SAME ARTIFACT, TWO FACES: WHY BLACK-BOX LINEAGE FINGERPRINTS LOOK BOTH ROBUST AND FRAGILE

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowing which training branch a deployed model is produced from matters for model auditing and policy. We ask to what extent the post-training branch of same-family models can be identified from black-box fingerprints alone, i.e. output texts. Qwen3 base/IT pairs are the most comparable case: under a pre-training continuation format, both branches generate readable text. We collect 100 to 200 MATH500 problems per model and train a 16-feature output classifier. Within a single scale the separation is near-perfect (0.941 / 0.947), but only when the reading is uncontrolled: it is not stable across the five scales we ran, and it moves further than the headline effect we report when we change the prefix length we keep, chosen arbitrarily. Once length and problem difficulty are controlled together, the leave-one-model-out transfer is no longer distinguishable from chance, and it does not collapse to a clean negative either. Disguise, prompt and answer rewriting, template and temperature perturbation, and cross-lingual attacks all leave the reading without a discernible decrease, because the signal carrying it is termination and length; the one arm that moved the reading moved it by shortening the output, which is that channel.

1 INTRODUCTION

1.1 THE SETTING

A deployed model rarely comes with its training history, yet auditors, safety teams and regulators often need to decide which post-training branch it came from. They may hold the weights or other model statistics such as gradients while generating, or in the hardest case, only the text it generated. This separates the white-box and black-box questions. The strictly conducted black-box version of this question already has a quantified answer, and it is not a satisfying one. On LeaFBench the strongest black-box method reaches TPR@1%FPR = 0.356, against 0.992 / 0.943 for white-box static methods (Appendix A); which means more than sixty percent of derived models are missed at the operating point a real audit would use.

We work in the easiest possible regime available for the black-box question, and ask what the resulting numbers actually measure. The base and instruction-tuned counterparts of Qwen3 are same-family, same-scale pairs that share an architecture, a tokenizer and a pre-training corpus, so everything that separates them is introduced by post-training. If a black-box lineage signal exists anywhere, it should be here.

1.2 TWO FACES OF ONE READING

We collect 4 variants $\times$ 100 MATH500 problems from each of five Qwen3 scales, extract 16 surface features from the output text alone, and ask a classifier to tell the base branch from the instruction-tuned one. Within a single scale this reads 0.941 and 0.947 at $0.6b$ and $1.7b$: with nothing but a bag of length, lexical and self-repetition statistics, the two branches come apart almost completely (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.

The second face appears when we change the apparatus around the same comparison, and it appears on two axes. The first is transfer: asked to carry the classifier from one scale to the neighbouring one, the reading starts at 0.885 and 0.892, and every control we add lowers it; in the most tightly controlled cell — the problems paired and the visible text truncated to its first 200 characters — it is 0.487 and 0.458, at or below chance. The second axis is within a single scale, where the headline number itself moves with the truncation length we keep. §4.2 gives the ladder control by control, against a within-scale baseline of 0.741 and 0.744.

The third face is what happens under attack — disguise of the outputs, rewriting of the questions, rewriting of the answers, the prompt template, the sampling temperature, and translation — and it is the face on which the instability matters more than the effect itself, because not one of these produces a discernible decrease. Taken alone that would read as robustness; taken with the first two faces it says something else — none of the attacks bites because none of them touches the channel that carries the reading.

Putting the three faces together gives the claim of this paper. The black-box lineage reading we obtain is not a property of the model pair but of the pair *as seen through a particular measuring device*, and in our device the dominant channel is termination, not style: how often, and how early, the instruction-tuned branch stops generating.

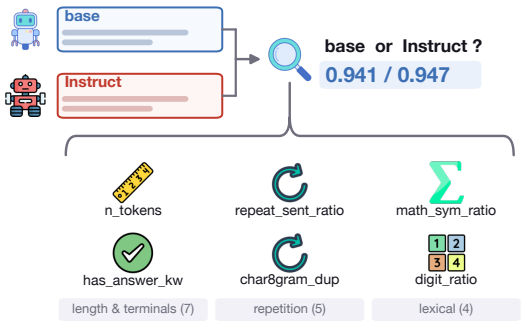

Figure 1: The classifier and what it reads: two branches of one pair, read through sixteen statistics of the output text (six shown).

1.3 Claims and contributions

We make four claims.

First, under the continuation protocol the base and IT branches are separated mainly by a termination channel: the IT branch stops at the generation cap on most problems, the base branch rarely does. Re-collecting at different caps leaves the separation in place at almost every scale (Appendix E). The length and lexical differences between the generated texts are real but are downstream of the same behaviour. Every attack we ran leaves the reading without a discernible decrease, and each fails for the same structural reason: it does not perturb the termination and length channel.

Second, the reading is not a stable quantity: it moves with the instrument we chose, and it does not survive the controls. Within a single scale it depends on how many characters of the output we keep, a parameter the previous protocol neither fixes nor reports. Once length and problem difficulty are controlled together, cross-scale transfer degrades to about chance in the most tightly controlled cell.

Third, the separability is family-dependent: in Mistral-7B the cap-rate difference reverses sign and the lexical block carries what separation remains in both of our cells; in OLMo-2-1B, every post-training recipe caps *less* than its base and the lexical block carries the residual separation in all four. "The degeneration block carries the fingerprint" is therefore a property of Qwen3 pairs under this protocol, not of base/IT pairs in general.

Fourth, the classifier is not specific to mathematics: on HumanEval a single bit flagging the arrival of generation cap separates the branches at AUC 0.965 and 0.990 on the two scales we ran. That bit has almost no variance on the IT side, so the score is a constant rather than a learned one, which is the point: what separates the branches is the degenerate behaviour itself.

Contributions. Our contribution is a measurement audit rather than a new detector: we take the easiest case in the black-box lineage problem and vary one setting of the apparatus at a time — same-family, same-scale, same-tokenizer pairs across five scales, the generation protocol held fixed. The result decomposes the reading into a channel the protocol determines and a residue that might belong to the models.

The artifacts we release are the per-cell readings, not the models: every reading in this paper is traceable to a cell of a table, with its sample size and its protocol attached.

2 Related Work

The nearest neighbours to our question operate at three granularities and on the attack-robustness axis; Appendix A gives each the paragraph it deserves. AttnDiff (Zhang et al., 2026a) is the closest white-box neighbour: a compact descriptor built from attention weights over minimally perturbed prompt pairs, same-family similarity above 0.98 under a CKA readout, tokenizing the prompt string itself and performing no decoding. Every number in this paper comes from decoded output text only, and that difference in access is what our whole measurement turns on.

