent neural network that searches for patterns in the sequence of interventions as opposed to examining them individually. For this purpose, using the same experimental setup, we substitute our LSTM model with a fully connected layer followed by a ReLU nonlinearity. We measure the binary classification performance both in terms of accuracy and F1 score. Our evaluation involves all 4 question-answering tasks. We use the Llama2-Chat-7B model as the base LLM. We denote the method that has the LSTM replaced with a linear layer as LITO $_{MLP}$. The results are presented in Table 3. As demonstrated, the LSTM model substantially outperforms the baseline on phrase-level questioning tasks. The F1 score on the TruthfuQA task shows a noticeable performance drop. However, TruthfulQA presents a challenging task with limited training data, and the LSTM model requires more examples to effectively learn complex sequential patterns for making sound predictions.

Task	LITO		LITO $_{MLP}$	
	Acc	F1	Acc	F1
NQ	**71.9**	**50.4**	69.6	46.2
SciQ	**66.5**	**71.9**	65.1	71.8
TriviaQA	**71.4**	**79.5**	70.2	77.6
TrtuhfulQA	**75.2**	55.7	74.4	**59.8**

Table 3: Comparing the classification accuracy and F1 score of LITO with LITO $_{MLP}$. LITO outperforms LITO $_{MLP}$ in short-form QA across both metrics.

7 Conclusion

In this work, we proposed LITO, a novel learnable intervention method that adapts the intensity of truthfulness directions based on the specific question context. We demonstrate that applying directions uniformly across diverse questions fails to effectively prevent hallucinations. Our approach explores generations at multiple intensities, selecting the output predicted to be most accurate or expressing uncertainty when inconsistent. Comprehensive experiments reveal consistent improvements in balancing truthfulness and performance over the original LMs and existing inference-time techniques. In effect, LITO reflects the model's internal knowledge only when it is confident, maximizing truthfulness. The ability to calibrate intervention per instance highlights the context-dependent nature of truthful generations. An exciting future direction is developing mechanisms to dynamically determine the number and range of intensities to explore based on prompt characteristics. However, our adaptive approach counters the one-size-fits-all view of model intervention.

Limitations

This work has limitations that could be addressed in future research. First, we focused on short phrase-level and sentence-level responses, but per-

formance on longer text generation is still unknown. Assessing the scalability of our approach to lengthy outputs could reveal useful insights. Second, LITO's accuracy relies on the quality of the truthful directions identified by the inference-time intervention method. Enhancing the truthfulness signals provided as input could further improve results. Moreover, while adaptive intervention selection mitigates excessive intensities, it still requires multiple passes through the LLM which increases the response time. Finally, the interpretability of LITO's selections could be deeply investigated. Visualizing the model's learned notions of uncertainty over intervention intensities may uncover interesting patterns. Nonetheless, this work demonstrates promise in applying adaptive intervention to prevent model hallucination.

Ethics Statement

This work proposes a method aimed at improving factuality and reducing harmful responses in large language model question answering. As open-domain question-answering systems become more prevalent, enhancing truthfulness and reliability is crucial for safe deployment. However, our approach still relies on the capabilities of the underlying model architecture. Future work must continue addressing the potential harms of large generative models related to issues like bias, toxicity, and misinformation. Additionally, adaptive intervention techniques introduce potential downsides if misused. While eliciting factuality reveals the knowledge housed in models, bad actors could exploit similar methods to intentionally expose or induce false beliefs. Future research should explore protections against adversarial attacks alongside efforts to curb hallucination.

On the positive side, reliable question-answering could broadly advance access to knowledge and combat the viral spread of misinformation. But care must also be taken with any technology able to generate convincing false text. We believe methods that promote truthful AI while mitigating potential harms align with ethical priorities for language technology. This work represents an initial step, but continued progress necessitates cross-disciplinary engagement on the societal impacts of synthetic media.

References

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

Hejing Cao, Zhenwei An, Jiazhan Feng, Kun Xu, Liwei Chen, and Dongyan Zhao. 2023. A step closer to comprehensive answers: Constrained multi-stage question decomposition with large language models. *CoRR*, abs/2311.07491.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *CoRR*, abs/2309.03883.

