IDENTITY-PRESERVING ATTENTION READOUTS FOR DEDUCTIVE VALIDITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Whether a conclusion follows from its premises can be read from where the conclusion's own attention goes, nearly as well as from the hidden states that a supervised probe reads. We train a logistic classifier, a readout, on a few attention values per head, such as the share of the conclusion's attention that lands on the premises. It reaches at least 89% of the probe's ProofWriter AUC in every one of the thirteen models where the readout is extracted, up to 27 billion parameters, without access to the residual stream. On Qwen3-1.7B it scores above LapEigvals, a trained detector of attention Laplacian features, on every benchmark we test, while reading fewer features per head. The reason is that statistics which stay the same when the tokens are relabeled, such as attention Laplacian spectra, cannot represent which token attends to which. We call this limit permutation blindness and state it as an elementary lemma. Spectra of the residual stream, the very states the probe reads, also stay near chance, so the limit lies in the summary and not in the tensor it reads. Even the entropy and the maximum of the conclusion's attention row, which ignore which token receives which share, beat every spectral summary of the whole attention matrix in each of those models, so detectors built on such summaries should read that row instead. The classifier also attributes the verdict to single heads, at about the per-item storage of a probe at one fixed layer, and needs more training labels than the probe. We also find that a single negation feature reads the labels of ProofWriter far above chance, invisibly to the usual length control, and release rebalanced splits and the code that regenerates every number.

Code: https://anonymous.4open.science/r/latent-validity-8074/

1 INTRODUCTION

Detectors of hallucination and of invalid reasoning increasingly read the model rather than its output: a linear probe on the hidden states of the residual stream, or a statistic of the attention map, computed from a single forward pass before anything is generated. For deductive validity the signal is present: across 14 open-weight models up to 31 billion parameters, a linear probe reads whether a conclusion follows from its premises at up to 0.992 AUC, above a hypothesis-only control. The attention statistics proposed for the same task, spectra of attention graphs, eigenvalue trajectories across depth and the geometry of pooled hidden states, share a property that has not been examined: each gives the same value when the tokens are renumbered in the same way as queries and as keys, so none of them records which token attended to which.

Which summaries of a forward pass can carry the verdict, and why? Supervised probes (Azaria & Mitchell, 2023; Marks & Tegmark, 2024) show that it is linearly decodable but read the residual stream. LLM-Check (Sriramanan et al., 2024) scores attention without training, LapEigvals (Binkowski et al., 2025) trains a probe on attention Laplacian features, and the per-head context ratios of Lookback Lens (Chuang et al., 2024) match a hidden-state detector on contextual hallucination. We score the first two against a probe on identical forward passes under identical controls, and ask what a relabeling-invariant statistic can represent.

We use three terms throughout. A readout is a logistic regression fitted on labels to features of one forward pass. A summary is invariant when it is unchanged under a joint relabeling of the tokens in queries and keys, $A \mapsto PAP^\top$ for a permutation P. A readout is anchored when its features

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.

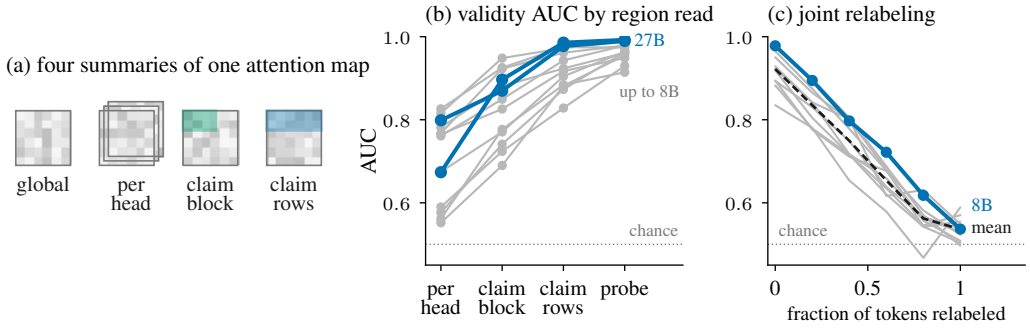

Figure 1: **Permutation blindness and the anchoring of a readout.** (a) Four summaries of the same attention maps, ordered by the region each is computed from: the eigenspectrum of the whole graph, per-head spectra, the singular spectrum of the claim→premise block, and scalar masses of the claim's own rows. (b) ProofWriter AUC of each summary and of the supervised probe, the 27B checkpoints in blue, other models in grey. (c) AUC of the anchored readout when a fraction of the tokens is relabeled in queries and keys together, the group action of Lemma 1, for the 11 models up to 8B, the largest in blue, the mean dashed. In (b) AUC rises with how closely the region is anchored to the conclusion's rows. The fall in (c) follows from the construction, since the relabeling moves the row the readout reads. Balanced ProofWriter, grouped cross-validation.

are read outward from one distinguished position, here the conclusion's own attention rows, so that which token attends to which survives the summary.

The anchored readout we propose is six scalars per head that record where the conclusion's attention rows send their mass: onto the premises, onto the claim itself, onto the sink. It reaches at least 89% of the probe's ProofWriter AUC in every one of the thirteen models where the readout is extracted, from attention alone (Figure 1). On the anchor model it scores above LapEigvals on all six benchmarks of Table 1, with six scalars per head against the ten LapEigvals reads.

The limit on invariant summaries is elementary. Which token attends to which is the relation that carries the verdict, and an invariant summary cannot represent a relation of that form (Lemma 1), a limit we call permutation blindness. One experiment separates the candidate causes. Eigenspectra of the residual stream, the exact states the probe reads, stay near chance, like the sequence-level attention spectra, which places the limit in the invariance of the summary rather than in the tensor. The entropy and the maximum of the conclusion's row, which ignore which key receives which share, already beat every spectral summary of the whole matrix in every one of the thirteen models where both are extracted.

Every number in this paper is computed on rebalanced splits. ProofWriter as distributed couples negation to its labels so strongly that a single "not"-token feature reads them at 0.927 AUC, while the usual length control reads 0.509. Three recommendations follow (Section 8): read the conclusion's own attention row rather than an invariant summary, prefer the probe where hidden states are available and labels are few, and balance every surface marker a benchmark's labels couple to.

Contributions.

1. **Anchored readouts** (§4): six scalars per head from the conclusion's own rows, at least 89% of the probe on ProofWriter from attention alone, above every spectral and LLM-Check readout on all 6 benchmarks with fewer features per head than LapEigvals, above the Lookback Lens ratio on 5 of 6, and with its storage and label cost measured against a fixed-layer probe.
2. **Permutation blindness** (§5): the elementary lemma, a two-branch experiment on residual-stream spectra that locates the failure in the summary's invariance rather than in the tensor, and a decomposition across the 13 models with per-head readouts (§6) that attributes the ordering of readouts to the region each is computed from.
3. **Heads and redundancy** (§7): a claim→premise head between about a quarter and three quarters of model depth in each of the thirteen models swept per head, and a matched-control ablation

that measures redundancy: the network absorbs the head in 10 of 13 ablated models, and the head separates from its controls in the rest.

4. **A negation artifact in ProofWriter** (§3, Appendix A): a label coupling that a one-word feature reads and the standard length control cannot see, its propagation into probes, rollout, spectra and a sequence-level spectral detector, and the splits that remove it.

2 Related work

Hidden-state probes. Linear probes (Alain & Bengio, 2017; Belinkov, 2022) recover factuality from activations, through contrast pairs (Burns et al., 2023), supervised statement probes (Azaria & Mitchell, 2023) and mass-mean directions that transfer across datasets (Marks & Tegmark, 2024; Li et al., 2023; Zou et al., 2023). Models also encode their own uncertainty and hallucinations (Kadavath et al., 2022; Orgad et al., 2025), reasoning models the correctness of their intermediate answers (Zhang et al., 2025), and structured quantities (Gurnee & Tegmark, 2024), though probes can learn the task (Hewitt & Liang, 2019) and lie detectors fail to generalize (Levinstein & Herrmann, 2024). We score probes, output likelihood (Farquhar et al., 2024; Manakul et al., 2023) and geometric summaries on the same forward passes under those critics' controls.

Attention-based diagnostics. Whether attention explains behavior is disputed (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Serrano & Smith, 2019). Analyses catalogue what heads attend to (Clark et al., 2019), rollout (Abnar & Zuidema, 2020) aggregates the maps into token relevance, and graph signal processing (Shuman et al., 2013; Ortega et al., 2018; Chung, 1997; Anonymous, 2025) summarizes each layer's attention graph by its Laplacian spectrum. Dahlem et al. (2026) prove every transpose-invariant spectral diagnostic of a head blind to the direction of flow, a limit independent of ours.

Attention-based detectors. LapEigvals (Binkowski et al., 2025) trains a probe on top Laplacian eigenvalues, the attention score of LLM-Check (Sriramanan et al., 2024) is a mean log eigenvalue used without training, and we port both from their released code. Lookback Lens (Chuang et al., 2024) fits a linear classifier to per-head ratios of attention on the context against attention on generated tokens, and finds it as effective for contextual hallucination as a hidden-state detector. Our premise and claim masses are of that kind, and this paper adds a different target, deductive validity, the conclusion's own row as the anchor, a lemma for what an invariant summary cannot represent, and controls that each remove one kind of information. TOHA (Bazarova et al., 2026) scores the topological divergence between prompt and response subgraphs of attention. Binkowski et al. (2026) argue that these detectors depend implicitly on attention sinks: under a causal mask the Laplacian eigenvalues are the sink scores less self-attention. Sorted, those scores no longer record which token holds each, the invariance of Lemma 1, and we read this as why two scalars of the conclusion's row beat every whole-matrix spectrum (§6). CHARM (Frasca et al., 2026) trains message passing on attention graphs whose nodes carry activations and whose edges mark prompt or response. Those edges mark each token's role, which places CHARM on the anchored side of the lemma, outside the invariant class our baselines are drawn from.

