CONTROLS BEFORE CLAIMS: AUDITING TRUTH AND CONTEXT-FAITHFULNESS STEERING DIRECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering adds a vector to a language model's hidden states to change its behaviour. Mean-difference directions are commonly used to make answers more truthful or to make a model follow retrieved context, and are usually judged by their gain on the metric used to select them. We test whether such gains come from the learned direction and what they cost. On five Llama, Qwen and Mistral models, we compare a truth direction and the stronger of two context directions with random vectors of the same length at the same layer, compare the two context directions at equal intervention size, and measure factual accuracy and abstention alongside the target metric. Context steering exceeds all matched random vectors on every model and raises context use on ConFiQA by up to 72 percentage points. However, the stronger context direction differs between models, the two largest context gains lower TruthfulQA accuracy, and on three models the stronger direction makes the model answer questions whose context contains no answer. The truth direction separates original from counterfactual contexts better than random projections on all five models, yet at the layer and scale chosen on validation data it does not raise TruthfulQA multiple-choice accuracy (MC1) beyond random vectors, and on Llama 3.1 8B Instruct it mainly suppresses answers. A truth direction built from affirmative and negated statements and applied at later layers raises MC1 from 0.51 to 0.59 on Llama Instruct and from 0.38 to 0.45 on Llama base, more than any of 30 random vectors, but not on the other three models. On Llama Instruct this gain comes with fewer correct ConFiQA answers when the context is correct. A gain on the selection metric is therefore weak evidence that steering improved the model's answers.

1 INTRODUCTION

Retrieval-augmented language models answer from two sources: the passages placed in the prompt and the knowledge stored in their weights. The two often disagree. A retrieved passage can be outdated or wrong, and a model that follows it gives an answer that is faithful to the passage but false; six strong models adopted incorrect retrieved content over their own correct prior answer more than 60% of the time (Wu et al., 2024). A passage can also lack the answer, and models that answer well from sufficient context often give a wrong answer instead of abstaining when the context is insufficient (Joren et al., 2025). Or the passage can be right while the model repeats what it memorised (Longpre et al., 2021; Xie et al., 2024). A method that changes how much a model relies on its context must therefore be judged on all three cases: whether the model uses the passage, whether its answers remain correct, and whether it still abstains when the passage holds no answer (Ming et al., 2025).

Activation steering is an attractive way to control these behaviours because it needs no retraining. A direction in the model's hidden-state space is estimated from contrasting inputs, for example as the mean activation difference between true and false statements, and a scaled copy of it is added to the hidden state during generation. Truth directions have been used to raise TruthfulQA accuracy (Li et al., 2023; Zou et al., 2023), and context directions to make models follow the supplied evidence (Anand et al., 2026; Goyal & Daumé III, 2026). Such a direction is normally evaluated on the metric it was selected for: its layer and scale are chosen on validation data, and the gain on the test split is reported.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

That report leaves three questions open. First, adding a random vector can itself change a model's behaviour (Korznikov et al., 2025; Malla et al., 2026), so a gain over the unsteered model does not show that the learned direction is responsible. Second, ContextFocus compared its passage-presence contrast with an answer-option contrast on only one model (Anand et al., 2026), and other steering methods often fail to transfer between model families (Da Silva et al., 2025). Third, a direction that improves one property can harm another. Factuality-enhancement methods can reduce faithfulness to edited context (Bi et al., 2025b), and steering one safety-relevant behaviour shifts others (Siu et al., 2026). A context direction that makes a model follow passages might also make it answer when the passage holds no answer, and a truth direction that raises a multiple-choice score might do so by making free-form answers less informative.

We test whether the gains of three standard mean-difference directions survive controls that earlier studies apply only in part, on five models from three families and on the same test items. The truth direction $\mathbf{v}_{\mathrm{truth}}$ contrasts true and false statements. The context direction $\mathbf{v}_{\mathrm{context}}$ contrasts prompts with and without a passage, as in ContextFocus (Anand et al., 2026), and $\mathbf{v}_{\mathrm{faithful}}$ contrasts the answer that a counterfactual passage supports with the original answer it replaced, as in the faithfulness steering of Goyal & Daumé III (2026). The truth direction and the stronger context direction are compared with random vectors of the same length at the same layer, and an effect is credited to the direction only if it exceeds what these vectors produce. We also compare the two context directions at equal intervention size, sweep the steering layer, and test whether $\mathbf{v}_{\mathrm{truth}}$ still ranks true above false when statements are negated. We measure context use on ConFiQA (Bi et al., 2025a), abstention on FaithEval questions whose passage contains no answer (Ming et al., 2025), and factual answering with TruthfulQA MC1 (Lin et al., 2022), the fraction of questions on which the model scores a true candidate answer above every false one (Figure 1).

The controls change several conclusions that the selection metric alone would support. Our contributions are:

- **Context steering beats random vectors, but the better construction varies across models.** On every model the stronger of the two context directions exceeds all matched random vectors, although these random vectors alone shift context use by up to 42 percentage points on Llama Instruct (Section 4). At equal intervention size $\mathbf{v}_{\mathrm{faithful}}$ is stronger on both Llama models and $\mathbf{v}_{\mathrm{context}}$ on Mistral, and on three models the stronger direction makes the model answer questions whose passage contains no answer (Section 5).
- **The standard truth direction detects counterfactual passages, but at its validation-selected setting it does not raise MC1.** Its projection separates original from counterfactual passages better than random projections on all five models, but at that setting its MC1 change never exceeds the largest random-vector change, and on Llama Instruct it mostly suppresses answers (Section 6). Validation selects layers where the direction fails a negation test; a polarity-balanced direction (Bürger et al., 2024) at later layers raises MC1 on both Llama models and on none of the other three (Section 7).
- **The conclusions survive audits of the benchmark and the judge.** We flag 14% of the ConFiQA test items as potentially defective and show that removing them changes no conclusion, audit the automatic judge against two blind annotators, and release code, per-item flags and run summaries (Appendices F and G).

2 Related work

Truth directions. Truth labels can be predicted from hidden states (Azaria & Mitchell, 2023; Burns et al., 2023), and Marks & Tegmark (2024) show that mean-difference directions both separate true from false statements and flip true/false judgements when added. Bürger et al. (2024) find that a direction from affirmative statements conflates truth with sentence polarity, propose a general truth direction that also covers negated statements, and leave interventions with it to future work. Truth probes transfer poorly across datasets (Orgad et al., 2025), and context changes the geometry of truth vectors (Adarsh et al., 2026). Inference-time intervention (ITI) raises TruthfulQA scores by shifting attention-head activations along directions probed on TruthfulQA itself, at some cost in helpfulness (Li et al., 2023). We ask what a residual-stream direction built from other true and false statements does under knowledge conflict, and whether its TruthfulQA gains survive random-vector controls and free-form scoring.

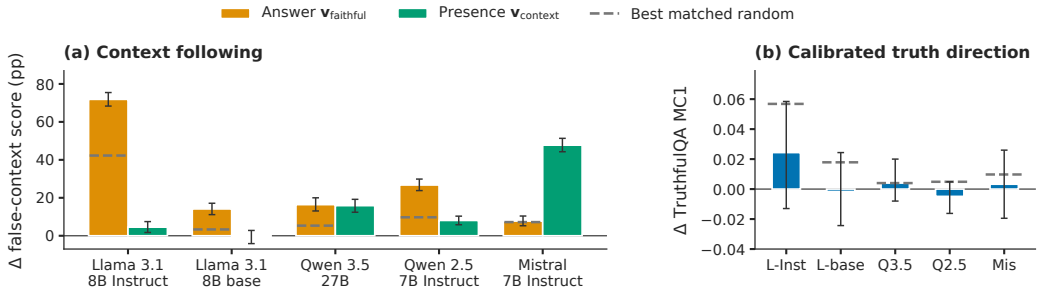


Figure 1: Steering effects at the validation-selected (calibrated) settings, with 95% paired bootstrap intervals. **Left:** change in the ConFiQA false-context score $P_c - P_o$ (the share of answers that give the passage's answer minus the share that give the original answer, in percentage points) under each model's two context directions. Dashed marks on the $\mathbf{v}_{\text{faithful}}$ bars give the largest change among random vectors of the same length at the same layer (50 for Llama Instruct, 30 for Llama base, 10 for each other model). **Right:** change in TruthfulQA MC1 under the truth direction, with the largest change among random vectors at its layer and length (50 and 25 for the two Llama models, 5 for each other model). Section 3 explains the item and vector counts.

