Development of Linear Truth Encodings in Language Models: A Replication Study

Mingchuan Zou*, Suchir Salhan, Paula Buttery

Department of Computer Science & Technology, University of Cambridge, U.K. · ALTA Institute, Cambridge, U.K.

BlackboxNLP @ EMNLP 2026 · Reproducibility Special Track

Motivation

🧠 Truth probes underpin hallucination detection and truthfulness interventions.

Ravfogel et al. (2025) propose that truth directions emerge because true statements co-occur, helping models predict the next token.

🔬 **Our replication study tests this mechanism, false-context effects on confidence, and cross-relation transfer at 8B and 70B to determine what truth probes actually reveal.**

Setup

Synthetic Toy Models. One-layer attention-only transformers; each sequence holds two facts, both true (probability ρ) or both false. **Ablation:** each fact's truth is drawn independently, so one fact's truth says nothing about the other's.
Pretrained. Llama-3.1-8B and 70B on six CounterFact relations; logistic-regression probes, 200 labelled statements per relation.

Without truth–fact co-occurrence, only the fully trainable model's truth probe drops to chance.

Training data

x *y* ***x′*** *y′*

Normal: one shared T

T = 1

Paris	France	Rome	Italy

T = 0

Paris	Germany	Rome	Spain

Ablated: independent T_1, T_2

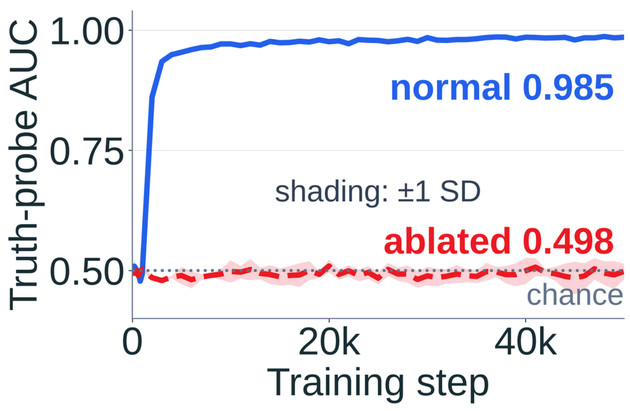
$T_1 = 1$, $T_2 = 0$

Paris	France	Rome	Spain

false attribute; probes read x′

Fully trainable

learned embeddings and attention
ρ = 0.8, 3 seeds



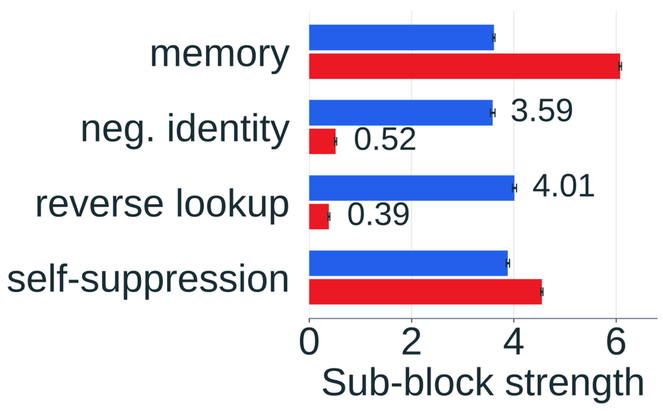
Constrained toy

frozen one-hot embeddings, uniform attention
ρ = 0.8, 5 seeds

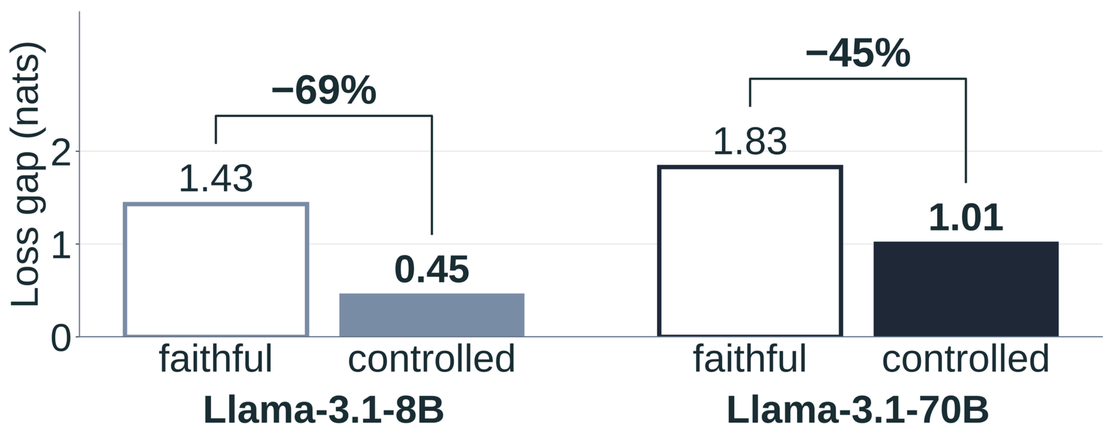


Constrained toy OV

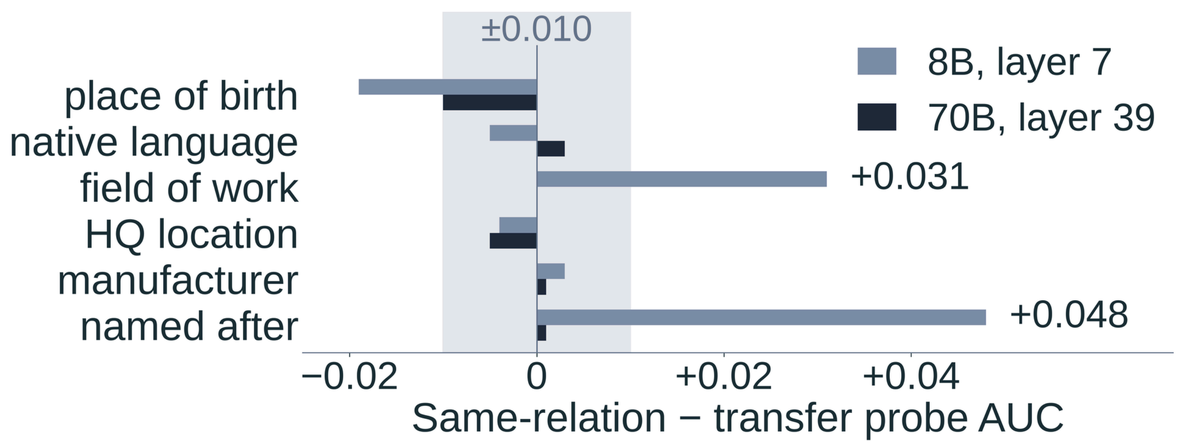
step 2000; mean ± SD, 5 seeds
blue: normal; red: ablated



False context makes true answers less likely, more so at 70B; controls shrink the effect 45–69%



Truth probes transfer to new relations at 70B (within 0.010 AUC) but lose up to 0.048 at 8B



Reproduction scorecard

- 🔁 **Reproducibility:** Reproduce the headline findings using released code, documenting two implementation changes: learning rate (10^{-3} vs 10^{-4}) and small random toy initialisation.
- 🔍 **Validation:** Recover the memorisation and decodability patterns, OV sub-blocks, and 8B false-context gap; investigate the confidence mismatch (0.48 vs 0.25) before claiming full reproduction.

Takeaways

🧠 *High probe accuracy can mislead.*
🔬 *Co-occurrence helps models predict truth.*
✅ *Ablations reveal what probes actually measure.*

***Corresponding author:** **mz541@cam.ac.uk**
Code: github.com/Samzmc/truth-encoding-blackboxnlp26