Heads, ablation and benchmark artifacts. Heads with identifiable functions (Voita et al., 2019; Michel et al., 2019; Olsson et al., 2022) and their circuits (Elhage et al., 2021; Wang et al., 2023; Conmy et al., 2023; Geva et al., 2021) are central to mechanistic interpretability, as is the warning that localization and causal importance diverge (Meng et al., 2022; Hase et al., 2023; Zhang & Nanda, 2024) and that ablations are compensated downstream (McGrath et al., 2023). Annotation artifact audits are standard for NLI (Poliak et al., 2018; Gururangan et al., 2018; McCoy et al., 2019) and argument comprehension (Niven & Kao, 2019), and to our knowledge we are the first to apply one to the deduction benchmarks that test reasoning (Clark et al., 2020; Tafjord et al., 2021; Han et al., 2024; Saparov & He, 2023).

3 Setup

Models and extraction. Fourteen instruction-tuned models from 8 families, one per model line and version, so Gemma-3 and Gemma-4 count as two (Appendix M): Qwen3 0.6B, 1.7B, 4B

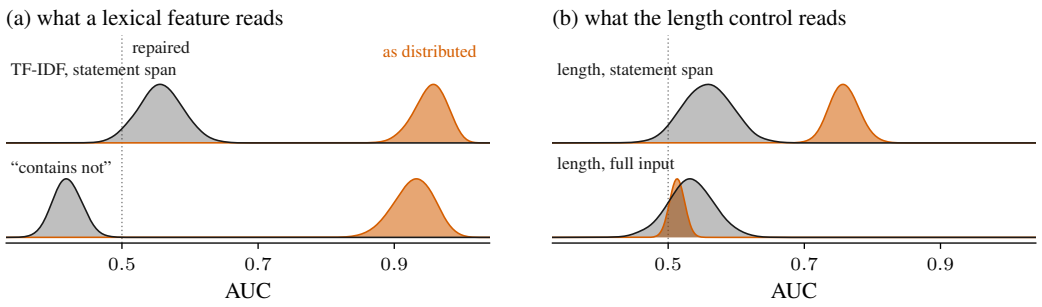

Figure 2: Text-only floors of ProofWriter before and after the negation repair. Each ridge is the bootstrap distribution of one floor under grouped resampling over theories (2000 replicates), orange as distributed and grey after the repair. (a) A binary "not"-token feature and TF-IDF on the statement span. (b) The length control on the full input and on the statement span.

and 8B (Yang et al., 2025), Qwen3.5 0.8B, 2B, 4B and 27B (Qwen Team, 2026a), Qwen3.8-27B (Qwen Team, 2026b), Gemma-3-1B (Gemma Team, 2025), Gemma-4-31B (Gemma Team, 2026), SmolLM3-3B (Bakouch et al., 2025), OLMo-2-1B (Team OLMo, 2025) and Phi-4-mini (Abouelenin et al., 2025). The Qwen3.5 and Qwen3.8 stacks are hybrids with softmax attention in one layer in four and linear attention elsewhere, so only their softmax layers have an attention matrix to read. Gemma-4 interleaves sliding-window and global attention, and every layer is read. Qwen3-1.7B is the anchor model of the single-model analyses, and Qwen3-1.7B and Qwen3-4B are the two exposition models of Section 5. From one bfloat16 forward pass per item, with nothing generated, we keep last-token hidden states for the probes, per-head attention Laplacian spectra (a graph signal processing library (Anonymous, 2025)), value-weighted and pooled hidden-state geometry, token-to-token Gram eigenshare spectra of the residual stream, and attention rollout (Abnar & Zuidema, 2020) onto the final token. The sink token is excluded from all geometry (Appendix E).

Benchmarks. ProofWriter (Tafjord et al., 2021): rule and fact theories with statements provably True or False under closed-world deduction, proof depths 1 to 5, $N = 240$, rendered *"Facts: ... Statement: ..."*. FOLIO (Han et al., 2024): expert-written first-order reasoning over natural language premises, $N = 200$. ProofWriter paired (120 theory-matched True/False pairs, $N = 240$) pins every pair to one theory, so a within-pair label flip is an exact permutation null. Four further benchmarks enter the comparison with published detectors (Appendix I).

A negation artifact in ProofWriter. ProofWriter generates statements under a closed-world assumption, and its negated statements are overwhelmingly the refutable ones: in a depth- and label-stratified sample, $P(\text{negation} \mid \text{False}) = 0.95$ against $P(\text{negation} \mid \text{True}) = 0.05$. A single binary "not"-token feature therefore reads the labels at 0.927 AUC, and TF-IDF on the statement alone, with no premises and no deduction possible, at 0.952 (Figure 2). The length control that reasoning papers report, computed on the full input, reads 0.509 and cannot see it. A sequence-level spectral detector, the uniform attention graph spectra of Section 6, loses most of its apparent signal after the repair (Appendix A). We resample ProofWriter and FOLIO so that negation is independent of the label within every cell, and every number in this paper is computed on those splits, whose post-balance floors sit at or below 0.558 and 0.547. In practice, a benchmark can reward a reader that never checks the deduction, so a readout's score means little until every such cue is removed.

Protocol. Every classifier is logistic regression on standardized features under grouped stratified 5-fold cross-validation (items sharing a theory or premise set are confined to one fold), averaged over five split seeds. Probes are reported at their best layer, an upper bound (Appendix E). Every benchmark carries four floors under the identical protocol: length, the presence of "not" in the claim span, TF-IDF on the claim span alone (the hypothesis-only control, Poliak et al., 2018) and TF-IDF on the full input. Every "A exceeds B" claim is a paired bootstrap on out-of-fold scores from shared splits.

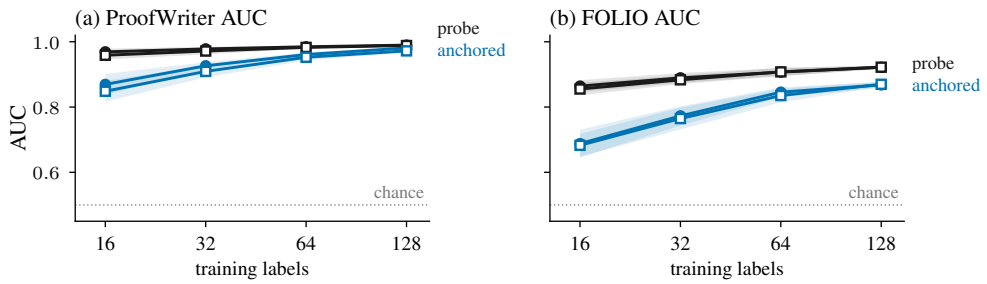

Figure 3: AUC against the number of class-balanced training labels on ProofWriter (a) and FOLIO (b), for Qwen3.5-27B (circles) and Qwen3.8-27B (squares). Black: the probe at one fixed layer. Blue: the anchored readout. Bands are one standard deviation over 20 draws. The evaluation folds are the same for every budget and both readouts.

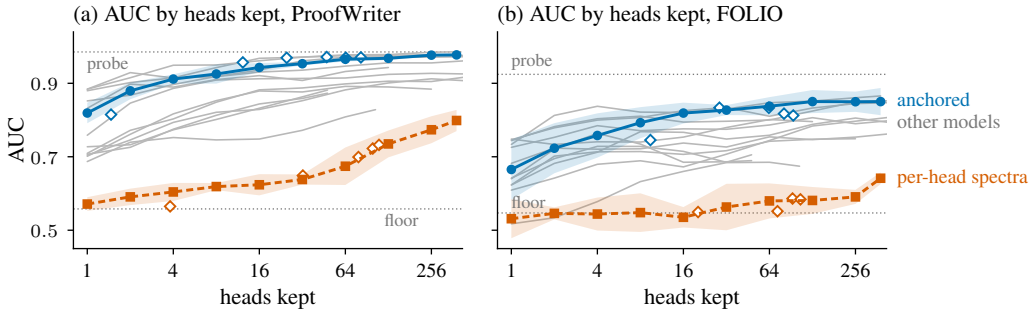

Figure 4: From one head to all heads on Qwen3.8-27B, heads ranked on training folds only. Blue: the anchored readout. Vermilion: per-head spectra, same protocol. Bands: ± 1.96 standard deviations over folds. Open marks: the ℓ_1 path. Grey: the anchored readout of the other models. Its fixed-layer probe and the lexical floor are drawn where named.

4 Anchored Readouts

An anchored readout summarizes a head from a distinguished position outward, so that which token attends to which survives the summary. We build it at two levels, spectra that keep span identity and scalars that keep the conclusion's own row, and compare both with the invariant spectra of Table 2.

Span-anchored spectra. The singular value spectrum of the bipartite block $B = A[\text{claim rows, premise columns}]$ is invariant only to permutations within each span, and it measures how coherently the claim reads the premises.

Anchored scalars. Let $A^{\ell h} \in \mathbb{R}^{N \times N}$ be the post-softmax attention of head h at layer ℓ, with the sink at position 0, premise positions $\mathcal{P}$, claim positions $\mathcal{C}$ (the conclusion and its marker), $a = A^{\ell h}_{N,\cdot}$ the row of the final claim token, and $\bar{a} = |\mathcal{C}|^{-1} \sum_{i \in \mathcal{C}} A^{\ell h}_{i,\cdot}$ the mean claim row. The readout of the head is

$$\phi_{\ell h} = \Big(\sum_{j \in \mathcal{P}} a_j, \;\; \sum_{j \in \mathcal{C}} a_j, \;\; a_0, \;\; -\sum_j a_j \log a_j, \;\; \max_j a_j, \;\; \sum_{j \in \mathcal{P}} \bar{a}_j \Big) \in \mathbb{R}^6, \tag{1}$$

premise, claim and sink mass, entropy and maximum of the final token row, and the premise mass of the mean claim row. A logistic regression on the concatenation over all attentive heads, fitted on labels under the grouped protocol of Section 3, is the *anchored readout*. Across the set it reads 0.828 to 0.986 on ProofWriter, at least 89% of the probe's AUC in every one of the thirteen models where the readout is extracted (Table 2). It ranks a provable statement above its theory-identical refutable twin in 0.958 of pairs at 4B (Appendix C). In practice, a monitor that sees only where the conclusion looks recovers most of what a probe on the hidden states finds.

