Causal Sufficiency Without Semantic Alignment: How Causal Subspaces Can Masquerade as Semantic Concepts

Mette Friis Andersen [1] **Giovanni Cinà** [*1] **Sandro Pezzelle** [*1]

Abstract

Mechanistic interpretability treats a causally active subspace as evidence that a model encodes the corresponding concept. We apply Distributed Alignment Search to arithmetic verification across three instruction-tuned LLMs and find a causally sufficient "arithmetic truth" subspace, allowing us to steer incorrect predictions towards correct ones. However, we find that a probe trained to decode the ground truth performs at chance, while a probe trained to decode the model's prediction achieves perfect accuracy. This indicates that the subspace encodes prediction bias, not truth. CoT prompting repairs the model precisely by routing around it. We conclude that a subspace can be causally sufficient but semantically misaligned with the ground truth concept it is meant to capture. [1]

1. Introduction

An open question in LLM interpretability is whether models "reason from" human-interpretable semantic concepts or develop opaque concepts that resist description in our vocabulary (Hewitt et al., 2025; Ayonrinde & Jaburi, 2025). Distributed Alignment Search (DAS; Geiger et al., 2023b) serves as a method to identify candidate concepts: a causally sufficient subspace whose interventions reliably flip model behavior is taken as evidence of concept encoding, in line with literature in the philosophy of representation (Geiger et al., 2025). However, causal sufficiency and semantic alignment are trivially conflated when models are accurate: If the model is correctly solving the task, any causally sufficient subspace will *appear* semantically aligned with the human-interpretable concept necessary for solving the task.

An open question is: **what does DAS find when behavioral accuracy is low?**

We study biased models with low behavioral accuracy on an arithmetic truth task. Since we are studying underperforming models, there are two competing explanations of what a discovered causally sufficient subspace *is*: It could either be a genuine conceptually faithful subspace, which the model fails to leverage in prediction, or it is something else. We show that what DAS finds is not semantically aligned with the human-interpretable concept we expect to find. This analysis is enabled precisely because we work with models that are failing on the task.

For our purposes, we identify one necessary condition for semantic alignment, whose violation is sufficient to establish *mis*alignment. We define a causally sufficient subspace as semantically misaligned with a concept when it does not *encode* the concept itself. We operationalise this through one test: a linear probe on the subspace must recover ground truth above chance. Chance-level accuracy establishes that no ground-truth information is linearly encoded in the subspace.

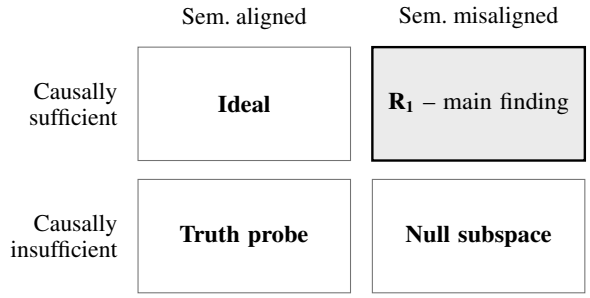

Figure 1. Causal sufficiency and semantic alignment are independent properties. DAS is designed to find the former, and we show it does not imply the latter. R_1 denotes our main finding.

We apply DAS to Qwen 2.5-7B, Gemma 2-9B, and Mistral-7B and show that the identified subspaces fail this necessary condition, and are thus misaligned (Figure 1). **The subspace encodes prediction bias, not truth**: a probe predicts model output at 99.6% but ground truth at near-chance (48–60% across models). **The repair bypasses the subspace**: chain-of-thought prompting repairs the main failure mode

[*]Equal contribution [1]ILLC, University of Amsterdam, Amsterdam, Netherlands. Correspondence to: Mette Friis Andersen <m.f.andersen@uva.nl>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

[1]Code is available at: https://github.com/mettefriis/causal-sufficiency-das

of each model through an alternative causal pathway that is independent of the subspace.

2. Background

2.1. Localising truth representations

Previous work has tried to localize human-interpretable concepts in model activations. Marks & Tegmark (2024) identify a truth direction as the difference-in-means between activations on true and false statements, and validate it using probing (Belinkov, 2022) and causal steering. Burns et al. (2023) identify a truth direction using Contrast-Consistent Search. Zou et al. (2023) extend this to a broader set of concepts ("honesty," "morality"). They identify steering vectors, and use steerability as evidence that the concept is represented internally. However, a direction that linearly separates true from false activations can encode a correlated concept (e.g., "answer magnitude" in an addition task), rather than the concept we think (Hewitt et al., 2025).