Context steering and knowledge conflict. ContextFocus (Anand et al., 2026) steers with the difference between prompts with and without an instruction and a passage, and on Llama-3.1-8B compares it with a contrast between answer options. Goyal & Daumé III (2026) steer faithfulness with the answer-text contrast we call $\mathbf{v}_{\text{faithful}}$ and find that it fails under distractor contexts. Minder et al. (2025) locate a one-dimensional subspace that controls context sensitivity through fine-tuning and a subspace search. Probes detect knowledge conflicts in hidden states before generation (Zhao et al., 2024; Gao et al., 2026), and conflicts have also been studied through attention heads (Jin et al., 2024; Li et al., 2025) and residual-stream editing (Zhao et al., 2025). SteeringSafety (Siu et al., 2026) steers FaithEval behaviour and reports side effects on TruthfulQA and other safety behaviours in three families, without random-vector controls. None of these steering studies compares $\mathbf{v}_{\text{context}}$ and $\mathbf{v}_{\text{faithful}}$ across families or measures truthfulness and abstention with random-vector controls.

Controls for steering and probing. Pres et al. (2024) argue that steering evaluations need baseline comparisons; random baselines test sparse autoencoders (Korznikov et al., 2026) and control tasks test probes (Hewitt & Liang, 2019; Belinkov, 2022). Random steering vectors can change safety behaviour (Korznikov et al., 2025), and concurrent work finds that norm-matched random directions move the same off-target behaviours as learned ones (Malla et al., 2026). Ferrando et al. (2025) compare entity latents with random directions of the same norm, and ITI reports a nearly inert random-direction baseline at the head level (Li et al., 2023). Steering effects also vary across inputs and can reverse sign (Tan et al., 2024; Braun et al., 2025). We apply these controls to truth and context directions together and measure side effects on the other property and on abstention.

3 Setup

Directions. For each contrasting pair of inputs we record the residual-stream hidden state at the last input token. The direction at layer ℓ is the mean difference $\mathbf{v}^{(\ell)} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{h}_i^{(\ell),+} - \mathbf{h}_i^{(\ell),-})$ between the hidden states of the positive and negative members of N pairs (Marks & Tegmark, 2024; Rimsky et al., 2024). Directions, calibration and controls are computed separately for each model and layer; no direction is transferred between checkpoints.

- $\mathbf{v}_{\text{truth}}$: 1000 pairs of true and false statements, 500 from the Marks–Tegmark cities and counterfactual sets (Marks & Tegmark, 2024) and 500 support and refute claims from VitaminC (Schuster et al., 2021). We keep only minimal pairs, whose two statements concern the same entity or the same evidence passage, and weight the two sources equally. The first source has short statements with unambiguous truth values, the second natural claims written against evidence. Neither overlaps TruthfulQA, which rules out test leakage.

- $\mathbf{v}_{\text{context}}$, the *presence* contrast: an instruction, a passage and a question, against the question alone, read at the token before generation. ContextFocus builds the same contrast from 1501 NQ-Swap examples (Longpre et al., 2021); we use 1000 NQ-Swap examples and our own 20 instruction paraphrases.

- $\mathbf{v}_{\text{faithful}}$, the *answer* contrast: NQ-Swap replaces the answer entity in a Natural Questions passage. For each of the same 1000 examples we append either the substituted answer, which the edited passage supports, or the original Natural Questions answer to the same passage and question, and read the last token of the answer turn. The direction therefore contrasts which answer the model commits to while the passage is held fixed.

Steering adds $\alpha\mathbf{v}^{(\ell)}$ to the hidden state at every token position of one layer; ContextFocus steers generated tokens only, and Appendix D reproduces its exact setting. $\mathbf{v}_{\text{truth}}$ is normalised to unit length before scaling; the context directions keep their extracted length.

Models and scope. Llama 3.1 8B Instruct is the primary model, and its base variant tests the role of instruction tuning (Grattafiori et al., 2024). Qwen 3.5 27B (Qwen Team, 2026), Qwen 2.5 7B Instruct (Qwen Team, 2024) and Mistral 7B Instruct v0.2 (Jiang et al., 2023) add two families and a larger model. The primary model receives the most extensive checks: 50 random vectors at the length of each of $\mathbf{v}_{\text{faithful}}$ and $\mathbf{v}_{\text{truth}}$, layer sweeps, the ContextFocus replication, judged answers for random vectors and the human audit of the judge. The other four models are evaluated with fewer random vectors, because each random vector needs its own generation pass.

Measures. ConFiQA pairs each question with the original passage and a counterfactual passage that supports a different answer (Bi et al., 2025a); we use the counterfactual passages and instruct the chat models, steered or not, to answer only from the passage. P_c is the percentage of responses that contain only the counterfactual (context) answer and P_o the percentage that contain only the original answer, found by matching normalised answer strings and aliases (Appendix F). We call their difference $P_c - P_o$ the *false-context score*. The score measures reliance on the false passage rather than accuracy, and it also rises when the model stops giving the original answer. We report changes in percentage points (pp). The test set has 1500 held-out items, 500 each from the single-hop, multi-hop and multi-conflict subsets. Each random vector is evaluated on a stratified slice of 750 of these items, drawn with its run's seed, and compared with the unsteered model on the same slice. Qwen 3.5, the largest model, uses the first 250 items of each subset in all calibrated runs.

Factual answering is measured on 616 held-out TruthfulQA questions (250 for Qwen 3.5) (Lin et al., 2022). MC1 counts a question as correct when the model gives a higher summed log-probability to some true candidate answer than to every false one. We use the candidate sets of the MC2 format, which contain several true answers, so random ranking already scores about 0.45 and our values are not comparable with published single-answer MC1 scores. For 80 of the 616 questions, "I have no comment." is a true candidate and never a false one, so a preference for declining can raise MC1. Because MC1 cannot show whether the model's own answers change, we also score free-form ConFiQA answers with a Qwen 3.5 27B judge, which labels each answer as faithful to the passage or not and as informative or not, where informative means that the answer commits to relevant content, right or wrong (Appendix G). Abstention is the percentage of FaithEval's 500 unanswerable items (300 for Qwen 3.5) on which the model says that the passage does not contain the answer, the correct response for these items (Ming et al., 2025). Unlike FaithEval's prompt, ours does not offer "unknown" as an answer, and a fixed phrase list detects abstention (Appendix F).

Calibration and intervention size. The *calibrated* setting of a direction is the layer and scale chosen on validation data: for $\mathbf{v}_{\text{truth}}$ the pair that maximises MC1 on 200 validation questions, and for the context directions the pair that maximises $P_c - P_o$ on 100 NQ-Swap validation items, among settings that raise WikiText-2 perplexity by less than 20% (Merity et al., 2017). Appendix A lists the chosen settings. The same scale adds very different amounts at different layers, so we describe every intervention by its *relative dose*: the length of the added vector divided by the mean hidden-state length at that layer. A dose of 0.5 adds a vector half as long as a typical hidden state. Calibrated doses range from 0.02 to 0.79. When two directions are compared at equal dose, each stays at its own calibrated layer and only its scale changes, so that both add vectors of the same relative length.

Table 1: Context steering at the calibrated settings. Scores are the ConFiQA false-context score $P_c - P_o$; Δ is its change from the unsteered model in percentage points. The stronger context direction is in bold, with its 95% paired bootstrap interval. The last two columns summarise random vectors of the same length at the layer of $\mathbf{v}_{\text{faithful}}$; each random change is measured against the unsteered model on its own 750-item slice. For Mistral, whose stronger direction is $\mathbf{v}_{\text{context}}$, the matched comparison is in Figure 2. Learned effects use 1500 test items (750 for Qwen 3.5).

	Unsteered			95% interval	Random vectors	
Model	score	Δ $\mathbf{v}_{\text{faithful}}$	Δ $\mathbf{v}_{\text{context}}$	(bold)	mean Δ	max Δ (n)
Llama 3.1 8B Instruct	-6.1	$\mathbf{+71.9}$	$+4.5$	$[+68.3, +75.5]$	$+12.5$	$+42.3$ (50)
Llama 3.1 8B base	44.7	$\mathbf{+14.1}$	-0.5	$[+11.1, +17.1]$	-1.4	$+3.3$ (30)
Qwen 3.5 27B	32.7	$\mathbf{+16.4}$	$+15.9$	$[+13.1, +20.0]$	0.0	$+5.3$ (10)
Qwen 2.5 7B Instruct	8.5	$\mathbf{+26.7}$	$+8.1$	$[+23.8, +29.9]$	$+2.2$	$+9.7$ (10)
Mistral 7B Instruct	5.7	$+7.7$	$\mathbf{+47.7}$	$[+44.3, +51.3]$	$+1.9$	$+7.2$ (10)

Controls. *Random vectors* test whether an effect is due to the learned direction. We scale random directions to the length of the learned intervention, add them at the same layer, and credit an effect to the direction only if it exceeds what they produce. The *equal-dose* comparison runs $\mathbf{v}_{\text{context}}$ and $\mathbf{v}_{\text{faithful}}$ at doses 0.25 and 0.5, so that neither context direction can win only because calibration gave it a larger intervention. *Layer sweeps* move a direction across layers at a fixed dose, with random vectors at every layer, to locate where the effect arises. *Extraction-source* ablations test whether $\mathbf{v}_{\text{faithful}}$ depends on its examples by re-extracting it from 200 NQ-Swap pairs, from bootstrap resamples and from 1000 ConFiQA items outside the test set. The *negation* test checks whether $\mathbf{v}_{\text{truth}}$ tracks truth or sentence polarity by counting how often the true statement of 748 negated true–false pairs projects higher on it (Bürger et al., 2024).