Sparsity. The signal is distributed but compressible (Figure 4). With heads ranked on training folds only, one head reads 0.884 at 1.7B, 0.880 at 4B and 0.819 to 0.884 on the 2 largest models swept, at 27B. On the anchor model four heads reach 0.914.

readout, Qwen3-1.7B	reads	fitted on	ProofWriter	FOLIO	FLD	LogicNLI	ConTRoL	ProntoQA
supervised probe	hidden	labels	0.982	0.828	0.875	0.831	0.728	0.862
SAPLMA	hidden	labels	0.979	0.868	0.866	0.843	0.717	0.821
anchored readout	attention	labels	0.965	0.801	0.820	0.752	0.640	0.679
Lookback Lens ratio	attention	labels	0.958	0.753	0.746	0.732	0.695	0.675
LapEigvals	attention	labels	0.937	0.680	0.715	0.735	0.535	0.490
LLM-Check, claim span	attention	labels	0.905	0.630	0.682	0.674	0.516	0.598
per-head spectra	attention	labels	0.781	0.583	0.575	0.628	0.478	0.497
LLM-Check, hidden	hidden	labels	0.580	0.603	0.500	0.676	0.539	0.531
LLM-Check, sequence†	attention	labels	0.538	0.519	0.404	0.520	0.520	0.471
CCS	hidden	unlabeled*	0.517	0.537	0.663	0.539	0.519	0.510
$P(\text{True})$	output	nothing	0.793	0.827	0.754	0.684	0.695	0.506

Table 1: Published detectors, ported from their authors' code, and the anchored readout on six benchmarks (anchor model, grouped CV, AUC of the out-of-fold score averaged over five seeds, slightly above the per-seed mean of Table 2): SAPLMA (Azaria & Mitchell, 2023), LapEigvals (Binkowski et al., 2025), LLM-Check (Sriramanan et al., 2024), CCS (Burns et al., 2023). Reads: the tensor a readout is computed from. Fitted on: every labeled readout is a logistic regression (SAPLMA a perceptron) on the same grouped folds, which the LLM-Check scores, summed over heads in their authors' code, receive as a per-head vector. *CCS is fitted without labels and its sign is set with them. †Exactly relabeling-invariant. Appendix J gives the invariance measurement for each row.

Storage, time and labels. Per item in bfloat16 the readout keeps 5.2 kB on the anchor model, a probe at one fixed layer 4.0 kB, and a best-layer probe 112 kB until its layer is chosen. Across the 14 models the readout needs 0.14 to 2.70 times the storage of the fixed-layer probe, less in 7 of them (Table 6). On the anchor model a bare pass takes 30 ms per item, the anchored pass 49 ms and the probe path 46 ms. We fit both on 16 to 128 class-balanced labels for ProofWriter and FOLIO on Qwen3.5-27B and Qwen3.8-27B, where both feature sets are stored (Figure 3). Both use standardized features and the fixed ℓ_2 penalty of every readout here, with no inner search. The fixed-layer probe leads in all 16 comparisons, most widely at 16 labels, and the readout reaches the probe's 16-label AUC only at 128 labels in every cell. Its advantage is therefore access, a verdict read without the residual stream and attributable to single heads. The case it serves is a monitor inside an inference stack that already materializes the attention rows of a few heads, for instance eager kernels on selected layers, and that does not export hidden states.

Comparison with published detectors. Table 1 scores every ported detector and the anchored readout on six benchmarks on the anchor model. The relabeling-invariant LLM-Check score reads chance on all six. Restricted to the claim span, a use of identity, it reads 0.905 on ProofWriter. Among the spectral and LLM-Check readouts of attention the anchored readout is the highest on 6 of 6 benchmarks, by 0.017 (LogicNLI) to 0.121 AUC, above LapEigvals everywhere with six scalars per head against its ten. The Lookback Lens ratio (Chuang et al., 2024) is, per head, premise mass over premise plus claim mass of the final claim row, two of those six scalars. It falls below the full readout on 5 of 6 benchmarks and above it on ConTRoL (0.695 against 0.640), where it is the strongest readout of attention in the table. The probe's lead is widest on ProntoQA and next widest on ConTRoL, the one benchmark of real passages (0.640 against 0.728). The ConTRoL gap is not premise mass diluted over long passages, since a window near the conclusion barely moves the readout (Appendix G). SAPLMA, a supervised probe on hidden states, leads on all six by 0.014 to 0.142 AUC, and $P(\text{True})$, which fits nothing, leads on FOLIO and ConTRoL (0.827, 0.695): where states or outputs can be read, those are the instruments of choice.

5 Permutation blindness

What limits geometric summaries on states that provably contain the signal? Two mechanisms act. Head averaging is a lossy projection: any functional of the averaged graph is constant on head configurations sharing a mean, and the average can only overstate algebraic connectivity, since λ_2 is a pointwise minimum of linear functionals of the Laplacian and hence concave (Chung, 1997). That explains the head-averaged spectra at their floor but not the per-head spectra, which keep every head and still trail the probe (Table 2). There the mechanism is the summary itself.

Lemma 1. *Let s be a function of a graph-signal pair (W, X) with $s(PWP^\top, PX) = s(W, X)$ for every permutation matrix P. Then s is constant on each relabeling orbit, so it cannot represent a predicate that distinguishes points of one orbit, such as "the conclusion token attends to the third premise". Eigenvalue functionals of the Laplacian (Fiedler value, spectral entropy, eigenshares) and every energy-type diagnostic $\mathrm{tr}(X^\top \Pi_B X)$ against a spectral projector (smoothness, HFER, band and Dirichlet energies) have this invariance.*

The second claim holds because the spectrum is unchanged by conjugation and each projector transports as $\Pi_B \mapsto P\Pi_B P^\top$, which is basis free, so no eigenvector convention enters (Appendix F). The lemma bounds what these summaries can represent and not what they can score: a provable and a refutable item are different graphs, so whether an invariant correlates with validity is an empirical question.

The two mechanisms make opposite predictions for one experiment, whose decision rule was fixed in the extraction code before it was run (Reproducibility statement). Compute the eigenvalue spectrum of the token-to-token Gram matrix of the residual stream: the states the probe reads, with attention nowhere in the computation, summarized invariantly. If the substrate were the problem, these spectra should carry the signal the probe finds in the same states, and if invariance is the problem, they should fail as the attention spectra do. The decision rule: the substrate branch fires if the Gram spectra reach 0.70 AUC on balanced FOLIO, the invariance branch if they stay below while the fixed-layer probe on identical states reads at least 0.75.

The invariance branch holds. In the larger exposition model, on balanced FOLIO, Gram eigenshare spectra with their layer velocities read 0.518 ± 0.024 where the fixed-layer probe on identical states reads 0.866, and on ProofWriter 0.485 ± 0.014 against 0.978. Applied to the Gram eigenvalue functionals stored for every 27 to 31B checkpoint, the rule returns the invariance branch in 2 of 3 (Qwen3.5-27B and Qwen3.8-27B, FOLIO 0.511 to 0.512 against a probe of 0.922 to 0.924). In Gemma-4-31B the fixed-layer probe itself reads 0.687 on FOLIO, short of the bar the rule sets for it, so the rule returns neither branch, with the Gram functionals at 0.554. In other words, the states hold the verdict and the invariant summary discards it: the fault lies in how the summary is taken, not in what the model computes.

6 Readouts across models and scales

Table 2 reports every readout under one protocol for each model where it is extracted. The probe ceiling of 0.992 on ProofWriter at $N = 240$ (Qwen3.5-27B) is not leakage: items sharing a theory sit in one fold, and the hypothesis-only control, which sees no premises, reads chance on ProofWriter and 0.580 to 0.736 on FOLIO (last column). The model's own answer likelihood reads ProofWriter at 0.655 to 0.936, below the probe in 13 of the 13 models scored.

Readout ordering by region. Reading the ProofWriter columns orders the readouts. Sequence-level invariant summaries read chance (0.541 at 4B, Appendix A). Per-head spectra read 0.552 to 0.826, span-anchored spectra 0.690 to 0.948 and the anchored scalars 0.828 to 0.986, and 13 of the 13 models order the three exactly so. What orders them is the region each summary is computed from rather than any use of key identity. The entropy and the maximum of the conclusion's row are symmetric in its keys, so they do not use which token receives which share, yet scored alone they sit above the five spectral summaries of the whole matrix in 13 of 13 models (Figure 5, Table 7), and handing the spectra that same row lifts them by 0.123 to 0.295. Put simply, the closer a readout sits to the conclusion's own row, the more of the verdict it recovers.

Scale. The 27 to 31B checkpoints extend each column without changing its order (Figure 5). On ProofWriter the anchored readout reads 0.977 to 0.986 and the per-head spectra 0.674 to 0.799 on the 2 of the 3 where per-head readouts are extracted, against a probe of 0.962 to 0.992 on all 3.