2.1 Our position

Our position is a scope statement rather than a claim to have occupied a cell: four conditions have to hold together for a reading to bear on the classification question, and each is individually occupied already. They are branch granularity within one family, the output text alone with zero internal access, a cross-scale check that has to be passed rather than merely run, and a confound control matrix that levels length, pairs the problems, truncates the prefix and disguises the format. Appendix A takes each in turn and names the neighbours that occupy it.

What the four have in common is a measurement surface produced in quantity which no one has audited in the sense of asking what its readings are measuring. The closest external echo is a probe study (Sahoo et al., 2026), where a hidden-state probe separating reasoning modes at perfect accuracy on Qwen3-$14b$ turned out to be a format artifact, collapsing to chance once the format was projected out — the same structure one level up, in output text, and the shape of the correction we apply to our own headline number in §1.3. Appendix A carries this positioning in full.

3 Protocol

3.1 Comparability is a prerequisite, not a result

A base checkpoint has never been instruction-tuned: it has no trained response to a chat template, and under one it degenerates into repeated special tokens and produces no usable text, which would make any comparison vacuous. We therefore adopt a pre-training continuation format, rendering the prompt as `Problem:\n{question}\nSolution:` and continuing generation from there, under which both branches emit coherent mathematical text. It is a controlled factor, held fixed across every condition in §4.1 to §4.5.

The format is not a neutral container either, and the size of its effect can be stated exactly: under the continuation format the IT branch hits the cap on every problem and the base branch on 29% and 35% of them, a difference of +71 and +65 points, while under the same chat template with thinking disabled the difference is -61 and -2 points, so the protocol does not shrink it but reverses its sign at $0.6b$ and removes it at $1.7b$ (Table A10). Two qualifications travel with the reading: the base side of the chat arm is out of distribution for base, and the criterion pre-registered for this arm landed inside its undetermined band. We therefore claim neither that the continuation format is required for the reading to exist nor that it is neutral.

3.2 Models, data and collection

The primary evidence is Qwen3 at five scales, $0.6b$ / $1.7b$ / $4b$ / $8b$ / $14b$, each a same-family base/IT pair sharing an architecture, tokenizer and pre-training corpus. Two second families serve as controls: OLMo-2-1B, four post-training recipes on one shared base (sft, dpo, instruct, rlvr), and Mistral-7B-v0.3 with its Instruct variant, on an independent tokenizer.

All problems come from MATH500, except §4.7, which repeats the collection on HumanEval; generation is greedy with `max_new_tokens = 512`. The question count is not uniform (200 per model for §4.1, 100 from §4.2 onward) and the two are not interchangeable, so we print the sample size in every table cell.

The five scales were collected as a single batch, never mixed with the earlier two-scale directories; within a scale both branches were generated on the same device, with one accelerator type to avoid turning any device difference into a branch difference.

One setting of that collection is swept since it is the setting our reading is most exposed to: if the channel we report were an artifact of the 512-token budget, the comparison would be measuring our instrument rather than the pair. We therefore re-collected both branches of all five scales at caps of 1024 and 2048 on the same problems. The separation survives at four of the five scales — 94% to 100% against 6% to 16% at cap 2048. The cap rate of base model falling from 26–47% to 2–16% as the budget grows, while the post-trained branch's length is the instrument's, its median exactly the current cap and its cap rate 94–100% except at $14b$, where it is 75%. Appendix E prints all fifteen cells.

3.3 Features and classifier

We use 16 pure output-side features, one caliber for the whole paper, in four families: five for length and density, two terminal markers, four for lexical and five for self-repetition features. Nothing here is model-specific, and nothing requires access beyond the text. Appendix B names all sixteen.

The classifier is a standardized logistic regression with `max_iter = 2000`, its scaler fit on the training split only. Within-scale readings use 5-fold stratified out-of-fold prediction, cross-scale readings use leave-one-model-out, and every confidence interval comes from 2000 bootstrap resamples.

One caliber change has to be declared, in two dimensions. Earlier scripts in this line used 17 features, the extra one being `has_formula_kw`, and the two sets differed by at most 0.0075; separately, an earlier pass estimated output tokens from the word count, where this paper uses the count the generator recorded, and the two rulers differ by 0.017 at $0.6b$ and 0.010 at $1.7b$. Every reading here is the 16-feature caliber on the recorded token count, the first row of Table A16 carries that cell, and the first row of Table 1 is 17-caliber by construction and must not be compared against the rows beneath it, which also use a different scaler fit.

3.4 Controls

Four controls carry the paper (Figure 2), and they are applied cumulatively.

The first is length levelling: we keep only outputs that reached the 512-token cap, which removes the termination difference at the sample level but leaves it in the feature vector, where `n_tokens`, `n_chars`, `n_words` and `n_sentences` still carry it; only the truncation below and the block decomposition of §4.4 take the length channel out of the reading.

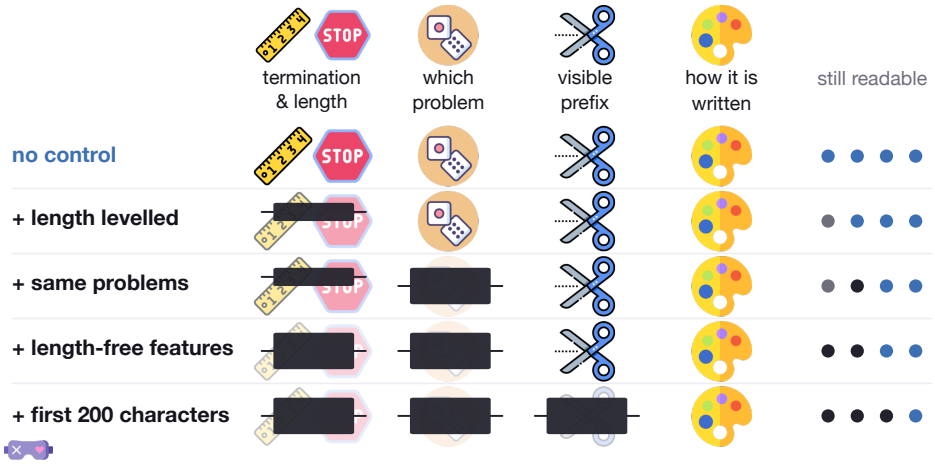

Figure 2: Adding the controls one at a time: each rung takes one more channel away from the reader, and the dots on the right show what it can still read.

The second is question-id pairing: both classes are restricted to the question ids on which the base branch hit the cap, so they cannot differ in problem difficulty. This is the control a first version lacked, and the one that changes the cross-scale conclusion; Appendix F measures what the length control selects in its place.