Readouts. To test whether a direction carries information about the context, we take the unsteered hidden state at the last prompt token under the original and under the counterfactual passage, project it onto the direction, which gives one number per prompt, and fit a logistic regression that predicts which passage the prompt contains. Items are split 70/15/15 into training, validation and test sets, with both versions of an item in the same split. The regression is fitted on the training set, the layer is chosen by the validation accuracy of the $\mathbf{v}_{\text{truth}}$ readout, and balanced accuracy, the mean of the per-class accuracies, is reported once on the test set. Repeating the procedure with random directions gives the reference that a learned direction must beat.

Statistics. Generation is greedy, with at most 64 or 100 new tokens depending on the run. Runs also differ in item slice, so each change is measured against the unsteered model of the same run. Confidence intervals are 95% bootstrap intervals over test items for the per-item difference between steered and unsteered outputs (2000 resamples), so they describe how much a result depends on the particular items. The spread of the random-vector effects answers a different question: how far an arbitrary vector of the same size moves the score.

4 Context steering depends on the learned direction

On every model, the stronger context direction moves context use further than any matched random vector. On Llama Instruct, $\mathbf{v}_{\text{faithful}}$ raises the false-context score by 71.9 pp: the share of answers that give the passage's answer rises from 20.9% to 70.5%, and the share that give the original answer falls from 26.9% to 4.7%. The largest effect among 50 random vectors of the same length at the same layer is $+42.3$ pp. On the other models the gains are smaller, but each exceeds the largest random change by more than 10 pp (Table 1; Figure 2 for Mistral).

Random vectors of the calibrated sizes can move the score by tens of points, so the unsteered model is not a sufficient reference. On Llama Instruct, random vectors at the length of $\mathbf{v}_{\text{faithful}}$ raise the score by $+12.5$ pp on average, and those at the larger length of the calibrated $\mathbf{v}_{\text{truth}}$ intervention shift it by -25.5 to $+55.9$ pp (Appendix B).

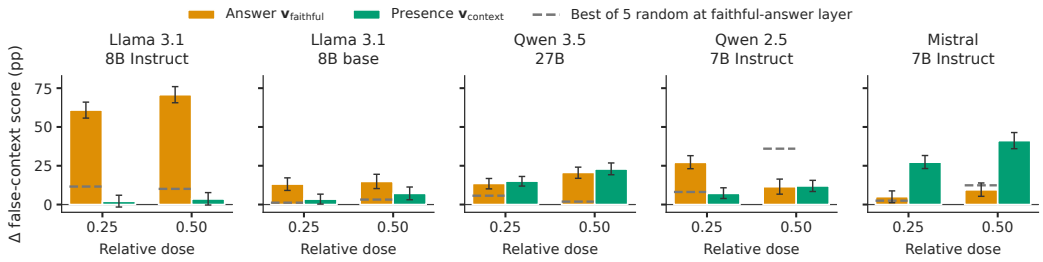


Figure 2: The two context directions at equal relative dose. Bars are changes in the false-context score on a 750-item slice, with 95% paired bootstrap intervals; each direction stays at its calibrated layer. Dashed marks are the largest change among 5 random vectors of the same length at the layer of $\mathbf{v}_{\text{faithful}}$, which on Mistral is also the layer of $\mathbf{v}_{\text{context}}$ (layer 14). The answer contrast dominates on both Llama models, the presence contrast on Mistral.

The context effect does not come from a component shared with $\mathbf{v}_{\text{truth}}$. Removing from $\mathbf{v}_{\text{faithful}}$ its component along $\mathbf{v}_{\text{truth}}$ and keeping its length changes the Llama Instruct effect from $+71.9$ to $+71.6$ pp and its MC1 cost from -0.169 to -0.167, and the effect on each other model by 3 pp or less. Because the two directions are nearly orthogonal (Appendix H), the removed component is small, and the result does not show that the two behaviours use separate mechanisms.

The effect does not depend on the examples used to extract $\mathbf{v}_{\text{faithful}}$, and it arises only in a narrow band of layers. Re-extracting $\mathbf{v}_{\text{faithful}}$ from 200 NQ-Swap pairs, from bootstrap resamples, or from 1000 ConFiQA items outside the test set gives Llama Instruct effects of $+68.9$ to $+70.4$ pp at matched length, against $+69.9$ pp for the original direction in the same run; Qwen 3.5 behaves the same way. Holding the dose fixed and moving $\mathbf{v}_{\text{faithful}}$ through Llama Instruct's layers, the effect is $+42$ pp at layer 8 and $+70$ pp at layer 12 but at most $+9$ pp at the five other tested layers (Figure 3a). On Qwen 3.5 the effective layers are 32 and 40 of 64.

5 The better context direction varies and lowers abstention

Which context contrast works depends on the model. Figure 2 compares the two directions at equal doses of 0.25 and 0.5. On both Llama models the answer contrast $\mathbf{v}_{\text{faithful}}$ is clearly stronger. On Mistral the order is reversed: at dose 0.5 $\mathbf{v}_{\text{context}}$ gives $+41.2$ pp and $\mathbf{v}_{\text{faithful}}$ $+9.5$ pp, both at layer 14, and a sweep of $\mathbf{v}_{\text{faithful}}$ over layers 4 to 24 at its calibrated dose never exceeds $+14$ pp, less than the $+27.3$ pp of $\mathbf{v}_{\text{context}}$ at dose 0.25, so the reversal is not a matter of layer choice. On Qwen 3.5 the two directions are similar. Qwen 2.5 favours $\mathbf{v}_{\text{faithful}}$ at dose 0.25, but at dose 0.5 two of its five random vectors already move the score by more than 28 pp and neither learned direction stands out. On Llama, our reconstruction of ContextFocus at its published setting raises the score by 18.4 pp, far less than $\mathbf{v}_{\text{faithful}}$, and does not reproduce the published gain (Appendix D), so the size of the Llama gap between the two contrasts may partly reflect our implementation.

On three of the four models that abstain, the stronger context direction also makes the model answer when the passage has none. On Llama Instruct and Qwen 2.5, the stronger context direction lowers abstention on FaithEval's unanswerable items below the range of random vectors of the same length, and on Mistral it lowers abstention from 29.6% to 3.8%, although random vectors were run there only at the smaller length of $\mathbf{v}_{\text{faithful}}$ (Table 2). Abstention is itself linearly steerable (Lavi et al., 2026), and our context directions move it as a side effect. On Qwen 3.5 the steered rate lies just below the random range, but the interval of the change includes zero ($[-8.7, +1.0]$ points), and unsteered Llama base never abstains. The false-context score cannot register this cost, because it is computed only on items whose passage contains an answer. Under the original passage, accuracy changes by at most 2.2 points on four models and stays within the random range on Qwen 3.5.

The truthfulness cost depends on the model and on the dose. MC1 falls on Llama Instruct and Mistral, the two models with the largest context gains, and its interval includes zero on the other three (Table 2). The cost is not fixed to the direction: halving Llama Instruct's dose from 0.5 to 0.25 keeps most of the context effect ($+60.9$ instead of $+70.8$ pp; Figure 2) and, in the same run, cuts

Table 2: Side effects of the stronger context direction at its calibrated setting. ΔMC1 is the change in TruthfulQA MC1 with its 95% paired bootstrap interval. Abstention is the percentage of FaithEval unanswerable items on which the model says the passage lacks the answer, unsteered → steered; brackets give the range over 5 random vectors at the layer and length of $\mathbf{v}_{\text{faithful}}$ (3 for Qwen 3.5). Judged columns are the percentages of the 1500 false-context ConFiQA answers that the judge labels informative and faithful to the passage, unsteered → steered; no judged run exists for Qwen 3.5.