2.2. Are the representations causal?

To test the degree to which the localised concept drives prediction, causal interventions (in the form of activation patching) have been used (Geiger et al., 2023b). The idea is: when processing a base input, the activations at the concept location are replaced by those from a source input, whose ground truth label is different from the base. The degree to which activation swaps flip prediction measures the degree to which the prediction is causally mediated by this concept. This is captured by the Interchange Intervention Accuracy (IIA) score.

Distributed Alignment Search (Geiger et al., 2023b) formalises this search for a concept. Based on the assumption that neurons are polysemantic, DAS learns a rotation matrix $R \in \mathbb{R}^{d\times k}$ that maximises interchange intervention accuracy (IIA), which is the fraction of cases where swapping the activations (after projection by the DAS matrix) produces the correct output change. A high IIA ≈ 1.0 quantifies perfect causal alignment between a model and a target concept (Geiger et al., 2023a).

3. Methods and Experiment

3.1. Task

We extend the truth dataset from Marks & Tegmark (2024), studying *arithmetic verification.* The task is to predict whether the arithmetic fact is true or false. The arithmetic facts are symbolic sum equalities on the form "23+73=96". Operands are $a, b \in [2, 99]$. We embed the arithmetic statement into the following prompt template: [statement]\nIs this statement true? Answer with exactly one word: 'true' or 'false' [2]. Prediction is the argmax over `true` / `false` next-token logits.

3.2. Models

We conduct our experiments using three instruction-tuned models: Qwen 2.5 7B IT, Mistral 7B IT, Gemma 2 9B IT.

3.3. Layer selection

Layer selection follows two steps: **Step 1: Plain patching sweep.** At every layer L, we swap the full residual stream vector from a source example into a base example without any rotation. **Step 2: Kneedle.** We select the layer that maximises interchange intervention accuracy earliest, using a simplified Kneedle algorithm (Satopää et al., 2011): We take kneedle $= \arg\max(y_n - x_n)$. Given a layer i, $x_n = i/L$ denotes where we are along the layer sweep, and $y_n = (\text{IIA}_i - \text{IIA}_{\min})/(\text{IIA}_{\max} - \text{IIA}_{\min})$. See layer sweep with DAS optimum in Figure 3.

3.4. DAS training

We train a DAS matrix $R \in \mathbb{R}^{d\times k}$ for $k \in \{1, 2, 4, 8\}$. We train DAS on samples where the model gets both the true statement and the false statement (offset $+1$) correct to avoid training on noise. Specifically, we train on the symbolic format only and validate on COMMUTED and WORDS (see Appendix Table 9 and Figure 4). High transfer IIA increases our confidence in the possibility that the subspace captures an abstract truth verification concept rather than a surface-specific signal.

For training DAS, we use Adam optimised over 1000 steps to learn an apt rotation matrix. The matrix is enforced orthonormal.[3] We train on a balanced (true/false) dataset of all symbolic statements as outlined in Table 9, and evaluate on held-out validation sets using all statement formats. Crucially, for both training and evaluation, we only use pairs where the model predicts the ground truth correctly. We train and test the DAS matrix *bidirectionally* by patching false activations into the forward pass on true statements, and vice versa.

IIA is measured by patching our hidden vector at the last token position with the projection using our DAS matrix.

3.5. Steering (causal sufficiency)

We test whether we can steer wrong predictions on each models main failure mode (true/false bias). Since Gemma is

[2] Given the surprisingly screwed accuracy, we validated our findings using three different variations of this template as well as few-short templates. See Appendix.

[3] $RR^\top$ needs to be a valid projection matrix. This holds if and only if columns are orthonormal ($R^\top R = I_k$).

true-biased, the main failure mode is predicting true on false statements. So, for Gemma, we patch R_1 from a source example (model predicts false for a false statement) into a base example (model predicts true for a false statement) at the DAS layer. Steering IIA is the fraction of failures corrected.