The third is prefix truncation: each output is cut to its first 200 characters, with sensitivities at 400 and 800, which makes the visible text length exactly equal across the two classes. Truncation also changes how many samples survive, and the cell that resolves that ambiguity combines cap bucket, pairing and truncation, holding everything fixed but the amount of visible text.

The fourth is symmetric disguise: decoys are added to both classes rather than to one, since the earlier one-sided version produced a leakage construction rather than an attack (§4.5).

3.5 Criteria

We use two criteria on the readings. **H1** is the strong negative claim: the cross-scale confidence interval lies entirely below 0.5, the right criterion for saying that a reading has no discriminative content left. **H2** is the difference claim: the paired bootstrap difference between the within-scale and cross-scale readings has an interval entirely above zero, corrected over the pre-registered family of 20 ordered pairs by Holm over the pooled 40 one-sided p-values. Read as questions, **H1** asks whether the classifier has anything left to read — whether even the top of its interval sits at or below a coin flip — and **H2** asks whether a control actually moved the reading, judged on the same problems, rather than whether two separately drawn intervals happen to overlap.

4 Results

4.1 Within a single scale the reading looks almost complete

Holding the scale fixed and asking a classifier to separate the two branches inside one model pair, the reading is near ceiling: 0.941 at $0.6b$ and 0.947 at $1.7b$, from 16 surface features and nothing else. Table A16 scores the same cell at 0.944 across the two scales, and Table A18 gives the family profile behind the five scales; the column later sections use is the cap rate of the channel §4.4 is about, and under the continuation protocol the IT branch reaches the 512-token cap on 100% of problems at both scales, the base branch on 29% and 35%.

Table A19 asks what is left of the within-scale reading once the ladder's controls are applied inside one scale: 0.746 and 0.747 with the full feature set at $L = 200$, and 0.719 and 0.696 with the length-free set. It does not fall to chance, but it falls a long way from ceiling and it moves with the instrument: at $4b$ the length-free column travels from 0.615 to 0.852 between $L = 200$ and $L = 400$, a shift of 0.237, larger than the effect the column is measuring. Figure 3 draws both columns across the five scales and the three lengths, and L is a setting with no principled value that nothing in the protocol fixes.

So at fixed scale the reading is near ceiling only while the apparatus is left alone.

4.2 Transfer degrades as the controls are added

Table 1 is the central table of the paper: it carries the same classifier from $0.6b$ to $1.7b$ and back, adding one control per row (Figure 5).

At full length and unpaired, transfer reads 0.885 and 0.892, the number our earlier version reported as evidence of a scale-general lineage signal. Levelling the length by keeping only capped outputs takes it to 0.838 and 0.672, pairing the questions as well takes it to 0.744 and 0.625, and truncating the visible text to its first 200 characters to 0.487 and 0.458, at or below chance in the most tightly controlled cell we have.

Row $3'$ isolates the feature group at full length and reads 0.724 and 0.598, so the group alone costs 0.020 and 0.027 in the two directions and the rest of the drop to 0.487 and 0.458 is the truncation.

The within-scale baseline at the foot of the table reads 0.741 and 0.744 against the transfer cells above it, and it comes from the ladder run's smaller cap bucket rather than from the corresponding

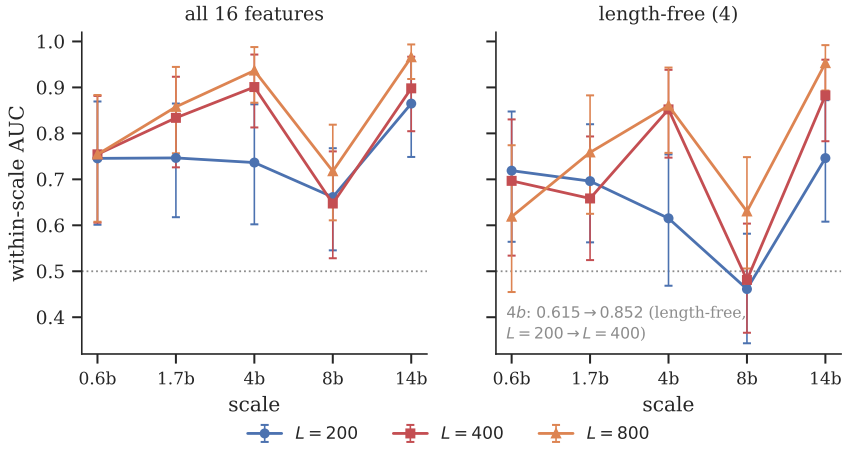

Figure 3: The within-scale reading of Table A19 across five scales and three truncation lengths, with 95% bootstrap intervals. The length-free panel is the one that moves: at $4b$ it travels from 0.615 to 0.852 between $L = 200$ and $L = 400$.

Table 1: The transfer ladder, transfer in both directions between $0.6b$ and $1.7b$. n is printed per cell because the controls change the sample as well.

#	control	feat.	$0.6b \to 1.7b$	n	$1.7b \to 0.6b$	n
1	full length, unpaired	17	0.885 [0.836, 0.926]	200	0.892 [0.835, 0.940]	200
2	length levelled (cap bucket)	16	0.838 [0.764, 0.909]	135	0.672 [0.528, 0.810]	129
3	+ question-id pairing	16	0.744 [0.627, 0.852]	70	0.625 [0.472, 0.767]	58
3′	row 3, length-free features only	6	0.724 [0.605, 0.836]	70	0.598 [0.440, 0.743]	58
4	+ truncation $L = 200$	4	0.487 [0.346, 0.618]	70	0.458 [0.301, 0.617]	58
5	$L = 400$ (row 4, L only)	4	0.544 [0.408, 0.682]	70	0.508 [0.353, 0.666]	58
6	$L = 800$ (row 4, L only)	4	0.637 [0.503, 0.765]	70	0.598 [0.432, 0.758]	55
—	*within-scale baseline*, $L = 200$	4	0.741 [0.606, 0.859]	58	0.744 [0.624, 0.853]	70

Row 1 reproduces the original pipeline (all 17 features, scaler fitted on the test set) ; rows 2–6 fit the scaler on train only. Rows 2–3 use the 16-feature caliber, row 3′ the six features that carry neither length nor repetition, and rows 4–6 the four that also drop the two terminal proxies. So $3 \to 3'$ isolates the feature group alone, and $3' \to 4$ the truncation alone, holding the feature group fixed. Rows 4–6 hold everything fixed and vary only L. Transfer does not fall to chance on any row, and it does not decay monotonically in L either. **The within-scale baseline at the foot of this table is not the same cell as the corresponding column of Table A19.**

column of Table A19; the sample sizes differ and we print both rather than merging them. Once length and difficulty are both controlled, then, the cross-scale reading is no higher than the within-scale one, and in the tightest cell its point estimate falls below chance, at 0.487 and 0.458 with intervals that still cover it.