Model	Direction (dose)	ΔMC1 [95% CI]	Abstention (%) [random]	Informative (%)	Faithful (%)
Llama Inst.	$\mathbf{v}_{\text{faithful}}$ (0.49)	-0.169 [-0.205, -0.135]	$32.6 \rightarrow 9.0$ [25.0, 33.0]	$61.3 \rightarrow 94.9$	$26.9 \rightarrow 84.7$
Llama base	$\mathbf{v}_{\text{faithful}}$ (0.37)	$+0.018$ [-0.003, $+0.041$]	$0.0 \rightarrow 1.0$ [0.2, 0.6]	$84.5 \rightarrow 78.4$	$68.7 \rightarrow 75.5$
Qwen 3.5	$\mathbf{v}_{\text{faithful}}$ (0.26)	$+0.016$ [-0.020, $+0.052$]	$25.7 \rightarrow 21.7$ [22.7, 25.3]	–	–
Qwen 2.5	$\mathbf{v}_{\text{faithful}}$ (0.19)	-0.015 [-0.036, $+0.006$]	$23.2 \rightarrow 16.6$ [20.4, 25.0]	$76.3 \rightarrow 90.7$	$41.1 \rightarrow 63.0$
Mistral	$\mathbf{v}_{\text{context}}$ (0.63)	-0.104 [-0.138, -0.071]	$29.6 \rightarrow 3.8$ [22.4, 34.0]†	$74.1 \rightarrow 93.5$	$39.1 \rightarrow 73.0$

†Random vectors at the smaller length of $\mathbf{v}_{\text{faithful}}$ (dose 0.34), not of $\mathbf{v}_{\text{context}}$.

Table 3: Reading the context label from the last prompt token before generation. Each cell gives the test balanced accuracy of a logistic readout that predicts whether the prompt holds the original or the counterfactual passage, followed by the 95th percentile of the same readout on random directions (300 for the Llama models, 200 for the others, 100 for all-layer rows). Bold cells beat that percentile. One-layer rows use the layer chosen for the $\mathbf{v}_{\text{truth}}$ readout on validation data, and all-layer rows use every layer. The last row uses the whole hidden state and has no random reference.

Readout	Llama Instruct	Llama base	Qwen 3.5 27B	Qwen 2.5 7B	Mistral 7B
1-D $\mathbf{v}_{\text{truth}}$, one layer	**0.818** / 0.767	**0.580** / 0.565	**0.898** / 0.801	**0.724** / 0.632	**0.638** / 0.593
1-D $\mathbf{v}_{\text{context}}$, one layer	**0.776** / 0.765	0.562 / 0.567	0.637 / 0.784	0.544 / 0.647	0.524 / 0.602
1-D $\mathbf{v}_{\text{faithful}}$, one layer	0.507 / 0.760	0.480 / 0.569	0.783 / 0.805	0.558 / 0.631	**0.622** / 0.596
all-layer $\mathbf{v}_{\text{truth}}$	**0.851** / 0.847	**0.680** / 0.658	**0.938** / 0.929	**0.756** / 0.731	**0.718** / 0.711
all-layer $\mathbf{v}_{\text{truth}}$+$\mathbf{v}_{\text{faithful}}$	0.856 / 0.862	**0.731** / 0.698	0.942 / 0.942	0.764 / 0.769	**0.751** / 0.747
full hidden state, one layer	0.909	0.818	0.947	0.884	0.864

the MC1 loss from 0.166 to 0.041. On Qwen 3.5, $\mathbf{v}_{\text{context}}$ at dose 0.5 raises MC1 by 0.177 on all 616 questions, but only two random vectors were run at its layer and dose (up to $+0.075$), too few to credit this gain to the direction.

The context gains are not produced by degenerate text. The judge labels more steered answers as informative on Llama Instruct, Qwen 2.5 and Mistral (Table 2). On Llama Instruct, five random vectors of the same length leave informativeness near the unsteered level of their 750-item slice (56.0% to 69.2%, against 60.0%; Appendix G). Only on Llama base does informativeness fall.

6 Calibrated truth steering fails despite a usable readout

A single projection onto $\mathbf{v}_{\text{truth}}$ identifies the counterfactual passage better than random projections on all five models. The random reference matters because the label is easy to read: on Llama Instruct a random direction reaches 0.63 balanced accuracy on average, and the best 5% reach 0.767. The truth direction's 0.818 is a real but modest advantage, well below the 0.909 of the whole hidden state (Table 3). Combining projections from all layers raises every accuracy, the random ones included, and the advantage of $\mathbf{v}_{\text{truth}}$ is then at most 0.025.

Reading the context and steering it are different properties of a direction. At the shared readout layer, $\mathbf{v}_{\text{faithful}}$, the strongest steering direction, reads the label at chance on Llama Instruct (0.507), whereas $\mathbf{v}_{\text{truth}}$ reads it well but moves the false-context score by only $+2.3$ pp. At its own steering layer, 12, $\mathbf{v}_{\text{faithful}}$ does not beat random directions either (AUROC 0.573, against a 95th percentile of 0.658). The two properties also come apart under steering. When $\mathbf{v}_{\text{faithful}}$ makes Llama Instruct follow the false passage, the projection onto $\mathbf{v}_{\text{truth}}$ still separates the two passages almost as well as before (best-layer AUROC 0.86, against 0.88 unsteered and 0.85 to 0.91 under random vectors of the same length). Probes with similar accuracy can likewise differ in causal effect (Marks & Tegmark, 2024), and detection and steering rank methods differently (Wu et al., 2025).

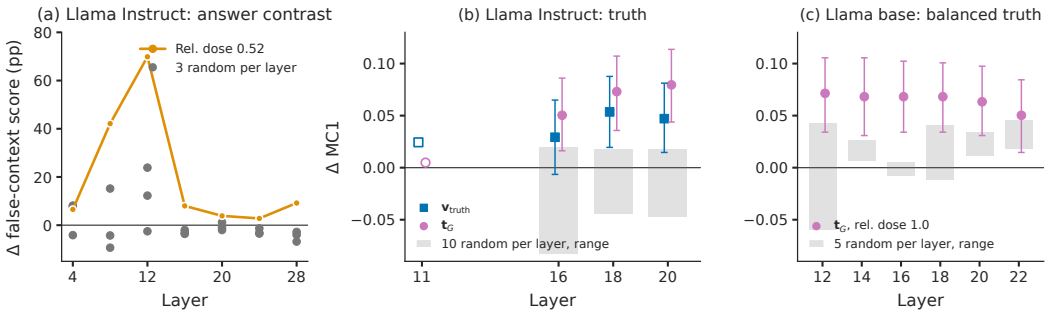


Figure 3: Steering across layers. (a) Change in the false-context score under Llama Instruct's $\mathbf{v}_{\text{faithful}}$, held at its calibrated dose at every layer (relative dose 0.52 measured on ConFiQA prompts; Appendix A) on a 750-item slice, with 3 random vectors per layer. (b) Change in TruthfulQA MC1 on Llama Instruct under $\mathbf{v}_{\text{truth}}$ and the polarity-balanced direction $\mathbf{t}_G$ at relative dose 1.0, with 95% paired bootstrap intervals and the observed range of 10 random vectors per layer (grey). Open markers at layer 11 are the calibrated $\mathbf{v}_{\text{truth}}$ (dose 0.79) and $\mathbf{t}_G$ at dose 0.5. (c) The same for $\mathbf{t}_G$ on Llama base, with 5 random vectors per layer. MC1 uses 616 questions.

At its calibrated setting, steering with $\mathbf{v}_{\text{truth}}$ does not improve truthfulness on any model. On Llama Instruct the MC1 change is $+0.024$ with an interval that includes zero, and the steered MC1 of 0.536 lies below the 95th percentile of the 50 matched random vectors (0.544). On the other four models the change lies between -0.005 and $+0.004$ (Table 4), but on Qwen 3.5 and Qwen 2.5 validation chose doses of only 0.02 and 0.09, so for them the later-layer tests of Section 7 are more informative. On Llama Instruct the calibrated direction mainly suppresses answers: under the original passage, the share of answers that give the passage's answer falls from 80.7% to 29.5% by string matching, and under a false passage the judge labels 24.1% of answers informative, against 61.3% without steering. Because MC1 only ranks the candidates it is given and can reward declining, calibrating on it selected a setting that mainly makes the model decline to answer.

7 Later-layer truth steering raises MC1 only on Llama

Validation selects layers at which the truth direction fails the negation test. A direction estimated from affirmative statements can encode whether a sentence is negated instead of whether it is true (Bürger et al., 2024). On Llama Instruct, $\mathbf{v}_{\text{truth}}$ ranks the true member of a negated pair above the false one in fewer than 5% of pairs at layers 5 to 11, and in 83% to 100% of pairs at layers 16 to 29 (Appendix C). The calibrated layer, 11, lies in the first range, and the calibrated layers of the other four models also rank fewer than half of the negated pairs correctly (Table 4).

