

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.

Abstract

Probing studies consistently find that language models (LMs) encode the truth of a statement along a low-dimensional linear direction of the residual stream. The Truth Co-occurrence Hypothesis (TCH) (Ravfogel et al., 2025) provides a mechanistic interpretation: tracking a latent truth variable is loss-reducing because true statements co-occur with true statements in natural text, evaluated using toy models and base LLMs. We replicate and extend the toy setting. In the tested fully trainable single-layer transformers, removing truth co-occurrence while preserving fact frequency eliminates linear truth decodability, supporting the necessity of co-occurrence in this setting. However, the same manipulation leaves the original toy model largely unchanged, revealing that probe-based evidence can arise from architectural artefacts rather than a functional truth-tracking mechanism. In pretrained Llama-3.1 models (8B and 70B), we reproduce the reported behavioural effect that false context reduces confidence in true facts, but show that nearly half of the original effect is explained by surface confounds. After controlling for these confounds, the effect remains robust across six factual relations. Under our probing protocol, cross-relation performance is closer to within-relation performance at 70B than at 8B. Our work characterises the emergence conditions for linear truth representations, and highlights the limitations of probe-based interpretability studies. Reproducibility Code is available here: <https://github.com/Samzmc/truth-encoding-blackboxnlp26>

1 Introduction

Large language models (LMs) appear to represent whether statements are true in a surprisingly simple way. Across models and probing methods, true and

false statements become linearly separable in the residual stream, allowing a low-dimensional direction to predict factual correctness across diverse domains (Burns et al., 2023; Azaria and Mitchell, 2023; Li et al., 2023; Marks and Tegmark, 2024; Bürger et al., 2024; Orgad et al., 2025; Slobodkin et al., 2023; Farquhar et al., 2024; Ferrando et al., 2025). This has enabled methods for hallucination detection, factuality estimation, and truthfulness interventions that exploit internal representations rather than surface correlations.

Prior work shows that hidden states contain linearly decodable factual knowledge (Gekhman et al., 2025) and that interventions along learned truth directions can steer generations towards greater truthfulness (Li et al., 2023), even when surface predictions disagree with internal representations (Liu et al., 2023). However, the mechanism behind these representations remains unclear. Ravfogel et al. (2025) propose the *Truth Co-occurrence Hypothesis* (TCH): because reliable sources tend to contain many true statements and unreliable sources many false ones, the truth value of one statement provides predictive information about neighbouring statements. Inferring a latent truth variable from context therefore reduces next-token prediction loss by adjusting confidence in candidate factual attributes. Ravfogel et al. (2025) theoretically show that truth co-occurrence is sufficient for a linear truth representation because encoding the latent variable improves the training objective. Empirically, they observe a two-stage learning process in toy and fully trainable transformers, where factual associations are memorised before a linearly decodable truth representation emerges. In the toy model, they identify the learned OV circuit and LayerNorm as important components, and show that increasingly long false contexts reduce confidence in subsequent true facts in Llama-3-8B.

In this paper, we reproduce and extend these experiments through three research questions to

Corresponding Author: mz541@cam.ac.uk. Code and data available: <https://github.com/Samzmc/truth-encoding-blackboxnlp26>

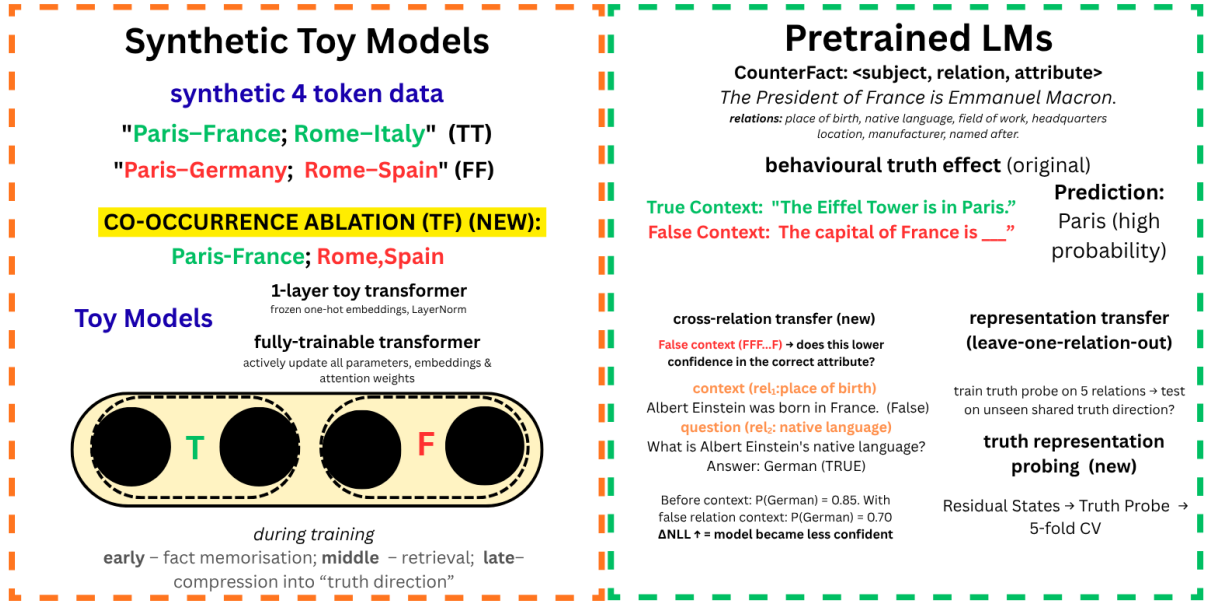


Figure 1: **Replication and Extension of Ravfogel et al. (2025)**: Our work investigates the emergence of linear truth representations in (a) a toy model setting and (b) a base LM setting. We extend the toy setting by isolating truth co-occurrence from fact frequency to test the mechanism proposed by the Truth Co-occurrence Hypothesis. In pretrained language models, we reproduce and extend the original behavioural and probing experiments, examining robustness to confounds, transfer across factual relations, and scaling from 8B to 70B models.

investigate the robustness of linear truth encodings. **RQ₁ (Necessity)**: Is truth co-occurrence necessary for linear truth encoding, or can attribute frequency alone explain the effect? We test this by removing cross-fact truth co-occurrence while preserving attribute frequencies, finding that probe performance drops to chance in fully trainable transformers without impairing factual recall, implying that shared truth structure is necessary for linear truth encoding to emerge in this setting, although constrained architectures can still produce misleading residual signals. Behavioural truth effects have been reported in language models, but their robustness at larger scales and susceptibility to surface artefacts remain unclear. **RQ₂ (Behaviour at Scale)**: Does the behavioural effect replicate at larger scale, survive surface-confound controls, and generalise across factual relations? We evaluate larger pretrained models across six factual relations, finding that the effect reproduces in Llama-3.1-8B, strengthens in Llama-3.1-70B, and remains robust after controlling for surface confounds that explain approximately 45% of the original effect, implying that behavioural truth signals are genuine but require careful validation. Even when truth can be decoded, it is unclear whether models learn a general truth representation or relation-specific patterns. **RQ₃ (Representation at Scale)**: Does the

learned truth representation generalise across relations, and does this improve with model scale? We test cross-relation transfer using within-relation and leave-one-out probes. Their performance is more similar at 70B than at 8B under our evaluation protocol. This finding is limited to the two tested models and six relations. High probe accuracy alone can still be misleading without controls for surface confounds and validation of the underlying mechanism.