4.3 Over twenty ordered pairs, transfer is not uniformly below within-scale

Two scales and one direction each is a small basis for a claim about transfer, so we ran the design over all five scales in both directions as a pre-registered extension: 20 ordered pairs on the same double-paired cell (Table A15).

Within-scale reads 0.751 on average over the five scales, cross-scale 0.654 over the twenty pairs, with a range from 0.355 to 0.843. The strong negative criterion, a cross-scale interval lying entirely below 0.5, is met by **0 of 20 pairs** on the full feature set and 0 of 20 on the length-free one, and the nearest miss is $1.7b \to 0.6b$ on the full set, upper bound 0.5120.

The difference criterion behaves the same way. Uncorrected, 6 of 20 pairs on the full set and 2 of 20 on the length-free set have a paired difference whose interval lies entirely above zero; after Holm correction over the pre-registered family of twenty, **one** pair survives on the full set and **none**

does on the length-free set. One pair goes the other way, $1.7b \rightarrow 8b$ on the length-free features, significantly *higher* across scales than within, by 0.230, and Figure 7 plots the twenty differences with their intervals.

Transfer is therefore not, as a rule, below within-scale: it is unreliably close to chance in the cells we control most tightly, and sometimes higher. We report the full twenty because the two-scale version is what made the over-claim in the first place.

4.4 The channel that carries the reading

The 16 features divide into three mutually exclusive and exhaustive blocks: a degenerate block of seven (five length and density features, two terminal markers), a lexical block of four, what remains once length, repetition and the terminal markers are removed; and a repetition block of five (Table A16).

At full length the degenerate block leads, and the ordering looks like the story we told in §4.1: the separation is a termination effect. Truncating reverses the order — the degenerate block falls to 0.640 while the lexical block holds at 0.719 at $L = 200$ — and the strong-attack blocks, none of which shortens the output, behave like full length, while the real-rewriting block is ordered like the truncated ones. Figure 8, in the appendix, shows the same reversal as the difference between the two blocks, condition by condition.

The zero-rewrite control at the foot of the table is the cell where our instrument, not the models, is doing the moving: 3 of its 6 cells have the degenerate block below the lexical one, and we return to it in §4.5.

One further test is needed, because "the degenerate block carries the strong face" is a claim about seven specific features, and any seven of the sixteen might do: Table A14 calibrates it against a cardinality-matched null over all 11,440 seven-feature subsets, with Figure 9 showing the distributions block by block. On no block does the degenerate block come closer to the full set than chance would explain (the closest is a p of 0.0738, which does not survive the seven-way correction), and on the truncated and zero-rewrite blocks it sits *further* from it than a typical subset. That table therefore does not support "the strong face is carried by the degenerate block": what supports the reading is the flip in Table A16.

4.5 Every attack fails, and each fails for the same reason

Table A17 lists the attack surface: disguise, question rewriting, answer rewriting, prompt template, sampling temperature, and translation with the answer language pinned (Figure 4), and not one produces a decrease that survives the criterion of §3.5.

The first is a withdrawal: an earlier version reported a disguise attack that retained the reading at 1.000, but that number was a construction error, since the decoys were added to one class only and "was this perturbed" became a perfect proxy for the label; the control measures that leakage at or above 0.997 in all four cells, and the symmetrised version reads 0.950 and 0.938.

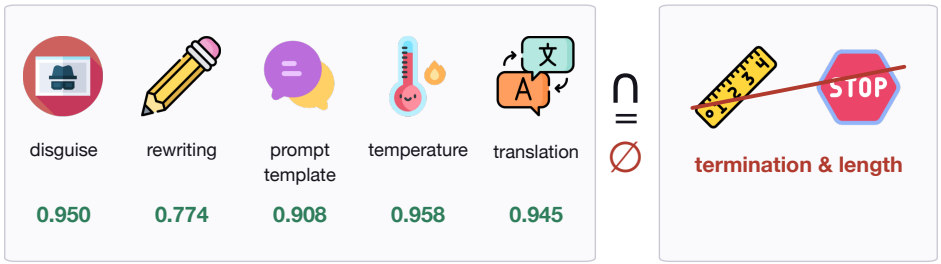

Figure 4: Five attacks perturb the text surface, and none of them touches the termination and length channel that carries the reading.

The second is the rewriting arm, read against its own zero-rewrite control, which applies the same transformation to text that is already identical: in four of the six cells the two agree to within 0.013, and in the exception the control itself collapses to 0.342, below chance, because style transfer compresses the output to a median of 0.32 of its length and the control no longer controls anything. What is load-bearing there is that the control collapses, not that the attack succeeded.

The third arm changes the question language while pinning the answer language, and leaves the reading at 0.945 and 0.960; without the pinning the two changes are confounded — the output-side Chinese fraction only reaches 0.13 and 0.15 (Table A17) — and the arm is not readable as an attack. Taken alone, "no attack works" reads as robustness, but taken with §4.2 it reads differently: the attacks fail because the channel that carries the reading, termination and length, is not the channel any of them perturbs.

4.6 Two more families

Three further controls test whether what we have described is a property of Qwen3 base/IT pairs or of this measurement. Mistral-7B, with an independent tokenizer, reverses the cap-channel sign (base 62% against Instruct 23%, against +65 to +71 on Qwen3) and its lexical block carries what separation remains; OLMo-2-1B is the stronger version, with four post-training recipes on one shared base, every recipe capping less than its base and the degenerate block below the lexical block in all four arms. Appendix L prints the arms and Figure 10 collects the cap rates, while the protocol arm of §3.1 is a scope limit rather than a result. The correction is specific: the degenerate channel carries the reading for Qwen3 base/IT pairs under the continuation protocol, and in the two other families we can test it does not.

4.7 The classifier is not specific to mathematics

One further control checks that we have not mistaken a property of one task for a property of the reading: on HumanEval, code generation rather than competition mathematics, a single bit — whether generation hit the token cap — separates the branches with AUC 0.965 and 0.990 (Appendix K).

