TOKEN-LEVEL EVIDENTIAL SUPPORT IS LINEARLY DECODABLE FROM THE RESIDUAL STREAM

Anonymous authors
Paper under double-blind review

ABSTRACT

We replace hallucination detection with a relaxation that is decidable by construction: for each token a model generates, was the evidence for it present in the context, and does the model's internal state reflect that? Each item is a two-part question about a 10-K filing, with the supporting sentence for one part deleted and replaced by a length-matched sentence from another filing, alternating which part is affected. Because we delete it ourselves the label is exact rather than annotated, and because both parts are answered in one generation the two classes share prompt, document, question and output. A linear probe on mid-network activations ranks the supported part's tokens above the withheld part's at 0.862 AUC, bootstrap interval $[0.841, 0.882]$ over 493 answers, against a bar of 0.70 fixed before the data was collected.

The probe also localizes, which a single score for a whole response cannot do. Thresholding its per-token score recovers the supported part's tokens at a token-level intersection over union of 0.545, against 0.461 for next-token entropy read at the same position. At a matched scoring unit the probe leads entropy by 0.164 of AUC per token and 0.088 per part.

Three pre-declared controls come out at chance, as does a randomly initialized placebo, and two positive controls come out high, which is what makes the battery falsifiable. One leak is real: a bag-of-words classifier on the answer text alone separates the conditions across answers at 71.8%, so the answer-level gate is benchmarked against that rather than against chance. Within one answer the contrast is not recoverable from which tokens were generated, and that is the control the main result rests on.

The probe reads support and not accuracy: answers whose evidence was present but whose value was wrong are indistinguishable from correctly answered supported ones, while entropy places those same answers at its own midpoint. This matters most on long documents, where the model stops flagging the problem itself. Padding the document from 300 to $3{,}200$ tokens with the question and the removed evidence held fixed, the rate at which the model declines to answer falls from 57.3% to 9.3%. The depth at which the signal appears does not move with length, on a grid of four probed layers, refuting a hypothesis we had registered.

What the probe provides is a ranking within an answer, not a calibrated absolute score, and a pre-registered test of locality failed. Every result here comes from a single 4.2B-parameter model, with the headline replicated on one other.

1 INTRODUCTION

We replace *hallucination detection* with a relaxation that is decidable by construction, and show that a linear readout of the residual stream recovers it token by token. Hallucination labels track the output rather than the process. A statement can be marked a hallucination when the model faithfully repeated something false from its context, and a correct statement can be one the model reached by hallucinating gracefully. We ask instead: for each token the model generates, was the evidence for it present in the context, and do the activations reflect that? We build two-part questions in which the supporting sentence for one part stays in the document and the sentence for the other is removed, so the label is a property of the input rather than a judgment about the output (Figure 1).

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.

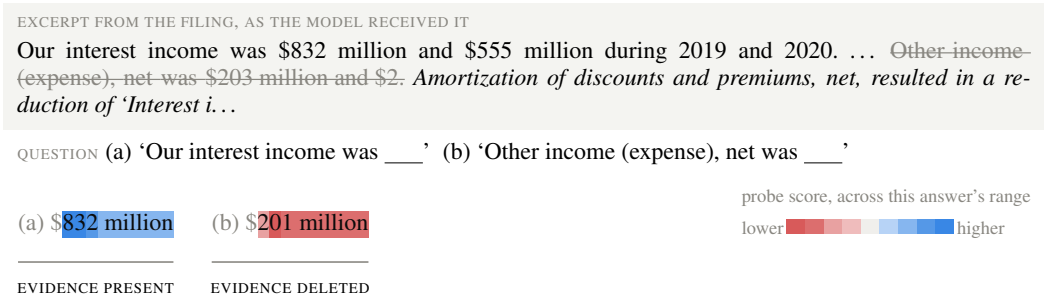


Figure 1: One generated answer to a two-part question about AMZN's 2020 10-K (item 29, a 962-token prompt shown here in excerpt). The supporting sentence for part (b) was removed and replaced by a passage of equal length from another filing; the original is struck through and its *replacement* follows it. Scored tokens are shaded on a diverging scale spanning this answer's score range, with the neutral step at its midpoint; format and label tokens are not scored and are left unshaded. The scale is within-answer, so shades are not comparable across items. The model answers (a) correctly at $832 million and answers (b) with $201 million against a gold of $203 million. This item separates its two parts perfectly (1.000), as 51% of the 493 items do, against a mean of 0.862.

The closed-domain framing that makes such a question askable is not ours. It appears in Ridder et al. (2026) as a distinction between contextual and parametric knowledge, and in Yeh et al. (2026) as a decomposition of generation into external-context and internal-knowledge utilization. Our design differs in three ways. The label is exact by construction rather than annotated: Obeso et al. (2025) annotates with a language model and discards annotations it cannot map back to the completion, and Ridder et al. (2026) annotates with Gemini, validated against one human annotator on 100 samples. We separate support from correctness: a label defined as whether a token is hallucinated fuses the two, so a corpus built that way cannot test which of them a detector reads. And both of our classes come from a single answer, sharing prompt, document, question and output, whereas those comparisons are made across separate responses.

The probe is fit on answers whose document retained both supporting sentences and on answers that retained neither, then evaluated on answers where exactly one was removed, ranking tokens inside each single answer. Activations come from the residual stream at layer 16 of 32 in Minitron-4B-Base (Muralidharan et al., 2024). The probe reaches 0.862 AUC with a bootstrap interval of [0.841, 0.882] over 493 answers, clearing a bar of 0.70 fixed before any data was collected. The per-token score also points at the supporting sentence, more closely than next-token entropy does, and a system built on the measure needs that localization. What the probe provides is a ranking within one answer; it is not a calibrated absolute score, and a pre-registered test of locality failed.

Three pre-declared controls and a randomly initialized placebo come out at chance (Section 4.2). One check does not: a bag-of-words classifier on the answer text alone separates the conditions across answers, so the answer-level gate is benchmarked against that classifier rather than against chance (Appendix C). Two positive controls come out high, which is what makes the battery falsifiable. Within one answer the contrast is not recoverable from which tokens were generated, and that is the control the within-answer result depends on.

Answers that were wrong but whose supporting evidence was present still score as supported, so the probe recovers whether evidence was available rather than whether the answer was right.

Section 2 places the work against uncertainty-based detectors, the probing literature, and corpora that deliberately remove supporting evidence. Section 3 gives the corpus and the probe. Section 4 reports the result and then bounds it. Section 5 states the limitations, calibration first. Four appendices follow: the definitions (Appendix A), the pre-registration record and the findings we withdrew (Appendix B), the corpus and its reconciliations (Appendix C), and the depth and geometry results (Appendix D).

2 Related work

Our relaxation stands in for the hallucination detection problem, whose established solutions read the model's own uncertainty and supply our first baseline. The line spans length-normalized entropy and perplexity (Malinin & Gales, 2020), semantic entropy (Farquhar et al., 2024) and probes that recover it in one forward pass (Kossen et al., 2024), sampling-based agreement (Bouamor et al., 2023; Chen et al., 2024), attention-weighted variants (Zhang et al., 2023), and elicited confidence (Kadavath et al., 2022). None of them is defined over a single token: the unit is a response or one of its sentences. A token-level comparison therefore rests on a choice about aggregation; we state ours and report at both units, per token and per part (Section 4.3).

A second line supplies our method rather than a bar to clear. Truthfulness is linearly decodable: from a single forward pass a linear readout recovers whether a statement is true (Burns et al., 2022; Marks & Tegmark, 2024; Li et al., 2023). A small nonlinear readout does the same on topics held out of its training (Azaria & Mitchell, 2023). Dubanowska et al. (2025) find that once a dataset confound is controlled, more elaborate detectors fall back to the level of a supervised linear probe, so the simplicity of the readout is not a concession. Two cases matter for our design. Sky et al. (2024) is the closest prior work to a streaming token-level claim, and it reads the same thing we do: probes on internal states at generated-token positions, predicting whether the generated content is grounded in the context the model was given. Its label is the difference. Three annotators mark the minimal spans that would have to change to make a response grounded, reconciled to individual tokens by majority, so what the probe is fit against is a judgment about the output. Obeso et al. (2025) trains per token and scores per entity by taking the maximum over the entity's tokens, an aggregation that is neither response-level nor per-token. What this line lacks is a construction in which the evidence is the only thing that moves: a statement can be true whether or not its evidence was in the context.

A third line reads how the model uses its context rather than how confident it is. The Lookback Lens trains a linear classifier on per-head ratios of attention paid to the context against attention paid to the generation (Chuang et al., 2024). Two recent systems quantify context use itself: Yeh et al. (2026) scores a token by how far the model's predictions move when the retrieved documents are swapped for random ones, and Ridder et al. (2026) attaches a detection head to an intermediate hidden state and fine-tunes the model to report on itself. Our instrument belongs to this line, and the Lookback Lens is our second baseline. Scored per token on the same within-answer task, fit on the same answers and transferred alike, the probe leads it by a margin whose interval excludes zero (Section 4.3). Both recent systems also score per token, so the unit is not what separates us from them: one needs a second forward pass over substituted documents and the other needs the model retrained, where ours reads a frozen model once, and both are scored against annotated hallucination labels. Sahoo et al. (2026) show that a linear probe on hidden states can separate task format rather than the construct it was fit for, and recommend reporting source-prediction accuracy alongside any probe result, which we do (Section 4.2).

3 Setup

We replace the question of whether a generated statement is true with the question of whether the evidence for it was present in the context the model was given. This changes the label's status rather than its difficulty. Truth requires a fact about the world to check against; evidence presence is a property of an input by construction, so it is known exactly for every token with no annotation pass and no judge model. The price is scope, and Section 5 states what the narrower question leaves out. What we get back is a corpus in which one thing varies and everything else is held fixed, outlined in Figure 2 and specified below.