2 Methodology: Replication and Extension of Ravfogel et al. (2025)

We investigate the emergence of linear truth representations in two settings introduced by Ravfogel et al. (2025): synthetic transformers trained on controlled data and pretrained language models. The synthetic setting isolates the mechanism proposed by the Truth Co-occurrence Hypothesis (TCH), while the pretrained setting tests whether the resulting behavioural and representational phenomena extend to larger models and diverse factual relations. We reproduce the original experiments and introduce controlled interventions addressing three questions: whether truth co-occurrence is necessary, whether behavioural truth effects generalise at scale, and whether learned truth directions become relation-general.

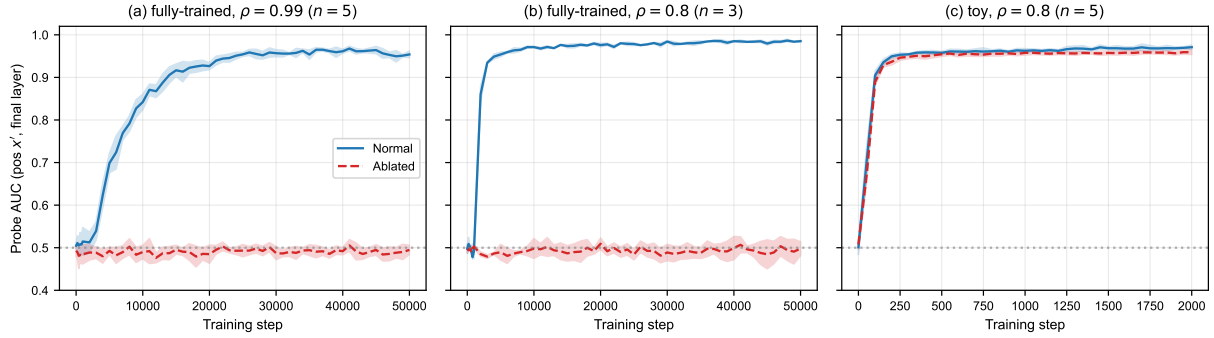


Figure 2: Our Reproduction of Probe AUC across training for Toy Models, reproduced from Ravfogel et al. (2025) (blue), compared to **new ablations** (red) that investigate new truth settings that break truth co-occurrence, while keeping truth frequency controlled. We evaluate two architectures across three settings: a *fully trainable synthetic transformer* with learned embeddings and attention, and a *constrained toy transformer* with frozen embeddings and uniform attention. **(a)** fully trainable, $\rho = 0.99$ ($n = 5$); **(b)** fully trainable, $\rho = 0.8$ ($n = 3$); **(c)** toy, $\rho = 0.8$ ($n = 5$). Shading: ± 1 SD across seeds; dotted line: chance.

Synthetic Setting: Truth Co-occurrence Ablation (RQ1)

The original synthetic experiments of Ravfogel et al. (2025) train transformers on four-token sequences $x y x' y'$ containing two subject-attribute pairs. Subjects and attributes use frozen, orthogonal one-hot embeddings, with a ground truth mapping g defining the correct attribute for each subject: $y = g(x)$. A latent truth variable $T \sim \text{Ber}(\rho)$ determines whether both pairs are correct: with probability ρ , $y = g(x)$ and $y' = g(x')$; otherwise both attributes are replaced by random incorrect values. Thus, when $\rho > 0.5$, correct attributes are also the most frequent attributes, but truth frequency and truth cooccurrence are confounded. We replicate the toy model from Ravfogel et al. (2025): a constrained toy model with frozen one-hot embeddings, uniform attention, and only the OV matrix $W = W_o W_v$ trainable.

Extensions We introduce two extensions – *extending the synthetic dataset* and *testing the toy model architecture*, as shown in Figure 1. First, we systematically investigate the relationship between truth frequency and truth cooccurrence by introducing a **co-occurrence ablation** for synthetic data. We replace the shared truth variable with independent samples $T_1, T_2 \sim \text{Ber}(\rho)$ for the two statements. This preserves the mapping g , sequence structure, and marginal truth frequencies, while removing statistical dependence between the statements. Thus, any learned truth representation cannot rely on cross-statement co-occurrence. For example, Paris–France; Rome–Italy may be replaced by Paris–France; Rome–Spain, breaking the shared truth pattern while keeping the same format. Sec-

ondly, we compare two architectures from Ravfogel et al. (2025): the *constrained toy transformer* and the *fully trainable synthetic transformer*. Both are one-layer, single-head attention-only models. By applying the same co-occurrence ablation, we assess the roles of truth co-occurrence and architectural constraints in linear truth decodability.

Extended Analysis Following Ravfogel et al. (2025), we measure truth decodability using logistic probes at the second subject position x' , and distinguish truth encoding from factual recall using mapping accuracy. We further **extend the analysis with OV-circuit measurements** by decomposing $W = W_o W_v$ into functional sub-blocks and comparing each block B to its theoretical target T using signed projection $a_B = \langle B, T \rangle / \|T\|^2$, Frobenius norm, and cosine similarity. These measures distinguish changes in circuit amplitude from changes in orientation.

Pretrained LM Setting: Behaviour and Generalisation (RQ2–RQ3)

For pretrained models, Ravfogel et al. (2025) evaluate whether false contextual information changes confidence in subsequent factual predictions using COUNTERFACT (Meng et al., 2022). COUNTERFACT contains factual statements represented as (subject, relation, attribute) triples rendered into natural-language templates. We first reproduce the **behavioural truth effect** experiment of Ravfogel et al. (2025) using a single relation (P103) in COUNTERFACT. This relation is NATIVELANGUAGE (P103), which produces statements such as “The mother tongue of Danielle Darrieux is French”, with false variants generated by replacing the attribute with another value from the

same relation. The experiment (using Llama-8B) prepends false preceding context to reduce confidence in a correct factual answer without changing the model’s stored knowledge. For example, while “*The Eiffel Tower is in ____*” still retrieves *Paris*, surrounding false statements activate an unreliable-context signal, lowering $P(\text{Paris})$. This shows that the model uses contextual truthfulness information to calibrate confidence in factual predictions.

Extensions to Pretrained Setting We use six relations in COUNTERFACT: place of birth, native language, field of work, headquarters location, manufacturer, and named after. This allows us to test whether truth effects are specific to individual relations or reflect a shared contextual truth signal. We introduce three extensions to the work of [Ravfogel et al. \(2025\)](#) (as shown in [Figure 1](#)): (i) **truth representation probing**, which evaluates whether residual states contain linearly decodable truth information using 3-fold cross-validated probes; (ii) **cross-relation transfer**, which tests whether contextual falsehoods from one relation (e.g., place of birth) alter confidence in predictions for another relation (e.g., native language), measured by changes in answer likelihood and negative log-likelihood (NLL); and (iii) **representation transfer via leave-one-relation-out probing**, which tests whether a truth direction learned across five relations generalises to an unseen relation, indicating a shared truth representation across factual domains. Our extension experiments use larger pretrained models (Llama-3.1-70B), in addition to Llama-3.1-8B, which substitutes the Llama-3-8B used by [Ravfogel et al. \(2025\)](#).