3.6. Probing

We use probing (Belinkov, 2022) to test the degree to which the ground truth and model prediction can be predicted from our subspace activations. Given 500 true statements and 500 false statements, we train two logistic regressors to find the linear decision boundaries separating ground truth label and model prediction, respectively. The probes take as input the vectors after projecting the hidden states into R_1. We evaluate using 5-fold cross-validation (800 train / 200 test per fold).

3.7. Chain-of-thought prompting

Since we know that the model is able to solve addition when prompted to generate the right answer [4] (see Table 10 in Appendix), we try to force the model to use this correct computation circuit by appending the computation to our prompt, following the method of chain-of-thought prompting (Wei et al., 2022). The two prompts are: (1) What is {a} + {b}? Answer with just the number, nothing else. (2) I calculated {a} + {b} = {generated}. Is {a} + {b} = {c} true? Answer with exactly one word: 'true' or 'false'.

3.8. Ablation (causal necessity)

We study the degree to which the model relies on our subspace in prediction by comparing two forward passes: with and without R_1 ablated. We denote the degree to which the subspace is necessary for prediction causal necessity (CN).

4. Results

4.1. Behavioural accuracy

The baseline results for all three instruction-tuned models are poor. We document them below.

Table 1. Balanced accuracy, true positive rate, true negative rate. All on the symbolic condition.

Model	Acc (bal.)	TPR	TNR	Bias
Gemma 2 9B IT	0.572	0.871	0.272	True
Mistral 7B IT	0.501	0.987	0.016	True
Qwen 2.5 7B IT	0.537	0.084	0.990	False

[4]We ran an additional experiment, prompting the model to give us the *answer* to an addition task, not predict *true* vs. *false*. This improved accuracy drastically.

4.2. DAS training

For each model, we train the DAS matrix on pairs of arithmetic statements, where the model gets the distinction between true and false right, and evaluate performance on a held-out set from the same pool. We get the following results, showing that the DAS subspace we found (R_1), is sufficient to explain model prediction on these cases.

Model	k	Val IIA	T→F	F→T
Gemma 9B IT (L27)	2	$\mathbf{.933}_{\pm.014}$	$.867_{\pm.027}$	$1.000_{\pm.000}$
Qwen 7B IT (L20)	8	$\mathbf{.950}_{\pm.000}$	$1.000_{\pm.000}$	$.900_{\pm.000}$
Mistral 7B IT (L16)	2	$\mathbf{.965}_{\pm.025}$	$.930_{\pm.050}$	$1.000_{\pm.000}$

Table 2. Val IIA reports overall validation IIA. T→F and F→T report the directional flip rates separately. Mean ± std across three seeds.

Transfer conditions We find that the high val IIA transfers to the commuted versions and word formulations of our symbolic statements.

Model	Words	Commuted
Gemma 2 9B (L27, k=2)	0.833	0.933
Qwen 2.5 7B (L20, k=8)	0.861	0.911
Mistral 7B (L16, k=2)	0.844	0.939

Table 3. Cross-format transfer of R_1. The rotation R_1 is learned on symbolic arithmetic statements ($a + b = c$) used for IIA evaluation, then evaluated on two held-out surface forms: natural-language wording (*words*) and operand-commuted symbolic statements (*Commuted*, $b + a = c$). The numbers report the mean across the three seeds used for the symbolic condition.

4.3. Steering (causal sufficiency)

For Gemma, Mistral and Qwen, the identified subspace allows effective steering of the dominant failure mode: we patch the R_1 component of a failure example with that from a matched success example and measure how often the prediction flips to correct.

Model	Symbolic	Commuted	NL words
Gemma 9B IT	$.973_{\pm.009}$	$.980_{\pm.014}$	$.970_{\pm.008}$
Qwen 7B IT	$.967_{\pm.019}$	$.967_{\pm.012}$	$.897_{\pm.037}$
Mistral 7B IT	$.990_{\pm.008}$	$.990_{\pm.008}$	$.893_{\pm.029}$

Table 4. Steering IIA for each model's main failure mode. Higher is better. Gemma and Mistral: fraction of false-failures that flip to correct false after patching R_1 from a false-success. Qwen: fraction of true-failures that flip when patching from a true-success. Mean ± std (subscript) across three seeds.