5 Discussion

5.1 The apparatus decides the reading

Four times in this paper the same classifier over the same features on the same two branches gives a different answer, and only a setting of the instrument changes: transfer reads 0.885 and 0.892 at full length and unpaired, and 0.487 and 0.458 once the problems and the visible length are held fixed. More budget makes such a fingerprint look better and no more informative: over the eight ordered adjacent pairs the uncontrolled transfer rises from 0.932 at cap 512 to 0.982 at 2048, while the cell that removes the channel shrinks from 52–94 problems to 4–32, because the base branch's cap bucket is exactly where the hard problems were (Appendix E). Nor is the ladder a property of our classifier: three text-only detectors that a lineage claim would plausibly use — our sixteen features, TF-IDF over word and character n-grams, and Burrows' Delta — read 0.86 to 0.96 uncontrolled and lose the transfer reading under the same controls, to 0.36 to 0.68 (Appendix G). A number in this literature is a property of a pair together with an apparatus, and the apparatus is usually what goes unreported.

5.2 Why the two faces are one

Both faces are one channel — whether the model terminates, and how much text it emits before it does: at full length a block of length plus terminal markers scores 0.913 against the full set's 0.944, and pairing the questions and levelling the length removes what that channel was encoding. Within a scale, at a fixed length, a residual lexical signal separates the branches at 0.719 and 0.696, so a reading that does not control this channel measures the channel.

5.3 What we would ask a lineage claim to report

An apparatus that is not reported is not reproducible, and the fault is not in the reader. Four things carry most of the weight: the termination and cap rate for both branches under the collection format; whether the classes were scored on the same problems; the visible length at which the reading was taken, and whether more than one was tried; and whether the reading survives on features carrying neither length nor repetition. §3.4 lists the controls we would ask to be reported with any black-box lineage claim; Table 1 is what happens when they are; and Appendix N reproduces this section in full, including the caveat that an adversary aimed at the termination channel would likely succeed.

6 Limitations

Three families, five recipes. The family-level claim rests on Qwen3, Mistral and OLMo, five recipes in all: enough to show the Qwen3 decomposition does not generalise, not enough to say what decides it.

The chat arm is undetermined. The pre-registered criterion asked whether the IT branch still caps; the measured rate fell inside the open band.

Two tasks. MATH500 and HumanEval; the second confirms that the classifier travels and not much more.

Surface features, and a reconstruction. The classifier reads 16 statistics of the visible text, and our comparison against a representation-level method preserves the ordering of readings without their thresholds.

Instrument settings that we chose. The truncation length L has no principled value and the reading moves with it (§4.1); the cap bucket used by rows 2 and 3 of Table 1 conditions on the base branch having failed at full length; and the generation cap itself we did sweep, with the channel surviving at four of the five scales and weakening without dissolving at the fifth (Appendix E). That sweep also shrinks the comparable sample: the paired cell that resolves the channel falls from 52–94 problems at cap 512 to 4–32 at 2048, so the setting which makes the uncontrolled reading strongest is the setting in which least can be compared. Appendix O reproduces these limits in full.

7 Conclusion

A lineage fingerprint is usually reported as one number, and we have tried to show what that number is made of: across five scales, twenty ordered pairs, three families and a second task, the readings move further under changes to the instrument than under attacks on the text. The apparatus is not a nuisance to be averaged away, since the same pair of models separates at 0.941 within a scale and at chance across scales, only because each branch was asked different problems and wrote a different amount of text.

What survives is worth having, and worth stating carefully. Within a scale, at a fixed length, on the same problems, a residual signal separates the branches and no surface attack we ran removes it, and that is the claim we are prepared to defend; what we are not prepared to defend is that this signal carries across scale, across protocol, or across families.

Reproducibility statement

Every reading is traceable to a cell of a table, and each cell carries its sample size and its protocol. The appendix specifies the prompt string, the greedy decoding, the 512-token cap together with its 1024 and 2048 sensitivities (Appendix E), the sixteen features (Appendix B), the classifier and its validation protocol (5-fold out-of-fold within a scale, leave-one-model-out across scales, 2000 bootstrap resamples). The 512-token arm of the cap sweep reproduces the original collection exactly, cap rate and within-scale reading included, which is the sense in which this pipeline is reproducible across passes. We release the per-cell readings with their generation records and the analysis code, so that every number here can be re-derived without re-running a model.

ETHICS STATEMENT

Lineage fingerprinting carries a dual use: the audit question is one that safety teams and regulators legitimately need answered, and a negative result of this kind makes the black-box version harder to overstate, but the same reading can be offered as provenance evidence where the apparatus has not been controlled, which is the failure mode this paper is about. We therefore recommend reporting the four items of §5.3 rather than a single number.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods, namely the feature extraction, sweep and analysis code. Additionally, we used generative AI tools for creating figures, creating and editing software code, writing the appendix of the manuscript, and editing the manuscript for readability. We have not used generative AI tools to generate data, to develop the theoretical framing, to formulate or prove claims, or for the drafting of the main body of the paper.

REFERENCES

Zhiyuan Fu, Junfan Chen, Lan Zhang, Ting Yang, Jun Niu, Hongyu Sun, Ruidong Li, Peng Liu, Jice Wang, Fannv He, Qiuling Yue, and Yuqing Zhang. FDLLM: A Dedicated Detector for Black-Box LLMs Fingerprinting, 2025. URL https://arxiv.org/abs/2501.16029. arXiv:2501.16029v4.

Hongyu Liang, Yuting Zheng, Yihan Li, Yiran Zhang, and Shiyu Liang. Origin Tracer: A Method for Detecting LoRA Fine-Tuning Origins in LLMs, 2025. URL https://arxiv.org/abs/2505.19466. arXiv:2505.19466v1.

Mingyi Liu. The Alignment Tax: Response Homogenization in Aligned LLMs and Its Implications for Uncertainty Estimation, 2026. URL https://arxiv.org/abs/2603.24124. arXiv:2603.24124v2.

Cen Lu, Yung-Chen Tang, and Andrea Cavallaro. Similar Models Learn Differently: Final-Window Pretraining Shapes Post-Training Beyond SFT, 2026. URL https://arxiv.org/abs/2607.25063. arXiv:2607.25063v1.

Anshul Nasery, Edoardo Contente, Alkin Kaz, Pramod Viswanath, and Sewoong Oh. Are Robust LLM Fingerprints Adversarially Robust?, 2025. URL https://arxiv.org/abs/2509.26598. arXiv:2509.26598v1.

Subramanyam Sahoo, Vinija Jain, Aman Chadha, and Divya Chaudhary. Linear Probes Detect Task Format, Not Reasoning Mode in Language Model Hidden States. In *The 6th Workshop on Trustworthy NLP at ACL 2026*, 2026. URL https://arxiv.org/abs/2606.02907. arXiv:2606.02907v2.

Shuo Shao, Yiming Li, Yu He, Hongwei Yao, Wenyuan Yang, Dacheng Tao, and Zhan Qin. SoK: Large Language Model Copyright Auditing via Fingerprinting, 2025. URL https://arxiv.org/abs/2508.19843. arXiv:2508.19843v3.