For **behavioural evaluation (RQ2)**, we compare matched true and false contexts while controlling for surface-level confounds, including subject overlap, answer leakage, and attribute priming. The faithful replication setting uses 896 native language (P103) prompts per model. The controlled setting forms a 6×6 grid, with context relations as rows and target relations as columns. Each cell contains 64 matched TT/FF prompt pairs, with two context sentences per prompt ($k = 2$). Within each pair, the target and the two context subjects are identical; only the context attributes change from their true values (TT) to false alternatives (FF). We quantify contextual effects as the paired change in target attribute negative log-likelihood: $\Delta\text{NLL} = \text{NLL}_{\text{false}} - \text{NLL}_{\text{true}}$, where positive values indicate reduced confidence un-

der false contextual information. Statistical significance is assessed using one-sided Wilcoxon signed-rank tests independently for each context-target relation pair (6/6 within-relation for both models; 29/30 off-diagonal at 70B vs 27/30 at 8B, $p < 0.05$). For **representational evaluation (RQ3)**, we extract residual-stream activations and output logits from Llama-3.1-8B and Llama-3.1-70B using NDIF/NNsight ([Fiotto-Kaufman et al., 2025](#)). We evaluate truth representations using logistic-regression probes with class-balanced three-fold stratified cross-validation and balanced true/false evaluation sets. We compare within-relation probes, trained and tested on the same factual relation, with leave-one-relation-out probes, trained on five relations and evaluated on an unseen relation, to measure whether a shared, transferable truth direction exists across factual domains. Full hyperparameters and experimental settings required to reproduce the released implementation are provided in Appendices [A](#), [B](#) and [E](#).

3 Results

We first examine how truth co-occurrence and architecture affect synthetic truth decodability, then evaluate behavioural truth effects and probe transfer in pretrained models.

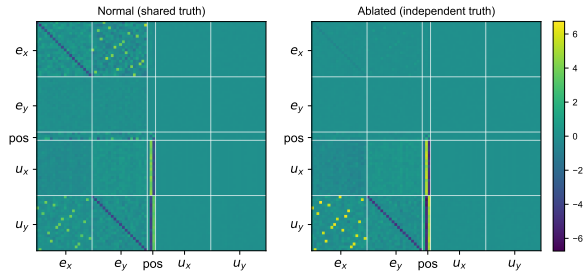


Figure 3: Toy-model OV circuit ($W = W_o W_v$, step 2000, seed 0, shared color scale) under normal versus ablated conditions. The embedding-space blocks (negative identity, reverse lookup) collapse under ablation; the unembedding-space blocks (memory, self-suppression) persist or strengthen.

3.1 Reproduction

We reproduce the synthetic experiments of [Ravfogel et al. \(2025\)](#) across both the constrained toy and fully trainable models. The two settings differ substantially in their inductive biases: the toy model uses frozen one-hot embeddings and uniform attention, whereas the fully trainable model

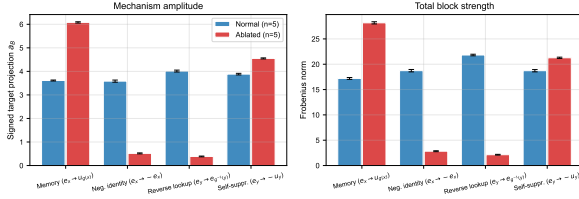


Figure 4: The ablation shrinks toy OV sub-blocks without rotating them. Signed target projection a_B (left) and Frobenius norm (right) of each sub-block against its theoretical target at step 2000 (mean $\pm$ SD, $n = 5$ per condition). Negative identity and reverse lookup collapse; memory and self-suppression persist (orientation: Table 4).

learns embeddings, attention, and all parameters end-to-end (Appendix A).

The normal condition recovers the reported two-stage behaviour: early memorisation followed by gradual emergence of linear truth decodability. Final probe AUCs match the original findings, and the learned OV circuit exhibits the same four sub-block structure identified in the original analysis (Figure 3; Appendix C). Table 3 summarises implementation differences resolved during reproduction, including learning-rate and initialisation choices required to recover the reported behaviour.

3.2 Truth Co-occurrence Drives Learned Encoding in the Synthetic Setting (RQ1)

RQ1 tests whether truth co-occurrence is required for linear truth encoding or whether attribute frequency alone is sufficient. We remove mutual information between paired truth variables while preserving the mapping, sequence structure, and marginal frequencies (Appendix A).

Removing statistical dependence between facts eliminates truth decodability in the fully trainable model: probe AUC remains at chance (maximum 0.536) compared with 0.985 ± 0.002 and 0.954 ± 0.009 in normal settings, while factual recall remains perfect (Figure 2a,b; Tables 2 and 2). The ablated model instead learns a Bayes-optimal frequency-based strategy, predicting the correct attribute with confidence ρ regardless of context truth (Appendix Figure 8). Thus, attribute frequency alone cannot explain truth encoding; shared truth co-occurrence provides the predictive signal driving the learned representation.

3.3 The Constrained Toy Transformer Remains Decodable Without Co-occurrence

In contrast, the constrained toy model remains linearly decodable after removing truth co-occurrence: AUC decreases only from 0.971 to 0.959 and remains above 0.95 throughout training (Figure 2c; Table 2).

Unlike the fully trainable model, where truth representations emerge in response to predictive signals, the toy model’s linear separability is primarily induced by its fixed architectural structure. Detailed OV analysis confirms that the four sub-block circuit decomposition remains aligned after ablation (Table 4), demonstrating that the representation is structurally induced rather than learned because truth is functionally useful (Appendix C).

3.4 The Ablation Shrinks the Circuit Without Rotating It

OV circuit analysis shows that the toy model’s truth decodability survives because the underlying circuit structure is preserved despite changes in magnitude. As shown in Figures 3 and 4, embedding-space components (negative identity and reverse lookup) weaken substantially in amplitude ($7\text{--}10\times$ decrease overall; negative identity $3.59 \rightarrow 0.52$; reverse lookup $4.01 \rightarrow 0.39$), while memory and self-suppression persist or increase.

However, Table 4 shows that all four circuit blocks maintain alignment with their theoretical targets. This demonstrates that cosine similarity and probe accuracy alone fail to capture the intervention’s effect; amplitude-sensitive analyses are required to reveal the change (Appendix C).

3.5 The Behavioural Truth Effect Replicates at Scale but Is Partially Explained by Confounds (RQ2)

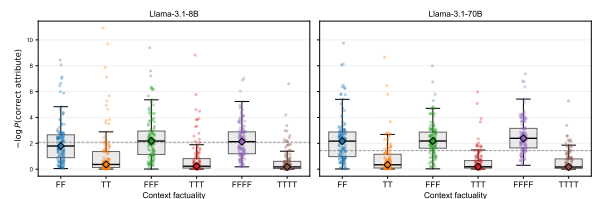


Figure 5: **Generalisation of Ravfogel et al. (2025) to scales:** NLL of the correct attribute after k false (F^k) or true (T^k) context sentences (CounterFact P103; 896 prompts per model). Results report the paper’s original pipeline, despite its known issues (see Appendix E).

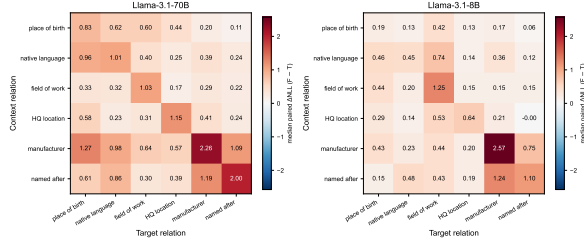


Figure 6: **Transfer of Context-Truthfulness across relations:** Median paired ΔNLL (false – true context) over the controlled 6×6 context–target relation grid (64 pairs per cell). Positive values indicate that false context reduces recall of the target fact.

RQ2 evaluates whether behavioural truth effects persist at scale and under controlled conditions. The faithful replication recovers the original finding: false contexts reduce confidence in subsequent factual predictions, with FF–TT NLL gaps of 1.43 nats at 8B and 1.83 nats at 70B (Figure 5; full condition-level distributions are reported in Table 7).