Items are two-part questions over real annual reports, each part asking for a numeric value that appears verbatim in one sentence of the document, so a part's evidence is a single identifiable sentence rather than a collection. Throughout, a part is one half of a question as the model sees it, labeled (a) or (b) by the order the two are rendered in; a span is the tokens generated in answer to a part; and a sentence is what was removed from the document. A part is not the same as a slot: each item fixes two slots, the sub-questions themselves, before anything is rendered, and which of them is rendered first alternates by item. The build discarded 5,225 candidate values to keep the 1,112 that make up

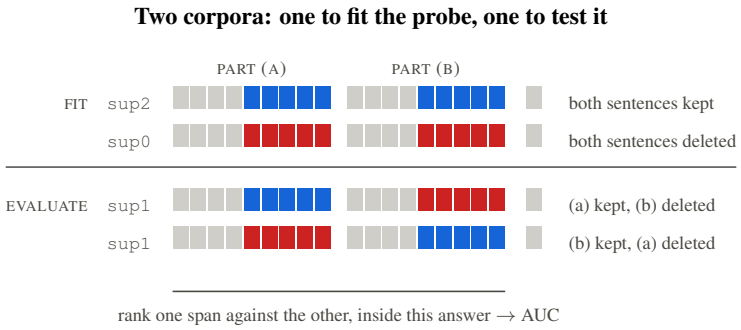


Figure 2: All three conditions share the document and the two-part question; only the supporting sentences change, and a deleted one is replaced by a donor matched to within one character , so the document is always present in full. Color marks whether a span's evidence survived, which Figure 1 shades the probe's estimate of. Grey cells are generated but not scored, 55 token ids in all: four label and currency tokens lead each part, and one closes the answer and belongs to neither. Fitting where a whole answer is one class and evaluating where one holds both is what makes the comparison within-answer: the ranked spans share prompt, document, question and decoder pass, and part order alternates exactly so position cannot predict which part kept its evidence. Fit: 1,112 answers, 12,352 tokens, `GroupKFold(5)` by item. Evaluate: 536 answers, 6,269 tokens, 3,195 kept against 3,074 deleted, 493 scorable. The cells are schematic.

556 items across 65 issuers. A value had to carry at least three digits, and no item pairs two equal values, since a model answering one part would then answer the other correctly by accident.

For one part of each question the sentence carrying its value is replaced by a donor from another filing, matched on character length rather than token count. That yields 1,668 generations, three per item: one with both parts supported, one with neither, and one with exactly one (`sup2`, `sup0` and `sup1`), each capped at 48 new tokens. Scoring excludes format and label tokens, leaving a median of 9 scored tokens in answers that run 13 to 47 generated tokens, so no answer reached the cap. Of the 556 `sup1` answers, 536 retain at least one scored token, and within-answer AUC needs both classes present among the kept tokens, so 493 are scorable (Appendix C).

The model is `nvidia/Minitron-4B-Base` (Muralidharan et al., 2024), a 4.2B-parameter base model with 32 layers, a hidden width of 3,072 and a 4,096-token context window, so the corpus's 307 to 1,439-token prompts are nowhere near truncation. We decode greedily and read the residual stream after each block. Results are reported at layer 16, the first layer reaching 95% of the peak's excess over chance; the peak itself sits deeper (Section 4.5). Appendix D carries the depth curve. Greedy decoding on this hardware is not bit-reproducible, so it also reports how often a re-run changes the generated tokens and how far the headline moves when it does.

The probe is an L2-penalized logistic regression at $C = 1.0$, fit under five-fold cross-validation grouped by item so no item contributes to both a fit and its own evaluation. A token's score is its signed distance to the decision boundary. The five folds are formed over the fitting rows alone, grouped by item, and balanced so that each fold holds an equal number of rows rather than an equal number of items (Appendix A). That precision is not decoration: seven readings of five-fold grouping by item, each of them defensible, span 0.0165 of AUC on identical data (Appendix C). Inside one `sup1` answer we rank the kept tokens of the supported span against those of the withheld span, giving an AUC computed within a single answer, which we average across answers. There is one `sup1` answer per item, so averaging across answers and resampling items are the same unit. Intervals are bootstrapped over that unit rather than over rows, since tokens inside an answer are not independent; 2,000 draws unless a figure states otherwise.

4 Results

Our headline result is that a linear readout at layer 16 separates supported from withheld tokens inside a single answer at 0.862 AUC over 493 answers (Figure 3). The rest of this section bounds it. Section 4.1 reports where in an answer the score falls, Section 4.2 the nulls and positive controls,

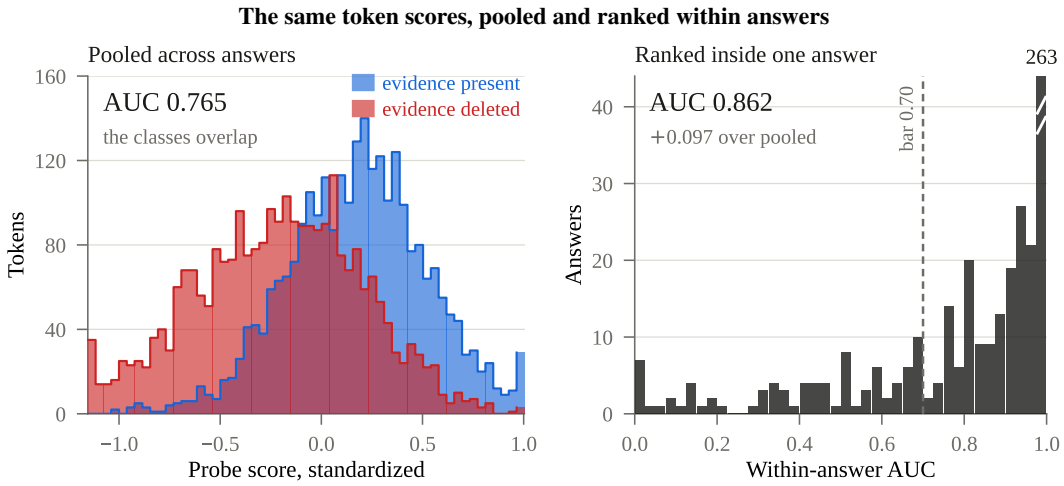


Figure 3: Both panels describe the same 493 answers, the ones carrying both support classes among their scored tokens: 5,270 tokens, 2,635 with evidence present and 2,635 with evidence deleted. Left: every one of those tokens pooled across answers and ranked together, AUC 0.765; scores are drawn over the central 99% with the tails collected into the end bins. Right: the same scores ranked inside each answer, one AUC per answer, AUC 0.862 with a bootstrap interval over answers of [0.841, 0.882], against the pre-declared bar of 0.70 that 83% of answers clear. The rightmost bar is clipped by the axis and holds 263 answers, 253 of them at exactly 1.000.

	probe	next-token entropy
token IoU	0.545	0.461
token precision	0.772	—
token recall	0.696	—

Table 1: Recovery of the supported span. Token IoU: tokens both selected and in the span, over tokens either selected or in the span. It is a set overlap, not an AUC, so 0.5 is not a chance value (Appendix A). Token precision: of the tokens selected, the fraction in the span. Token recall: of the tokens in the span, the fraction selected. The threshold is set per fold so that the number of predicted tokens matches the number of gold tokens on the training split, and is not tuned (Appendix B). IoU is paired over the 488 answers both instruments score; precision and recall are the probe's alone, on 446 and 493 answers, precision being undefined on the 47 answers where the threshold selected no tokens.

Section 4.3 the dissociation from correctness and confidence, Section 4.4 what does and does not change at length, and Section 4.5 where in the network the signal sits.

4.1 The score localizes to the supported span

Figure 3 gives the result and its interval. The probe's scores are more comparable within an answer than across them: pooled across those same answers rather than ranked inside each, they fall to 0.765.

The construction was built to ask where the score falls inside an answer. Thresholding it recovers more of the supported span than next-token entropy does, the entropy of the model's next-token distribution read at the same position (Table 1, Figure 1). Both spans come from one generation, so the comparison is local rather than a difference between answers.

Two caveats. Short answers give coarse scores, which is why so many land at exactly 1.000; the median answer has only 9 scored tokens (Appendix C). Longer answers score lower, by about 0.08 of AUC per ten generated tokens. That belongs to the measure rather than to the ceiling: restricting to answers with more scored tokens makes the slope steeper rather than flatter, reaching 0.097 on the 118 answers with at least six scored tokens per class, which a discreteness artifact would not do.

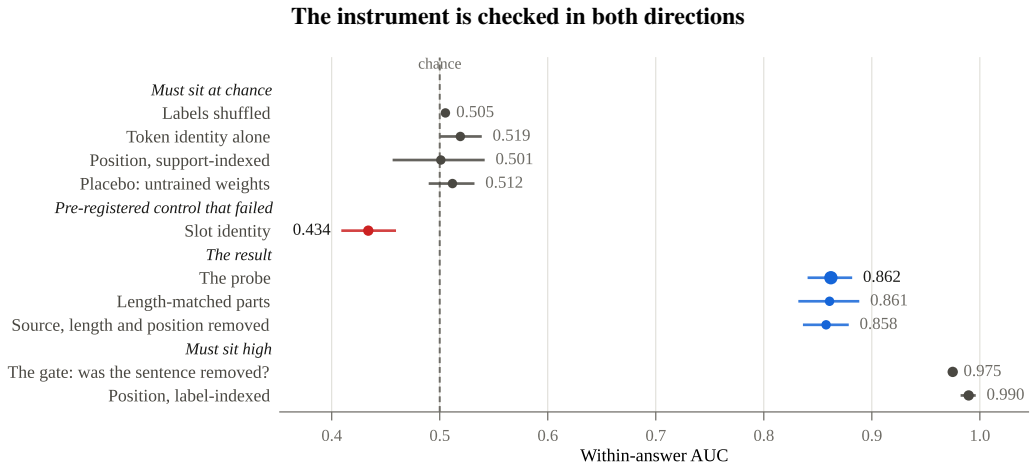


Figure 4: Every control on the within-answer AUC axis, grouped by what passing looks like. The two position rows are a pair: support-indexed asks a position-only probe which part kept its evidence and must fail, label-indexed asks which part is labeled (a) and must succeed, since position predicts the label by construction. Slot identity is the support probe asked to separate an item's two semantic slots on held-out `sup2` answers, where both slots are supported, against a pre-declared target of 0.5. Length-matched parts restricts the result to answers whose two parts hold the same number of scored tokens; the row below it residualizes source, length and position out of the activations. The gate is answer-level rather than within-answer, over 1,112 answers, against a bag-of-words AUC of 0.781 on the answer text alone. Intervals are item-level bootstraps over 493 answers except the placebo (151), slot identity and label-indexed position (511), and length-matched parts (257). Two rows carry no interval on disk and are drawn as points.