Graded relabeling. Relabeling a fraction of the tokens in queries and keys together, the group action of Lemma 1, is the change an invariant summary cannot see. On the 11 models up to 8B it falls by 0.246 to 0.478 AUC, to within the interval of chance in 10 of them (Figure 1c). This fall holds by construction, because the relabeling moves the conclusion's row away from the position the readout reads, and the empirical evidence is the region decomposition above (Table 2). Relabeling

	anchored (ours)		bip.	spectra	pair	transfer	probe (ceiling)		qonly
Model	PW	FO	PW	PW	acc.	PW→FO	PW	FO	FO
Qwen3-0.6B	0.917	0.808	0.826	0.765	0.892	0.717	0.949	0.848	*0.684*
OLMo-2-1B	0.884	0.749	0.777	0.561	0.850	0.728	0.914	0.830	*0.580*
Gemma-3-1B	0.828	0.674	0.724	0.590	0.842	0.746	0.930	0.773	*0.628*
Qwen3.5-0.8B	0.882	0.689	0.690	0.552	0.858	0.760	0.955	0.804	*0.670*
Qwen3-1.7B	0.961	0.793	0.923	0.767	0.917	0.760	0.979	0.818	*0.723*
Qwen3.5-2B	0.873	0.705	0.741	0.577	0.867	0.781	0.963	0.857	*0.692*
SmolLM3-3B	0.904	0.767	0.771	0.675	0.917	0.832	0.964	0.863	*0.730*
Phi-4-mini	0.924	0.807	0.880	0.781	0.925	0.853	0.961	0.869	*0.693*
Qwen3-4B	0.973	0.857	**0.948**	0.816	0.958	0.877	0.988	0.875	*0.715*
Qwen3.5-4B	0.943	0.796	0.850	0.761	0.983	0.893	0.979	0.880	*0.712*
Qwen3-8B	0.975	0.852	0.926	**0.826**	0.950	0.888	0.974	0.865	*0.736*
Qwen3.5-27B	**0.986**	**0.866**	0.897	0.674	**0.988**	**0.902**	0.992	0.922	*0.736*
Qwen3.8-27B	0.977	0.850	0.870	0.799	0.980	0.887	0.990	0.924	*0.725*

Table 2: The thirteen models with per-head extractions, one protocol, all cells AUC except pair accuracy, each the mean of five per-seed AUCs. PW = balanced ProofWriter, FO = balanced FOLIO. **Bold**/underline: best/second per column among one-pass readouts. Anchored = per-head anchored readouts (§4), Bip. = span-anchored bipartite spectra, Spectra = per-head attention graph spectra, Pair = within-pair accuracy on theory-identical (True, False) pairs, Transfer = zero-shot AUC of the ProofWriter mass-mean direction on FOLIO (Appendix H), Probe = best-layer linear probe on last-token states, Qonly = hypothesis-only probe. Column-wise 95% grouped bootstrap CI half widths: probe PW ±0.007 to 0.037, probe FO ±0.035 to 0.067, anchored PW ±0.015 to 0.049, anchored FO ±0.050 to 0.070 (Table 5). Gemma-4-31B, extracted without per-head readouts: bip. 0.820, transfer 0.730, probe 0.962 and 0.816, qonly 0.620.

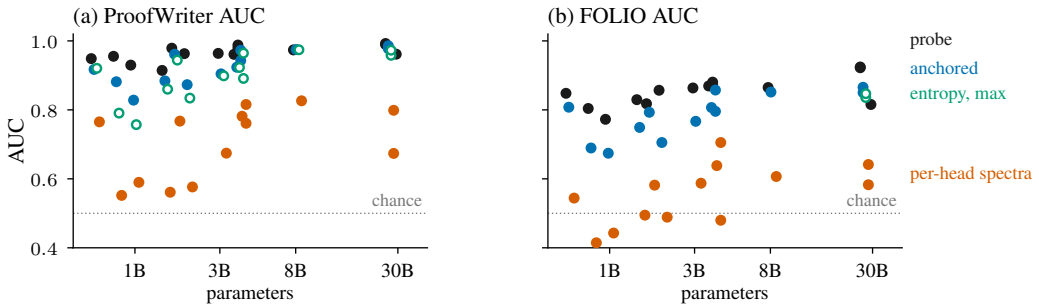

Figure 5: AUC against parameter count, one point per model, on ProofWriter (a) and FOLIO (b). Black: best-layer probe on hidden states. Blue: the anchored readout. Green, open: the entropy and maximum of the conclusion's row alone, which ignore which key receives which share. Vermilion: per-head spectra of the whole attention matrix. Points of equal size are offset slightly.

the claim span alone or the keys alone lowers it only slightly on the anchor model (Appendix G). The readout uses the correspondence between a row and the token that owns it.

Transfer. Fitted on ProofWriter, the anchored readout reads FOLIO zero shot at 0.511 to 0.818 and the probe's mass-mean direction at 0.717 to 0.902 (Appendix H).

7 Head Selection and Ablation

Circuit discovery has found a syllogistic circuit built on middle-term suppression (Kim et al., 2025). Here heads are located by selection on the readout and tested by matched ablation. Nested head selection concentrates on the same kind of head in every model swept per head: the claim→premise head with the highest combined selection frequency across both benchmarks sits at 24 to 75% of model depth in every one of the 13 models (Figure 6, Appendix E). In 10 of the 13 models swept on both benchmarks the same head is selected on both, and in Qwen3-0.6B, Gemma-3-1B and SmolLM3-3B no head enters both nested top lists, so the target is the selection rule's fallback. The

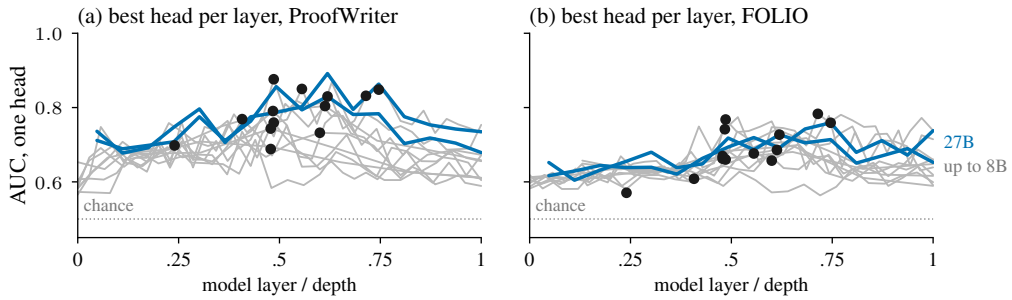

Figure 6: Univariate AUC of the best anchored head in each layer against the relative depth of that layer in the model, one curve per model with a per-head extraction, on ProofWriter (a) and FOLIO (b). Blue: the 27B checkpoints. Grey: the models up to 8B. Hybrid stacks are drawn at their softmax layers only. Black dots: the head the selection rule picks in each model.

best single head of a layer is strongest in mid-stack at every scale (Figure 6), above the floor that a maximum over heads sets even without signal.

Whether the network needs that head varies by model. We zero ablate each model's target head against matched random-head controls on identical items, from the same layer and from depth-matched layers, and test the target against the model's own controls with one two-sided rule shared by every model (Table 4 in Appendix D). In 10 of the 13 ablated models the target's behavioral effect sits inside the control range: zeroing the head is indistinguishable from zeroing a depth-matched random head, the self-repair pattern known as the hydra effect. In Qwen3-0.6B the ablation improves behavior beyond every control. Qwen3.5-0.8B and SmolLM3-3B separate: the target is the worst of each cohort on the behavioral readout, the two cases where a single head carries behavioral competence. Replacing the head's output by its dataset mean returns the same verdict on the two models up to 8B where zeroing is required (Appendix D). In practice, the claim→premise head is easy to locate, but most networks route around its removal, so it marks where the verdict is read rather than a single point of failure.

8 Conclusion

Reading the conclusion's own attention row recovers most of what a probe reads, because that row keeps what an invariant summary of the whole matrix loses. Three recommendations follow. First, a detector that reads attention should read that row. Per-head spectra of the whole matrix average 0.703 ProofWriter AUC where two scalars of the row read 0.899 (Table 7). Second, where hidden states can be read and labels are few, the probe is the better choice. At 16 labels it leads the anchored readout by 0.100 to 0.176 AUC (Figure 3). Third, a deduction benchmark should balance every surface marker its labels couple to, not only length, and ship each item's negative twin. On ProofWriter as distributed a one-word negation feature reads the labels at 0.927 AUC while length reads 0.509 (Figure 2).

9 Limitations

Label cost. The anchored readout is a supervised instrument that needs more labels than a fixed-layer probe at about the same storage, measured on ProofWriter and FOLIO on Qwen3.5-27B and Qwen3.8-27B. **Numerical precision.** Attention in bfloat16 drifts by up to 5×10^{-2} per feature between environments while the anchored AUC moves by 10^{-4}. Best-layer probe values are upper bounds, which favors the probe. **Benchmark realism.** Five of the six benchmarks are constructed deduction sets, and on the one open-world benchmark the probe's lead is the second widest of the six (§4). **Coverage.** English text only, with N per benchmark of 200 to 240 items. Graded relabeling covers the models up to 8B and the head ablation 13 of the 14 models, scored behaviorally alone on the hybrid stacks. **Confounding.** Residual confounding moves every readout up together and the invariant summaries most, so it cannot produce the ordering by region, though it can inflate its top.

AI USE STATEMENT

Large language model assistants contributed to this paper in three roles, and in no other. First, they polished the writing, tightening wording and keeping terms consistent. Second, they took part in research ideation and execution: they helped formulate and sharpen hypotheses and experimental designs, wrote and debugged the code that extracts activations, ports the baselines and runs the analyses, and helped interpret what the results mean. Third, they drafted sections of the paper, which we then rewrote. Each hypothesis, design, analysis and claim was checked by the authors, and every reported number is produced by the released code from the released result files. The authors are accountable for everything in the final paper, including any text, claims or artifacts generated with AI assistance.

ETHICS STATEMENT

The work uses public benchmarks and public checkpoints and involves no human subjects. The ProofWriter coupling of Section 3 bears on results that others have reported on that benchmark, so the audit code and the rebalanced splits are released for anyone who wishes to rescore them.

REPRODUCIBILITY STATEMENT

The released code regenerates each table, figure and quoted value from the stored extractions and result files. All fourteen checkpoints are public and were evaluated in bfloat16 (Appendix M). Activations were extracted on NVIDIA A100 GPUs and, for the smaller models, on a single 16 GB GPU. Folds are grouped, so that items sharing a theory, a premise set or a source theorem never straddle a split, and comparisons use a paired bootstrap over shared out-of-fold scores, in most cases across five split seeds. For the Gram experiment of Section 5, the statistic, the datasets and both decision branches were committed to the repository, in the extraction code, before the experiment was run on the two exposition models. That record is not independently timestamped. The same rule was applied afterwards to the larger checkpoints. The repository also holds the rebalanced splits, the leakage audits, a log of corrections to intermediate results and every extraction and analysis script: https://anonymous.4open.science/r/latent-validity-8074/.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of ACL*, 2020. URL https://doi.org/10.18653/v1/2020.acl-main.385.