However, controlled experiments remove eight identified confounds, including attribute priming, answer leakage, and unpaired comparisons (Appendix E; Table 6). The resulting analysis, summarised in Table 8, shows that approximately 45% of the original effect is explained by these surface confounds. After correction, Figure 6 shows that a smaller but robust truth-context effect remains across relations, demonstrating that contextual truth signals generalise beyond individual facts, although their magnitude was previously overstated in Ravfogel et al. (2025).

3.6 The Truth Direction Becomes Relation-General with Scale (RQ3)

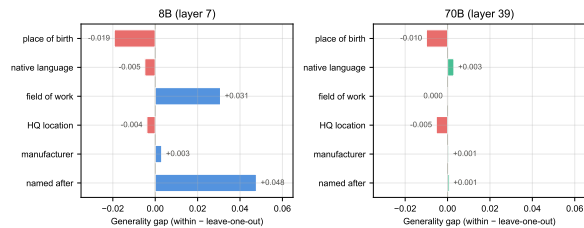


Figure 7: **Generality Gap (difference in probe performance within and across relations) for pretrained models:** We compare within-relation and leave-one-out probes for Llama-3.1 8B and 70B. Positive values favour within-relation probes. At 70B every relation has a gap within ± 0.010 .

RQ3 tests whether learned truth representations

generalise across factual relations and whether this improves with scale. We train linear probes using balanced relation-specific datasets and evaluate both within-relation and leave-one-relation-out transfer (Appendix F).

At the selected layers, the within-minus-LOO AUC gap reaches $+0.048$ at 8B, whereas all six gaps at 70B lie within approximately ± 0.010 (Figure 7; Table 11; Appendix F). Within-relation and LOO performance are therefore more similar at 70B under our evaluation protocol. However, LOO probes use 1,000 training examples, compared with 132–134 per within-relation CV fold, so this comparison does not control for training-set size.

However, probe transfer and behavioural transfer capture different properties. Probe transfer measures whether truth information is linearly recoverable across relations, whereas behavioural evaluation measures whether contextual truth changes predictions. We detect no statistically significant correlation between the two measures (Figure 13). Given the limited sample ($n = 36$ per model), this may reflect insufficient statistical power and does not establish that representing and using truth are independent.

4 Conclusion

Beyond reproducing and extending recent results, our study clarifies what linear truth representations reveal about language model internals. Across replications, ablations, and extensions of Ravfogel et al. (2025), we find that the emergence of linear truth decodability depends on truth co-occurrence and architecture in the tested synthetic settings. In our single-layer synthetic setting, truth co-occurrence is necessary for fully trainable transformers to develop linear truth decodability under the tested training conditions, while constrained toy transformers remain decodable after the ablation. In pretrained models, behavioural truth effects are robust across relations and scales, but surface confounds explain approximately 45% of the originally observed signal, highlighting the importance of controlling for statistical shortcuts in factuality evaluation. Under our probing protocol, cross-relation performance is closer to within-relation performance at 70B than at 8B, although training-set sizes differ between the two probe protocols. We detect no statistically significant association between behavioural and representational transfer, which does not establish independent capabilities.

More broadly, our findings suggest that high probe accuracy can reveal recoverable truth-related structure even without a functional truth-tracking mechanism, motivating amplitude-sensitive circuit analyses and causal interventions alongside probing methods. Thus, evaluating whether models represent factual truth requires not only asking what information is encoded in activations, but whether and how that information is used to guide model behaviour.

5 Limitations

The scope of our experiments and the limitations of behavioural and probing measurements constrain the conclusions we can draw.

Synthetic necessity is shown only in a toy setting. Our necessity result is established in a single-layer synthetic transformer, enabling causal isolation but differing substantially from modern LLMs. It should therefore be interpreted as evidence about the mechanism underlying linear truth probes rather than necessity in pretrained models.

Behavioural confounds are controlled jointly. Our behavioural audit jointly controls multiple confounds, so the observed $\approx 45\text{--}69\%$ reduction reflects their combined effect rather than individual contributions. The remaining $\approx 1\text{-nat}$ effect is best viewed as a residual signal.

Transfer is partly confounded with probe training size. Leave-one-relation-out probes are trained on substantially more data than within-relation probes, so improved transfer at 70B may reflect both more general representations and additional training data.

Probe accuracy limits resolution. Near-ceiling probe performance limits comparisons across conditions, and behavioural–representational correlations are descriptive because they are computed from related, non-independent observations.

Limited model and relation coverage. Experiments cover only two Llama-3.1 models and six CounterFact relations. The observed scaling trend therefore provides evidence for these settings rather than a general scaling law.

Linear probes do not establish causal use. Linear decodability does not prove that models use the recovered direction during generation. Our

conclusions are therefore limited to behaviourally relevant, linearly decodable truth signals.

Frozen activations and compute constraints. Analyses reuse a single set of saved activations and logits collected via remote inference, improving reproducibility but conditioning results on the sampled prompts rather than repeated resampling.

Acknowledgements

This work is supported by Cambridge University Press & Assessment. With thanks to Prof Andrew Caines, Dr Diana Galvan-Sosa and Lily Goulder. This work used the National Deep Inference Fabric (NDIF) and the NNsight library (Fiotto-Kaufman et al., 2025). NDIF is supported by the U.S. National Science Foundation under Award IIS-2408455. We thank the NDIF team for access to the shared inference infrastructure that enabled the Llama-3.1-70B experiments.

References

- Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it’s lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. 2024. Truth is universal: Robust detection of lies in LLMs. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*.
- Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations (ICLR)*.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630.
- Javier Ferrando, Oscar Obeso, Senthoran Rajamanoharan, and Neel Nanda. 2025. Do I know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Jaden Fried Fiotto-Kaufman, Alexander Russell Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Nikhil Prakash, Carla E. Brodley, Arjun Guha, Jonathan Bell, Byron C Wallace, and David Bau. 2025. *NNsight and NDIF: Democratizing access to foundation model internals*. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

- Zorik Gekhman, Eyal Ben-David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. 2025. [Inside-out: Hidden factual knowledge in LLMs](#).
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.
- Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. 2023. [Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness?](#)
- Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling (COLM)*.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*.
- Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Shauli Ravfogel, Gilad Yehudai, Tal Linzen, Joan Bruna, and Alberto Bietti. 2025. [Emergence of linear truth encodings in language models](#). NeurIPS 2025.
- Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

A Synthetic Model Configurations

The synthetic experiments compare two transformers that differ in how much of the model is learned. The **constrained toy model** reproduces the architecture of [Ravfogel et al. \(2025\)](#). Its token embeddings are fixed one-hot (hence orthogonal) vectors. Its query and key projections are zeroed and frozen, so attention is uniform over positions. Only the value and output projections W_v, W_o are trained. The **fully trainable model** removes all three constraints: embeddings and attention are learned, and every parameter is optimised end-to-end. One-hot, frozen embeddings are therefore specific to the constrained toy model. The contrast between the two models is what RQ1 tests (Appendix B).

Table 1 lists both configurations. Two points matter when interpreting the comparison. First, the two models also differ in scale, optimiser, learning rate and training length, so the comparison is not a single-factor ablation. Second, the only training setting run for both models is $\rho=0.8$, where ρ is the probability that a training statement is true. The fully trainable model is also run at $\rho=0.99$. Both models are evaluated on balanced data ($p_{\text{eval}}=0.5$), and a probe at chance has AUC 0.5.

B Reproducing the Synthetic Results

We report the synthetic reproduction and ablation results, relate probe decodability to model predictions, and document differences from the reference implementation.

