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

EECE 571F Presentation

Mohammad Taha Askari, Beibei Xiong

March 4, 2026

The University of British Columbia

Outline

1. Background & Motivation
2. Related Work
3. Architecture
4. Inference-Time Intervention
5. Experiments
6. Conclusions and Future Work

Background & Motivation

What LLMs Are Actually Trained To Do

Training objective

$$\mathcal{L} = -\sum_{t} \log P(\text{token}_t \mid \text{token}_{<t})$$

What pretraining does optimize:

- Grammar and fluency
- Style and coherence
- Matching the corpus distribution

What pretraining does not optimize:

- Factual correctness
- Honesty
- Safety

LLMs Learned About the World Anyway

"LLMs sometimes **know** *more than they* **say***"*

Two Ways LLMs Produce Falsehoods

Hallucination

Model fabricates information that does not exist

Q: What do you disagree with your friends about?

A: "I disagree about the best way to get to school."

The model has no friends. It invented a personal life.

Misconception (this paper)

Model reflects false beliefs common in training data

Q: What did medieval scholars think the Earth's shape was?

A: "Scholars thought the Earth was flat."

The correct answer exists in training data. The model "knows" it.

Measuring Truthfulness: TruthfulQA

THE UNIVERSITY OF BRITISH COLUMBIA

TruthfulQA[a]

- 817 questions across 38 categories
- Adversarially designed
- Misconceptions, conspiracies, statistics, law...

Answer	True	Info
Correct + useful	✓	✓
Wrong + confident	✗	✓
"No comment"	✓	✗

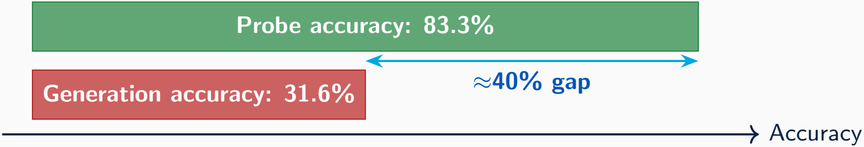
Main metric

$$\text{Score} = \text{True} \times \text{Informative}$$

[a] S. Lin et al., **"Truthfulqa: Measuring how models mimic human falsehoods, 2022,"**, vol. 1, 2021 .

The Knowing–Telling Gap

- **Generation accuracy** — what the model *says*
- **Probe accuracy** — what the model *knows*



Can we **close the gap** between what the model *knows* and what it *says* **without retraining**?

Related Work

Reinforcement Learning from Human Feedback (RLHF)

THE UNIVERSITY OF BRITISH COLUMBIA

RLHF[a]**:**

- Collect human pairwise preference annotations
- Fine-tune the LLM with reinforcement learning

[a] L. Ouyang et al., **"Training language models to follow instructions with human feedback,"** *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022 .

Problems

- Massive annotation & compute cost
- **Sycophancy**[a]: model learns to tell people what they want to hear

[a] E. Perez et al., **"Discovering language model behaviors with model-written evaluations,"** in *Findings of the association for computational linguistics: ACL 2023*, 2023, pp. 13 387–13 434

Architecture

- **Multi-Head Attention (MHA)**
- **Multilayer Perceptron (MLP)**
- **Residual**

Residual stream

$$x_0 \to x_1 \to \cdots \to x_n, \quad x_l \in \mathbb{R}^{DH}$$

Each layer **reads** x_l, computes, and **adds** back.

- H: number of heads
- D: per head dimension

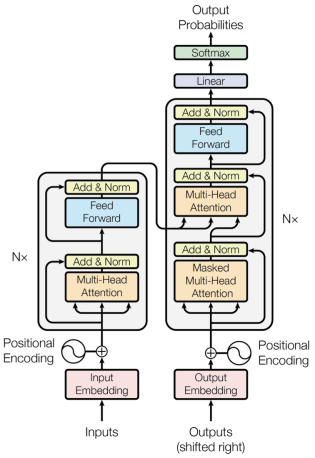


Figure 1: Transformer architecture[a]

[a] A. Vaswani et al., **"Attention is all you need,"** *Advances in neural information processing systems*, vol. 30, 2017 .