Abdelrahman Abouelenin et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025. URL https://arxiv.org/abs/2503.01743.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2017. URL https://arxiv.org/abs/1610.01644.

Anonymous. Reference suppressed for double-blind review. Software library, 2025.

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*, 2023. URL https://doi.org/10.18653/v1/2023.findings-emnlp.68.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: Smol, multilingual, long-context reasoner. Hugging Face blog, 2025. URL https://huggingface.co/blog/smollm3.

Alexandra Bazarova, Andrei Volodichev, Aleksandr Yugay, Andrey Shulga, Alina Ermilova, Konstantin Polev, Julia Belikova, Rauf Parchiev, Dmitry Simakov, Maxim Savchenko, Andrey

Savchenko, Serguei Barannikov, and Alexey Zaytsev. Hallucination detection in LLMs with topological divergence on attention graphs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15449–15470, 2026. URL https://doi.org/10.18653/v1/2026.acl-long.704.

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

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Kajdanowicz. Hallucination detection in LLMs using spectral features of attention maps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://doi.org/10.18653/v1/2025.emnlp-main.1239.

Jakub Binkowski, Kamil Adamczewski, and Tomasz Kajdanowicz. Attention sinks as internal signals for hallucination detection in large language models. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2604.10697.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2212.03827.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. Lookback Lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1419–1436, 2024. URL https://doi.org/10.18653/v1/2024.emnlp-main.84.

Fan R. K. Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1997. URL https://doi.org/10.1090/cbms/092.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the ACL Workshop BlackboxNLP*, 2019. URL https://doi.org/10.18653/v1/W19-4828.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In *Proceedings of IJCAI*, 2020. URL https://doi.org/10.24963/ijcai.2020/537.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2304.14997.

Dominik Dahlem, Diego Maniloff, and Mac Misiura. Self-attention as transport: Limits of symmetric spectral diagnostics. *arXiv preprint arXiv:2605.04893*, 2026. URL https://arxiv.org/abs/2605.04893.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. URL https://doi.org/10.1038/s41586-024-07421-0.

Fabrizio Frasca, Guy Bar-Shalom, Yftah Ziser, and Haggai Maron. Neural message-passing on attention graphs for hallucination detection. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2509.24770.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of EMNLP*, 2021. URL https://doi.org/10.18653/v1/2021.emnlp-main.446.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.02207.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT*, 2018. URL https://doi.org/10.18653/v1/N18-2017.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, et al. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22017–22031, 2024. URL https://doi.org/10.18653/v1/2024.emnlp-main.1229.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2301.04213.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, 2019. URL https://doi.org/10.18653/v1/D19-1275.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, 2019. URL https://doi.org/10.18653/v1/N19-1357.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Geonhee Kim, Marco Valentino, and André Freitas. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10074–10095, 2025. URL https://doi.org/10.18653/v1/2025.findings-acl.525.

Benjamin A. Levinstein and Daniel A. Herrmann. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, 2024. URL https://doi.org/10.1007/s11098-023-02094-3.

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 (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.03341.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. Natural language inference in context: Investigating contextual reasoning over long texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13388–13396, 2021. URL https://doi.org/10.1609/aaai.v35i15.17580.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of EMNLP*, 2023. URL https://doi.org/10.18653/v1/2023.emnlp-main.557.

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://arxiv.org/abs/2310.06824.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of ACL*, 2019. URL https://doi.org/10.18653/v1/P19-1334.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023. URL https://arxiv.org/abs/2307.15771.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2202.05262.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1905.10650.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. Learning deductive reasoning from synthetic corpus based on formal logic. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/morishita23a.html.

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *Proceedings of ACL*, 2019. URL https://doi.org/10.18653/v1/P19-1459.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://arxiv.org/abs/2209.11895.

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 *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.02707.

Antonio Ortega, Pascal Frossard, Jelena Kovačević, José M. F. Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106 (5):808–828, 2018. URL https://doi.org/10.1109/JPROC.2018.2820126.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, 2018. URL https://doi.org/10.18653/v1/S18-2023.

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

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.01240.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of ACL*, 2019. URL https://doi.org/10.18653/v1/P19-1282.

David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013. URL https://doi.org/10.1109/MSP.2012.2235192.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. LLM-Check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3c1e1fdf305195cd620c118aaa9717ad-Abstract-Conference.html.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2402.17762.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021. URL https://doi.org/10.18653/v1/2021.findings-acl.317.

Team OLMo. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025. URL https://arxiv.org/abs/2501.00656.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://doi.org/10.18653/v1/2021.emnlp-main.303.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of ACL*, 2019. URL https://doi.org/10.18653/v1/P19-1580.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2211.00593.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP*, 2019. URL https://doi.org/10.18653/v1/D19-1002.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.17453.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. In *Second Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2504.05419.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.16042.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

APPENDIX CONTENTS

A The label artifact in ProofWriter . . . 15
B Benchmark decontamination details . . . 15
C The pair-exact rollout control . . . 16
D Ablation controls per model . . . 17
E Per-model details . . . 17
F Proof of Lemma 1 . . . 19
G Graded relabeling and the windowed readout . . . 20
H Cross-benchmark transfer . . . 21
I Benchmarks beyond ProofWriter and FOLIO . . . 21
J Published detectors as baselines . . . 22
K The claim relabeling control . . . 22
L Readout construction and prompt strings . . . 23
M Checkpoints . . . 23
N One item . . . 23
O Implications for practice . . . 23

A THE LABEL ARTIFACT IN PROOFWRITER

Section 3 states the repair every later number depends on, and this appendix gives the full account of the artifact and of what the repair removes. The artifact is exposed by a lexical audit of the distributed benchmark. On the distributed benchmark the uniform attention graph spectra read 0.728 at 1.7B and 0.662 at 4B against a length floor of 0.509, a margin that is compatible with an artifact. Only a floor that length cannot proxy exposes it.

The coupling. ProofWriter generates statements under a closed-world assumption, and negated statements (*"Charlie is not blue"*) are overwhelmingly the provably false ones. In a (depth, label) stratified sample of the validation split, $P(\text{negation in statement} \mid \text{False}) = 0.95$ against $P(\text{negation} \mid \text{True}) = 0.05$. Under our standard grouped protocol, a single binary "not"-token feature reads the labels at 0.927. TF-IDF on the statement span alone (no theory, no deduction possible) reads 0.952. The length floor on the full input of the same sample reads 0.509 and is blind to it. The coupling comes from the generator rather than from our sampling: among theories offering a (True, False) statement pair at depth ≥ 1, 3317 phrase the False member with negation and the True member without, against 32 phrasing both positively.

Coupling by proof depth. The (label $\times$ negation) cell counts by depth (Table 3, Appendix B) shrink monotonically: depth 4 offers 6 True negated against 5 False positive statements (it cannot be negation balanced at all) and depth 5 barely can (42 per cell). Any depth-stratified result on unbalanced ProofWriter inherits a shortcut that strengthens with nominal difficulty.

Rebalancing. We resample both benchmarks so that regex negation is independent of the label by construction. FOLIO, which also negates lexically (*fails to*, *lacks*), keeps a reported residual floor. Post-balance floors sit at or below 0.558 (ProofWriter) and 0.547 (FOLIO), the pair-exact split's floors are 0.499 and 0.411 by construction. All cell designs and floors are in Appendix B.

Propagation to each readout. Figure 2 shows the text-only floors before and after the repair. On the repaired split at 4B the supervised probe at its fixed layer still reads 0.978 where the text-only floors sit near chance, so it reads deduction rather than the cue, and Section 6 shows its statement-only control at chance. Attention rollout reads 0.595 and value-weighted spectra 0.593 on the same split. The uniform attention graph spectra, a sequence-level spectral detector, scored on both versions of the benchmark, collapse onto their floor, from 0.662 to 0.541. Their result on the distributed benchmark was the negation cue in its entirety. The margin over the length floor offered no protection because the artifact resides in the statement text, which length does not proxy.

B BENCHMARK DECONTAMINATION DETAILS

ProofWriter. From the validation split of Tafjord et al. (2021) we sample statements provably True or False under closed-world deduction across proof depths 1 to 5, in (depth $\times$ label $\times$ regex

depth	0	1	2	3	4	5
True, positive	8282	5401	4010	2822	552	446
True, negated	750	391	200	297	6	42
False, positive	700	369	200	300	5	43
False, negated	8332	5423	4010	2819	553	445
balanceable/cell	700	369	200	297	5	42

Table 3: ProofWriter validation (label × regex negation) cell counts by proof depth. The negation-to-label coupling intensifies with depth. Depth 4 (red) is unbalanceable, depth 5 (orange) nearly so.

negation) cells of equal size, giving $N = 240$ with negation independent of the label by construction. Post-balance floors under the full grouped protocol: length 0.536, contains *not* 0.420, TF-IDF on the statement span 0.558. The pair-exact split draws one (True, False) statement pair per theory, negation balanced at the pair level (120 pairs, $N = 240$): its length and one-feature floors are 0.499 and 0.411 exactly, and the within-pair label-flip permutation null is exact on this design. The coupling before the repair, reconstructed deterministically from the source data: $P(\text{negation} \mid \text{False}) = 0.95$ vs. $P(\text{negation} \mid \text{True}) = 0.05$. Among theories offering a (True, False) pair, 3317 phrase the False member with negation against 32 phrasing both positively.

FOLIO. The raw pool of Han et al. (2024) couples negation to the label at $P(\text{neg} \mid \text{False}) = 0.40$ vs. $P(\text{neg} \mid \text{True}) = 0.22$. We rebalance label × regex negation cells ($N = 200$). Because FOLIO also negates lexically (*fails to*, *lacks*), the regex is a lower bound and the residual floor must be quoted next to any FOLIO number: post-balance TF-IDF on the conclusion span reads 0.547, contains *not* 0.432, length 0.546.