4.2 Nothing else in the construction accounts for the result

Figure 4 reports the battery: three pre-declared nulls and a randomly initialized placebo that must sit at chance, two positive controls that must sit high, and the result under two restrictions. The positive controls are the point, since a set of checks that can only return all-clear cannot fail. One of them is the answer-level gate, which asks of a whole answer whether its document kept the supporting sentences or kept neither. A third positive check sits outside the figure, since it is a paired score difference rather than an AUC: restoring a withheld slot's evidence raises that slot's score in the declared direction, with an interval excluding zero (Appendix B).

One pre-registered control failed. Which of an item's two slots a span answers is decodable from the same activations, which we had declared should not be the case. The separation runs below chance rather than above, which is equally a failure: only its distance from 0.5 is informative here, not its direction. It does not appear to carry the result: projecting out the direction that encodes it moves the headline indistinguishably from a random direction of the same norm, on a diagnostic with no pre-declared margin (Appendix B).

A related check comes from Sahoo et al. (2026), who show that probes on hidden states can separate task format rather than the construct they were fit for, and recommend reporting how well the source document can be predicted from the same activations. Ours can be identified 2.8% of the time against a chance rate of 0.3% across 296 documents. That cannot separate our two classes, because both spans come from one generation and so from one document. Appendix B reports their staged protocol in full, and names the arm of ours most exposed to it. The battery answers the same concern from the other side: residualizing source, length and position out of the activations moves the headline within its own interval, and the two intervals overlap heavily.

4.3 The probe reads support, not correctness or confidence

Answers that had their evidence and still got the value wrong are indistinguishable, under the probe, from correctly answered supported ones: their intervals overlap, and both sit far from the withheld

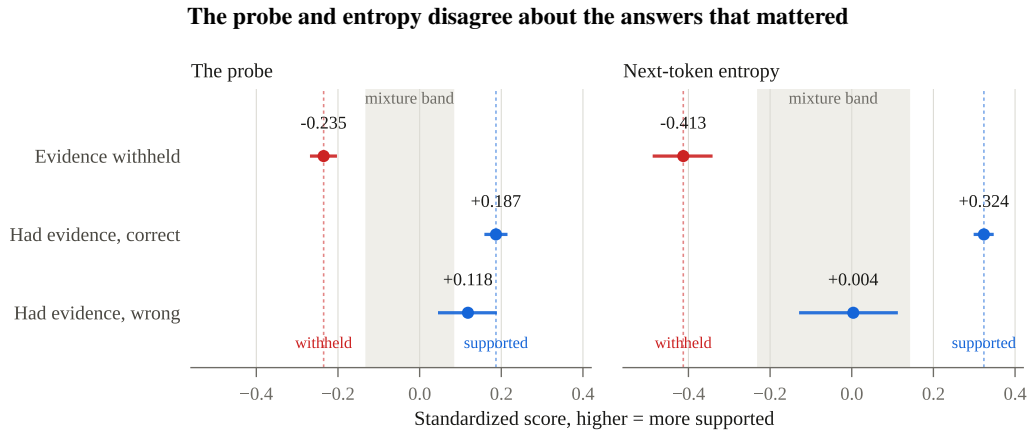


Figure 5: Mean score per span for three classes of answer under each instrument, with 10,000-draw bootstrap intervals over spans. Evidence withheld: the span whose supporting sentence was replaced. Had evidence, correct: the span whose sentence survived and whose value the model got right. Had evidence, wrong: the span whose sentence survived and whose value the model got wrong. The mixture band is the middle half of the gap between each instrument's own two endpoint classes. Entropy is sign-flipped so that higher means more supported in both panels, and the panels share an x axis. The three classes hold 492, 423 and 70 spans.

	scored per token	scored per part
probe	0.865	0.904
next-token entropy	0.701	0.816
difference	+0.164	+0.088

Table 2: Within-answer AUC for both instruments at both scoring units, on one identical set of 488 answers (Appendix C). Scored per token ranks individual generated tokens; scored per part averages each span's kept rows first, then ranks the two spans, so a per-part AUC takes one of three values per answer and its mean is the fraction of answers whose supported span outranks its withheld one. Both differences have intervals excluding zero. Comparing across the diagonal instead, a per-token probe against per-part entropy, mismatches the units and gives +0.023 with an interval that includes zero.

class (Figure 5). Entropy puts those same answers at its own midpoint. A pre-registered test of that disagreement gives a difference in placement of 0.272, interval [0.036, 0.552] (Appendix A). It rests on 70 spans, which bounds the claim (Appendix B).

Comparing against entropy requires a choice about aggregation. Entropy is defined over a response and the probe over a token, so neither can be compared to the other until a scoring unit is fixed for both. Table 2 fixes it both ways. At either matched unit the probe leads with an interval excluding zero. One unmatched comparison is worth naming: the probe per token against entropy per part, whose interval includes zero. We report it because it is the comparison a reader makes by default, since entropy is normally reported per response, not because it is a result. The per-part unit is coarse by construction, which is why its interval is the wider of the two rather than the effect being weaker there.

Our second baseline admits the same comparison at one unit. The Lookback Lens, scored per token on these same 488 answers and transferred as the probe is, reaches 0.788, which the probe leads by a paired difference of +0.077 whose interval excludes zero. Its per-part figure does not exist on this population under the same protocol, so the table does not carry it.

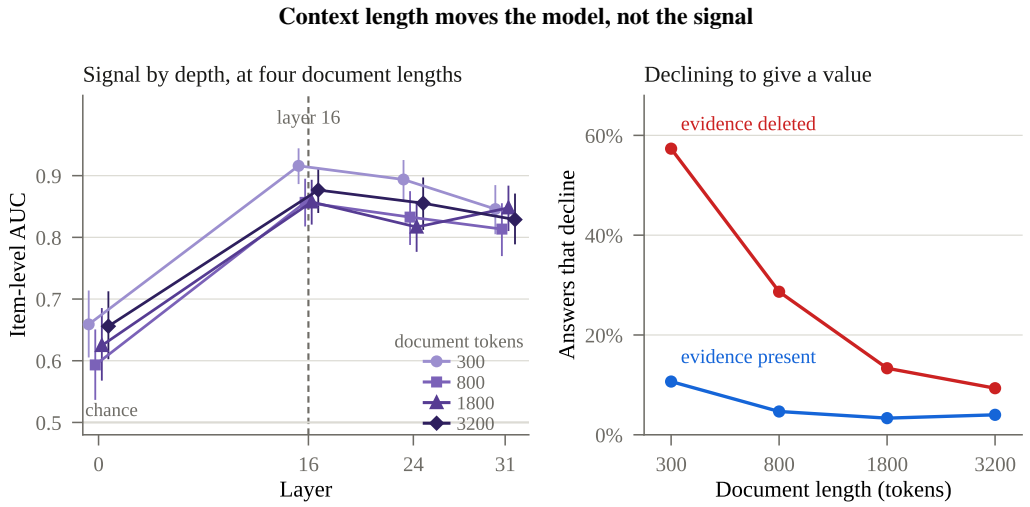


Figure 6: Both panels use a padded corpus, 150 items paired across all four lengths, with 300 tokens being the median unpadded document. Left: the probe's AUC by layer at each document length, with 2,000-draw bootstrap intervals paired across lengths. The label here is item-level, evidence present against substituted, rather than the within-answer label used elsewhere in this paper, so these values are not comparable to the headline. Argmax and the first layer reaching 95% of the peak's excess over chance are both layer 16 at every length, with intervals of [16, 16] at 300 and at 3,200 tokens. Only four layers were probed here, so each argmax is resolved to the nearest probed layer; the denser sweep of Section 4.5 puts the peak at layer 24. Right: the share of answers containing a phrase declining to give a value, on the same items, 150 per point, plotted separately for answers whose evidence was deleted and for the supported control.

4.4 Document length moves the model, not the signal

Our primary corpus tops out at 1,439 prompt tokens, so length is tested on a separate padded corpus (Figure 6), where filler is added to the document and the prompt grows with it. Its label is item-level, evidence present against substituted (Appendix A), so its AUCs are not comparable to the within-answer headline. We pre-registered that the signal should appear deeper as the document grows: a probe reading whether a supporting sentence exists anywhere cannot settle that until the document has been integrated. It is refuted. Across an eleven-fold change in document length the depth at which the signal appears does not move on the four layers probed there, and its strength stays flat within interval at every length.

The model's behavior does move. Answers that decline to give a value fall from 57.3% to 9.3% as the document grows, while the supported control stays low throughout (Appendix B). A detector that runs during generation is most useful where the model has stopped hedging on its own.

4.5 The signal is mid-network

Depth is swept over 17 layers on a second, independently collected set of activations over the same items, with the within-answer label rather than the item-level one of Section 4.4. Decodability rises into the middle of the network and falls before the output. The layer of highest AUC is 24, but its bootstrap interval spans nine layers, so the peak is not resolvable; the first layer reaching 95% of the peak's excess over chance is 16, with an interval of [16, 20], and that is the statistic we report. It lands on 16 under either of the two ways that crossing can be computed, and it is selected on this second collection rather than on the answers the headline is measured over. The curve, both computations, and where in activation space the signal sits are in Appendix D, which also sets our crossing against the closest prior report on depth.

5 LIMITATIONS

The largest limitation is one we tested for and failed. We pre-registered the question of whether the score depends only on the evidence for the span being scored, and both arms failed. Holding a span's own evidence present and adding the other part's raises that span's score, at 0.938 within-answer AUC against a declared target of 0.5; holding its own evidence absent and adding the other part's raises it as well, at 0.899. The score therefore carries an offset from the global evidence state of the prompt, so a single threshold will not transfer across prompts of differing evidence density. That rules out a calibrated absolute score and leaves the within-answer ranking as the claim. One caveat bounds the size of the failure without changing its direction: both arms compare conditions, so the generated text differs between them and the effect conflates the other part's evidence state with different generated content. The full record, including a statistic we specified badly and retracted, is in Appendix B.