Yao Tong, Haonan Wang, Siquan Li, Kenji Kawaguchi, and Tianyang Hu. SeedPrints: Fingerprints Can Even Tell Which Seed Your Large Language Model Was Trained From. In *ICLR 2026*, 2026. URL https://arxiv.org/abs/2509.26404. arXiv:2509.26404v2.

Suqing Wang, Ziyang Ma, Xinyi Li, and Zuchao Li. Ghost in the Transformer: Detecting Model Reuse with Invariant Spectral Signatures. In *AAAI 2026*, 2026. URL https://arxiv.org/abs/2511.06390. arXiv:2511.06390v3.

Yuqi Wu, Shengming Zhao, and Jie Chen. TokenPrint: A Calibrated Token-Space Fingerprint for Language-Model Provenance, 2026. URL https://arxiv.org/abs/2608.08139. arXiv:2608.08139v1.

Zehao Wu, Yanjie Zhao, and Haoyu Wang. Gradient-Based Model Fingerprinting for LLM Similarity Detection and Family Classification, 2025. URL https://arxiv.org/abs/2506.01631. arXiv:2506.01631v2.

Ziqing Yang, Yixin Wu, Yun Shen, Wei Dai, Michael Backes, and Yang Zhang. The Challenge of Identifying the Origin of Black-Box Large Language Models, 2025. URL https://arxiv.org/abs/2503.04332. arXiv:2503.04332v1.

Haobo Zhang, Zhenhua Xu, Junxian Li, Shangfeng Sheng, Dezhang Kong, and Meng Han. AttnDiff: Attention-based Differential Fingerprinting for Large Language Models. In *ACL 2026*, 2026a. URL https://arxiv.org/abs/2604.05502. arXiv:2604.05502v1.

Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. REEF: Representation Encoding Fingerprints for Large Language Models, 2024. URL https://arxiv.org/abs/2410.14273. arXiv:2410.14273v1.

Keyu Zhang, Vadim Safronov, and Andrew Martin. Stemma: Induced Decision Regions Reveal LLM Provenance, 2026b. URL https://arxiv.org/abs/2607.25880. arXiv:2607.25880v1.

The body of the paper reports readings; this appendix reports the instruments and the cells those readings came from. It carries the extended related work of §2, the sixteen features by name, the collection procedure down to the prompt string, the collection-environment drift and the cap sweep and what length matching selects and the external detectors under the same audit, the twenty ordered scale pairs one at a time, the disguise arm and the chat-template arm and the second task in full, the two further protocols and two further families, and the cardinality-matched null. Sections 5 and 6 are reproduced in full in Appendices N and O. Where a table in the body aggregates, the appendix prints the cells it aggregated, so that a reader who wants to check whether an average is representative of its members can do that without rerunning anything.

A Extended related work

The nearest neighbours to our question operate at three granularities, and it is worth separating them before we say where we stand.

At the model level, FDLLM (Fu et al., 2025) trains a dedicated detector on a 90K-sample, 20-model corpus with a LoRA decoder, and it is the direct methodological template for the classifier we use in §3.3: a supervised model over statistics of the generated text. TensorGuard (Wu et al., 2025) works instead from gradient responses, over 8 base models and 50 derivatives. TokenPrint (Wu et al., 2026) needs no training at all — it projects the output onto a top-vocabulary subspace, and its AUC rises from 0.72 to 0.90 as the readout moves toward the output layer. All of these require something beyond the text, whether gradients or a layer's activations, which is exactly what the setting of §1.1 withholds. Two further works read the *base* lineage rather than the branch, Origin Tracer (Liang et al., 2025) by detecting LoRA fine-tuning origins and the black-box probe of (Yang et al., 2025) by adversarial triggers, and neither separates two branches of one family from text alone.

SeedPrints (Tong et al., 2026) is the strongest near neighbour on the granularity axis: it fingerprints variation at the level of the training seed and follows it across a training lifecycle. It reads logits or final-layer representations, so it is not a pure black-box method, and that is why its numbers and the black-box column of the survey discussed below do not conflict — they are measuring different access regimes. REEF (Zhang et al., 2024) sits on the same representation side; it is a training-free CKA similarity between hidden representations rather than a classifier, and we cite it only for that reading.

The benchmark anchor comes from the copyright-auditing survey (Shao et al., 2025). OOn its LeaFBench Table 2 the strongest black-box method, MET, reaches TPR@1%FPR = 0.356, and LLMmap, the black-box method most often quoted, reaches 0.253; the two white-box static methods reach 0.992 and 0.943. The black-box column is the one that matters for a deployed model, and it is routinely the first column dropped when the table is quoted.

The most direct precedent for our attack experiments is the survey of Nasery et al. (2025), which tests ten fingerprinting schemes that claim robustness and reports near-complete attack success across

four broad classes of shared vulnerability. It is our template for how the attacks are built. The same survey states that output perturbations and "rephrasing of prompts" remain under-explored for model fingerprinting, which is the external basis for the rewriting experiments in §4.5. Its scope also excludes thinking-style models, so the regime we work in lies in its complement.

On the provenance side, Stemma (Zhang et al., 2026b) is the nearest claim to rewrite-robust provenance, reading provenance off induced decision regions. §4.5 is our test of that claim in the black-box lineage setting, and the answer turns out to depend on which side of the pipeline is rewritten. A recent spotlight paper supplies a floor of a different kind, that detectability, edit-resistance and stealth cannot all be had at once, although its setting is watermarking rather than lineage [OpenReview 097IlaUfQY].

The distinction that matters for us is that each of these works attacks a fingerprint's *carrier*. Our result is that in the black-box lineage case the carrier is a degenerate behaviour, so the attack surface and the signal surface do not overlap at all (§4.6).

The question we ask only makes sense if the two branches differ in behaviour to begin with, and that is established. The Alignment Tax (Liu, 2026) measures response homogenization on Qwen3-$14b$: across ten sampled responses per question, the share of questions whose samples collapse into a single semantic cluster is about 1% for the base branch against 28.5% for the instruct branch, a difference no auditor would want to miss. It also reports a token-entropy readout reaching AUROC 0.724 on GSM8K, and attributes the effect it studies to DPO rather than SFT. We cite it for the existence of the branch difference, which is the premise of everything this paper measures.

Similar Models Learn Differently (Lu et al., 2026) is the closest structural neighbour on the branch axis, and we cite it for the boundary rather than as support. It varies the pre-training data window, roughly 1% of total tokens, and then applies identical SFT and DPO; the finding is that the post-SFT score is blind to the window difference while later DPO and GRPO reveal it. Our independent variable is the post-training recipe itself, not the window it is applied to, so the two designs measure different things.