C THE PAIR-EXACT ROLLOUT CONTROL

Head-averaged attention rollout (Abnar & Zuidema, 2020) is the one classical aggregate that keeps a distinguished position, and its signal survives decontamination, so it tests whether a distinguished position alone carries the verdict. On the pair-exact split it ranks the provable statement above its theory-identical refutable twin in 0.775 of pairs at 4B. The within-pair label-flip null, exact on this design, centers at 0.495 ± 0.032 ($p = 0.005$), and a full label permutation confirms that the test returns chance under broken labels (Figure 7). At 1.7B the effect is marginal (0.725 $[0.645, 0.805]$). Anchored readouts pass the same control more strongly, in all 13 models where the pair split is extracted, over the same 120 pairs: 0.842 at the least of the set and 0.958 $[0.923, 0.994]$ at 4B.

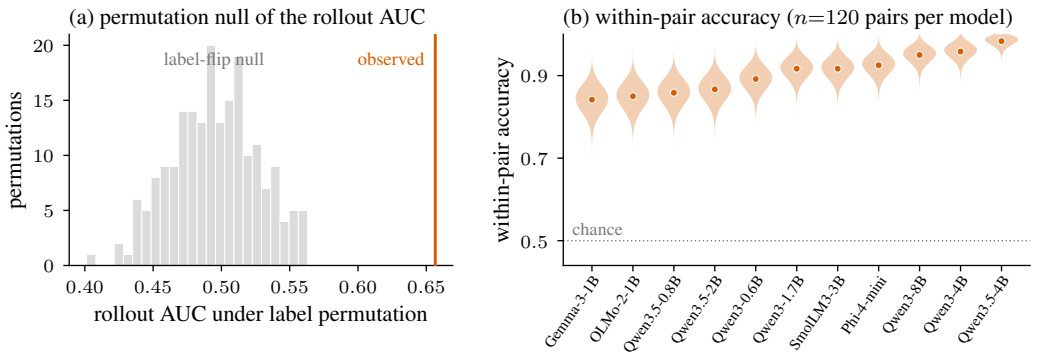

Figure 7: The pair-exact control. Left: the exact within-pair label-flip null for head-averaged rollout at 4B against its observed AUC. Right: within-pair accuracy of the anchored readout for each of the 11 models up to 8B, each drawn as its exact binomial sampling distribution over the 120 theory-identical pairs. Every one separates the provable statement from its twin, and the weakest of them sits well clear of chance.

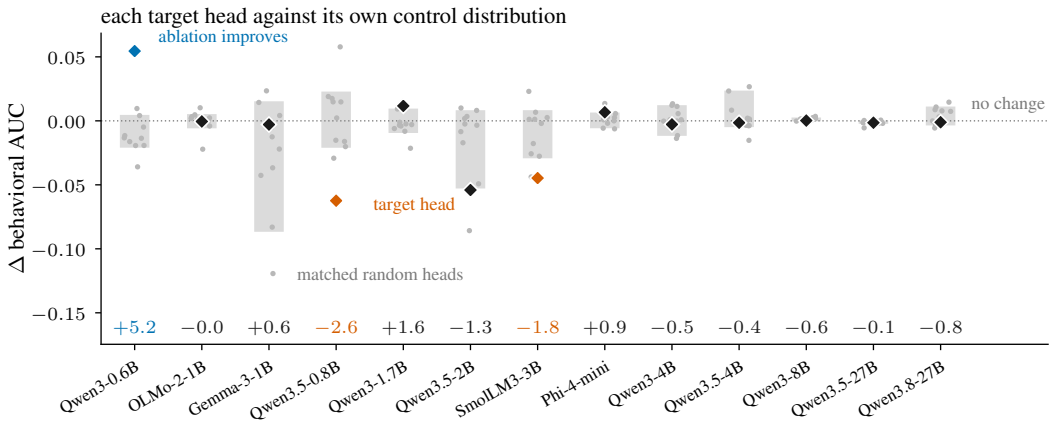

Figure 8: Zero ablation of each model's target deduction head against its own matched random-head controls, every control shown. The band spans the central 80% of the controls. Color marks direction: below every control, above every control, or inside them. The width of the control distribution differs by an order of magnitude across models, so an effect of one size can fall inside the controls of one model and below those of another. Qwen3.5-0.8B has the most negative z of the set (-2.6).

D Ablation controls per model

Table 4 gives each target head's verdict against its controls, including Qwen3.5-27B and Qwen3.8-27B. The control distributions themselves are below, because their width decides how large an effect must be to fall outside them. Replacing the head's output by its dataset mean returns the zeroing verdict on Qwen3.5-0.8B and SmolLM3-3B. Resampling another item's output costs every control head as much as the target, so at this scale that lesion singles out no head. Control spreads differ tenfold across models, hence z beside the rank.

Model	Head	Δ behavioral	z	Rank/N	Verdict
Qwen3-0.6B	L11H2	+0.055	+5.2	11/11	IMPROVES
OLMo-2-1B	L9H15	−0.000	−0.0	3/11	ABSORBED
Gemma-3-1B	L6H3	−0.003	+0.6	8/11	ABSORBED
Qwen3.5-0.8B	L11H4	−0.062	−2.6	1/11	REQUIRED
Qwen3-1.7B	L15H0	+0.012	+1.6	10/11	ABSORBED
Qwen3.5-2B	L11H7	−0.054	−1.3	2/11	ABSORBED
SmolLM3-3B	L17H5	−0.045	−1.8	1/11	REQUIRED
Phi-4-mini	L15H5	+0.007	+0.9	10/11	ABSORBED
Qwen3-4B	L17H28	−0.003	−0.5	3/11	ABSORBED
Qwen3.5-4B	L19H15	−0.001	−0.4	5/11	ABSORBED
Qwen3-8B	L25H8	+0.000	−0.6	3/11	ABSORBED
Qwen3.5-27B	L47H7	−0.001	−0.1	6/11	ABSORBED
Qwen3.8-27B	L39H6	−0.001	−0.8	4/11	ABSORBED

Table 4: Zero ablation of each model's target head, named by model layer. Δ behavioral is the change in True/False readout AUC, and z expresses it in standard deviations of that model's own random-head controls. Rank/N ranks the target among itself and its controls. Verdict, one two-sided rule for every model: REQUIRED when the target falls below every control, IMPROVES when it exceeds every control, ABSORBED otherwise.

E Per-model details

Shared deduction heads. The head selected by nested sweeps for each model, its relative depth, and its selection counts out of the 25 grouped folds (5 folds × 5 seeds), as PW/FO, by model layer: Qwen3-0.6B L11·H2 (41%) 25/below the nested top ten, OLMo-2-1B L9·H15 (60%) 25/19, Gemma-3-1B L6·H3 (24%) 25/below the nested top ten, Qwen3.5-0.8B L11·H4 (48%) 25/25, Qwen3-1.7B L15·H0 (56%) 25/17, Qwen3.5-2B L11·H7 (48%) 23/22, SmolLM3-3B L17·H5 (49%) 25/below the nested top ten, Phi-4-mini L15·H5 (48%) 25/20, Qwen3-4B L17·H28 (49%) 25/23, Qwen3.5-4B L19·H15 (61%) 25/23, Qwen3-8B L25·H8 (71%) 25/25, Qwen3.5-27B L47·H7 (75%)

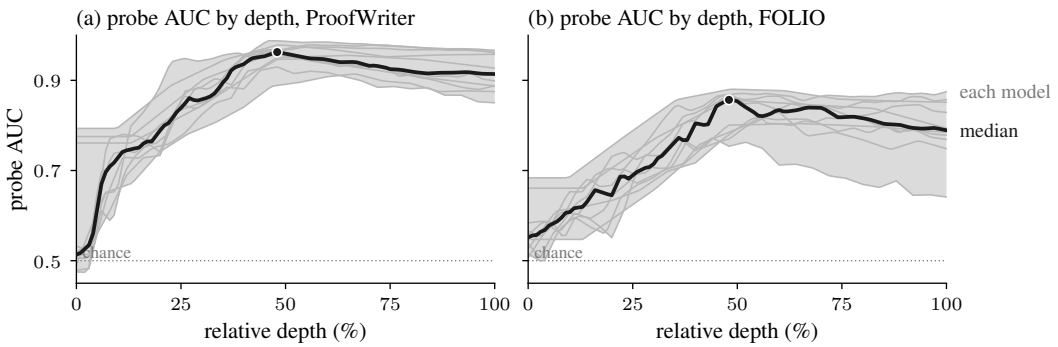

Figure 9: Per-layer linear probe AUC (last token, grouped CV), each of the 11 models up to 8B as a grey curve and the median in black. Validity is resolved by mid depth and partially lost afterwards.

25/25, Qwen3.8-27B L39·H6 (62%) 25/24. A target selected on one benchmark only lesions a benchmark-dominant readout rather than a shared one, and its verdict certifies correspondingly less.

	probe		anchored (ours)	
Model	PW	FO	PW	FO
Qwen3-0.6B	±0.024	±0.054	±0.032	±0.060
OLMo-2-1B	±0.037	±0.067	±0.042	±0.070
Gemma-3-1B	±0.032	±0.058	±0.049	±0.059
Qwen3.5-0.8B	±0.024	±0.059	±0.039	±0.065
Qwen3-1.7B	±0.013	±0.060	±0.018	±0.058
Qwen3.5-2B	±0.020	±0.058	±0.038	±0.063
SmolLM3-3B	±0.020	±0.053	±0.037	±0.061
Phi-4-mini	±0.020	±0.052	±0.034	±0.061
Qwen3-4B	±0.011	±0.049	±0.016	±0.055
Qwen3.5-4B	±0.013	±0.050	±0.026	±0.062
Qwen3-8B	±0.016	±0.053	±0.015	±0.057
Qwen3.5-27B	±0.007	±0.038	±0.015	±0.050
Qwen3.8-27B	±0.008	±0.035	±0.018	±0.055

Table 5: Per-cell 95% grouped bootstrap CI half widths for the probe and anchored columns of Table 2, computed on the out-of-fold scores. FOLIO intervals are systematically wider than ProofWriter's at the shared N. The caption of Table 2 reports the per-column ranges. Gemma-4-31B has no per-head extraction, and its probe half widths are ± 0.019 on ProofWriter and ± 0.061 on FOLIO.