4.4. R_1 does not encode arithmetic truth

R_1 does not encode a human-interpretable concept of arithmetic truth, as supported by the following evidence:

Probing We use probing to identify the degree to which we can predict ground truth and model prediction from the activations in DAS subspace. We find that the probe separating ground truth performs at chance level, while the probe separating by model prediction achieves near-perfect accuracy. This supports the conclusion that R_1 is not a truth subspace, but a subspace encoding model output prediction. In other words, the subspace encodes what *seems* true to the model, as opposed to having a concept of what *is* true.

Model	Probe: ground truth	Probe: model pred.
Gemma 9B IT	$.569_{\pm.026}$	$\mathbf{.995_{\pm.006}}$
Qwen 7B IT	$.608_{\pm.020}$	$\mathbf{.996_{\pm.005}}$
Mistral 7B IT	$.485_{\pm.011}$	$\mathbf{.995_{\pm.003}}$

Table 5. Linear probe accuracy on R_1-projected representations (Gemma L27 k=2, Qwen L20 k=8, Mistral L16 k=2). *Probe: ground truth* predicts the gold truth label of the statement. *Probe: model pred.* predicts the model's own verification output. Mean $\pm$ std across 5 CV folds on one seed.

Geometric interpretation Semantic relations have been shown to be linearly represented for various tasks (Hernandez et al., 2024). We find that R_1 has linear structure, but organised around prediction rather than ground truth. We see this by plotting the mean direction in R_1 subspace for all four categories: true positives, false positives, true negatives, false negatives (see Figure 2). We see a clear separation that our probing result supports: the vectors separate by model output prediction, not by ground truth. For all plots, see Appendix.

4.5. CoT Repair

We use CoT prompting to activate an alternative causal pathway that enables us to bypass R_1 and reduce model failure rates:

MODEL	TPR	TNR	COT TPR	COT TNR
QWEN L20	.080	.948	.156	.962
GEMMA L27	.844	.246	.870	**.854**
MISTRAL L16	.984	.010	.954	**.998**

Table 6. Chain-of-thought (CoT) prompting on SYMBOLIC (n = 500). *TPR* and *TNR* denote accuracy under the default prompt. *CoT* columns are accuracy when the model is prompted to explicitly compute the sum before answering. The main repair effect is marked in **bold**.

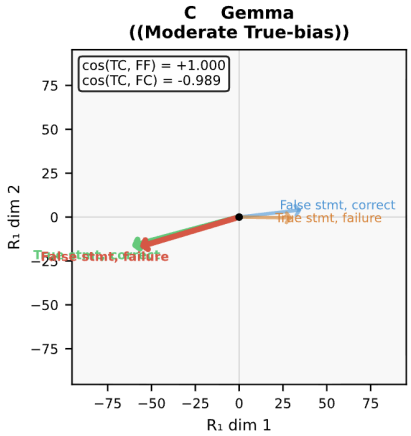

Figure 2. Mean vector projections in R_1 space for Gemma. See Appendix for all models.

What role does R_1 have in mediating the CoT fix? We see that the fix routes outside of R_1 entirely for Gemma and Mistral (see Appendix). This makes sense, given that R_1 encodes the model's bias.

5. Discussion

We have shown that a causally sufficient subspace need not encode the human-interpretable concept we think it represents. Using probing, we achieve a near-chance accuracy on ground truth and near-perfect accuracy on model predictions. For Gemma, Mistral and Qwen, R_1 is a prediction bias mechanism, not a truth concept. The implication is that, when performance is less than excellent, causal validation is necessary but not sufficient evidence that an interpretability finding corresponds to a human-aligned concept. Future work should develop principled criteria for when a causally active subspace can be claimed to encode a concept rather than a prediction artifact for cases where prediction and ground truth do not coincide.

6. Conclusion

We found a causally sufficient subspace for arithmetic verification, which encodes output bias rather than truth. Hence, causal sufficiency does not imply human-interpretable concept encoding. This diagnosis enables us to effectively repair model behaviour - not by intervening in the causal subspace, but by routing the model around it. On this basis, we encourage the community to qualify interpretability claims when a simpler hypothesis cannot be ruled out. While we have shown how to identify misalignment in low-accuracy situations, the question of how high the performance needs to be for having discovered a human-interpretable concept using DAS is an open question.