We have one corpus, so we cannot run the train-on-one-evaluate-on-another test that this class of detector is reported to fail (Dubanowska et al., 2025). A second model family reaches 0.852 on the same construction, with an interval overlapping the headline's, but that is a replication and not a transfer test. Both models are small, and we place no depth fraction on the second, which was not probed densely enough to support one (Appendix D).

The corpus is narrow in a way that bounds every result measured on it. Every gold answer is an integer of at least three digits with no decimal point, drawn from a 10-K filing dated 2014 to 2024 (Appendix C). Our provenance does not record the model's training cutoff, so we cannot rule out that an answer we deleted from the context remains available from the weights. The headline is also measured only on the answers the within-answer statistic can score, and pooled over the ones it cannot the same score sits close to chance (Appendix C).

Two defects in the corpus builder are named here and not fixed. The substitution always targets the same slot of the two, though which part of the question carries it is randomized. That slot's gold values are slightly shorter and smaller on several measures, so the two classes are not exchangeable by construction (Appendix C). Matching the two slots on length and magnitude, and alternating which of them is deleted, would fix both.

Two further limitations fall on the readout rather than the corpus. The fitted direction is sufficient but not unique (Appendix D). And one of the nulls does not reproduce: a check on whether the probe reads which position carried the supported slot gives different values under its stored protocol and under a later reconstruction of it, so that confound is recorded as unexcluded (Appendix B).

6 CONCLUSION

Replacing an annotated hallucination label with a relaxation we construct makes the question measurable. Because we delete the supporting sentence ourselves, the label is a property of the input rather than a judgment about the output, and both classes can be placed inside a single generated answer. On that construction, evidential support is linearly decodable per token, and at matched units it is distinct from output uncertainty rather than a restatement of it. The per-token score also localizes, which a single score for a whole response cannot do. What the construction needs is only that the support for a claim arrive as a unit we can withhold, which a retrieved passage or a tool's return also does.

What we do not have is calibration: the score ranks spans within one answer and carries an offset from the evidence state of the whole prompt, so the open problem is not whether support is readable but whether it can be read on an absolute scale. That bound is a control we pre-registered and failed, not a caveat we added afterward, and four further findings of our own were withdrawn on the same record (Appendix B).

STATEMENT ON THE USE OF GENERATIVE AI

Generative AI tools were used substantially in this work, and this statement is specific about where, because a brief description would understate it. The division of labor throughout was that the authors

set the design and made the decisions, and AI tools carried out much of the implementation and checking under that direction.

Research design and pre-registration. The relaxation this paper substitutes for hallucination detection, the two-part construction that makes it decidable, the pre-registered implications VF1 through VF5, and the threshold of 0.70 fixed before data collection were specified by the authors. LLM-based assistants were used to iterate on their wording and to argue against them before they were registered. The decisions that determine what this paper measures are the authors'.

Implementation. Most of the code was written with assistance from an LLM-based coding agent operating on the repository: the corpus builder, the activation capture, the probe, and the analysis scripts. Every load-bearing path was then checked by the authors, by reading it and against unit tests the authors wrote. The construction's label does not depend on a model's judgment at any point. A supporting sentence is deleted and replaced by a real sentence taken from another filing, so the label is a property of the input; no part of the corpus text is model-generated, and no annotation model was used. The role of AI here was to implement a specification, not to produce or to label data.

Analysis and verification. Much of the numerical verification reported in Appendices B and C was carried out with AI assistance, including the reconciliation of the pooled and within-answer figures over one population, the enumeration of the seven fold partitions and their spread, the counts behind the n chain, and the audit that found the mismatched row sets behind Table 2. Several corrections the paper now reports as its own were produced that way. The authors independently verified the corpus's reported properties with their own exploratory analysis.

Writing, figures and literature. Drafting and editing were iterative between the authors and LLM-based assistants across every section, and the authors edited the whole paper by hand. Figures were designed with AI assistance and their LaTeX produced the same way. AI assistance was used for literature analysis, including checking this paper's characterizations of cited work against the sources, which produced two corrections to Section 2. The bibliography was compiled and checked by the authors against each publisher's record.

What generative AI was not used for. No synthetic data was generated for any purpose. No theoretical model, mathematical claim, or proof was developed with AI assistance; the paper makes none beyond the definitions in Appendix A. No translation was involved.

Responsibility. The authors are responsible for the contents of this submission. Because AI tools were used in the analysis, the points at which a fabricated number or a confidently wrong claim would have been load-bearing were treated as the ones to verify by hand, and each was checked against the stored artifacts. The pre-registration record, the retractions, and the reconciliations in Appendix B and Appendix C are the audit trail for that checking rather than a separate exercise.

REPRODUCIBILITY STATEMENT

We do not release code, activations, or the corpus. The corpus is built from public annual reports by way of a third-party curated collection that we do not name here, because naming it would compromise anonymity; we will name it in the camera-ready. The activations are internal states of a released base model and we are not distributing them. What we do claim is that the experiment can be rebuilt from this document: Section 3 specifies the corpus, the model, the layer, the probe and the fitting discipline; Appendix A gives every statistic in closed form, including the ones that are easy to compute a different way; Appendix C gives the construction, the filter chain from answers built to answers scored, the two builder defects we did not fix, and the exclusions and their direction of bias; and Appendix D identifies which collection each depth and geometry result comes from, since they are not all the same collection.

Two things in Appendix C are addressed to a reimplementer rather than to a reader. The Reproduction subsection reports what a specification has to say to be reproducible: seven readings of "five-fold cross-validation grouped by item," each of them defensible, span 0.0165 of AUC on byte-identical data, which is wider than several differences this paper treats as meaningful elsewhere, so the fold partition is stated exactly rather than by name. The same subsection separates two pooled AUCs that are computed over different populations, and names a stored field whose obvious reading returns a plausible number for the wrong quantity. Appendix B carries the pre-registration record, the three

implications that failed, the statistic we retired, and the four findings we withdrew under our own checks, which is the material a replication would otherwise have to rediscover. One provenance gap is stated where it belongs and repeated here: no manifest in this project records the library version that produced the original run, though the published value reproduces under the versions named in Appendix C.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Houda Bouamor, Juan Pino, and Kalika Bali. Proceedings of the 2023 conference on empirical methods in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, volume 2024, pp. 3056–3076, 2024.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James 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.

Zuzanna Dubanowska, Maciej Żelaszczyk, Michał Brzozowski, Paolo Mandica, and Michal P. Karpowicz. Representation-based broad hallucination detectors fail to generalize out of distribution. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17563–17575, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.952. URL https://aclanthology.org/2025.findings-emnlp.952/. arXiv:2509.19372.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2310.06824.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation. *Advances in Neural Information Processing Systems*, 37:41076–41102, 2024.

Oscar Obeso, Andy Arditi, Javier Ferrando, Joshua Freeman, Cameron Holmes, and Neel Nanda. Real-time detection of hallucinated entities in long-form generation. *arXiv preprint arXiv:2509.03531*, 2025.

Fabian Ridder, Laurin Lessel, and Malte Schilling. Ragognizer: Hallucination-aware fine-tuning via detection head integration. *arXiv preprint arXiv:2604.15945*, 2026.

Subramanyam Sahoo, Vinija Jain, Aman Chadha, and Divya Chaudhary. Linear probes detect task format, not reasoning mode in language model hidden states. In *Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026)*, pp. 227–239, San Diego, California, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.trustnlp-main.12. arXiv:2606.02907.

CH-Wang Sky, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4401–4420, 2024.

Min-Hsuan Yeh, Sharon Li, and Tanwi Mallick. Lumina: Detecting hallucinations in rag system with context–knowledge signals. In *International Conference on Learning Representations*, volume 2026, pp. 151430–151448, 2026.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 915–932, 2023.

A Definitions

Within-answer AUC. For one answer holding n_+ supported and n_- withheld scored tokens, $(R_+ - n_+(n_+ + 1)/2) \,/\, (n_+ n_-)$, where R_+ sums the ranks of the supported tokens when that answer's $n_+ + n_-$ scored tokens are ranked in ascending order of score. The ranks are argsort ranks rather than average ranks, and ties do not arise for continuous probe scores. The primary statistic is the mean of this over answers; an answer with $n_+ = 0$ or $n_- = 0$ has no value and is excluded (Appendix C).

Pooled AUC. The same expression evaluated once over the scored tokens of all answers together, so a token is ranked against tokens from other answers. Differences in an answer's overall score level, which the within-answer form removes by construction, count against it.

Item-level AUC. The item's generated rows are mean-pooled into one activation vector; the classifier is fit and scored on those vectors, and the scores are ranked across items against a label of whether the item's document retained the supporting evidence or had it substituted. Used only for the padded-corpus sweep of Section 4.4. Its direction is fit on pooled activations rather than on rows, so it is neither the probe below nor a pooling of that probe's scores, and its values are not comparable to the within-answer primary.

The probe. L2-penalized logistic regression, $C = 1.0$, lbfgs solver (scikit-learn's default, not set explicitly), at most 2,000 iterations, fit on the kept rows of `sup0` and `sup2` answers and evaluated on the kept rows of `sup1` answers at layer 16. The five folds are formed over the fitting rows alone, grouped by item, and balanced so that each fold holds an equal number of rows rather than an equal number of items.

The answer-level gate. The same probe family asked whether a whole answer came from a document that retained its supporting sentences or from one that retained neither, fit on `sup2` against `sup0` rows and evaluated out of fold on those same rows. Both an AUC and an accuracy at a decision boundary of zero exist for it; Figure 4 carries the AUC, matching that figure's axis. It is benchmarked against a bag-of-words classifier on the answer text rather than against chance (Appendix C).

Next-token entropy. The entropy of the model's next-token distribution, $-\sum_v p_v \log p_v$ in nats, evaluated at the position whose activation the probe reads for that token. It is negated before use, so that higher means more supported and its orientation matches the probe's.