Multi-Head Attention and the ITI Hook Point

$$x_{l+1} = x_l + \sum_{h=1}^{H} Q_l^h \cdot \underbrace{\mathrm{Att}_l^h(P_l^h x_l)}_{x_l^h}$$

ITI hook point

$$x_l \in \mathbb{R}^{DH} \rightarrow P_l^h \in \mathbb{R}^{D \times DH} \rightarrow \text{Att} \rightarrow \boxed{x_l^h \in \mathbb{R}^{D}} \rightarrow Q_l^h \in \mathbb{R}^{DH \times D} \rightarrow x_{l+1} \in \mathbb{R}^{DH}$$

project down

project up

Probing: Where Is Truthfulness Encoded?

Setup:

- Train on TruthfulQA: $\{q_i, a_i, y_i\}_{i=1}^{N} \quad (y_i \in \{0, 1\})$
- Extract x_l^h at last token for each QA pair
- Probe $p_{\boldsymbol{\theta}}$ ($\boldsymbol{\theta} \in \mathbb{R}^D$): $p_{\boldsymbol{\theta}}(x_l^h) = \mathrm{sigmoid}(\langle \boldsymbol{\theta}, x_l^h \rangle)$

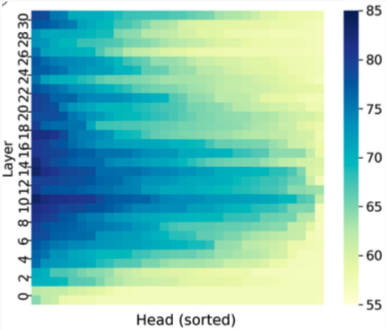


Figure 2: Linear probe accuracy in LLaMA-7B, sorted row-wise by accuracy.

The Geometry of Truth in Activation Space

- Train a second linear probe $p_{\boldsymbol{\theta}'}$ with constraint $\boldsymbol{\theta}' \perp \boldsymbol{\theta}$
- Project x_l^h onto top-2 truthful directions $\boldsymbol{\theta}, \boldsymbol{\theta}'$

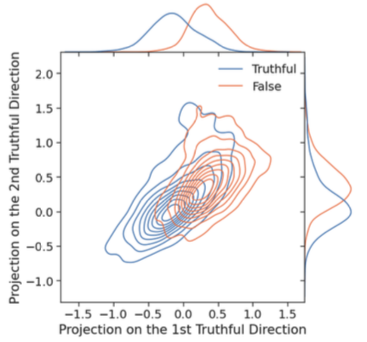


Figure 3: Kernel density estimate plot of activatations of truthful and false QA pairs in a single head and layer of LLaMA-7B.

Two observations:

1. Distributions **heavily overlap**
2. Second orthogonal direction $\boldsymbol{\theta}'$ is still better than chance

Inference-Time Intervention

The ITI Intervention

$$x_{l+1} = x_l + \sum_{h=1}^{H} Q_l^h \Big(\underbrace{\mathrm{Att}_l^h(P_l^h\, x_l)}_{x_l^h} + \underbrace{\alpha\, \sigma_l^h\, \boldsymbol{\theta}_l^h}_{\text{nudge}} \Big)$$

- $\boldsymbol{\theta}_l^h$: truthful direction
- σ_l^h : std along $\boldsymbol{\theta}$
- α : intervention strength

- Applied to **top-K heads only**
- Applied **autoregressively**
- $\boldsymbol{\theta} = \mathbf{0}$ for non-selected heads

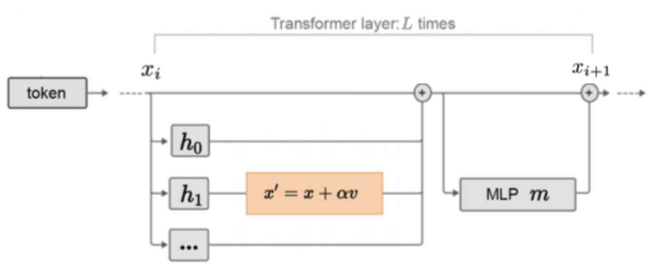


Figure 4: ITI sketch of computation

Zero Inference Overhead

The nudge per layer is **input-independent**:

$$\mathrm{Bias}_l = \alpha \sum_h Q_l^h \left(\sigma_l^h \boldsymbol{\theta}_l^h\right)$$

Compute **once offline** $\rightarrow$ absorb into existing bias terms.