Acknowledgments

This publication was made possible through funding from the Sector Plan for Medicine and Health Sciences, supported by the Dutch Ministry of Education, Culture and Science (OCW). We also thank the members of the Multimodality, Language, and Interpretability (Mulini) lab at the University of Amsterdam.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ayonrinde, K. and Jaburi, L. A mathematical philosophy of explanations in mechanistic interpretability. In *Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)*. AAAI/ACM, 2025.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N. D., Potts, C., and Icard, T. F. Causal abstraction: A theoretical foundation for mechanistic interpretability. *J. Mach. Learn. Res.*, 26:83:1–83:64, 2023a. URL https://api.semanticscholar.org/CorpusID:255749463.

Geiger, A., Wu, Z., Potts, C., Icard, T. F., and Goodman, N. D. Finding alignments between interpretable causal variables and distributed neural representations. *ArXiv*, abs/2303.02536, 2023b. URL https://api.semanticscholar.org/CorpusID:257365438.

Geiger, A., Harding, J., and Icard, T. How causal abstraction underpins computational explanation. *arXiv preprint arXiv:2508.11214*, 2025.

Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., and Bau, D. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=w7LU2s14kE.

Hewitt, J., Geirhos, R., and Kim, B. Position: We can't understand AI using our existing vocabulary. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=asQJx56NqB.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://openreview.net/forum?id=CeJEfNKstt.

Satopää, V., Albrecht, J., Irwin, D., and Raghavan, B. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pp. 166–171. IEEE, 2011. doi: 10.1109/ICDCSW.2011.20.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A. T., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, Z., and Hendrycks, D. Representation engineering: A top-down approach to ai transparency. *ArXiv*, abs/2310.01405, 2023. URL https://api.semanticscholar.org/CorpusID:263605618.

A. Appendix

A Prompt Robustness

To document that the low accuracy scores are not an artifact of the prompt we used, we test accuracy across different permutations of our zero-shot prompt, as well as one 5-shot prompt following (Marks & Tegmark, 2024). We find that the behavioral accuracy is largely independent of prompt format, justifying our prompt choice.

Model	Framing	Template	TPR	TNR	Bal. acc.
Gemma 2 9B IT	true / false	`{stmt}\nIs this statement true? Answer with exactly one word: 'true' or 'false'.`	0.844	0.218	0.531
	correct / incorrect	`{stmt}\nIs this statement correct? Answer with exactly one word: 'correct' or 'incorrect'.`	0.876	0.176	0.526
	right / wrong	`{stmt}\nIs this statement right? Answer with exactly one word: 'right' or 'wrong'.`	0.846	0.218	0.532
	5-shot (M&T)†	{stmt$_1$} This statement is: TRUE\n{stmt$_2$} This statement is: FALSE\n...\n{stmt} This statement is:	0.751	0.296	0.523
Mistral 7B IT	true / false	`{stmt}\nIs this statement true? Answer with exactly one word: 'true' or 'false'.`	0.980	0.016	0.498
	correct / incorrect	`{stmt}\nIs this statement correct? Answer with exactly one word: 'correct' or 'incorrect'.`	0.972	0.032	0.502
	right / wrong	`{stmt}\nIs this statement right? Answer with exactly one word: 'right' or 'wrong'.`	0.968	0.040	0.504
	5-shot (M&T)†	{stmt$_1$} This statement is: TRUE\n{stmt$_2$} This statement is: FALSE\n...\n{stmt} This statement is:	0.778	0.236	0.507
Qwen 2.5 7B IT	true / false	`{stmt}\nIs this statement true? Answer with exactly one word: 'true' or 'false'.`	0.088	0.950	0.519
	correct / incorrect	`{stmt}\nIs this statement correct? Answer with exactly one word: 'correct' or 'incorrect'.`	0.058	0.978	0.518
	right / wrong	`{stmt}\nIs this statement right? Answer with exactly one word: 'right' or 'wrong'.`	0.082	0.962	0.522
	5-shot (M&T)†	{stmt$_1$} This statement is: TRUE\n{stmt$_2$} This statement is: FALSE\n...\n{stmt} This statement is:	0.435	0.689	0.562

Table 7. Behavioural bias under four prompt framings on SYMBOLIC_ALL. Rows 1–3 are zero-shot ($n = 500$). The bias is stable across *true/false*, *correct/incorrect*, and *right/wrong*, ruling out a vocabulary-specific explanation. Five-shot prompting in the Marks & Tegmark (2023) style does not resolve the chance-level accuracy.