Per-token and per-part scoring. Scored per token, each kept row is one ranked unit. Scored per part, a part's kept rows are averaged first and the part is one ranked unit, so an answer contributes two units rather than one per token. A per-part within-answer AUC therefore has $n_+ = n_- = 1$ and takes a value in $\{0, 0.5, 1\}$, so its mean over answers is the fraction whose supported part outranks its withheld one.

Token IoU. Within one answer, $|S \cap G| \,/\, |S \cup G|$ for S the token set the threshold selects and G the supported span. It is a set overlap and not an AUC, so 0.5 is not a chance value for it, and there is no clean chance value: the reference depends on the sizes of both sets.

Token precision and recall. With S and G as above, precision is $|S \cap G| \,/\, |S|$ and recall $|S \cap G| \,/\, |G|$, each taken within one answer and then averaged. Precision has no value when the threshold selects nothing, so its mean runs over fewer answers than recall's, whose denominator is at least 1 for every scorable answer by construction.

The f placement statistic. Anchored on the mean of the withheld class $\bar{w}$ and the mean of the supported-and-correct class $\bar{c}$, $f(x) = (\bar{x} - \bar{w}) \,/\, (\bar{c} - \bar{w})$, so $f = 1$ places a class at the supported-and-correct endpoint of its own instrument's range. A difference in placement compares f for one class under two instruments. For the supported-but-wrong class the interval on f includes 1, which is why the claim is indistinguishability rather than placement.

The g statistic, retired. $g(X) = (\bar{X} - \overline{\texttt{sup0}}) \,/\, (\overline{\texttt{sup2}} - \overline{\texttt{sup0}})$: the same form as f, but anchored on two different conditions rather than on two classes of span inside one generation. Its denominator therefore mixes the effect of the evidence with a difference in generated text, which f's does not, and within-answer AUC replaced it (Appendix B).

The bootstrap. Intervals resample the answer, never the row, because tokens inside one answer are not independent; the interval is the 2.5 and 97.5 percentiles of the resampled means. Draw counts differ by experiment and are given at each use.

Depth fraction. The layer index divided by the model's layer count, 32 here. It is the axis for any comparison across models or against another report, since a layer index alone is not comparable between networks of different depth.

Ninety-five percent of peak excess over chance. The first probed layer at which $\mathrm{AUC} - 0.5 \geq 0.95\,(\mathrm{peak} - 0.5)$. The excess over chance is taken rather than the raw AUC because an AUC's floor is 0.5 and not 0, so a fixed fraction of the raw value would cross far too early.

Squared projection onto the top-k principal subspace. For a unit direction u and the matrix P_k of the top k principal components, $\|P_k^\top u\|^2$. A uniformly random unit direction in d dimensions has expectation exactly k/d, which is analytic rather than estimated.

B Validity: the pre-registered record and the findings we withdrew

B.1 The pre-registered items

Five implications were declared before any of them was run, and a sixth was declared after the first of them failed. Table 3 reports all six: the three that failed, the two never run, and the one that passed. The labels VF1 to VF5 are the pre-registration's own, kept here so that each row can be checked against the registered document rather than against our description of it.

B.2 The placebo was not pre-declared

The abstract, the introduction, Section 4.2 and Figure 4 all separate three pre-declared nulls from a randomly initialized placebo, and the separation is deliberate. The pre-registration declares three nulls and one positive control. The placebo was added afterward, because a comparison paper reported one. It is a null in the sense that it must sit at chance, and it is not part of the pre-registered record.

item	declared in advance	outcome
VF5	Part identity is not decodable. Target 0.5; pass only if the interval contains 0.5 and its upper bound does not exceed 0.60.	**FAIL.** 0.434, interval $[0.409, 0.460]$, $n = 511$. Below chance, and the interval does not contain the target.
VF1	Dose response and locality, measured by g. Equivalence, margin 15% of the anchor gap; pass only if the whole interval falls inside the band.	**FAIL** on both probe arms. Entropy passes one arm and fails the other.
VF2	Restoration. Positive direction, no margin; the interval must exclude zero.	**PASS.** Mean paired difference $+12.64$, interval $[12.13, 13.17]$, $n = 506$.
VF3	Irrelevant evidence present: g inside $[-0.15, 0.15]$ on a condition that adds an unrelated sentence to `sup0`. A rise would be a serious finding against us.	**NOT RUN.** Needs a corpus and activations that do not exist. Declared build-only.
VF4	Evidence position within the chunk: $\lvert g(\text{early}) - g(\text{late})\rvert < 0.15$. Presence, not salience.	**NOT RUN.** Same reason.
VF1b	Locality without a ratio. Target 0.5 on both arms, band $[0.40, 0.60]$; pass only if the whole interval sits inside the band.	**FAIL** on both arms.

Table 3: The six pre-registered validity implications and their outcomes. VF1b was declared after VF1 failed, which the pre-registration states in its own words rather than leaving to be inferred. VF3 and VF4 are stated in terms of g, which was retired before either could be run.

B.3 Slots and rendered labels

Two things in this construction are easy to conflate, and three of the items below turn on the difference. Each item fixes two semantic slots before anything is rendered. The substitution under `sup1` always removes the sentence supporting the second slot, so in every `sup1` item that slot is the withheld one and the other is the supported one: on `sup1`, slot identity and support status are one partition, not two. Which slot is rendered first, and so which one the model sees labeled (a), alternates by item. The rendered label therefore follows position and the slot does not, and the two disagree for half the corpus. Figure 2 shows the alternation, and its kept and deleted counts are the slot split, which on `sup1` is the same thing as the support split.

This is why VF5 is posed on `sup2`: there both slots are supported, so slot identity is the only thing left for a probe to separate. It is also why the two position rows of Figure 4 are a separate pair rather than the same check under another name. VF5 asks about the slot; they ask about the rendered position.

B.4 Dose response and locality (VF1)

Against declared bands of $[-0.15, 0.15]$ for the span whose evidence was removed and $[0.85, 1.15]$ for the span whose evidence survived, the probe at layer 16 gives 0.376, interval $[0.342, 0.408]$, and 0.669, interval $[0.631, 0.716]$. Both intervals fall outside their bands, so both arms fail. Entropy gives 0.279, interval $[0.210, 0.345]$, which fails, and 0.903, interval $[0.872, 0.930]$, which passes. The ordering is monotone and was reported, but the pre-registration states that ordering is not the claim.

A sensitivity check was declared on the subset of items whose answer is byte-identical across all three conditions, with the rule that a subset below 30 items be marked underpowered rather than reported as a number. Two items qualify. It is marked underpowered and no number is given.

B.5 LOCALITY WITHOUT A RATIO (VF1B), DECLARED AFTER VF1 FAILED

VF1b asks VF1's question with no ratio in it: the same span, the same local evidence state, and only the other part's evidence differing. Arm 1 compares the supported slot's spans under `sup1` against the same slot's spans under `sup2`, so the span's own evidence is present in both. Arm 2 compares the withheld slot's spans under `sup1` against `sup0`, so the span's own evidence is absent in both. A score depending only on local evidence gives 0.5 on both.

Arm 1 gives 0.938, interval $[0.927, 0.947]$, $n = 508$. Arm 2 gives 0.899, interval $[0.885, 0.913]$, $n = 503$. Both intervals lie far above the declared band, so both arms fail. Arm 1 was declared as likely to fail before it was run, on the stated grounds that the size and direction of the miss was the quantity wanted.

One caveat bounds the size of this failure without changing its direction, and it is large. Both arms compare two conditions, and the model regenerates its answer under each condition, so the text being scored is not the same text. The effect therefore conflates the other part's evidence state with a difference in generated content, and this design cannot separate them. Separating them needs a contrast that varies the other part's evidence inside a single generation, which this corpus does not contain.

B.6 RESTORATION AND SLOT IDENTITY (VF2 AND VF5)

VF2 asks whether restoring a span's evidence raises its score. Scoring the withheld slot's spans under `sup1` against the same slot's spans under `sup2`, paired by item, the mean difference is $+12.64$ in probe units, interval $[12.13, 13.17]$, over 506 items. The interval excludes zero in the declared direction, so it passes. It is paired at the item level only: the model regenerates, so tokens are not matched across the two conditions.

VF5 asks whether the support probe, without refitting, separates the two slots on held-out `sup2` answers, where both are supported. It gives 0.434, interval $[0.409, 0.460]$, over 511 answers against a target of 0.5. The interval does not contain the target and sits below chance, so the control fails; only its distance from 0.5 is informative, not its direction. Nothing was dropped from any other result because of it: the headline is computed over the population it would have been computed over had the control passed. In the artifacts this is the contrast between `semantic_a` and `semantic_b`, which are the slots and not the rendered labels.

B.7 THE FOUR-STAGE CONFOUND CHECK

Sahoo et al. (2026) give a staged protocol for establishing that a probe reads its construct rather than the task's format. We ran it at layer 16 against the primary, 0.862 with an interval of $[0.841, 0.882]$ over 493 answers.

Stage 1 asks whether the source document is identifiable from the activations. It is, at 2.8% against a chance rate of 0.3% over 296 sources, and the source is constant within every evaluation comparison, in every item, because both spans come from one generation.

Stage 2 asks about answer options. We free-generate, so there are none and the stage has no analogue here. Two substitute scalars are reported instead: rendered position separates at 0.5010, interval $[0.456, 0.542]$, and an answer's own generated length takes one value per item, so ranking on it inside an answer gives exactly 0.5. That is not the length Stage 3 controls, which is the count of scored tokens in each of the two parts and differs between them in most answers.

Stage 3 asks for the result under format control. Restricting to answers whose two parts hold the same number of scored tokens gives 0.8609, interval $[0.832, 0.888]$, over 257 answers; within one token, 0.8748 over 353; within two, 0.8738 over 437.

That restriction does two things at once and cannot separate them. Inside a `sup1` answer one rendered part keeps its evidence and the other loses it, so the two parts are the two classes; matching their scored-token counts is therefore also a restriction to the answers whose rank grid is balanced, which is where the statistic is least coarse (Appendix C). Landing at 0.8609 against 0.8621 unrestricted is consistent with both readings and distinguishes neither.