At later layers, a polarity-balanced direction raises MC1 on both Llama models. We build $\mathbf{t}_G$ from 600 pairs, half affirmative and half negated city statements (Bürger et al., 2024), and apply it at later layers, where $\mathbf{v}_{\text{truth}}$ passes the negation test on the Llama models. On Llama Instruct at layer 20, $\mathbf{t}_G$ raises MC1 from 0.511 to 0.591, while none of 30 random vectors at the same dose in layers 16 to 20 gains more than 0.019 (Table 4, Figure 3b). On Llama base it raises MC1 at layer 18 from 0.383 to 0.451, about the level of random ranking (Figure 3c). These settings were chosen on the test questions, but at dose 1.0 every tested layer on both models gains at least 0.050, more than any random vector at that dose. The original $\mathbf{v}_{\text{truth}}$ also raises MC1 at layers 18 and 20, by 0.054 and 0.047 on Llama Instruct but by at most 0.021 on Llama base. Because $\mathbf{t}_G$, unlike $\mathbf{v}_{\text{truth}}$, is built only from city statements, its larger gains cannot be credited to polarity balancing alone.

Part of the Llama Instruct gain is bought with fewer correct answers from correct passages. On a separate 750-item slice with answers of at most 64 new tokens, $\mathbf{t}_G$ at layer 20 and dose 1.0 lowers the share of correct answers under the original passage from 78.7% to 60.5%, while 10 random vectors at the same layer and dose keep it at 77.1% to 79.7%. Under a false passage, the judge labels 34.7–36.5% of $\mathbf{t}_G$ answers at dose 1.0 informative, against 49.9% unsteered. The original $\mathbf{v}_{\text{truth}}$ at layer 18 and dose 0.75 raises MC1 by 0.047 but lowers the share of correct answers to 67.6%.

Table 4: Truth steering and TruthfulQA MC1. Left: $\mathbf{v}_{\text{truth}}$ at its calibrated layer; "Neg." is the fraction of 748 negated city pairs in which the true statement projects higher onto $\mathbf{v}_{\text{truth}}$ at that layer (values near 0 reverse the ordering). Right: the polarity-balanced direction $\mathbf{t}_G$ at later layers, at the setting with the largest MC1 change among doses for which random vectors were run (for Llama base, see note a). ΔMC1 is the change from the unsteered model with its 95% paired bootstrap interval; "Rand." is the largest change among n random vectors of the same size, at the same layer (left) or the same dose over the tested layers (right). MC1 uses 616 questions, except for Qwen 3.5 (250 left, 400 right).

	Calibrated $\mathbf{v}_{\text{truth}}$				$\mathbf{t}_G$ at later layers		
Model	Layer	Neg.	ΔMC1 [95% CI]	Rand. (n)	Layer, dose	ΔMC1 [95% CI]	Rand. (n)
Llama Inst.	11	0.004	$+0.024$ [-0.013, $+0.058$]	$+0.057$ (50)	20, 1.0	$+0.080$ [$+0.044$, $+0.114$]	$+0.019$ (30)
Llama base	5	0.039	-0.002 [-0.024, $+0.024$]	$+0.018$ (25)	18, 1.0[a]	$+0.068$ [$+0.034$, $+0.101$]	$+0.045$ (30)
Qwen 3.5	24	0.364	$+0.004$ [-0.008, $+0.020$]	$+0.004$ (5)	40, 0.75	$+0.018$ [-0.030, $+0.062$]	$+0.107$ (6)
Qwen 2.5	6	0.170	-0.005 [-0.016, $+0.005$]	$+0.005$ (5)	10, 0.5	-0.010 [-0.039, $+0.019$]	$+0.019$ (30)
Mistral	18	0.418	$+0.003$ [-0.019, $+0.026$]	$+0.010$ (5)	18, 0.5	$+0.010$ [-0.018, $+0.039$]	$+0.026$ (30)

[a] Layer 12 gives $+0.071$, but 85% of its false-context answers then match neither answer (18% unsteered, 24% at layer 18).

On the Qwen and Mistral models, $\mathbf{t}_G$ does not beat random vectors. On Mistral and Qwen 2.5, $\mathbf{t}_G$ was tested at six layers and four doses from 0.25 to 1.0, including two layers per model at which $\mathbf{v}_{\text{truth}}$ ranks at least 75% of negated pairs correctly. Mistral's best MC1 change is $+0.010$, with an interval that includes zero, and every setting lowers MC1 on Qwen 2.5. On Qwen 3.5, three of six random vectors raise MC1 by 0.09 to 0.11, more than any $\mathbf{t}_G$ setting (Table 4). We do not know why the Llama models respond differently, and the tested grids leave room for untried settings.

8 Discussion

What the results mean for evaluating steering. The unsteered model is too weak a reference for steering effects. On Llama Instruct, random vectors at the calibrated lengths of $\mathbf{v}_{\text{faithful}}$ and $\mathbf{v}_{\text{truth}}$ move the false-context score by up to 42 and 56 pp, so learned directions should be compared with random vectors of equal length at the same layer. Reports should also cover what the selection metric does not measure. Calibration on MC1 chose a $\mathbf{v}_{\text{truth}}$ setting that mainly suppresses Llama Instruct's answers, and calibration for context use chose settings that lower abstention on three models. Halving the dose of Llama Instruct's $\mathbf{v}_{\text{faithful}}$ kept most of its effect at a much smaller MC1 cost, a trade-off that a single calibrated setting hides. Finally, the better context construction reversed between Llama and Mistral, and later-layer truth steering helped only the Llama models, so constructions and layers should be re-derived for each model.

Limitations. Generation uses one seed and greedy decoding. The random vectors are isotropic, and we did not test covariance-matched random directions. The judge agrees with our annotators on 67% to 92% of informativeness labels, less often than on faithfulness, and the audit covers three Llama Instruct conditions and no $\mathbf{t}_G$ answers (Appendix G). Where only $n \leq 10$ random vectors were run, exceeding all of them gives a rank-test p-value of $1/(n+1)$, so those comparisons rest on the size of the margin. Qwen 3.5 has smaller test sets, extraction-source ablations cover only Llama Instruct and Qwen 3.5, and the Mistral and Qwen 2.5 calibration grids are reduced. The ConFiQA defect filters were written after we saw results, so the main text reports all items. We study mean-difference directions only.

9 Conclusion

Learned context directions beat matched random vectors on all five models, but the better construction depends on the model, and on three of them the gains come with answers to unanswerable questions. The standard truth direction reads counterfactual contexts but does not steer truthfulness at its validation-selected setting. Steering gains should be reported with random-vector controls and with the behaviour the selection metric cannot see.

Reproducibility statement

All experiments use publicly available checkpoints (Appendix I), the public ConFiQA, TruthfulQA, NQ-Swap, FaithEval, Marks–Tegmark, VitaminC and WikiText-2 data, greedy decoding with seed 42 (random-direction families also use seeds 7 and 23), and 2000-resample paired bootstraps. ConFiQA answers at the calibrated settings and under their random-vector controls, and at the ContextFocus setting, are limited to 100 new tokens (64 for Qwen 3.5). All other generation runs, including those behind the readout diagnostics of Appendix E, are limited to 64. Appendices A to H give, in order, the calibrated layers, scales and relative doses for every model and direction, the random-vector statistics, the negation control, the ContextFocus replication, the judge-labelled readout diagnostics, the prompts, answer matching and ConFiQA data-quality check, the judge audit and the geometry of the directions. The supplementary material contains the code for extraction, calibration, steering, every control, the judge (with its prompts) and the figures; the exact data splits; the summary table written by every run, with a map from each result in the paper to its script and table; the per-item ConFiQA defect flags (Appendix F); and the human audit sample, both label sets and the scoring script (Appendix G). All new runs were made on a single 96 GB GPU in about 130 GPU-hours; the original calibration runs used a single 80 GB A100.

Ethics statement

This work studies interventions on open-weight models using public benchmarks. Context steering can increase reliance on false supplied information and reduce appropriate abstention on unanswerable questions. These effects should be measured when evaluating such interventions. Two authors performed the human annotation described in Appendix G; no external participants or private personal data were involved.

AI use statement

Qwen 3.5 27B served as an automatic judge of answer faithfulness and informativeness (Section 3). An AI assistant helped write and debug scripts, manage experiment queues, build the human-audit form, aggregate results, and draft and edit text. The authors designed the experiments, checked reported values against retained outputs, and take responsibility for the manuscript. They produced the human audit labels without model assistance. No model was used to fabricate or alter experimental results, human labels or references.

References

Shivam Adarsh, Maria Maistro, and Christina Lioma. How context shapes truth: Geometric transformations of statement-level truth representations in LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36570–36590. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1695. URL https://aclanthology.org/2026.acl-long.1695/.