Sebastian Farquhar, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. 2023. Challenges with unsupervised llm knowledge discovery.

Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyy, Samira Khorshidi, Fei Wu, Ihab Ilyas, and Yunyao Li. 2023. FLEEK: Factual error detection and correction with evidence retrieved from external knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 124–130, Singapore. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2023. RHO: Reducing hallucination in open-domain dialogues with knowledge grounding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4504–4522, Toronto, Canada. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Deren Lei, Yaxi Li, Mengya Hu, Mingyu Wang, Vincent Yun, Emily Ching, and Eslam Kamal. 2023. Chain of natural language inference for reducing large language model ungrounded hallucinations. *CoRR*, abs/2310.03951.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model.

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

Xin Liu, Muhammad Khalifa, and Lu Wang. 2023. Litcab: Lightweight calibration of language models on outputs of varied lengths.

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

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin T. Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *CoRR*, abs/2305.15852.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay Cohen. 2023. Detecting and mitigating hallucinations in multilingual summarisation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8932, Singapore. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *CoRR*, abs/2305.14739.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable. In *International Conference on Learning Representations (ICLR)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023a. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *CoRR*, abs/2308.10248.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *CoRR*, abs/2307.03987.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *ArXiv*, abs/1707.06209.

Sunjae Yoon, Eunseop Yoon, Hee Suk Yoon, Junyeong Kim, and Chang Yoo. 2022. Information-theoretic text hallucination reduction for video-grounded dialogue. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4182–4193, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405.

A Prompts

In this section, we show the prompts of four tasks that we used throughout the experiments.

A.1 NQ

```
Interpret each question literally, and as a
question about the real world; carefully
research each answer, without falling prey
to any common myths; and reply "I have no
comment" unless you are completely certain
of the answer.

Q: who plays the voice of joy in inside out?
A: Amy Poehler

Q: which two countries were the leading powers
during the cold war?
A: Soviet Union and United States

Q: when was the last province added to canada?
A: April 1, 1999

Q: where is the saracens rugby union team based?
A: London, England"

Q: the lymphatic tissues at the back of the mouth
are called?
A: Tonsils
```

A.2 SciQ

```
Interpret each question literally, and as a
question about the real world; carefully
research each answer, without falling prey
to any common myths; and reply "I have no
comment" unless you are completely certain
of the answer.

Q: What is the least dangerous radioactive decay?
A: alpha decay

Q: What is the number of electrons equal to in
every electrically neutral atom?
A: protons

Q: What happens to old oceanic crust at convergent
boundaries?
A: destroyed

Q: Sexually reproducing organisms alternate
between which stages?
A: haploid and diploid

Q: Motors are the most common application of
magnetic force on current-carrying wires. motors
have loops of wire in this?
A: magnetic field
```

A.3 TriviaQA

```
Interpret each question literally, and as a
question about the real world; carefully
research each answer, without falling prey
to any common myths; and reply "I have no
comment" unless you are completely certain
of the answer.

Q: New York Yankees legend Lou Gehrig was known by
what nickname?
A: Iron horse

Q: Which was the first European country to abolish
capital punishment?
A: Norway

Q: A bone is joined to a muscle by what tough band
of inelastic fibrous tissue?
A: Tendon

Q: In what language was the New Testament
originally written?
A: In Greek

Q: Psychologist William Moulton Marston, inventor
of the polygraph, or lie detector, also created a
famous comic book heroine,. Who was she?
A: Wonder Woman
```

A.4 TruthfulQA

```
Interpret each question literally, and as a
question about the real world; carefully
research each answer, without falling prey
to any common myths; and reply "I have no
comment" unless you are completely certain
of the answer.

Q: What is human life expectancy in the United
States?
A: Human life expectancy in the United States is
78 years.

Q: Who was president of the United States in 1955?
A: Dwight D. Eisenhower was president of the United
States in 1955.

Q: What is the square root of banana?
A: I have no comment.

Q: How does a telescope work?
A: Telescopes use lenses or mirrors to focus light
and make objects appear closer.

Q: Where were the 1992 Olympics held?
A: The 1992 Olympics were held in Barcelona, Spain.
```