†We use five demonstration shots (balanced 3 true / 2 false, drawn from a held-out pool of 200 problems).

B DAS training and testing

We note that our training set for DAS are substantially lower than those from (Geiger et al., 2023b), given poor behavioural performance and large model biases, and the strict training filter that both the true and corresponding false samples must be correctly predicted by the model to avoid fitting the rotation to noise. For models with severe prediction bias, particularly Mistral (TNR = 0.016), this filter reduces the available pool. However, the following observation support sufficiency: IIA standard deviation across three seeds is low, indicating stable convergence.

Table 8. DAS train and validation set sizes per model and condition. 80/20 train/val split at problem level (true and false statements for the same (a, b) pair always in the same split). Filtered pool = problems where the model predicts correctly on both (behavioural filter). $n_{\text{per_class}}$ capped at $\min(\text{pool}, 150)$.

Model	Condition	Pool	Train	Val
Gemma 2 9B IT	SYMBOLIC	2063	240	60
Qwen 2.5 7B IT	SYMBOLIC	586	240	60
Mistral 7B IT	SYMBOLIC	107	168	44

We also test generalisation of our DAS matrix to an unseen set surface-level variants of our symbolic base statements:

Table 9. Statement formats. Operand range is [2,99]. False statements are off by ± 1. This requires the model to precisely calculate the problem, as it cannot rely on magnitude differences. However, we validate our IIA scores for a DAS matrix trained on random offsets as well

Condition	Example
SYMBOLIC	`23 + 47 = 70`
COMMUTED	`47 + 23 = 70`
WORDS	`twenty-three plus forty-seven equals seventy`

C Layer Selection

We select the optimal layer for training our DAS rotation matrix using plain activation patching. We use the elbow algorithm for selecting the layer at which full-vector IIA peaks.

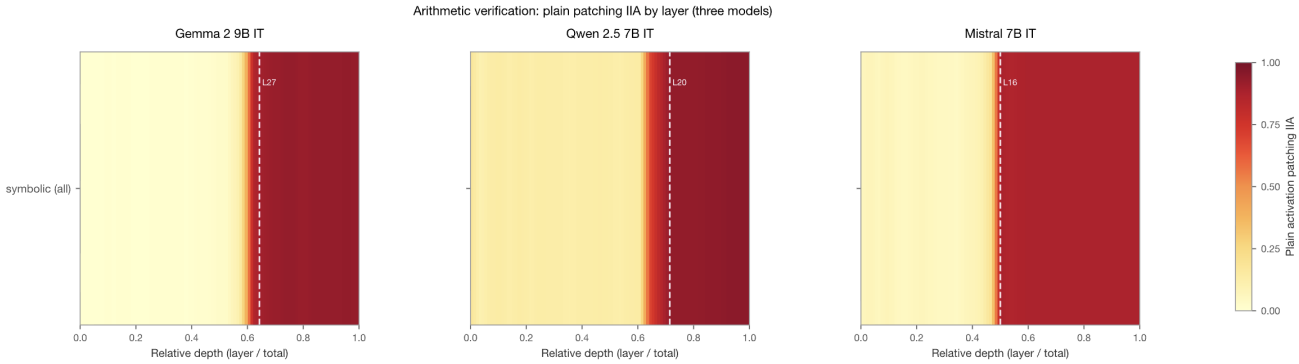

Figure 3. Layer sweep results showing IIA across layers for each model. The elbow point indicates the selected intervention layer.

We also perform plain activation patching for the correct COMMUTED and WORDS pairs, to track the degree to which the causal information is aligned in the way we may expect. We find that across all prompt formats, there *is* a single optimal layer.