Nikhil Anand, Shwetha Somasundaram, Anirudh Phukan, Apoorv Saxena, and Koyel Mukherjee. ContextFocus: Activation steering for contextual faithfulness in large language models. *arXiv preprint arXiv:2601.04131*, 2026. URL https://arxiv.org/abs/2601.04131.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. Context-DPO: Aligning language models for context-faithfulness. In *Findings of the Association*

for Computational Linguistics: ACL 2025, pp. 10280–10300. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-acl.536. URL https://aclanthology.org/2025.findings-acl.536/.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Junfeng Fang, Hongcheng Gao, Shiyu Ni, and Xueqi Cheng. Is factuality enhancement a free lunch for LLMs? Better factuality can lead to worse context-faithfulness. In *The Thirteenth International Conference on Learning Representations (ICLR)*, pp. 41266–41297, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/660d0ed5885662219244b6e44aba8fe3-Abstract-Conference.html.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models. *arXiv preprint arXiv:2505.22637*, 2025. URL https://arxiv.org/abs/2505.22637. Presented at the ICLR 2025 Workshop on Foundation Models in the Wild.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. Truth is universal: Robust detection of lies in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 138393–138431. Curran Associates, Inc., 2024. doi: 10.52202/079017-4392. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f9f54762cbb4fe4dbffdd4f792c31221-Abstract-Conference.html.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. Steering off course: Reliability challenges in steering language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19856–19882. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.974. URL https://aclanthology.org/2025.acl-long.974/.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do I know this entity? Knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, pp. 77965–78000, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c1c44e46358e0fb94dc94ec495a7fb1a-Abstract-Conference.html.

Linfeng Gao, Qinggang Zhang, Baolong Bi, Bo Zeng, Zheng Yuan, Zerui Chen, Zhimin Wei, Shenghua Liu, Linlong Xu, Longyue Wang, Weihua Luo, and Jinsong Su. Beyond black-box interventions: Latent probing for faithful retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29981–30000. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1499. URL https://aclanthology.org/2026.findings-acl.1499/.

Navita Goyal and Hal Daumé III. Steering safely or off a cliff? Rethinking specificity and robustness in inference-time interventions. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5723–5738. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.268. URL https://aclanthology.org/2026.eacl-long.268/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1193–1215. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.70. URL https://aclanthology.org/2024.findings-acl.70/.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations (ICLR)*, pp. 20310–20334, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/33dffa2e3d2ab74a783d1a8c292f66d9-Abstract-Conference.html.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Y. Rogov, Ivan Oseledets, and Elena Tutubalina. The rogue scalpel: Activation steering compromises LLM safety. *arXiv preprint arXiv:2509.22067*, 2025. URL https://arxiv.org/abs/2509.22067.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Sanity checks for sparse autoencoders: Do SAEs beat random baselines? *arXiv preprint arXiv:2602.14111*, 2026. URL https://arxiv.org/abs/2602.14111.

Maor Juliet Lavi, Tova Milo, and Mor Geva. Detecting (un)answerability in large language models with linear directions. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 682–699. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.29. URL https://aclanthology.org/2026.eacl-long.29/.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming knowledge conflicts in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34074–34104. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25c.html.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Srikanth Malla, Chiho Choi, and Joon Hee Choi. Steering interference reflects the model's defaults, not the behavior directions. *arXiv preprint arXiv:2609.06951*, 2026. URL https://arxiv.org/abs/2609.06951.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

Julian Minder, Kevin Du, Niklas Stoehr, Giovanni Monea, Chris Wendler, Robert West, and Ryan Cotterell. Controllable context sensitivity and the knob behind it. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=Igm9bbkzHC.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. FaithEval: Can your language model stay faithful to context, even if "The moon is made of marshmallows". In *The Thirteenth International Conference on Learning Representations (ICLR)*, pp. 29430–29456, 2025. URL https://openreview.net/forum?id=UeVx6L59fg.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. Towards reliable evaluation of behavior steering interventions in LLMs. *arXiv preprint arXiv:2410.17245*, 2024. URL https://arxiv.org/abs/2410.17245. NeurIPS 2024 Workshop on Foundation Model Interventions (MINT).

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3.5: Towards native multimodal agents. Blog post, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Tal Schuster, Adam Fisch, and Regina Barzilay. Get your vitamin C! Robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 624–643. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.52. URL https://aclanthology.org/2021.naacl-main.52/.

Vincent Siu, Nicholas Crispino, David Park, Nathan W. Henry, Zhun Wang, Yang Liu, Dawn Song, and Chenguang Wang. SteeringSafety: Benchmarking representation steering in LLMs across safety perspectives. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2509.13450. arXiv:2509.13450.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, volume 37, pp. 139179–139212, 2024. doi: 10.52202/079017-4417. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fb3ad59a84799bfb8d700e56d19c231b-Abstract-Conference.html.

Kevin Wu, Eric Wu, and James Zou. ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 37, pp. 33402–33422, 2024. doi: 10.52202/079017-1053. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3aa291abc426d7a29fb08418c1244177-Abstract-Datasets_and_Benchmarks_Track.html.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080. PMLR, 2025. URL https://proceedings.mlr.press/v267/wu25a.html.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

Yu Zhao, Xiaotang Du, Giwon Hong, Aryo Pradipta Gema, Alessio Devoto, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. Analysing the residual stream of language models under knowledge conflicts. *arXiv preprint arXiv:2410.16090*, 2024. URL https://arxiv.org/abs/2410.16090. NeurIPS 2024 Workshop on Foundation Model Interventions (MINT).

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5117–5136. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.264. URL https://aclanthology.org/2025.naacl-long.264/.

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. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

Appendix

The appendices below are supplementary text. Code, data splits, per-run result tables and the human audit files are in the separate supplementary material.

A Calibrated settings and relative doses

Table 5: Validation-selected layer and scale for each direction and model, with the relative dose $\|\alpha\mathbf{v}\|/\overline{\|\mathbf{h}^{(\ell)}\|}$ at that setting. The scale α multiplies the unit-length $\mathbf{v}_{\text{truth}}$, and the multiplier m multiplies a context direction at its extracted length. All context directions use all-position steering. The Mistral and Qwen 2.5 grids were reduced (truth: layers 6 to 24 step 2, $\alpha \in \{1, 2, 3, 5\}$; context: layers 8 to 20 step 2, multipliers 0.5 to 3).

Model	$\mathbf{v}_{\text{truth}}$ (layer, α; dose)	$\mathbf{v}_{\text{context}}$ (layer, m; dose)	$\mathbf{v}_{\text{faithful}}$ (layer, m; dose)
Llama 3.1 8B Instruct	11, 5.0; 0.79	16, 1.5; 0.70	12, 1.5; 0.49
Llama 3.1 8B base	5, 2.0; 0.42	12, 2.0; 0.70	16, 1.5; 0.37
Qwen 3.5 27B	24, 1.0; 0.02	40, 0.5; 0.27	32, 1.0; 0.26
Qwen 2.5 7B Instruct	6, 3.0; 0.09	20, 0.5; 0.27	16, 1.0; 0.19
Mistral 7B Instruct	18, 3.0; 0.37	14, 1.0; 0.63	14, 3.0; 0.34

The doses in this table divide by the mean hidden-state length over the extraction inputs. The equal-dose comparison (Figure 2), the layer sweeps (Figure 3) and the $\mathbf{t}_G$ runs divide by the mean length over the unsteered ConFiQA test prompts instead, which gives different values for the same vector: Llama Instruct's calibrated $\mathbf{v}_{\text{faithful}}$ is 0.49 here and 0.52 against ConFiQA prompts, Llama base's 0.37 and 0.43, Qwen 3.5's 0.26 and 0.31, Qwen 2.5's 0.19 and 0.20, and Mistral's 0.34 and 0.31. The Qwen 3.5 truth intervention is a 2% perturbation of the residual norm and the Llama Instruct one

a 79% perturbation. Cross-model statements about "the effect of the truth direction" at calibrated settings compare different dose regimes, which is why Sections 5 and 7 work at fixed relative dose.

B RANDOM-DIRECTION BANDS

Table 6: Norm-matched random-direction families, all seeds pooled. Every random value is a change in false-context score from its own run's 750-item baseline; sd is the sample standard deviation. Learned changes use the main evaluation set (1500 items; 750 for Qwen 3.5). Control slices differ across seeds, so these are descriptive comparisons.