Does truth decodability require truth to be useful? In the normal condition, both statements in a training sequence share one truth value $T \sim \text{Ber}(\rho)$, so the truth of the first pair predicts the truth of the second. The ablation draws the two truth values independently, $T_1, T_2 \sim \text{Ber}(\rho)$. The model still has to learn the same subject–attribute mapping, but knowing whether the first statement is true no longer helps predict the second. If a truth probe succeeds only because the model has learned to use truth, its AUC should fall to chance under the ablation. If the probe succeeds for other reasons, its AUC should survive.

Table 2 shows that the answer differs between the two models. In the fully trainable model, the ablation reduces probe AUC to chance at both training settings: from 0.985 to 0.498 at $\rho=0.8$, and from 0.954 to 0.495 at $\rho=0.99$. No ablated seed exceeds 0.536. In the constrained toy model at $\rho=0.8$, AUC

	Constrained toy	Fully trainable
N / M	20 / 20	512 / 512
Vocabulary size	40	1025
d_{model}	83	256
Layers / heads	1 / 1	1 / 1
MLP	none	none
Normalisation	RMSNorm	RMSNorm
Embeddings	one-hot, frozen	learned
Attention	uniform (Q, K zeroed, frozen)	learned
Trainable parameters	W_v, W_o	all
Optimiser	SGD	AdamW
Learning rate	1.0	10^{-3}
Weight decay	10^{-4}	10^{-5}
Batch size	16	128
Training steps	2,000	50,000
Seeds ($\rho=0.8$)	$n=5$	$n=3$
Seeds ($\rho=0.99$)	—	$n=5$
ρ (p_{train})	0.8	0.8 / 0.99
p_{eval}	0.5	0.5

Table 1: Configurations of the two synthetic models.

falls only from 0.971 to 0.959. Mapping accuracy is 100% in every cell, so the fully trainable model’s collapse is not a failure to learn the mapping. In the fully trainable model, then, a linearly decodable truth signal appears only when truth is useful for prediction. In the toy model it appears either way.

This comparison shows that the two configurations differ in whether probe AUC depends on truth utility. It does not show which difference between them causes this. Besides the architectural constraints, the models also differ in the settings listed in Appendix A.

B.1 Model Predictions Under the Ablation

The model’s predictions show the same pattern as the probe. Figure 8 tracks the probability that the fully trainable model ($\rho=0.99$) assigns to the correct attribute $g(x')$. In the normal condition, the model learns to lower this probability on false sequences, ending at ≈ 0.49 . In the ablated condition, it assigns $\approx \rho$ on both true and false sequences. When the first pair carries no information about the second, this constant prediction minimises the loss. The ablated model therefore shows no linearly decodable truth signal and makes predictions that do not depend on truth.

Differences from the reference implementation.

Reproducing the synthetic pipeline required choosing between the paper and its released code where they disagree. Table 3 lists these choices and one result we did not reproduce.

- *Learning rate.* We follow the released code rather than the paper’s Appendix C. With this

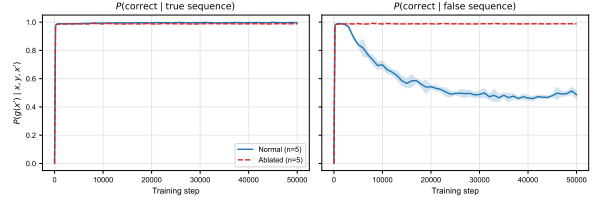


Figure 8: Probability of the correct attribute $g(x')$ on true (left) and false (right) sequences over training. Fully trainable model, $\rho=0.99$, normal vs. ablated condition (mean $\pm$ SD, $n=5$ per condition).

setting, the fully trainable model reaches a final mean AUC of 0.954 at $\rho=0.99$.

- *Initialisation.* The paper fixes $W_o=I$ and initialises W_v at zero. We train both factors. When both factors are trainable, initialising both at zero is a stationary point: the gradient with respect to each factor is proportional to the other factor, so both gradients are zero. We therefore initialise both with small Gaussian noise. The problem does not arise in the paper’s parameterisation, where W_o is a fixed identity.
- *Uniform attention.* We zero and freeze both Q and K . Zeroing either one already makes attention uniform, so this choice matches the original notebook, which zeroes only Q .

One discrepancy remains. At the end of training, Ravfogel et al. (2025) report $P(\text{correct} \mid \text{false}) \approx 0.25$ (their Figure 3b), whereas our reproduction gives ≈ 0.48 . We report this as an unresolved

Model	ρ	Final probe AUC		Best
		Shared T	Independent T_1, T_2	
Fully trainable	0.99	.954 $\pm$.009	.495 $\pm$.011	.536
Fully trainable	0.80	.985 $\pm$.002	.498 $\pm$.016	.536
Constrained toy	0.80	.971 $\pm$.005	.959 $\pm$.006	.967

Table 2: Final truth-probe AUC (mean $\pm$ SD over seeds) in the normal (shared T) and ablated (independent T_1, T_2) conditions. The last column is the highest AUC among ablated seeds. Mapping accuracy is 100% in every setting.

difference. It concerns output probabilities, which do not enter the comparisons in Table 2.

C OV-Circuit Structure Under the Ablation

The toy model keeps its truth-probe signal when truth is useless (Appendix B). Here we ask whether its learned weights change under the ablation. Ravfogel et al. (2025) describe the toy model’s OV matrix $W = W_o W_v$ as four sub-blocks. If these blocks formed because truth is useful, they should degrade under the ablation. If they follow from the constrained architecture, they should survive. Figure 9 confirms that the four blocks appear in our reproduction under the normal condition.

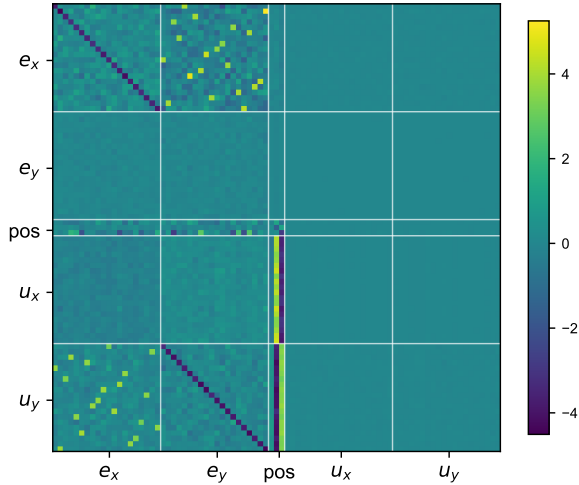


Figure 9: OV circuit $E W_o W_v E^\top$ of the constrained toy model (normal condition, step 2000), showing the four sub-blocks reported by Ravfogel et al. (2025).

To measure each block, we reorder the attribute dimensions by the ground-truth mapping g . Each block then has a signed-identity target. Write e_t and u_t for the input and output directions of token t . The four blocks are:

- *memory*: $e_x \rightarrow u_{g(x)}$, target $+I$;
- *negative identity*: $e_x \rightarrow -e_x$, target $-I$;

- *reverse lookup*: $e_y \rightarrow e_{g^{-1}(y)}$, target $+I$;
- *self-suppression*: $e_y \rightarrow -u_y$, target $-I$.

For a block B with target T , we report the cosine similarity $\langle B, T \rangle / (\|B\| \|T\|)$. This measures orientation only. We track magnitude separately with the signed projection amplitude and the Frobenius norm.

Table 4 shows that all four blocks remain well aligned with their targets under the ablation, with cosine at least 0.81 in both conditions. The changes are small and go in both directions. Memory and self-suppression move slightly closer to their targets ($+0.026$ and $+0.029$). Negative identity and reverse lookup move slightly away (-0.040 and -0.015). In the toy model, therefore, neither the probe signal nor the four-block OV structure depends on truth being useful. This is consistent with both being produced by the constrained architecture. The analysis covers only the constrained toy model at $\rho=0.8$.