Selection in the probe column. The probe values of Table 2 are the maximum over layers of a grouped out-of-fold curve, with the argmax taken on the curve that is then reported. Choosing the layer inside each training fold instead, on the same states and folds, lowers them: by 0.006 on average on ProofWriter (maximum 0.016) and 0.015 on FOLIO (maximum 0.023). On the expanded ProofWriter split, where N grows 4.5×, the inflation falls to 0.001, the shrinkage with sample size that selection bias predicts. On expanded FOLIO, where N grows only 1.7×, it does not fall (0.011), so we report the shrinkage as observed on the larger increase and not as a general claim. The anchored column is not exposed to this selection, because its head choice is nested inside the folds (§7), and the hypothesis-only probe is exposed by the same argmax. Table 2 therefore favors the probe by roughly these margins, so the anchored readout's shortfall against it is if anything overstated. The inflation is measured on 11 models, with per-model values in the released results.

Probe layers. Per-layer probe curves peak in the middle layers and decline afterwards (Figure 9), and on the hybrid stacks the residual stream is read at the softmax layers. Fixed-layer (75% depth) probe values, reported alongside the best-layer upper bounds of Table 2, are in the released results. On the Qwen3 pair, the fixed-layer value never changes an ordering.

Storage per item. Table 6 counts what each readout keeps per item in bfloat16, two bytes per value. The anchored readout keeps six values for every head of every layer it reads, the fixed-layer

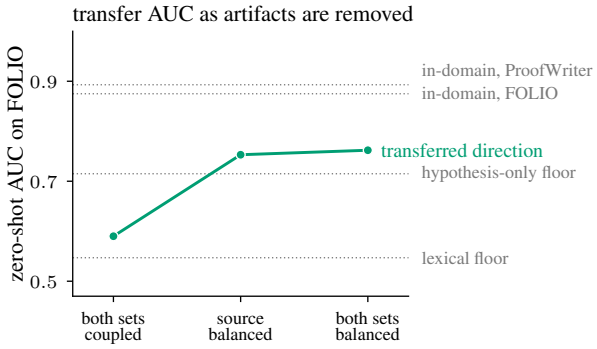

Figure 10: Zero-shot transfer of the ProofWriter mass-mean direction to FOLIO (4B) across the three decontamination stages, with in-domain ceilings and FOLIO's floors. Removing artifacts raises transfer, the signature of a real shared direction.

probe one last-token state of the hidden size, and a best-layer probe one such state per model layer until its layer is chosen. The anchored readout is the smaller of the two single-pass readouts in 7 of the 14 models, and no single-pass entry exceeds 22.5 kB.

		kB per item			
Model	layers × heads	anchored	fixed layer	best layer	ratio
Qwen3-0.6B	28×16	5.2	2.0	56	2.62
OLMo-2-1B	16×16	3.0	4.0	64	0.75
Gemma-3-1B	26×4	1.2	2.2	58	0.54
Qwen3.5-0.8B	6×8	0.6	2.0	48	0.28
Qwen3-1.7B	28×16	5.2	4.0	112	1.31
Qwen3.5-2B	6×8	0.6	4.0	96	0.14
SmolLM3-3B	36×16	6.8	4.0	144	1.69
Phi-4-mini	32×24	9.0	6.0	192	1.50
Qwen3-4B	36×32	13.5	5.0	180	2.70
Qwen3.5-4B	8×16	1.5	5.0	160	0.30
Qwen3-8B	36×32	13.5	8.0	288	1.69
Qwen3.5-27B	16×24	4.5	10.0	640	0.45
Qwen3.8-27B	16×24	4.5	10.0	640	0.45
Gemma-4-31B	60×32	22.5	10.5	630	2.14

Table 6: Storage per item at two bytes per value, with the ratio of the anchored readout to the fixed-layer probe. Layers are the attentive layers, the softmax layers alone on the hybrid stacks. Gemma-4-31B is counted from the shape of its span-anchored extraction.

Spectral configuration. Attention graph spectra use the symmetric normalized Laplacian. The random walk normalization common in graph signal processing toolkits has a nonorthonormal eigenbasis under which energy-based diagnostics are invalid. The BOS/attention sink token (Xiao et al., 2024; Sun et al., 2024) is excluded from all representation geometry: with it included, the top Gram eigenvalue carries 0.999 of the trace and every spectral summary measures the sink. Per-head spectra are computed with the per-head routine of the same library (Anonymous, 2025): symmetric $(A+A^\top)/2$ adjacency, symmetric normalized Laplacian, the sink dropped by restricting to the token span $[1, N)$ before the graph is built, and one batched eigendecomposition per layer over all heads. Five scalars per head: Fiedler value, connectivity ratio $\lambda_2/\lambda_{\max}$, eigenvalue entropy normalized by $\log N$, eigenvalue mass in the upper half of the spectrum, and spectral radius.

F Proof of Lemma 1

Write a summary in the failing set canonically as a function of the graph-signal pair (W, X), where W is the symmetrized attention (or Gram) adjacency and X the per token signal, and let L be the corresponding normalized Laplacian. Joint relabeling by a permutation matrix P acts as $(W, X) \mapsto (PWP^\top, PX)$, hence $L \mapsto PLP^\top$.

Eigenvalue functionals. $PLP^\top$ is similar to L, so $\mathrm{spec}(PLP^\top) = \mathrm{spec}(L)$. Any function of the spectrum alone (the Fiedler value λ_2, spectral entropy, eigenshares, spectral gap) is therefore unchanged.

Energy-type diagnostics. Let $\Pi_B = \sum_{\lambda_i \in B} u_i u_i^\top$ be the orthogonal projector onto the eigenvalue band B. Under relabeling the eigenvectors map as $u_i \mapsto Pu_i$, so $\Pi_B \mapsto P\Pi_B P^\top$, and

$$\mathrm{tr}\big((PX)^\top P\Pi_B P^\top (PX)\big) = \mathrm{tr}\big(X^\top P^\top P\Pi_B P^\top PX\big) = \mathrm{tr}\big(X^\top \Pi_B X\big),$$

using $P^\top P = I$. Smoothness, HFER and band energies are of this form, and Dirichlet energy $\mathrm{tr}(X^\top LX)$ is the eigenvalue weighted case $\sum_i \lambda_i \,\mathrm{tr}(X^\top \Pi_{\{\lambda_i\}} X)$. Because the statement is made with projectors rather than with a chosen eigenbasis, no eigenvector sign or ordering convention enters, and degenerate eigenvalues are aggregated over rather than assumed away.

Consequence. Each such s is a class function of the joint permutation action and is constant on every orbit $\{(PWP^\top, PX) : P \in S_n\}$. Two configurations in one orbit therefore receive identical values, so any predicate that distinguishes points within an orbit, such as "the conclusion token attends to the third premise", is not representable. This is a statement about representation, not about scores: distinct items are generally distinct orbits, and an invariant may still correlate with a label without representing the relation that determines it. §6 and §4 settle that empirically.

G GRADED RELABELING AND THE WINDOWED READOUT

Both experiments here reduce attention inside the same forward pass as the readout of §4, under the grouped protocol of §3, so only the index sets differ from the numbers in the body.

readout, balanced ProofWriter	features per head	rows read	AUC
anchored, all six scalars	6	the conclusion's	0.925
masses on premises, claim and mean row	3	the conclusion's	0.919
sink, entropy and maximum	3	the conclusion's	0.911
entropy and maximum, symmetric in the keys	2	the conclusion's	0.899
per-head spectra with that pair added	7	all, and the conclusion's	0.880
per-head spectra	5	all	0.703

Table 7: Parts of the anchored readout and the per-head spectra scored alone, on the same forward passes, the same grouped folds and the same five seeds, mean over 13 models. Every row is a per-head feature set over the same layers and heads, so the rows differ in the region of the attention matrix they are computed from and in how many scalars they keep, not in depth or width. The entropy and the maximum of a row are unchanged by relabeling the keys of that row, so the fourth row uses no key identity and the second uses the span labeling alone. Dropping the span labeling from the full readout costs 0.014 and the sink a further 0.012.

Identity used by the readout. A readout of the conclusion's row can use three separable kinds of token identity, and the group action of Lemma 1 destroys only one of them. Relabeling a fraction of the tokens can mean three different things. Trading membership of the claim span for premise positions changes *which tokens are called the claim*. Permuting the keys of every row changes *which token receives a given share of the mass*. Permuting queries and keys together, the action of the lemma, additionally changes *whose row sits at the readout position*. At full strength on the anchor model the first lowers the anchored readout by 0.020 on ProofWriter and the second by 0.008, both inside the interval of the unperturbed readout. The third takes it from 0.965 to 0.523, inside the interval of chance. The identity the readout depends on is the correspondence between a row and the token that owns it, exactly the structure the lemma says an invariant cannot represent, and not the labeling of the span or of the keys. Appendix K reports the binary form of the span control and a row-relocation control, which agree. A relabeled matrix $PAP^\top$ is generally not causal, and so is not itself the output of any forward pass. That is what the comparison is for: every summary in the failing set of Lemma 1 gives $PAP^\top$ and A the same value, so a detector built on one cannot tell them apart, and the curve measures how much discrimination rides on exactly the distinction such a detector cannot make.

Graded relabeling sweep. Sweeping the strength of the joint relabeling in five steps gives Figure 1(c). On ProofWriter every one of the 11 models falls, from 0.835 to 0.978 unperturbed, scored like Table 1 on the seed-averaged out-of-fold score from this experiment's own forward pass and so slightly above Table 2, down to 0.497 to 0.589 fully relabeled, and the endpoint lies inside the interval of chance in 10 of them. On FOLIO every model falls by 0.098 to 0.506, and with N smaller the endpoint scatters further: 5 inside the interval of chance, 3 below it and 3 above, between 0.310 and 0.596. The sweep has five steps whose intervals overlap between neighbors, so the fall is measured end to end rather than step by step: it is strictly monotone at every step in 7 of 11 models on ProofWriter and 3 on FOLIO. Position 0 is held fixed throughout, because the sink mass is one of the six features and BOS is dropped everywhere else in this paper, so each curve is an upper bound on what a fully relabeled readout retains.