Stage 4 asks for the result with the named confounds residualized out. Removing source, length and position gives 0.8577, interval $[0.836, 0.879]$, over 493 answers, with 2.4% of the variance removed; removing format alone gives 0.8627, interval $[0.841, 0.882]$, with 0.5% removed.

The artifact's own summary reads that the check passes all four stages. That is three passes and one substitution, and we report it as such.

B.8 Where this construction is most exposed

The same check names the arm most open to the critique, and it is not the primary. It is the gate. The gate classifies `sup2` against `sup0` at the answer level, where the two conditions have wholly different documents and nothing is held constant inside a comparison. That is the regime in which the critique bites: with nothing held constant, a high separation is not evidence that a probe reads support rather than format. So the gate is a sanity check on the instrument and not a support result, and Section 4.2 uses it as one, as a positive control that must come out high and is benchmarked against a classifier on the answer text rather than against chance. The within-answer primary is the arm the construction exists to protect, and the critique does not reach it, because both of its classes come from a single generation.

B.9 A statistic we specified badly and retracted

g (Appendix A) anchors on two different conditions, so its denominator is a difference between activations from two different generations. Part of that denominator is a difference in generated text rather than a difference in evidence, and the ratio inherits the confound. The pre-registration records the confound of the anchors before VF1 was run and the retirement after it.

g is retired and VF1 is not relabeled: its failure stands as declared against the band declared for it. VF1b replaces the question with a measurement that has no ratio, and within-answer AUC is used everywhere else. We report the retraction rather than substituting a better-behaved statistic and presenting the result as though that statistic had been declared in advance.

B.10 Findings we withdrew

Each row is a claim we made, the check that caught it, and what stands now.

The fourth row is a sequence rather than one check. The claim was withdrawn on the grounds that the two curves peak at different layers. That was not a comparison: the information curve used was the one for the gate, which asks whether a whole answer came from a document that kept its supporting sentences, while the AUC curve was the within-answer one. They measure different problems, and the gate is close to saturated (Figure 4). The withdrawal was itself retracted the same day and the comparison that was owed was run on one task. On the within-answer task both point estimates fall at layer 24, the information interval spanning layers 20 to 26 and the AUC interval spanning 18 to 26, but the two argmaxes coincide in only 29% of bootstrap draws. The permitted statement is that both point estimates fall at layer 24, not that two independent measures agree.

B.11 A number that is not evidence

Thresholding the probe's per-token score selects 1.017 times as many tokens as the supporting span contains, and that ratio is not evidence of calibration. The rule sets the number of predicted tokens equal to the number of gold tokens on the training split of each fold, so a ratio near one follows from the rule and would hold for any score it was applied to. It was removed from Section 4.1's table for that reason, and is recorded here because a reader who computes it would otherwise meet a suspiciously good number.

B.12 The f placement result

f (Appendix A) places the supported-but-wrong class at 0.837 under the probe, interval $[0.660, 1.022]$, and at 0.565 under entropy, interval $[0.382, 0.718]$. The probe's interval includes 1. We therefore cannot distinguish this class from the supported-and-correct class, and that is the

what we claimed	what the check found	status
The signal lives in the directions carrying most of the activation variance.	A principal component sweep at layers 16 and 24. The top component at layer 16 carries 47.5% of the variance and scores AUC 0.457, interval $[0.431, 0.483]$: below chance, with the interval excluding 0.5. The 16 components spanning 90% of the variance retain 36% of the probe's margin, so 64% lies outside them; at layer 24 the four components spanning 90% retain 13%. That 64% is measured against this sweep's 16-component grid point; the exact crossing is 11 components, which is the figure Appendix D quotes.	Withdrawn, and reversed. Appendix D states the opposite.
Span recovery at token IoU 0.677.	A threshold rule committed to in advance. The earlier rule tuned its threshold per fold to maximize IoU, scored against the largest gold span only, and discarded extra predicted spans. Under the committed rule token IoU is 0.545, interval $[0.518, 0.573]$; the share of answers above 0.5 falls from 86.8% to 65.9%; and the earlier operating point over-selects by $1.58\times$, which an F1 of 0.751 had concealed.	Withdrawn. Section 4.1 reports the committed value.
A baseline matches or beats the probe, at $+0.011$, also quoted as $+0.023$.	A twelve-cell matrix crossing instrument against scoring unit. Both figures pair a per-token probe against a per-part entropy, and they are the only two cells of the twelve whose interval includes zero. At a matched unit the difference is $+0.164$, interval $[0.143, 0.184]$, per token, and $+0.088$, interval $[0.051, 0.125]$, per part.	The claim is withdrawn. The number $+0.023$ still appears in Section 4.3, named there as the mismatched cross-unit cell whose interval includes zero, which is the opposite of the claim withdrawn here.
The within-answer information bound and the AUC independently select layer 24.	Withdrawn on a bad basis, then re-established weakly. The sequence is below.	Stated weakly.

Table 4: Findings withdrawn by our own checks.

claim the paper makes: the two classes' intervals overlap, and both sit far from the withheld class. No stronger reading is available from these numbers.

f is anchored on all three class means, so its basis is the withheld class at 492, the supported-and-correct class at 423 and the supported-but-wrong class at 70. The 70 is the class being placed, and it is what bounds the claim rather than the whole basis.

Section 4.3 reports the difference between the two placements as 0.272, which is the difference of the two point estimates above, with an interval of $[0.036, 0.552]$ bootstrapped over the difference itself. The key stored for that difference holds 0.274, the mean over the bootstrap draws rather than the difference of the estimates; because f is a ratio those are not the same quantity, and they differ by 0.002. We report the difference of the estimates.

B.13 Five refusal definitions that return one measurement

The refusal rate was recomputed under five definitions: the published rule; a stricter rule requiring no digit anywhere in the answer; a stricter rule again requiring the decline phrase in the first 30 characters; a looser rule admitting empty generations; and the loosest, admitting hedges carrying no explicit decline phrase. The five return numerically identical values in every cell of every sample, the supported control included.

The reason is that every declining generation in these samples is the same string, `NOT STATED`. Four of the five rules therefore have nothing to disagree about: a string carrying no digit, whose decline phrase begins at character zero, which is neither empty nor a hedge, satisfies all five at once. Across the three samples, 254 of $1{,}360$ generations decline, and every one of them is that string, in each sample separately and in the supported control. The five are one rule by construction on this

corpus. So the identity is not robustness evidence, and Section 4.4 makes no robustness claim from it. It is recorded because a reader who recomputes the rate under a variant definition will meet the same identity and should know what produces it.

B.14 Two further items of the same class

The control battery has caught a construction defect, and did so before any result in this paper was computed. An earlier version of the corpus rendered the two slots in a fixed order, so slot identity, rendered position and support status were one variable rather than three. A position-only probe separated the classes perfectly on that version, at 1.000; that store carries no record of a rendered order because there was nothing to record. The corpus was rebuilt with the order alternated, and a position probe indexed by label was kept as a positive control that must come out high, precisely because it is the check that fired. The earlier version's within-answer AUC was 0.853, close enough to the current headline that the defect was not visible from the result.

One control does not reproduce, and is recorded as unresolved rather than repaired. A null asking whether the probe reads which rendered position carried the supported slot is stored at 0.566, interval $[0.540, 0.591]$. A later reconstruction of its protocol returns 0.612, and the companion figure for a probe fitted directly on that target moves from 0.866 to 0.935. Both halves of the reconstruction reproduce bit for bit when it is re-run, which rules out noise and rules out the fold partition, so the stored pair came from a protocol we no longer hold: its script is absent. The stored values were kept and the reconstruction recorded as failed rather than adopted. An unreproducible result is a claim to drop; an unreproducible null is a confound never actually excluded, which is the worse of the two. This quantity is not VF5: VF5 separates the two slots, and this null asks about the rendered position.

C The corpus, and what it takes to reproduce these numbers

C.1 Construction

Table 5 is the chain from answers built to answers scored. Each row is a filter, and the counts are recomputed from the corpus file rather than carried forward from any earlier report.

count	stage	what the step removes
556	`sup1` answers built, one per item	
536	retain at least one scored token	20 answers contribute no scored token at all
493	carry both support classes among their scored tokens, and are therefore scorable by the within-answer statistic	43 are single-class: 23 all-supported and 20 all-withheld
488	also carry a next-token entropy score, and are the set on which the two instruments are compared	5 of the 493 are not covered by teacher forcing

Table 5: The n chain. A single-class answer has no within-answer AUC because the statistic needs both classes inside the answer, so the 43 are not scorable rather than discarded by a threshold.

Row counts follow the same chain. The 536 answers contribute 6,269 scored rows and the 493 contribute 5,270, split 2,635 supported against 2,635 withheld. What the construction fixes is the span count and not the row count: every scorable answer contributes exactly one supported span and one withheld span, because each `sup1` item keeps the supporting sentence for one part and deletes it for the other. Two spans of one answer need not hold the same number of scored tokens, and in 236 of the 493 they do not, by as much as 27 rows. The totals come out equal because the 149 supported-heavy answers and the 87 withheld-heavy ones cancel, at $+367$ against -367. We report the equality as a count and do not derive it.

The 43 answers the statistic cannot score are 8% of answers but 16% of scored rows, 999 of the 6,269, because they are the long ones: their median is 28 scored tokens against 9 for the answers

that are scorable. That exclusion is length-biased in the direction that favors us, since Section 4.1 finds longer answers scoring lower. What the same score column reaches on those rows is given below.

Teacher forcing covers $1{,}491$ of the $1{,}668$ item-conditions, the remaining 177 failing an exactness gate on prompt reconstruction. That exclusion is length-biased, and the bias runs the unhelpful way: the excluded items have a median prompt length of $1{,}072$ tokens against 721 for the corpus, so the subset on which the probe and entropy are compared is shorter-prompted than the corpus it is drawn from.

Part order alternates exactly rather than approximately: the builder swaps every even-numbered item, and the realized corpus confirms it at 278 of 556 both ways. Position therefore cannot predict which of the two slots kept its evidence (Appendix B).