D Data Construction

The behavioural and probing experiments on Llama-3.1-8B and Llama-3.1-70B use COUNTERFACT (Meng et al., 2022). Its (subject, relation, attribute) triples are rendered into natural-language templates. How the data are cleaned and how false statements are built determine which confounds the later experiments can rule out.

Relations and cleaning. We use six relations: place of birth (P19), native language (P103), field of work (P101), headquarters location (P159), manufacturer (P176), and named after (P138). We chose them for semantic diversity and low subject and target overlap, before inspecting any probe results. Records pass through four deterministic stages:

1. raw extraction;
2. removal of duplicate subjects within a relation;

Component	Reference	This study
Learning rate (fully trainable)	10^{-4} in the paper (Appendix C); 10^{-3} in the released code.	10^{-3} for 50,000 steps.
OV initialisation (toy)	W_o fixed to the identity; W_v initialised at zero.	W_o and W_v both trainable, initialised $\mathcal{N}(0, 0.02^2)$.
Uniform attention (toy)	Q zeroed and frozen in the original notebook.	Q and K both zeroed and frozen.
$P(\text{correct} \mid \text{false})$ at end of training	≈ 0.25 (original Figure 3b).	≈ 0.48 (not reproduced).

Table 3: Implementation choices and observations in the synthetic reproduction.

OV sub-block	Normal	Ablated
Memory ($e_x \rightarrow u_{g(x)}$)	$0.939 \pm .005$	$0.965 \pm .004$
Neg. identity ($e_x \rightarrow -e_x$)	$0.856 \pm .003$	$0.816 \pm .007$
Reverse lookup ($e_y \rightarrow e_{g^{-1}(y)}$)	$0.823 \pm .005$	$0.808 \pm .008$
Self-suppression ($e_y \rightarrow -u_y$)	$0.928 \pm .006$	$0.957 \pm .004$

Table 4: Cosine similarity between each OV sub-block and its signed-identity target. Constrained toy model, $\rho=0.8$, step 2000, mean $\pm$ SD over five seeds.

- removal of subjects that appear in more than one relation, to prevent leakage;
- removal of duplicate sentences.

As Table 5 shows, the last stage removes nothing further in any relation. Every relation keeps at least 270 records, more than the 200 balanced examples drawn per relation for probing. Cleaning does not fix the skew in P103, which remains 64.1% “French”. This skew motivates two of the controls below.

Rel.	Raw	Dedup	X-rel	Final	Top attr. (%)
P19	779	779	776	776	London (7.6)
P103	919	918	908	908	French (64.1)
P101	545	519	512	512	physics (13.4)
P159	756	754	753	753	London (11.4)
P176	911	900	899	899	Toyota (11.0)
P138	279	272	270	270	Victoria (3.6)

Table 5: Record counts after each cleaning stage. Dedup: within-relation subject deduplication. X-rel: cross-relation subject removal. Final: after duplicate-sentence removal. The last column gives the most frequent attribute after cleaning.

Negative construction. In the controlled sets, false sentences reassign attributes with a fixed-seed permutation σ such that $a_{\sigma(i)} \neq a_i$ for every record. Because a permutation preserves the multiset of attributes, true and false sentences have identical attribute frequencies. A probe or a behavioural

effect therefore cannot separate them by attribute frequency alone. Such a permutation exists only if no attribute accounts for more than half of the records. The 40% cap on “French” in P103 meets this condition. The faithful reproduction (Stage A, Appendix E) instead keeps the original COUNTERFACT negatives, without marginal matching or the other controls.

Prompt sets. The faithful behavioural reproduction uses 896 P103 prompts: 128 for each of the seven conditions in Table 7. The controlled design (Stage B) contains 4,992 prompts. These are built from 64 paired targets for each of the 36 context-relation $\times$ target-relation cells, each with a true and a false context variant. All prompt files are frozen before activation extraction, so both models see identical inputs.

E Behavioural Pipeline and Confound Audit

The behavioural effect reported by Ravfogel et al. (2025) (their Figure 6a) is a gap in target-attribute negative log-likelihood (NLL): the target is less likely after false context statements than after true ones. We measure this gap in two stages. The *faithful recipe* (Stage A) keeps the original design. The *controlled paired design* (Stage B) removes the design confounds. Comparing the two stages bounds how much of the original gap those confounds explain.

Audit of the original pipeline. Table 6 lists eight audit items, and they are of three kinds. Items 1–3 are design confounds: each can open a false–true gap even if the model does not track whether the context is true. Items 4–5 and 7–8 are implementation issues. Item 6 concerns the probe data, not the behavioural evaluation. Stage A keeps confounds 1–3 so that it stays comparable with the original

experiment, and applies fixes 4, 5, 7 and 8. Stage B also controls for 1–3. Item 7 is a risk in the code path, not an established error: we found no evidence that it excluded observations in the inspected notebook, and all our per-prompt NLLs are finite. Item 8 affects efficiency only.

How much of the gap do the confounds explain?

The faithful recipe reproduces the original effect. Table 7 gives the NLL in each condition. In the no-context baseline the mean is well above the median, because a few facts have very high NLL; we therefore compare medians.

At $k=2$, the median false–true gap for Llama-3.1-8B is 1.43 nats, close to the 1.52 that [Ravfogel et al. \(2025\)](#) report for Llama-3-8B. The gap widens with context length, reaching 1.95 nats for 8B and 2.22 nats for 70B at $k=4$ (Table 8). Under the controlled design, the same relation (P103) gives a gap of 0.45 nats at 8B and 1.01 at 70B. These are 69% and 45% smaller than the faithful $k=2$ gaps.

A gap survives the controls at both scales: about 1 nat at 70B, which makes the target about $2.7\times$ more probable after a true context than after its matched false context, and 0.45 nats ($\approx 1.6\times$) at 8B. Two caveats limit this comparison. First, the faithful gaps are differences between unpaired condition medians, whereas the controlled gaps are medians of paired differences. Second, the two designs differ in several choices at once. The reduction therefore bounds the effect of the confounds taken together. It does not attribute the reduction to any single confound.

Suppression or facilitation? The gap has two possible sources. False contexts could raise the target’s NLL above the no-context baseline (suppression). True contexts could lower it below the baseline (facilitation). Table 9 measures each against the baseline.

Under the faithful recipe, the 8B gap comes almost entirely from facilitation. False contexts leave the median NLL within 0.3 nats of the baseline, whereas true contexts lower it by 1.72–1.91 nats. At 70B, false contexts also suppress, by +0.73 to +0.95 nats, alongside facilitation of -1.10 to -1.27 .

Under the controls, false-context suppression is small: +0.14 (8B) and +0.20 (70B) across all cells. Within-relation, it is +0.16 at 70B and absent at 8B (-0.14). Most of the 70B suppression

seen under the faithful recipe therefore disappears once the confounds are removed. This is consistent with attribute priming (item 1), but because the controls are applied together they cannot isolate priming. Within-relation, true-context facilitation persists under the controls (-1.09 and -1.17), so it accounts for most of the remaining within-relation gap.

Does the effect transfer across relations? The controlled design pairs each false context with a matched true context across all six context relations and six target relations. For each pair we compute

$$\Delta\text{NLL} = \text{NLL}_{\text{false}} - \text{NLL}_{\text{true}},$$

so a positive value means the target is less likely after the false context than after its matched true context.