The four conditions of §2.1, one at a time, and the position they leave us with:

The first is branch granularity: separating a raw checkpoint from one or more of its post-trained derivatives within one family. The spectral method of (Wang et al., 2026) works in weight space, TensorGuard (Wu et al., 2025) at family level, and Origin Tracer (Liang et al., 2025) as a binary question about LoRA origins, so none of them asks for a discrimination among three or more branches of one family.

The second is the signal source, which for us is the output text alone. Weight statistics (Wang et al., 2026), trajectory structure, and representation-level readouts all sit outside that regime, and SeedPrints (Tong et al., 2026) is representation-side in particular, since it needs logits or final-layer representations. Our setting allows zero internal access.

The third is a cross-scale check, and here the distinction is between running the check and passing it. On the representation side SeedPrints does work across scales, but at seed granularity and with internal access, so it is not a comparison on the same axis. On the pure-output side we ran the check ourselves, and it does not hold (§4.2, §4.3).

The fourth is a confound control matrix: levelling the output length, pairing the problems, truncating the visible prefix, and disguising the format. This is the condition that no neighbour above satisfies.

That is the whole of our positioning. Taken one at a time, none of the four conditions is unoccupied, and we do not claim otherwise. What they have in common is that they describe a measurement surface which is produced in quantity — every neighbour occupies a corner of it — and which has not been audited, in the sense that no one has asked what the readings on it are measuring. The closest external echo of that question is a probe study (Sahoo et al., 2026), which shows that a hidden-state probe separating reasoning modes at perfect accuracy on Qwen3-$14b$ is entirely a format artifact and collapses to chance once the format is projected out. We find the same structure one level up, in output text rather than hidden states, and the correction we apply to our own headline number in §1.3 has the same shape.

Table A1: The sixteen features, in the order `extract_features` computes them, with the block each one belongs to. The names and the order are read out of the extractor rather than restated here, and the block assignment is the partition of Section 4.4.

feature	block	definition
`n_tokens`	DEG	output tokens as recorded by the generator, passed in as an argument rather than recomputed from the text
`n_chars`	DEG	number of characters
`n_words`	DEG	whitespace-separated tokens
`n_sentences`	DEG	segments after splitting on sentence-final punctuation followed by whitespace, or on newlines; floored at 1
`repeat_sent_ratio`	REP	duplicate sentence occurrences divided by the number of sentences
`uniq_sent_ratio`	REP	distinct sentences divided by the number of sentences
`max_sent_count`	REP	multiplicity of the most frequent sentence
`tail_repeat`	REP	1 if there are at least three sentences and the last three are identical
`char8gram_dup`	REP	1 minus distinct 8-grams over 8-grams, over the sliding 8-character windows
`ttr`	LEX	type-token ratio over the word tokens of the lowercased text
`n_newlines`	LEX	count of newline characters
`avg_line_len`	DEG	mean character length of the non-blank lines
`digit_ratio`	LEX	digit characters divided by length
`math_sym_ratio`	LEX	characters in `+-=*/%^ (){}[]` divided by length
`has_boxed`	DEG	1 if the string `boxed` occurs anywhere
`has_answer_kw`	DEG	1 if the lowercased text contains any of *answer is*, *the answer*, *final answer*, *therefore*

The seventeenth quantity the extractor computes, `has_formula_kw` (1 if the lowercased text contains *formula*, *distance* =, *speed* or = *speed*), is the feature dropped in the caliber change declared in Section 3.3. `n_tokens` is the one feature that is not read off the text: it is the count the generator recorded, which is why the estimated-token caliber of Section 3.3 is a different ruler rather than a different feature.

B THE SIXTEEN FEATURES, BY NAME

The body refers to feature blocks rather than to features, and this has a cost: the five features that carry self-repetition are never named there, so a reader cannot tell whether the repetition block is about repeated sentences, repeated n-grams, or repeated lines. Table A1 gives all sixteen, in the order the extractor computes them, with the block each one belongs to and a definition transcribed from the extractor itself.

Two things about the table are worth stating in words. The names and the order are read out of the extraction script when the table is generated, so the table cannot drift from the code that produced the numbers it describes. And `n_tokens` is the one feature that is not a property of the text at all: it is the count the generator recorded for the output, passed into the extractor as an argument. Every other feature is recomputed from the string. That is why the caliber change declared in §3.3 has a token dimension as well as a feature-count dimension, and why the two dimensions are not interchangeable.

C THE MEASUREMENT APPARATUS

Every reading in this paper comes from one of two collection scripts, and both build their prompt from the same string:

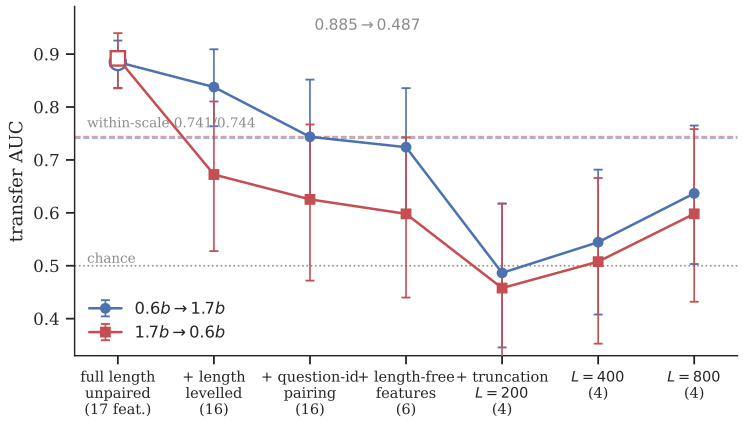

Figure 5: The transfer ladder of Table 1, with 95% bootstrap intervals. The open markers in the first column are the original 17-feature pipeline and are not on the same caliber as the six closed-marker columns to their right. The dashed lines are the within-scale reading at the same truncation. The reading falls from 0.885 to 0.487 as the controls are added, and the sample it is measured on falls with it, from 200 to 70 problems.

```
Problem:
{question}
Solution:
```

Generation is greedy, with `do_sample=False`, a cap of 512 new tokens, and the tokenizer's end-of-sequence token used as the padding token. The cap is the quantity the degeneracy channel is about, so it is a setting of the instrument rather than a safety limit: an output that reaches it is recorded as having reached it, and the recorded token count is the count the generator produced, not an estimate from the word count.

For the mathematics collection the questions are the first hundred entries of the loader's mathematics split, taken in the loader's order, and the same hundred are used at every scale so that a pair of scales can be scored on the problems both of them answered. The five scales are the Qwen3 base and instruction-tuned checkpoints at $0.6b$, $1.7b$, $4b$, $8b$ and $14b$. The protocol comparison of §4.6 uses only the two smallest of each branch, because a base checkpoint in this family is published only at those two scales.