The windowed readout. ConTRoL passages run an order of magnitude longer than the synthetic sets, which suggests that the premise mass the readout sums is spread too thin to separate. If that were so, summing the mass over the premise positions nearest the conclusion rather than over all of them would recover some of the gap. The measurement rules this out. Over windows of 16 to 128 positions the ConTRoL AUC moves by at most 0.013 on any of the 3 Qwen3 models, in both directions and inside the bootstrap interval, so the gap on real passages is not the dilution of premise mass over a long context.

Saturation. Fusing the bipartite block spectra with the anchored scalars moves the shared fold AUC by $\Delta = +0.000$, which is not significant, and likewise over the probe. The block spectrum is a lossy compression of information the anchored scalars already carry, which is what Table 7 measures for a summary read from the claim-to-premise block rather than from the conclusion's own row.

H Cross-benchmark transfer

If the probe were fitting benchmark-specific quirks, its direction should not travel. We take the mass-mean direction (Marks & Tegmark, 2024), the difference of class means of last-token states with nothing fitted beyond a train-side standardization, on ProofWriter at a mid-depth layer chosen by ProofWriter in-domain AUC alone, and apply it unchanged to FOLIO, a different generator and a different logic in expert-written text. It transfers in every model, at 0.717 to 0.902 zero shot (Table 2), far above FOLIO's lexical floor (0.547) and strongest in Qwen3.5-27B. The anchored readout transfers as a detector as well: fit on all of ProofWriter and applied to FOLIO unchanged, it reads 0.511 to 0.818 against an in-domain 0.674 to 0.866. As a specificity control (Figure 10, Appendix E), the 4B transfer read 0.590 with both benchmarks negation-coupled, 0.753 with the source balanced and 0.762 with both, so it improves once the shared artifact is gone.

I Benchmarks beyond ProofWriter and FOLIO

Four further deductive benchmarks were built with the protocol of Appendix B, so that the anchored readout is measured outside the two benchmarks it was developed on. Each is sampled to 240 items where the source allows, balanced on the label and on every surface marker the audit found coupled to the label, grouped so that items sharing a premise set never cross a fold, and audited with the same lexical floors, quoted as intervals.

FLD. Morishita et al. (2023) generate deductions from a different synthetic grammar than ProofWriter and ship a negative twin for most items, so the pair-exact control of Appendix C transfers. The audit finds a residual floor above chance that does not come from negation, because a truth frame tracks the label ("is wrong", "does not hold" on refuted items against "is true", "is right" on proved ones). Balancing that axis leaves a residual whose interval still clears chance, so FLD is quoted with its floor rather than called clean.

LogicNLI. Tian et al. (2021) pair each premise set with about twenty hypotheses, which gives the strongest group structure in the suite: a fold that sees one hypothesis of a premise set sees none of the others.

ConTRoL. Liu et al. (2021) is expert-written reasoning over real passages, news and examination text, and is the one open-world benchmark in the suite, with passages an order of magnitude longer than the synthetic sets. Section 4 reports what the anchored readout loses there.

ProntoQA. Saparov & He (2023) uses fictional ontologies, so world knowledge cannot stand in for deduction.

J PUBLISHED DETECTORS AS BASELINES

Every baseline is ported from its authors' released code and checked against that code numerically by a released script before its number is reported.

LLM-Check. Sriramanan et al. (2024) release two single-pass scores. The attention score is the mean log diagonal of each head's attention map over the span being judged. A causal map is triangular, so this is its mean log eigenvalue. Their implementation sums the per-head values. We keep the per-head vector and give it to the same grouped logistic regression as every other readout, which can only help the baseline. The hidden score transposes the span's hidden states before centering, so each token is centered across features and the score is the mean log singular value of the $N \times N$ token Gram, not of the $d \times d$ covariance the description suggests. Both agree with the reference functions to 10^{-6}.

LapEigvals. Binkowski et al. (2025) build a Laplacian over a multi-layer attention graph in which layer ℓ is joined to $\ell + 1$ through the self loops, and the feature is the top ten entries of its diagonal, per layer and head. The diagonal needs only each layer's column sums and self loops, which the extraction retains, and matches the reference to 10^{-7}.

Contrast consistent search. Burns et al. (2023) is run with the hyperparameters of the authors' notebook rather than the paper text: one thousand epochs, ten random restarts keeping the lowest-loss probe, learning rate 10^{-3}, weight decay 0.01, mean centering without variance normalization, a single linear layer, fitted per grouped fold. The objective is symmetric under swapping the two poles, so the sign is resolved with the labels, which is the favorable choice for the baseline.

Behavioral and probe baselines. $P(\text{True})$ is the likelihood difference between *Answer: True* and *Answer: False* appended to the item. Perplexity and mean entropy are computed over the claim span. SAPLMA (Azaria & Mitchell, 2023) is a three-layer perceptron on the states the linear probe reads.

Measured invariance of each readout. Membership of the class Lemma 1 constrains is measured, not assigned: the released script applies a random joint relabeling $A \mapsto PAP^\top$ and reports the change. The whole sequence LLM-Check attention score is exactly invariant (change 10^{-15}) and reads chance. The per-head spectra are invariant up to the excluded sink token: on the full matrix the change is 10^{-15}, and as run, with row and column zero dropped, a general relabeling moves a different token into that slot (change 2×10^{-4}). LapEigvals is not invariant, because its diagonal divides by a position-dependent out-edge count, and neither is the claim span LLM-Check score, because restricting to a span is a use of identity, and both score well. The table in Section 4 labels each readout by this measurement.

K THE CLAIM RELABELING CONTROL

Which identity the anchored readout depends on is tested by keeping every attention value and the feature count and destroying one kind of information at a time, from the same forward pass. Under *span relabeling* the readout row stays the final token, but the tokens called "claim" are a random window of the same length inside the premises and "premise" is the rest. Two of the six scalars (entropy and maximum of the final row) do not depend on the labeling and are unchanged by construction. Under *row relocation* the readout row moves to the last token of that window and premise is what precedes it, so span identity and readout position are destroyed together and the

relocated row also sees less context. On the anchor model and ProofWriter, relabeling costs 0.019 and relocation 0.445, and on FOLIO 0.009 and 0.250. The readout depends on the conclusion's own row, and only weakly on the partition of its context.

L READOUT CONSTRUCTION AND PROMPT STRINGS

Every item is one string. ProofWriter items read *Facts:* … *Statement:* …, FOLIO items *Premises:* … *Conclusion:* …, and the four further benchmarks use the marker their appendix paragraph names. The premise span is every token after the first (the sink) and before the marker, the claim span is the marker and everything after it, and the readout row is the final token. Spans are located by tokenizing the text up to and including the marker, so they are exact for every tokenizer. The behavioral readout appends *Answer: True* and *Answer: False* to the item and takes the difference of the two negative log likelihoods. Nothing is generated. The hypothesis-only control runs the claim span alone, with no facts and no marker.

M CHECKPOINTS

All models are public instruction-tuned checkpoints run in bfloat16 with the eager attention implementation, which materializes the attention probabilities the readouts need: Qwen/Qwen3-0.6B, Qwen/Qwen3-1.7B, Qwen/Qwen3-4B-Instruct-2507, Qwen/Qwen3-8B, Qwen/Qwen3.5-0.8B, Qwen/Qwen3.5-2B, Qwen/Qwen3.5-4B, Qwen/Qwen3.5-27B, Qwen/Qwen3.8-27B, google/gemma-3-1b-it, google/gemma-4-31b-it, HuggingFaceTB/SmolLM3-3B, allenai/OLMo-2-0425-1B-Instruct and microsoft/Phi-4-mini-instruct. The Qwen3.5 and Qwen3.8 stacks carry softmax attention in one layer in four (six of twenty-four layers up to 2B, eight of thirty-two at 4B, sixteen of sixty-four at 27B) and linear attention elsewhere, and only the softmax layers are read. Qwen3.8-27B is released as a vision-language model and is run on text alone. Gemma-4-31B alternates sliding-window and global attention layers, and all sixty are read. The extraction stores the model layer of every attentive index, and every layer and head index printed in this paper is that model layer. No model is quantized and no chat template is applied.

N ONE ITEM

A depth-one ProofWriter item, label refutable: *Facts: The cat does not need the tiger. The dog likes the tiger. The squirrel visits the cat. The tiger does not visit the cat. If someone likes the tiger then they do not like the squirrel. If someone is young and red then they need the dog. If the tiger is nice then the tiger needs the cat. Statement: The dog likes the squirrel.* The second fact and the first rule refute the statement in one step. The sink is token zero. The premise span runs from token one to the token before *Statement:*, the claim span is *Statement:* and the statement, and the six anchored scalars of a head are the final token's attention mass onto the premise span, onto the claim span and onto the sink, the entropy and maximum of that row, and the premise mass of the mean claim row. Items are grouped by their facts, so any other statement over the same facts sits in the same fold.

O IMPLICATIONS FOR PRACTICE

For anyone building a validity or hallucination detector. Read the conclusion's own attention row and keep the identity of what it attends to. A summary invariant to relabeling the tokens cannot represent which token attends to which, whatever tensor it reads. Fitted on the same labels as a probe but from attention alone, the six anchored scalars per head read no hidden state and recover at least 89% of the probe on ProofWriter in every one of the thirteen models where the readout is extracted. Where hidden states can be read, the probe still leads, and it needs fewer labels (Figure 3). **For anyone constructing a deduction benchmark.** Balance every surface marker the label couples to, not only length, quote the lexical floor as an interval, and ship each item's negative twin so the pair-exact control transfers. **For interpretability work on attention.** Head averaging discards the contrast between heads and permutation invariance discards which token a head attends to. Test which loss a summary suffers by applying the group action and measuring the change. **For deploying a check on model output.** The readout needs the conclusion's position and little else

about the prompt, and on long open-world passages the probe's lead is the second widest of the six benchmarks. If the inputs are of that kind, measure the probe against the anchored readout on a held-out set before choosing.