Table 10 shows that the effect is present in almost every cell but is concentrated within relations. The median ΔNLL is positive in 71 of the 72 cells; the remaining cell is 0.00. The median of the diagonal cells is 0.87 nats at 8B and 1.09 at 70B, compared with 0.21 and 0.40 for the off-diagonal cells. At 70B, the diagonal cell is the largest entry in every row. At 8B, this holds in only three of six rows (P101, P159, P176). The 70B value exceeds the 8B value in 29 of 36 cells, and four of the seven exceptions are in the P101 target column. Scale therefore increases the controlled effect in most, but not all, relation pairs.

F Probing Details

We describe how truth probes are trained, how layers are selected, and how performance is compared within and across factual relations.

Probe protocol. Truth probes on the pre-trained LMs are ℓ_2 -regularised logistic regressions (scikit-learn `LogisticRegression`, `lbfgs`, $C=1.0$, `max_iter= 1000`, tolerance 10^{-4} , with `intercept`). They are fit on residual activations normalised with `StandardScaler`. The configuration is the same for every relation, layer and model. Each relation contributes 200 balanced examples (100 true, 100 false). We use three kinds of probe, reporting ROC-AUC as the primary metric and balanced accuracy as the secondary metric:

- *Within-relation* probes use stratified, class-balanced 3-fold cross-validation (seed 42), giving about 132–134 training and 66–68 test examples per fold.

Audit item	Observation	Handling in this study
1. Attribute priming	P103 is 64% French, and context attributes can coincide with the target answer.	Cap French at 40%, match attribute marginals, and exclude context attributes equal to the target answer.
2. Subject overlap	Context and target are sampled independently, so the target subject can appear in the context.	Exclude the target subject from the context.
3. Unpaired conditions	FF and TT prompts use different targets and context facts.	Pair each TT and FF prompt on target and context facts; only the context attributes differ.
4. Tokenisation	A prompt ending in a space, scored against a leading-space target token, gives an inconsistent token boundary.	Build prompts without a trailing space and encode targets with a leading space.
5. Prompt construction	Deleting the last whitespace-delimited word can leave part of a multi-word attribute in the prompt.	Build the target stub from the template and metadata.
6. Probe data	The original layer-11 probe uses raw P103 representations with skewed attribute frequencies.	Fit probes on marginal-matched data.
7. Numerical handling	The notebook has a branch that maps infinite values to NA and drops them, but it returns FP32 probabilities; we could not establish that any observations were excluded.	Compute log-probabilities with FP32 log-softmax; all per-prompt NLLs are finite and no rows are dropped.
8. Forward passes	The probability helper runs two forward passes per measurement.	Return logits and activations from a single trace.

Table 6: Audit of the original behavioural pipeline. Fixes 4, 5, 7 and 8 apply to both stages. Controls 1–3 apply to Stage B only. Item 6 applies to probing.

Condition	8B		70B	
	median	mean	median	mean
FF ($k=2$)	1.79	2.08	2.17	2.31
TT ($k=2$)	0.37	1.09	0.34	0.82
FFF ($k=3$)	2.18	2.42	2.18	2.32
TTT ($k=3$)	0.23	0.76	0.19	0.58
FFFF ($k=4$)	2.13	2.25	2.39	2.55
TTTT ($k=4$)	0.18	0.59	0.17	0.55
none (baseline)	2.08	3.22	1.44	3.03

Table 7: Target-attribute NLL (nats) under the faithful recipe, P103, $n=128$ prompts per condition.

Median false–true gap (nats)	8B	70B
Faithful, $k=2$	1.43	1.83
Faithful, $k=3$	1.95	1.99
Faithful, $k=4$	1.95	2.22
Controlled, P103 within-relation	0.45	1.01
Reduction vs. faithful $k=2$	69%	45%

Table 8: False–true context gap under the faithful recipe and the controlled design. Faithful rows: difference between the FF^k and TT^k medians in Table 7. Controlled row: median paired Δ NLL with P103 as both context and target relation.

- *Single-source transfer* probes train on the 200 examples of one relation.
- *Leave-one-relation-out* (LOO) probes train on

	8B		70B	
	FF	TT	FF	TT
Faithful ($k=2$)	−0.29	−1.72	+0.73	−1.10
Faithful ($k=3$)	+0.09	−1.86	+0.74	−1.25
Faithful ($k=4$)	+0.05	−1.91	+0.95	−1.27
Ctrl. (all)	+0.14	−0.28	+0.20	−0.40
Ctrl. (within)	−0.14	−1.09	+0.16	−1.17

Table 9: Median NLL in each context condition minus the no-context baseline median (nats). Negative values indicate facilitation; positive values indicate suppression. Ctrl. (all) pools all context–target cells; Ctrl. (within) pools within-relation cells.

the 1,000 examples of the other five relations.

Transfer and LOO probes are evaluated on the 200 examples of a target relation that was not used in fitting. Probes on the synthetic models follow the original release: an SGDClassifier with 5-fold stratified cross-validation on 2,000 balanced samples at the second-subject position x' .

Layer selection. We extract residual activations at four normalised depths (25%, 50%, 75%, 100%). For each model we select the depth with the highest mean within-relation AUC. When depths are within 0.005 of each other, we take the shallower one. This gives layer 7 (25%; mean AUC 0.962)

Context	8B — target relation						70B — target relation					
	P19	P103	P101	P159	P176	P138	P19	P103	P101	P159	P176	P138
P19	0.19	0.13	0.42	0.13	0.17	0.06	0.83	0.62	0.60	0.44	0.20	0.11
P103	0.46	0.45	0.74	0.14	0.36	0.12	0.96	1.01	0.40	0.25	0.39	0.24
P101	0.44	0.20	1.25	0.15	0.15	0.15	0.33	0.32	1.03	0.17	0.29	0.22
P159	0.29	0.14	0.53	0.64	0.21	−0.00	0.58	0.23	0.31	1.15	0.41	0.24
P176	0.43	0.23	0.44	0.20	2.57	0.75	1.27	0.98	0.64	0.57	2.26	1.09
P138	0.15	0.48	0.43	0.19	1.24	1.10	0.61	0.86	0.30	0.39	1.19	2.00

Table 10: Median paired Δ NLL (nats) under the controlled design. Rows: context relation. Columns: target relation. Diagonal cells are within-relation.

for Llama-3.1-8B and layer 39 (50%; mean AUC 0.977) for Llama-3.1-70B (Figure 10). Because the layer is chosen using the same within-relation AUC that we report, those AUCs are slightly optimistic. LOO and transfer AUCs played no part in the choice.

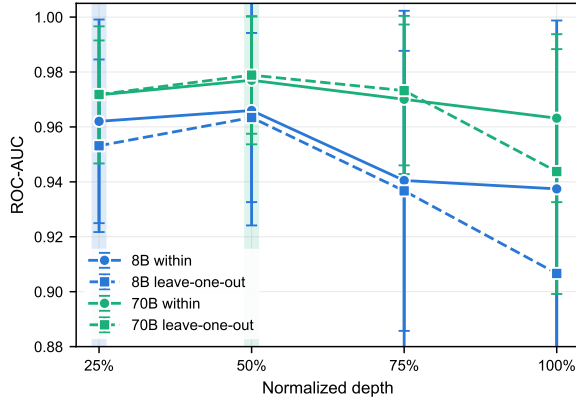


Figure 10: Truth-probe AUC at the four extracted depths. Marked layers (7 for 8B, 39 for 70B) are used in all relation-level analyses.