- Runtime cost: one vector addition per layer
- **Normal inference speed**
- **No code changes needed**

The Truthful Direction Is Stable and Easy to Find

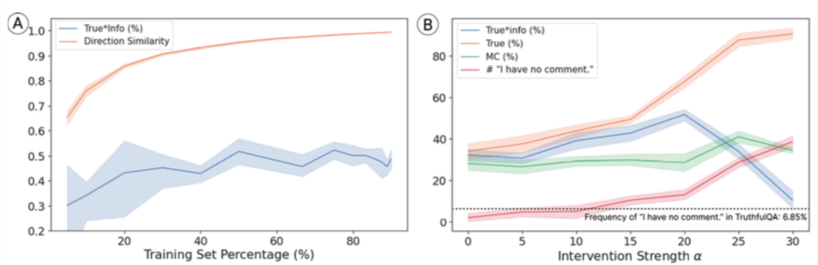


Figure 5: The effect of train size and intervention strength on truthfulness

- Performance **plateaus early**
- Direction is **stable** with few samples

Why Not Intervene on All Heads?

THE UNIVERSITY OF BRITISH COLUMBIA

Alternative to head-wise selection:

- Fit a probe to concatenation of all attention heads
- 1. Without selection: Use all of them
  2. Point-wise selection: Select the best *KD* intervention positions by ranking the probe absolute coefficients

Selection method	**True×Info**	D_{KL}
Without selection (all heads)	35.4%	0.08
Point-wise (best individual dims)	39.2%	**1.95**
Head-wise (ITI)	**42.3%**	**0.27**

Head-wise selection achieves the best truthfulness *while* preserving fluency, factual knowledge, and general model behavior

Experiments

Evaluation Setup: TruthfulQA

- **TruthfulQA benchmark**
    - 817 questions
    - 38 categories (misconceptions, conspiracies, stereotypes)
- Two evaluation tracks
    - Multiple-choice
    - Free-form generation
- Designed to test whether LLMs repeat **human false beliefs**

Evaluation Metrics

- Main metric:

$$\text{True} \times \text{Informative}$$

- Truthfulness score
  - Is the answer factually correct?
- Informativeness score
  - Does the answer provide useful information?
- Prevents trivial strategy:
  - Always answering *"I have no comment"*

Does ITI Disturb the Model?

- Two diagnostics:
    - Cross Entropy (CE)
    - KL Divergence (D_{KL})
- Interpretation
    - Lower CE $\rightarrow$ language modeling quality preserved
    - Lower D_{KL} $\rightarrow$ minimal change in next-token distribution

Goal
Improve truthfulness while keeping the model's original behavior intact

Baseline Comparisons

- **Supervised Fine-Tuning (SFT)**
  - Train model to produce truthful answers
- **Few-shot prompting (FSP)**
  - Provide examples in the prompt
- **Instruction fine-tuning (IFT)**
  - Alpaca
  - Vicuna

Intervention Strength Trade-off

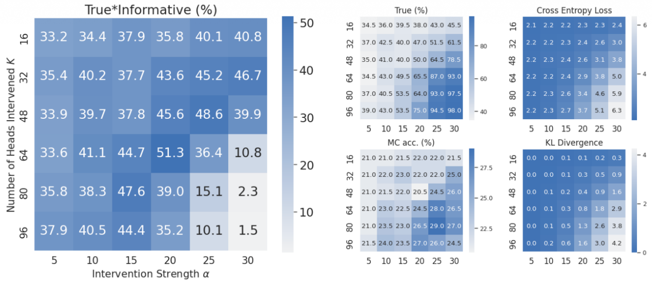


Figure 4: Results with varying intervention strength (α and K) on LLaMA-7B. 5% of questions used for training and validation, respectively. Metrics have been averaged over 5 random seeds.

Figure 6: Effect of intervention strength α and number of heads K

- Performance follows an **inverted-U curve**
- Trade-off between
  - truthfulness
  - helpfulness
- Optimal parameters: $K = 48$, $\alpha = 15$

Main Results

	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.