Model	Family (layer, norm of)	n	Mean Δ	sd	q95	Max	Learned Δ
Llama Instruct	$\mathbf{v}_{\text{faithful}}$	50	$+12.5$	12.4	$+34.1$	$+42.3$	$+71.9$
Llama Instruct	$\mathbf{v}_{\text{truth}}$	50	$+19.9$	18.0	$+52.1$	$+55.9$	$+2.3$
Llama base	$\mathbf{v}_{\text{faithful}}$	30	-1.4	2.3	$+2.1$	$+3.3$	$+14.1$
Llama base	$\mathbf{v}_{\text{truth}}$	25	-1.8	4.8	$+3.6$	$+4.5$	-6.5
Qwen 3.5 27B	$\mathbf{v}_{\text{faithful}}$	10	0.0	2.9	$+3.8$	$+5.3$	$+16.4$
Qwen 2.5 7B	$\mathbf{v}_{\text{faithful}}$	10	$+2.2$	5.0	$+8.0$	$+9.7$	$+26.7$
Mistral 7B	$\mathbf{v}_{\text{faithful}}$	10	$+1.9$	3.5	$+6.6$	$+7.2$	$+7.7$

Under the two Llama Instruct random families, at the lengths of $\mathbf{v}_{\text{faithful}}$ and $\mathbf{v}_{\text{truth}}$, MC1 has means of 0.489 and 0.490, 95th percentiles of 0.533 and 0.544, and maxima of 0.562 and 0.568 (baseline 0.511; $\mathbf{v}_{\text{truth}}$ 0.536). The smallest changes in the false-context score on Llama Instruct are -8.7 pp at the $\mathbf{v}_{\text{faithful}}$ length and -25.5 pp at the $\mathbf{v}_{\text{truth}}$ length. Mistral's calibrated random family matches $\mathbf{v}_{\text{faithful}}$, not the larger presence norm. Its presence direction has a separate matched comparison at relative dose 0.5: $+41.2$ against a best random $+12.4$ (Figure 2).

C NEGATION CONTROL PER LAYER

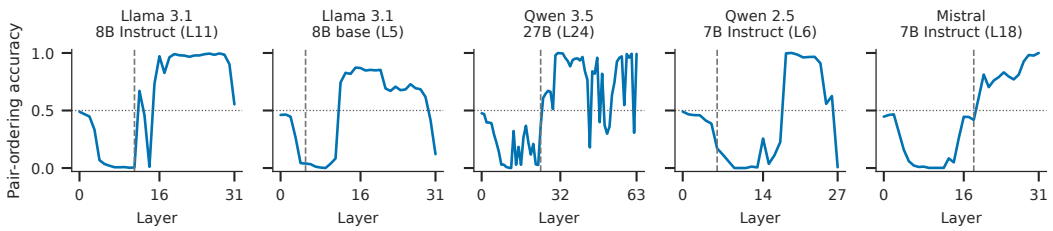


Figure 4: Negation control on 748 negated Marks–Tegmark pairs per model. Each value is the fraction of pairs where the true statement projects higher onto $\mathbf{v}_{\text{truth}}$ than the false statement. Values near 0 indicate reversed ordering; near 1, truth-consistent ordering. Dashed lines mark the validation-selected truth layers; dotted lines mark 0.5.

Figure 4 measures whether truth ordering transfers from affirmative to negated statements. The transition to higher accuracy occurs between layers 12 and 20 on the 32-layer models, around layer 18 on Qwen 2.5, and around layer 30 on Qwen 3.5. The selected truth layers precede, or lie near, these transitions. At some layers, a direction extracted from true and false statements reverses its ordering under negation, so its name alone does not show what it represents at a given layer. The check measures ordering on these paired statements only and does not establish a general truth representation or a steering mechanism.

D THE CONTEXTFOCUS REPLICATION GAP

We reconstruct ContextFocus from 1501 NQ-Swap examples, as in the original paper, and apply its published setting: layer 13, multiplier 2, steering generated tokens only. On all 1500 items

it raises the false-context score by $+18.4$ pp [$+15.5$, $+21.5$], mostly by lowering P_o; P_c reaches 26.3%. On the single-hop (QA) subset, P_c rises from 30.6% to 31.6%, against a published rise from 35.3% to 70.9%. On the multi-hop and multi-conflict subsets it rises from 23.6% to 32.4% and from 8.4% to 15.0%, against published rises from 30.2% to 54.5% and from 12.5% to 53.3%. A raw-prompt variant without the chat template reaches $P_c = 30.9\%$ on all 1500 items. Sweeping our presence direction $\mathbf{v}_{\text{context}}$ at its calibrated layer (16, all positions) over multipliers 0.5 to 3 changes the false-context score by at most $+4.5$ pp, and at multiplier 3, 90% of responses match neither answer. At the published setting, FaithEval abstention falls from 32.6% to 27.6%; no random vectors were run at this setting. These checks do not resolve the gap. The evaluation samples also differ: ContextFocus evaluates on 1500 randomly sampled items per ConFiQA subset (4500 in total) (Anand et al., 2026), whereas we use one fixed random sample of 500 items per subset (1500 in total). Sampling differences of this size can account for a few points, such as part of the gap between the unsteered single-hop rates (30.6% against 35.3%), but not for the 39-point gap between the steered rates. Evaluation prompts, decoding length and extraction splits remain possible causes, and the tested sweep does not exhaust all settings. We therefore report a replication gap without claiming to outperform the published method.

E Judge-labelled diagnostics on Llama Instruct

On Llama Instruct, we also generate answers without a passage and under the original and false passages, label them with the judge, and fit logistic readouts that predict five per-item diagnostics from all-layer projections onto $\mathbf{v}_{\text{truth}}$, alone or together with $\mathbf{v}_{\text{context}}$ or $\mathbf{v}_{\text{faithful}}$ (for correction of a wrong prior answer, differences of these projections between prompt conditions), with the feature set chosen on validation data. Each readout is compared with 300 readouts on random directions. The readout separates original from false context with balanced accuracy 0.873 ($p = 0.030$ against the random readouts). The other four diagnostics show no clear advantage over random readouts: correctness of the prior answer, given without a passage (0.744, $p = 0.20$), original versus false context given a correct prior answer (0.903, $p = 0.65$), following the false context (0.796, $p = 0.33$), and correction of a wrong prior answer by the original context (0.784, $p = 0.35$). These readouts use a different data split and selection rule from those in Table 3.

F Prompts, answer matching and ConFiQA data quality

Prompts. ConFiQA and FaithEval questions are posed with the system message "Answer the question based only on the provided context." and the user turn "Context: *passage* Question: *question*", rendered with each model's chat template. Llama base, which has no chat template, receives only the user-turn text followed by "Answer:", without the instruction. Unlike FaithEval's own prompt (Ming et al., 2025), ours does not tell the model to answer "unknown" when the passage lacks the answer. A FaithEval answer counts as an abstention if its lowercased text contains one of 24 phrases: unknown, cannot be determined, can't be determined, not enough information, insufficient information, does not provide, doesn't provide, no information, not mentioned, cannot answer, can't answer, unanswerable, not specified, not stated, i don't know, i do not know, does not say, doesn't say, not provided, no answer, does not mention, doesn't mention, not possible to determine, not clear from. We did not check this list against human labels, and it misses phrasings such as "does not contain".

Answer matching. We lowercase responses and answer aliases, remove articles and punctuation, and collapse whitespace. A match occurs when a normalised answer string is a substring of the normalised response. For the counterfactual answer, we reject the match if the response contains any token in the fixed list: no, not, never, none, cannot, nobody, nothing, nowhere, neither, nor, without, hardly. This heuristic can reject a response because of an unrelated negation. Our score counts context-only and original-only matches; responses matching both or neither contribute zero to $P_c - P_o$. The judge provides a separate check on generated answer content.

Item-quality flags. We flag four potential defects in the 1500 test items. First, 139 questions (9.3%) contain their own original answer or an alias. The multi-hop template can also ask for the wrong entity type. For example, a chain ending *Joe Biden* $\rightarrow$ *position held* $\rightarrow$ *President of the*

United States produces "Who is the President of the United States of the country of origin of the film that follows Predators?" The question includes the original target and asks "who" although the chain returns a position.

Second, 70 items (4.7%) have overlapping original and counterfactual answers or aliases, making the answer-choice distinction ambiguous. Third, 14 items (0.9%) omit the counterfactual answer from the counterfactual context. Fourth, 4 contexts contain repeated-word artefacts. The union is 210 items (14.0%), of which 98 are multi-hop, 80 multi-conflict and 32 single-hop items. A separate check finds 60 questions (4.0%) containing the counterfactual answer; these are not part of that union.

The flagged items have a lower false-context score on Llama Instruct: -59.0 at baseline, compared with -6.1 on the full set. Under $\mathbf{v}_{\text{faithful}}$, they reach -11.0. To test whether these items drive the main result, we rescore the retained outcomes on the 1290 unflagged items. Table 7 reports the checked conditions across four models. This uses existing outputs and requires no new generation.

Table 7: Change in the false-context score against the same model's baseline, on all 1500 items and on the 1290 clean items (95% paired bootstrap CI over the clean items). No conclusion changes.