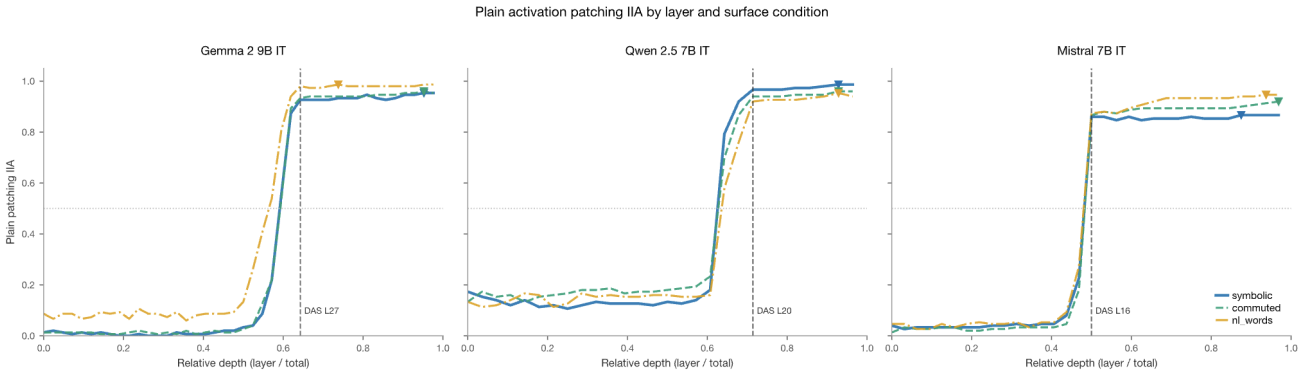

Figure 4. Layer sweep across conditions. Instead of using the original dataset where the sums for false statements are off-by-one, for this validation sweep, we choose a random offset (specifically, offset $\in \{-10, -5, -2, -1, +1, +2, +5, +10\}$).

D Geometry

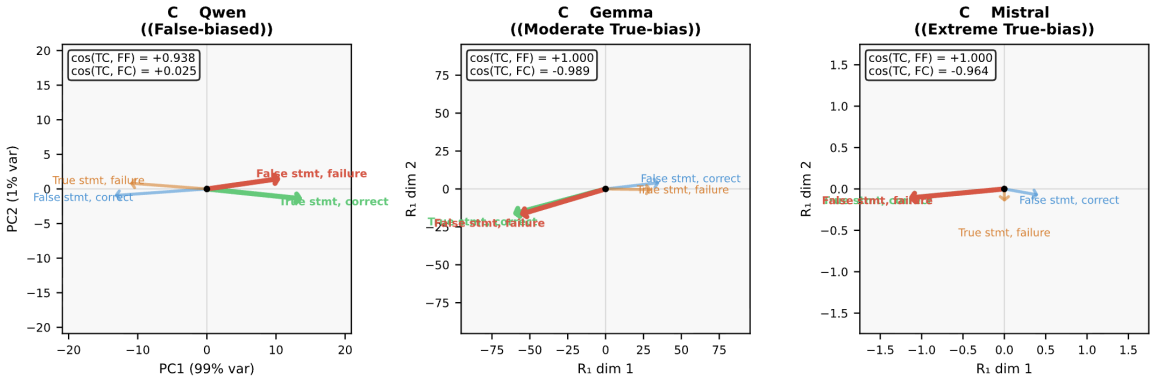

Figure 5. Mean vector projections in R_1-space. For Qwen (k=8), we project the four mean vectors (true correct, True failure, false correct, false failure) onto the first two principal components of their distribution in R_1.

E Computation Accuracy

We test model performance when prompted to compute the answer to our addition task, not decide whether it is true or false. We use this to inform the way we design the CoT prompt to reroute our model. We find that the model achieves near-perfect accuracy on the task when prompted to *compute* as opposed to *verify*.

Model	Accuracy
Gemma 2 9B IT	0.848
Mistral 7B IT	0.960
Qwen 2.5 7B IT	0.854

Table 10. Accuracy on the direct computation prompt ("What is a + b? Answer with just the number, nothing else."). n=500 problems per condition.

F R_1 is bypassed by CoT prompting

Model	Bias	Main failure	CN (standard)	CN (forced CoT)
Gemma	True-biased	Low TNR	$.402_{\pm.068}$	$.004_{\pm.004}$
Mistral	True-biased	Low TNR	$.695_{\pm.041}$	$.003_{\pm.001}$
Qwen	False-biased	Low TPR	$.987_{\pm.022}$	$.907_{\pm.041}$

Table 11. Causal necessity (CN) measures how ablating the DAS subspace changes prediction. CN=0 means "no change". For R_1 ablation under standard vs. forced-CoT inference, measured against each model's dominant failure mode (CN_{false} for True-biased models, CN_{true} for False-biased). Mean $\pm$ std across three seeds.