Scoring excludes 55 token ids, being those that decode to whitespace alone plus those that decode to one of the answer template's own characters, and drops every row preceding the first part.

The corpus is narrow, and the following bound every result measured on it. It draws on 65 distinct issuers and filing years 2014 to 2024. Every gold answer is an integer of at least three digits, and no gold anywhere in the corpus contains a decimal point: zero of the $3{,}336$ gold values in the file. Prompts run 307 to $1{,}439$ tokens with a median of 721. Over the 493 scorable answers the median generation is 18 tokens and the median number of scored tokens is 9, with quartiles 8 and 12 and a range of 5 to 39.

Two defects in the builder, named and not fixed

The substitution always targets the same semantic slot. Each item fixes two slots before anything is rendered, and `sup1` always removes the sentence supporting the second of them, so the withheld class is the same slot in every item (Appendix B). That is consistent with the alternation above rather than in tension with it: what alternates is which slot is rendered first, and so which one the model sees labeled (a), not which slot loses its evidence.

That slot's gold values are consistently the shorter and the smaller of the pair: paired by item over all 556, they are shorter by 0.30 characters, interval $[0.13, 0.47]$, with order of magnitude and token count running the same way and digit count not separating them. These come from an exploratory diagnostic with no pre-declared margin and are descriptive. The consequence is that the two classes are not exchangeable by construction, which matching the slots on length and magnitude, and alternating which of them is deleted, would fix.

The donor sentence is matched to the removed sentence in characters, not in tokens. The builder truncates the donor to the character length of the sentence it replaces and strips it, so the passage keeps its length to within one character and does so in every item. Token counts do not follow: they match exactly in 4.9% of items and to within one token in 10.6%, with a mean absolute difference of 7.9 tokens and a range of -27 to $+39$ against a mean of 37.6 tokens removed. Because the cut is made at a character offset rather than a token or sentence boundary, a donor can begin or end part way through a number or a word. Figure 1 shows one: the replaced passage ends `$2.` where the original continued. The paper's description of the donor as length-matched is accurate in characters and should be read that way.

The answer text alone carries the contrast

An item-grouped bag-of-words classifier on the generated answer text, with no access to activations, separates `sup2` from `sup1` at 61.7% accuracy and `sup2` from `sup0` at 71.8%, over $1{,}112$ answers at seed 0. The dispersions are ± 3.1 and ± 2.3, across-fold standard deviations over the five folds rather than intervals. The design document's conclusion is that the support contrast is visible in the output text, and we report it as such rather than as a null that passed.

Two consequences follow, and the paper depends on both. The first is that 71.8% is the benchmark the answer-level gate must clear, not chance: a gate scoring above 50% has shown nothing, because the wording alone reaches 71.8%. The gate clears it on both axes, at 93.0% accuracy against the classifier's 71.8% and at 0.975 AUC against its 0.781, the second of which is the pair Figure 4 plots. The two comparisons agree, so the choice of axis is not doing the work. What clearing the

benchmark establishes is bounded: the gate compares conditions with wholly different documents, so a high separation there is a sanity check on the instrument rather than evidence of a support representation (Appendix B). The second consequence is that the within-answer primary is unaffected, because its two classes are spans of one generation and the comparison holds item, prompt, context and answer text constant by construction.

One specification detail is free and appears nowhere a reader can see. The vectorizer's token pattern admits digits and periods, `[A-Za-z0-9.]+`, rather than scikit-learn's default, which requires two or more word characters and therefore drops currency symbols, splits on decimal points and discards single characters. The default gives 70.6% on the gate contrast against the published 71.8%. Every other free choice in the specification moves the figure by under a point, so this is the largest of them, but at 1.2 points it is inside the ± 2.3 above: a sensitivity rather than a determinant. It is recorded because the prose does not fix it and a reimplementation would have to guess, not because the result turns on it. The published pattern remains the defensible one, since the mechanism the design document proposes concerns the gold values themselves and a tokenizer that discards numerals cannot see them.

One figure must not be used. An ungrouped version of the same classifier reaches 70.0% on the gate contrast. It comes out lower than the grouped figure, which is the opposite of what leakage usually does, and the reason is structural: every item contributes exactly one answer per condition, so an item's two answers always carry opposite labels, and memorizing an item's vocabulary pushes its mate toward the wrong one. The ungrouped figure is not the conservative choice it appears to be; it is a differently biased one. The benchmark is 71.8%, item-grouped.

THE CEILING IS GRANULARITY

Of the 493 scorable answers, 253 score exactly 1.000. That is a property of the measure's resolution and not evidence of agreement: the median answer carries 9 scored tokens, with quartiles of 8 and 12, so a typical answer offers roughly four supported tokens to rank against five withheld ones, and a perfect ranking of so few is not rare.

C.2 REPRODUCTION

POOLED AGAINST WITHIN-ANSWER, ON ONE POPULATION

Two pooled AUCs exist and they are not interchangeable. The value over every scored row, 0.7289 across $6{,}269$ rows from 536 answers, is the one that ships in the headline artifact. The value the paper reports, 0.7648, is computed over the 493 answers that carry both classes, $5{,}270$ rows. The within-answer 0.8621 macro-averages over those same 493 answers, so only the second is like-for-like with it. Comparing the within-answer figure against the first mixes a change of unit with a change of population, and about a quarter of the resulting gap is population rather than unit.

Matched here means matched to the within-answer population, and that is a real restriction rather than a formality. A threshold deployed on this score would also have to rank the single-class answers, which the within-answer statistic never sees. Pooled over the 999 rows of the 43 excluded answers, where both classes are present across the group even though no individual answer carries both, the same score column reaches 0.5251, against 0.7648 on the matched population. The restriction is therefore worth stating: performance on the answers the primary statistic can score is not performance on all generations.

TWO SPLITS THAT LOOK ALIKE

One hazard has nothing to do with folds. The stored field `row_part` is the rendered position, and the semantic slot is `row_part` exclusive-or `swap`. On `sup1` the two disagree for half the corpus: $3{,}539$ rows against $2{,}730$ by position, $3{,}195$ against $3{,}074$ by slot, over the same $6{,}269$ rows. Both are legitimate partitions, and only the second is what the slot labels in the artifacts mean, so grouping on `row_part` returns a plausible number for a different quantity. Figure 2's kept and deleted counts are the slot split, which on `sup1` is also the support split.

What a specification has to say to be reproducible

The fit discipline as the paper states it elsewhere, five-fold grouping by item, is not sufficient to reproduce the headline to four decimals, and this is a finding about specification rather than an error in any run. Seven partitions, every one an honest reading of that phrase, were computed on identical data with identical per-fold standardization. Table 6 gives them.

how the five folds are formed	within-answer AUC
over the fitting rows, grouped by item (the specification we report)	0.8621
over all rows of all conditions, grouped by item	0.8595
over the evaluation rows, grouped by item	0.8586
equal item counts per fold, shuffled, one seed	0.8559
equal item counts per fold, shuffled, another seed	0.8507
contiguous blocks of sorted item ids	0.8504
item id modulo five	0.8456

Table 6: Seven fold partitions on byte-identical data, spanning 0.0165 of AUC. The gap between the widest and narrowest reading is larger than several differences the paper elsewhere treats as meaningful.

The mechanism is that `GroupKFold` balances folds by row count and not by item count, and which rows it is given decides the partition. Forming the folds over the fitting rows holds them within one row of each other, at 2,470 to 2,471 rows. Enumerating the items in sorted order and cutting them into five blocks spans 2,202 to 2,696, a spread of 494. Between those two schemes 442 of the 550 fitting items land in a different fold. Because the folds are formed over the fitting rows, three `sup1` items that have no fitting rows never enter a test fold and a first-fold fallback scores their 70 evaluation rows; removing that fallback leaves the headline bit-identical, because all three are single-class and the within-answer statistic skips them either way.

A second specification dependency, smaller and of the same kind. The twelve paired differences behind Section 4.3's unit matrix are drawn from one seeded generator in sequence, so each interval depends on how many cells were drawn before it. The point estimates do not, so a reimplementation that reports a different set of cells will reproduce the means and not the intervals.

What this is not. The data is not the cause: the activation file these checks were run against is byte-equivalent to what the pipeline reads, with the same row order, the same item, condition, part, support and token identifiers, the same keep mask, and values identical after the cast from half to single precision. Standardization accounts for at most 0.002, measured across fitting one scaler per fold, fitting one globally, and using none. The tie convention accounts for nothing here: argsort ranks and average ranks agree to machine precision on these scores, though they do not on a quantity that is constant within an item. The masks account for nothing. And the published value reproduces bit-identically under scikit-learn 1.9.0, numpy 2.5.2 and scipy 1.18.0.

No manifest, model record or environment lock in this project records the scikit-learn version that produced the original run. That is a provenance gap and we state it as one. It is not the cause of anything above, since the published value reproduces under the version named, but it is the kind of thing that would have to be recorded for a reader to rule out library drift without rerunning.

Provenance

Every figure in this subsection was computed against a single activation file, the one the probe is fit on, by three scripts written for the purpose rather than assembled from scattered runs. Neither the code nor the activations are released, and the reproducibility statement gives the reasons.