Model	Direction	All 1500	Clean 1290 [95% CI]
Llama 3.1 8B Instruct	$\mathbf{v}_{\text{faithful}}$	$+71.9$	$+75.7$ [+71.8, +79.7]
Llama 3.1 8B Instruct	$\mathbf{v}_{\text{context}}$	$+4.5$	$+4.6$ [+1.4, +7.6]
Llama 3.1 8B Instruct	$\mathbf{v}_{\text{truth}}$	$+2.3$	-0.9 [−4.7, +3.3]
Llama 3.1 8B Instruct	ContextFocus (published setting)	$+18.4$	$+19.7$ [+16.6, +22.9]
Llama 3.1 8B base	$\mathbf{v}_{\text{faithful}}$	$+14.1$	$+13.2$ [+10.0, +16.2]
Qwen 2.5 7B Instruct	$\mathbf{v}_{\text{faithful}}$	$+26.7$	$+26.6$ [+23.6, +29.7]
Mistral 7B Instruct	$\mathbf{v}_{\text{context}}$	$+47.7$	$+49.7$ [+46.0, +53.6]

The two largest steering effects increase by 2 to 4 points after filtering. Llama Instruct's calibrated truth effect changes from $+2.3$ to -0.9 pp, with a CI spanning zero. A narrower filter removes only 16 items with missing answers or repeated-word artefacts. Under that filter, the main context effects are $+72.5$, $+14.1$, $+27.0$ and $+48.1$ pp, within 0.6 points of the corresponding unfiltered values. The principal context effects therefore retain their size and sign under both filters.

We use all-item results in the main text to preserve the published benchmark and because the filters were designed after inspecting results. They use benchmark fields, never model outputs, and apply identically to all conditions. The question-overlap detector deliberately flags a broad set: a string match does not prove that every flagged item is defective. The filtered results test sensitivity to removing these potentially problematic items. We release the script and per-item flags.

G JUDGE

Free-form ConFiQA answers are scored by Qwen 3.5 27B with thinking disabled and a fixed prompt. It labels whether an answer follows the context (faithful) and whether it answers the question with relevant content (informative). Every judge output parses successfully, which establishes format compliance rather than label accuracy.

On Llama Instruct, calibrated $\mathbf{v}_{\text{truth}}$ lowers informativeness to 24.1% under false context and 22.1% under original context, from 61.3% and 91.8% unsteered.

The random-control slice has its own baseline: 25.6% context following and 60.0% informativeness. Five random interventions at the $\mathbf{v}_{\text{faithful}}$ norm give 30.1–43.2% context following and 56.0–69.2% informativeness. Two at the $\mathbf{v}_{\text{truth}}$ norm give 42.1% and 50.1% context following. These are different intervention sizes and should not be pooled into a single matched comparison. The labels show that random interventions can produce informative answers; they do not establish general answer quality.

Human audit. Two authors independently labelled 180 false-context answers from the Llama Instruct run judged in Table 2. We sampled 15 answers from each combination of three conditions (unsteered, $\mathbf{v}_{\text{faithful}}$, calibrated $\mathbf{v}_{\text{truth}}$) and four judge-label cells. The interface shuffled the answers

and hid the condition and judge labels. Annotators used the judge's definitions of faithfulness and informativeness. We reweight the sample to recover population-level agreement and corrected rates.

For faithfulness, judge agreement with the two annotators is 95% and 81% at baseline, 99% and 93% under $\mathbf{v}_{\text{faithful}}$, and 99% and 90% under $\mathbf{v}_{\text{truth}}$. Informativeness agreement is lower: 84% and 67%, 92% and 92%, and 92% and 74%, respectively. The annotators agree with each other on 86% of faithfulness labels ($\kappa = 0.72$) and 76% of informativeness labels ($\kappa = 0.45$). Thus informativeness is also the less consistent human label.

Correcting the rates separately with each annotator's labels changes $\mathbf{v}_{\text{faithful}}$ informativeness from 94.9% to 95% and 96%. It changes baseline informativeness from 61% to 68% and 56%, and calibrated-truth informativeness from 24% to 29% and 28%. Both corrections preserve the ordering of these conditions. At baseline, the direction of the correction differs between annotators.

The audit has two limitations. Of the 180 answers, 35 concern flagged ConFiQA items; agreement on these is within one point of the remaining items. The first interface version truncated 67 contexts at 1400 characters, and one annotator reported this on 6 rows. The 98 unflagged, untruncated rows show the same qualitative pattern. We release the sample, both label sets, flags and scoring script.

H GEOMETRY OF THE DIRECTIONS

Each direction is estimated reproducibly: on Llama Instruct, directions extracted from two halves of the same data have cosines of 0.89 for truth and of 0.997 and 0.95 for the two context directions, averaged over layers. Cosines between different directions average -0.03 to $+0.06$ over layers, and at single layers they range from -0.13 to $+0.18$ (Figure 5a). Because random directions in a high-dimensional space also have small cosines, each cosine is compared with a reference band. Raw cosines are dominated by the few high-variance directions of the activation space, so we also whiten the activations, transforming them so that their covariance becomes the identity matrix and every direction of variation has equal weight (Figure 5b). After whitening, the truth–context cosines stay within ± 0.05 on all five models, while the two context directions can align more strongly, up to 0.128 on Llama base. The truth direction is also less stable across its two data sources than across halves of the pooled data: estimated separately from the Marks–Tegmark and VitaminC statements, its layer-averaged cosines are only 0.03 to 0.27 across the five models, consistent with reports that truth probes transfer poorly across datasets (Orgad et al., 2025; Bürger et al., 2024). Small alignment does not show that the two behaviours use separate mechanisms, but together with the projection-removal test of Section 4 it argues against the context effects being a truth effect in disguise.

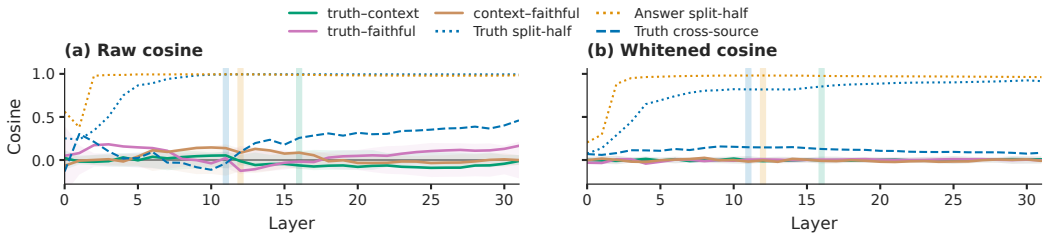


Figure 5: Cosines between directions across layers of Llama Instruct, (a) raw and (b) after whitening. Shaded regions are pointwise 95% reference bands, from directions extracted with shuffled labels in (a), one per pair, and from random directions in (b), where a single band applies to all pairs. Dotted lines are split-half cosines of $\mathbf{v}_{\text{truth}}$ and $\mathbf{v}_{\text{faithful}}$, the agreement between directions extracted from two halves of the same data; dashed lines are the cosine between truth directions extracted from Marks–Tegmark and from VitaminC alone. Vertical bands mark the calibrated layers. Some layers exceed their reference bands, so alignment is small rather than zero.

A subspace analysis extends this mean-direction comparison. At each tested layer, it takes the leading four singular vectors of the matrix of paired activation differences. Principal cosines measure alignment between two such subspaces: zero indicates orthogonality and one indicates a shared axis. Mean principal cosines between truth and presence subspaces are 0.001–0.005 after whitening, compared with about 0.025 for random subspaces.

This analysis whitens the paired differences with a regularised covariance pooled from centred difference groups. The mean-direction share is the sum of squared projections onto the mean direction divided by the total squared length of the differences. It measures how much of the contrast is captured by a single axis. On the two Llama models and Qwen 3.5, whitened shares are 0.5–1.7% for truth, 21–95% for presence and 5–26% for faithful answers. These descriptive quantities do not establish a causal mechanism for steering.

I MODEL AND JUDGE CHECKPOINTS

The experiments use the following public checkpoints without additional fine-tuning or quantisation, in bfloat16 or float16. All inputs in this study are text. The judge uses the same Qwen 3.5 27B checkpoint listed below; it was not trained on our human audit labels.

- Llama 3.1 8B Instruct: `meta-llama/Llama-3.1-8B-Instruct`.
- Llama 3.1 8B base: `meta-llama/Llama-3.1-8B`.
- Qwen 3.5 27B and the judge: `Qwen/Qwen3.5-27B`.
- Qwen 2.5 7B Instruct: `Qwen/Qwen2.5-7B-Instruct`.
- Mistral 7B Instruct: `mistralai/Mistral-7B-Instruct-v0.2`.