Supported by Cambridge University Press & Assessment; used NDIF and NNsight (NDIF: NSF Award IIS-2408455).

NOT FOR PRINT — presenter notes (export page 1 only)

Captions removed from the poster on 20 Oct 2026, kept here for Suchir and Andrew.

RQ1 (hero figure)
Fully trainable: without co-occurrence the probe drops to chance, yet the model still predicts every fact correctly (top-1 accuracy 100%). Constrained toy: its circuit survives the ablation. Two of the four output-value (OV) weight sub-blocks shrink 7–10× but keep their direction (cosine ≥ 0.81), as expected if the structure comes from the fixed architecture rather than from truth being useful. Caveat: the two models also differ in scale, optimiser, learning rate and training length, so we cannot tell which difference causes the contrast.

RQ2 (loss-gap bars)
Bars: extra loss on the true answer after false vs true context (two sentences, native language). Faithful: original design, difference of unpaired medians. Controlled: median of matched-pair differences, removing attribute priming, subject overlap and unmatched prompts together. The controlled effect is positive in 71 of 72 relation-pair cells.

RQ3 (transfer bars)
Bars: AUC of a probe trained on the same relation minus one trained on the other five (> 0: transfer is worse). Same-relation AUC averages 0.962 (8B) and 0.977 (70B), slightly optimistic (layers chosen on it). Transfer probes see ≈7.5× more data; some 70B relations are near ceiling.

Takeaways

- **High probe AUC can survive when truth no longer predicts the next fact** (constrained toy: 0.959 after ablation).
- In our one-layer fully trainable transformers, truth was not linearly decodable without co-occurrence.
- Probes are more informative alongside ablations, confound controls and checks of circuit magnitude.

- **Reproduced:** early memorisation, then truth decodability; the four OV sub-blocks; 8B false-context gap 1.43 nats (Llama-3.1) vs 1.52 reported (Llama-3).
- **Changed:** fully trainable learning rate 10^{-3} (released code), not 10^{-4} (paper text); toy initialised with small random weights, as zero initialisation gives zero gradients when both OV factors are trained.
- **Not reproduced:** final P(right answer | false sequence) ≈ 0.48 (ours) vs ≈ 0.25 (original); not used in AUC comparisons.

Limits: necessity shown only in synthetic models; confounds controlled jointly; two models, six relations; probes show decodability, not causal use.

Blue: reproduction of Ravfogel et al. (2025); red: our ablation. OV sub-blocks (x = subject, y = attribute): memory, x → its true attribute; neg. identity, x → −x; reverse lookup, y → its subject; self-suppression, y → −y at the output.