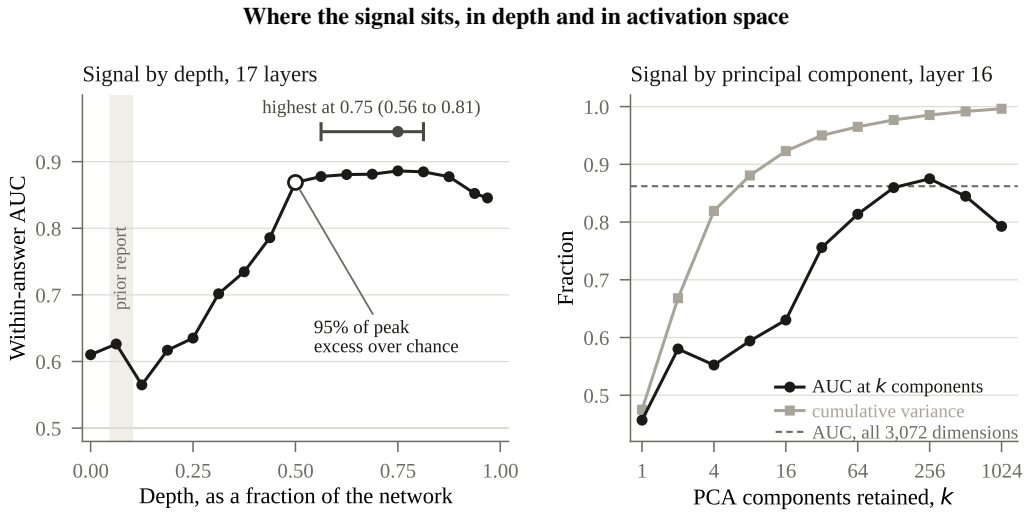


Figure 7: Left: within-answer AUC at each of 17 probed layers, with depth given as a fraction of the model's 32 layers. The bracket is the layer of highest AUC and its 2,000-draw bootstrap interval over the argmax. The open circle is the first layer reaching 95% of the peak's excess over chance. The shaded band is the depth at which the closest prior report places that same crossing, drawn as a band because their layer index is ambiguous between layers and sublayers. The source bootstraps the argmax and the crossing only, so the curve carries no per-layer intervals. Right: within-answer AUC at layer 16 when the activations are projected onto their top k principal components, against the cumulative activation variance those k components carry, with the AUC over all 3,072 dimensions marked. The components are fitted on the training rows without the label. The panels come from two separate captures of the same items, 545 answers on the left and 493 on the right, so the left panel's values are not comparable to the headline.

D Depth, the collections behind it, and where the signal sits

D.1 The layer sweep

Seventeen layers were probed on the dense collection, over 545 answers (Figure 7). Decodability runs from 0.610 at layer 0 through 0.635 at layer 8 and 0.869 at layer 16 to a highest value of 0.886 at layer 24, then falls to 0.846 at layer 31. Only two statistics on this curve are bootstrapped, over $2{,}000$ draws: the layer of highest AUC, at 24 with an interval spanning layers 18 to 26, and the first layer reaching 95% of the peak's excess over chance, at 16 with an interval spanning 16 to 20. No per-layer intervals exist, so no difference between two individual layers on this curve is resolved by anything reported here, and the flatness from layer 18 to layer 26 is why the peak is not resolvable.

The crossing can be computed two ways and both give layer 16. Taken literally, the first layer whose AUC reaches 95% of the peak crosses a threshold of 0.842, with an interval of 16 to 16. Taken on the scale an AUC lives on, the first layer whose excess over chance reaches 95% of the peak's excess crosses a threshold of 0.867, with an interval of 16 to 20. The second is the defensible one, because an AUC's floor is 0.5 and not 0: a fixed fraction of the raw value credits a layer with the half of the scale it gets for free, and would cross earlier on any curve. We report the second, and note that the first is the one a literal reading of Sky et al. (2024)'s phrasing would use.

D.2 The cross-paper comparison

Sky et al. (2024) report that hallucinations are detectable at many layers, reaching 95% of peak performance as early as their layer 4, on a 40-layer response model. Our crossing is at layer 16 of 32, a depth fraction of 0.50.

Their layer index cannot be resolved into a depth fraction without an assumption, and we do not resolve it. They train one probe per sublayer, over $2L$ sublayers, so their layer 4 is a depth fraction

of 0.10 if the index runs over layers and 0.05 if it runs over sublayers. Both readings are carried here; the source does not disambiguate them and neither do we.

Statistic choice accounts for part of the gap and not most of it. Our own argmax sits at depth 0.75 and our crossing at 0.50, so adopting their statistic moves us by 0.25, against a remaining difference of 0.40 to 0.45 depending on which reading of their index is taken. That is roughly a third of the apparent contrast, and what remains is a factor of five to ten in depth fraction. At our own layer 4, their crossing layer, our probe scores 0.565 against a peak of 0.886, which is 17% of the peak's excess over chance. Whatever is decodable at layer 4 in their setting is very nearly absent at the same relative depth in ours.

D.3 The position convention

Both read activations at generated tokens during decoding, so the difference in depth is not a difference in what was read. Their setup takes the hidden state corresponding to the i-th response token. Our primary collection holds generated-token rows only and never retained prompt positions at all, so the convention is forced rather than chosen.

The convention is not uniform across our collections. Six of the ten hold generated-token rows only and four retain prompt positions as well, so a result computed on one of the four is drawn from a collection holding a different population of rows. One result in this paper is: the document-length and refusal results of Section 4.4 come from a padded collection that retains prompt rows. The probe there pools generated rows only, matching the convention used everywhere else, so the difference is in what was recorded and not in what was read. The primary collection, the dense sweep and the second model all hold generated-token rows only.

D.4 The dense collection is a different sample

The seventeen-layer sweep comes from a separate collection, and it is a different sample rather than a refinement of the primary one. Of $1{,}500$ aligned generations, 167, or 11.1%, have a different token sequence from the primary despite the same seed, the same items and greedy decoding. The prefill is identical and decoding diverges from about the second token. No number from it is a better version of a number from the primary collection.

It is also truncated. It was killed at generation $1{,}579$ of $1{,}668$ under memory pressure, leaving $1{,}500$ aligned generations. The evaluation condition and the fully supported condition are complete at 556 of 556 each; the fully withheld condition is at 388 of 556, or 70%. The consequence is that the probe fitted on this collection sees an endpoint training set imbalanced 59 to 41 rather than evenly split. Item-level grouping is unaffected.

A re-run answers the question that divergence raises. Running the same $1{,}668$ item-conditions again from the same seed, 11.0% produce a different token sequence from the primary, with the first divergence at a median of the second generated token.

That figure and the 11.1% above are not two measurements. The $1{,}500$ generations this collection and the re-run share are identical string for string and token id for token id, this collection being the truncated one, and there is no case where it diverges from the primary while the re-run agrees. The two rates are one comparison against the primary reported at two denominators, and they could not have disagreed.

On the re-run the probe reaches a within-answer AUC of 0.865, interval $[0.846, 0.883]$, over the 545 answers scorable there, against 0.862, interval $[0.841, 0.882]$, over the 493 scorable in the primary. The two differ by 0.003 and the re-run's value falls inside the primary's interval. The answer counts differ because the regenerated answers are not the same answers, so which of them carry both support classes is not the same set, and the comparison is between two populations rather than within one. The 545 here is the same 545 as the layer sweep's, for the same reason: one set of regenerated answers rather than two. The pre-registered reading, recorded before the re-run, is that generation nondeterminism does not move the headline.

What that establishes and what it does not. The headline survives on a second set of generations, which is the claim we make. It is not corroborated by two independent recollections, because there is only one.

Its layer-16 value is 0.8688 and the headline is 0.8621. These are not the same measurement and the first is not an improvement on the second: they are computed on different samples of generations, over different answer counts, and with differently balanced training sets. The pair is given side by side here so that the resemblance cannot be mistaken for agreement.

D.5 The second model

The construction was rerun on `google/gemma-2-2b`, a second family at roughly half the parameter count, using the same corpus and the same template. Its within-answer AUC is 0.852, interval $[0.832, 0.872]$, over 550 answers, at layer 21. The headline's interval is $[0.841, 0.882]$, which overlaps it.

This is a replication on the same corpus, not a transfer test: the items, the substitutions and the questions are identical, and only the model changes. No cross-model depth claim is available from it. Its probed grid is layers 0, 4, 8, 13, 17, 21 and 25, and its two best layers, 17 at 0.849 and 21 at 0.852, are close enough that the argmax is resolved by the grid rather than by the data. Placing its peak as a fraction of depth would need a denser sweep.

D.6 Where the signal sits in activation space

Three quantities bear on this and they are not interchangeable. Each is reported separately below.

The principal components here are fitted on the training rows without any reference to the label. A low-dimensional activation spectrum is therefore a statement about where the activations vary and not about where the support signal lives, and nothing in this subsection should be read as evidence that the support signal is itself low-dimensional.

Squared projection of a fitted direction. The fitted support direction places 0.016 of its squared norm in the 32 components carrying 95% of the variance at layer 16, against 0.010 for a uniformly random unit direction, which is k/d exactly. The direction encoding slot identity places 0.002. These three come from an exploratory diagnostic with no pre-declared margin and are descriptive.

AUC retained under projection. Projecting the activations onto those 32 components and refitting gives 0.756, interval $[0.730, 0.781]$, against 0.862 over the full 3,072 dimensions. It does not recover until 128 components are kept, at 0.860, interval $[0.837, 0.880]$, and those 128 components carry 97.7% of the variance against 95.0% for the first 32.

Margin lying outside the high-variance subspace. At layer 16 the components spanning 90% of the variance retain 36% of the probe's margin, so 64% lies outside them. At layer 24 the effect is stronger: the four components spanning 90% retain 13%, so 87% lies outside. That 64% is measured against this sweep's 16-component grid point; the exact crossing is 11 components.

D.7 One direction is enough, and it is not the only one

Four arms were fitted at layer 16 on the primary collection. The fitted direction alone, as a one-dimensional readout, gives 0.867, interval $[0.847, 0.886]$. The residual after removing it gives 0.805, interval $[0.782, 0.827]$. The full space gives 0.861, interval $[0.839, 0.881]$, and the direction with its residual added back gives 0.862, interval $[0.841, 0.882]$, a change of -0.005 against the direction alone.

The reading is that the residual re-expresses the same distinction through correlated directions rather than carrying a second one, so the readout we report is sufficient and not unique. A reader wanting a single number for how much of the result one direction carries should take these four together rather than any one of them.

These four values are internally comparable and are computed on one arm set. None of them should be differenced against the headline reported elsewhere in this paper. Two invite that mistake. The full-space figure here is 0.861, close enough to the headline's 0.862 to look like a rounding difference when it is a different computation. Worse, the direction-with-residual arm prints as 0.862 with an interval of $[0.841, 0.882]$, which is the headline's printed value and the headline's printed interval,

digit for digit. That is a coincidence of rounding between two separate computations on two different arm sets, and it is not the headline reappearing.