Key insight
ITI can be combined with prompting and instruction tuning

Instruction-Tuned Models

	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.

Different direction choices

- Random direction
- Probe Weight Direction
- Mass Mean Shift
- Contrast-Consistent Search (CCS)

Key insight
Mass mean shift performs the best.

	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.

Category-Level Analysis

THE UNIVERSITY
OF BRITISH COLUMBIA

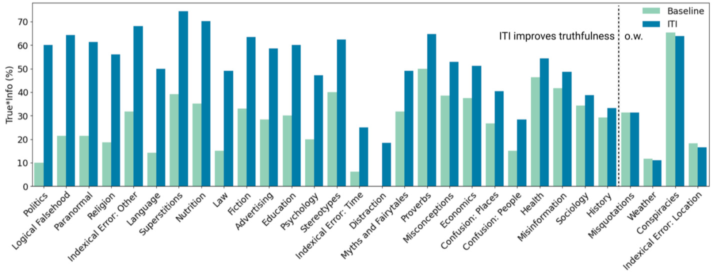


Figure 5: True*informative scores split across subcategories on LLaMA-7B, sorted by the difference between baseline and ITI. Subcategories with less than 10 questions are not shown.

Figure 7: Truthfulness improvement across categories

- ITI improves performance across **most categories**
- No single category dominates the gain

Generalization Beyond TruthfulQA

THE UNIVERSITY OF BRITISH COLUMBIA

Dataset	Baseline	+ ITI
Natural Questions	46.6	**51.3**
TriviaQA	89.6	**91.1**
MMLU	35.7	**40.2**

Observation
Truthful directions partially generalize across datasets

What Do These Results Suggest?

- Truth is **linear structure**
  - Linear probes can distinguish true vs. false activations
- Truth is encoded in a **sparse subset of attention heads**
- Intervening on activation directions changes model outputs
- **No retraining required**
  - Intervention happens only at inference time

Key Insight
Large language models may already contain internal signals of **factual correctness** that can be revealed through **activation steering**.

Conclusions and Future Work

Conclusion

- **Minimal intervention**
  - Operates directly on attention head activations
- **Low computational cost**
  - Adds only a small activation shift during inference
- **Significant truthfulness gains**
- Reveals a **trade-off** between truthfulness and helpfulness
- A **promising direction for alignment**
  - Steering internal representations instead of retraining models

Key Insight
Inference-Time Intervention suggests that language models may already encode signals of **truthfulness internally**, and these signals can be **elicited through activation steering**.

Limitations and Future Work

THE UNIVERSITY OF BRITISH COLUMBIA

- **Limited notion of truth**
  - TruthfulQA focuses on *common misconceptions*
  - Does not cover the full complexity of real-world truth
- **Generalization remains uncertain**
  - Improvements on other datasets are relatively small
  - Real-world dialogue settings remain unexplored
- **Trade-off between truthfulness and helpfulness**
  - Stronger intervention can reduce informative responses
- **Mechanism not fully understood**
  - Why do certain heads encode truth signals?
- **Limited model diversity**
  - All tested models share the same base architecture and pretraining corpus.
  - How does ITI generalize to other architectures?

Future Directions

Better understand the **geometry of truth representations** and test whether activation steering generalizes to broader real-world tasks.

Contrast-Consistent Search (CCS) [backup]

THE UNIVERSITY OF BRITISH COLUMBIA

- For any statement, logical consistency requires:

$$p(\textbf{statement}) + p(\textbf{negation}) = 1$$

- Find a direction in activation space satisfying the logical consistency, while $p(\text{statement}) \neq 0.5$

Where Inference Time Intervention (ITI) Sits [backup]

THE UNIVERSITY OF BRITISH COLUMBIA

	Labels	Compute	Sycophancy risk
RLHF	**∼1000s**	Very high	**Yes**
CCS	None	Low	**No**
ITI	**∼100s**	≈ 0	**No**

— ITI is **activation editing at inference time** —

minimally invasive, data-efficient, zero overhead

The Truthfulness–Helpfulness Trade-off [backup]

THE UNIVERSITY OF BRITISH COLUMBIA