Does a truth probe trained on one relation generalise to others? Table 11 reports within-relation and LOO AUC for each relation at the selected layers. We call their difference, within – LOO, the *generality gap*. Within-relation AUC is at least 0.896 at 8B (lowest for P19) and at least 0.948 at 70B. At 8B, the largest gaps are for P138 (+0.048) and P101 (+0.031). At 70B, no relation has a gap larger than 0.010 in magnitude, and the mean signed gap is −0.002. Figure 11 shows that the mean absolute gap is smaller for 70B than for 8B at every extracted depth.

Two factors limit how far this supports more general truth probes at 70B. First, the LOO probe trains on about $7.5\times$ more data than a within-relation fold, so small or negative gaps partly reflect training-set size. Second, several 70B relations are at or near ceiling ($\text{AUC} \geq 0.998$), which limits

how large a gap can be. The single-source transfer matrices (Figure 12) avoid most of the size difference: they train on 200 examples, compared with about 133 per within-relation fold.

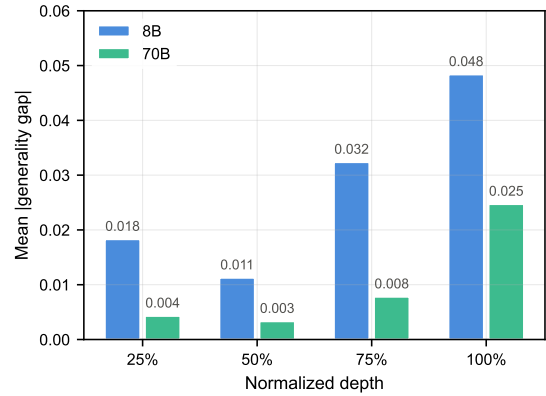


Figure 11: Mean absolute generality gap, $|\text{within} - \text{LOO}|$ averaged over the six relations, at each extracted depth.

G Behavioural vs. Representational Transfer

Probe transfer and behavioural transfer measure different things. Probe transfer asks whether a linear classifier trained on truth labels for one relation recovers truth labels for another. Behavioural transfer asks whether a false context in one relation changes the model’s prediction for a target in another. If the model’s response to context relied on the direction that the probes recover, relation pairs with better probe transfer should also show larger behavioural effects. We test this by comparing the two measures across the 36 relation pairs of each model (Figure 13).

We find no significant Spearman correlation at either scale. This null result is weak evidence. The 36 cells are not independent, because they share context and target relations, and 36 points give limited power. The result shows that, in these data,

Relation	8B (layer 7, depth 25%)				70B (layer 39, depth 50%)			
	within	SD	LOO	gap	within	SD	LOO	gap
P19 (place of birth)	0.896	0.023	0.915	−0.019	0.949	0.017	0.959	−0.010
P103 (native language)	0.963	0.007	0.968	−0.005	0.981	0.009	0.978	+0.003
P101 (field of work)	0.952	0.017	0.921	+0.031	0.948	0.004	0.948	0.000
P159 (HQ location)	0.970	0.023	0.974	−0.004	0.987	0.009	0.992	−0.005
P176 (manufacturer)	0.998	0.003	0.995	+0.003	0.999	0.002	0.999	+0.001
P138 (named after)	0.994	0.007	0.946	+0.048	0.999	0.001	0.998	+0.001
Mean	0.962	0.013	0.953	+0.009	0.977	0.007	0.979	−0.002

Table 11: Per-relation truth-probe AUC at each model’s selected layer. Within: 3-fold cross-validated AUC (SD across folds). LOO: probe trained on the other five relations. Gap = within − LOO; the mean row averages signed gaps.

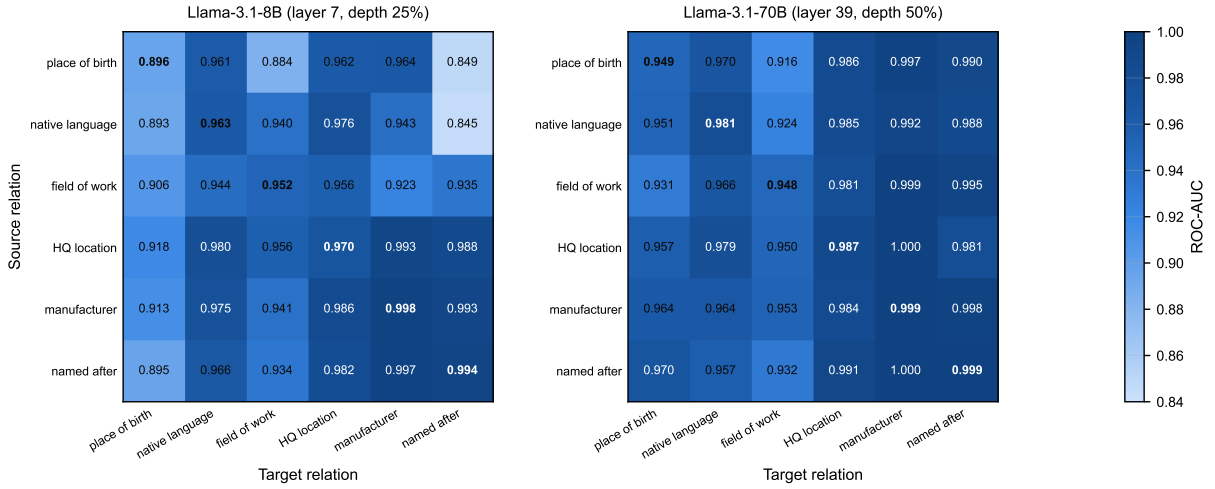


Figure 12: Single-source probe transfer AUC (training relation × test relation) at the selected layer, for both models.

relation-level variation in one measure does not detectably predict the other. It does not show that the two measures are unrelated.

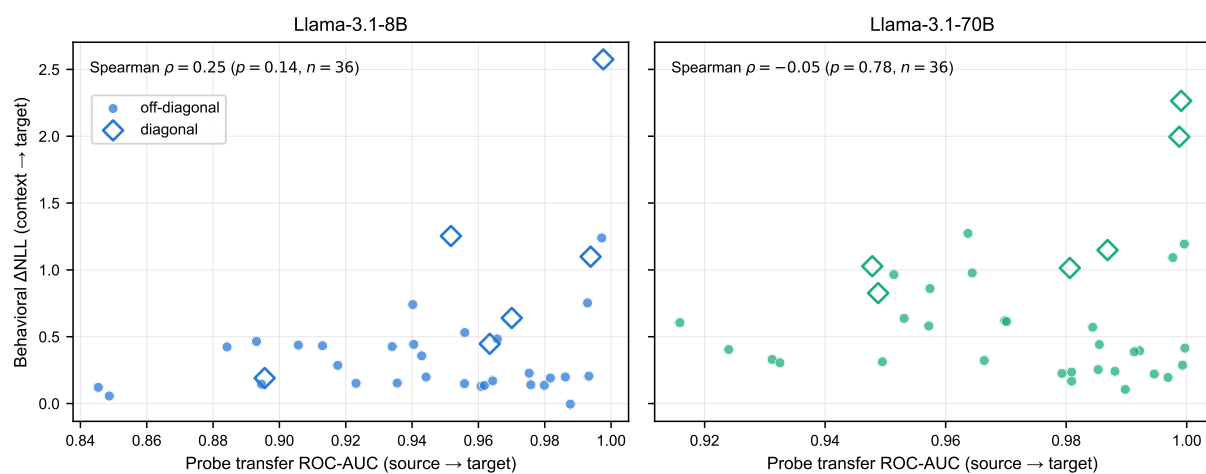


Figure 13: Probe transfer AUC against controlled behavioural Δ NLL for each relation pair ($n=36$ per model). Diamonds mark within-relation pairs.