The chat arm differs in one thing: the prompt is assembled by the model's own chat template, applied to a single user turn carrying the question, rather than by the continuation string above. For the no-thinking arm the template is asked for `enable_thinking=False`; where a template does not accept that argument it is dropped and the template is used unchanged. The instruction-tuned branch was additionally collected with thinking enabled, which the base branch cannot be, having no such mode. Both branches were collected under the chat template, so the comparison in §4.6 is between two branches under one format, and then between two formats under one branch.

The code collection that §4.7 reports repeats all of this on HumanEval: the same prompt string, the same greedy decoding, the same cap, and the same hundred-problem size. Its condition is recorded per output as the single bit of whether generation reached the cap, and that field is checked row by row against the recorded token count before any reading is computed from it, so that the indicator and the quantity it stands for are the same thing. The control ladder of §4.2 is drawn in Figure 5.

D Collection drift, measured

Every reading in this paper is a property of a checkpoint together with the environment that produced the text, and that environment is not bit-reproducible. Table A2 separates the two components by replaying the same thirty questions on the original architecture and comparing generations byte for byte. Within one environment, changing only the accelerator agrees on 26 of 30; changing the

Table A2: Byte-level agreement between generation passes for `qwen3-0.6b-base`, greedy decoding, cap 512, same prompt string. The first row covers all hundred questions; the other two replay the first thirty, which is the subset both passes share.

comparison	byte-identical	what it isolates
08-22 collection vs 09-11 strong-attack arm (A40)	88/100	period and accelerator together
08-22 collection vs 09-11 replay (A800)	14/30	the same, on the original accelerator
09-11 strong-attack arm (A40) vs 09-11 replay (A800)	26/30	one environment, accelerator only

A divergence here is not a change of prompt or tokenizer: no pair differs at the first character, and the median common prefix is 26 characters for the within-environment row and 134 for the cross-period rows, which is the signature of an argmax flip cascading through the decode rather than of a different input. Readings produced by `scripts/_p2_repro_probe.py` for the replay and by a byte comparison over the two stored collections for the first row; the replay's own generations are kept at `results/p2/probe_a800_20260925.json`. The third row is the one the attack criterion of §3.5 is written against.

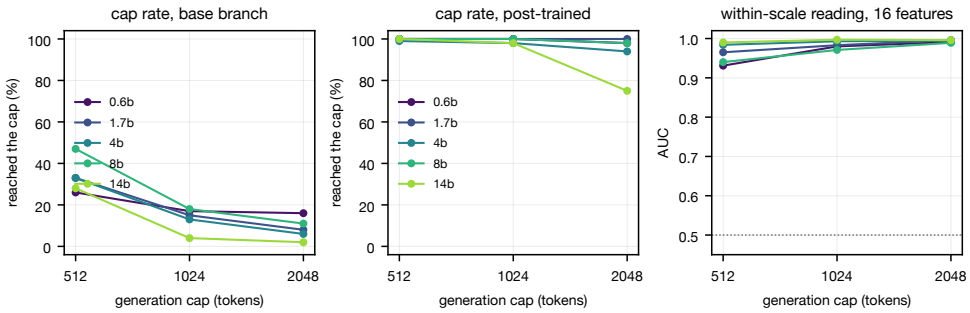

Figure 6: The cap sweep of Table A3. Left and centre: the share of outputs that reached the generation cap, for the base branch and for its post-trained counterpart. The base branch's share collapses as the budget grows — from 26–47% at 512 tokens to 2–16% at 2048 — while the post-trained branch's stays at 94–100% except at 14*b*, where it falls to 75%. Right: the within-scale reading rises with the budget rather than decaying, on the same fifteen cells.

collection period agrees on 14 of 30, and over the full hundred questions the earlier collection and the strong-attack pass agree on 88. The dominant component is therefore the library change between passes rather than the accelerator, which is why every arm here is compared only inside its own pass, and why the attack criterion of §3.5 is anchored on a within-run baseline rather than on a few points of decrease.

E CAP SENSITIVITY

The generation cap is the collection setting the reading is most obviously exposed to, because the channel we identify is whether generation reaches it. Table A3 re-collects every scale and both branches under caps of 1024 and 2048 on the same hundred problems, with the collection script otherwise unchanged, and reports the cap rate, the within-scale reading and the transfer average at each cap. Three things are worth reading off it. The channel is not specific to the cap we chose: the instruction-tuned branch reaches the cap on 94% or more of problems at four of the five scales even at 2048, where its base counterpart reaches it on 2% to 16%, and the exception is 14*b*, where the rate falls to 75% at cap 2048 while the base branch stays at 2%. The reading does not decay with the cap but grows with it, on the full feature set and on each block separately: the degenerate block leads the lexical block at every cap and scale, and the lexical block on its own rises as well (0.804 to 0.935 at 0.6*b*), so a longer output makes the residual signal, not only the terminal one, easier to read. And the uncontrolled transfer rises rather than falls, which is what the paper's thesis predicts: what a larger cap adds is more of the channel the ordinary protocol leaves uncontrolled, not more lineage.

Table A3: Cap sensitivity. The generation budget is swept from 512 to 2048 with the collection otherwise unchanged; every row is one scale at one budget, and the reading columns are the within-scale reading of the same cell. Over the eight ordered adjacent pairs the transfer average is 0.932 at cap 512, 0.969 at 1024 and 0.982 at 2048.

scale	cap	cap rate b/i	med. tok. b/i	16f	deg.	lex.	paired n
0.6b	512	0.26/1.00	233/512	0.931	0.935	0.804	52
0.6b	1024	0.17/1.00	233/1024	0.980	0.985	0.894	34
0.6b	2048	0.16/0.98	233/2048	0.991	0.991	0.935	32
1.7b	512	0.33/1.00	404/512	0.965	0.925	0.865	66
1.7b	1024	0.15/1.00	404/1024	0.983	0.964	0.956	30
1.7b	2048	0.08/1.00	404/2048	0.996	0.988	0.982	16
4b	512	0.33/0.99	284/512	0.984	0.963	0.885	66
4b	1024	0.13/0.98	284/1024	0.993	0.972	0.945	26
4b	2048	0.06/0.94	284/2048	0.995	0.993	0.975	12
8b	512	0.47/1.00	503/512	0.940	0.896	0.793	94
8b	1024	0.18/1.00	503/1024	0.971	0.948	0.897	36
8b	2048	0.11/0.98	503/2048	0.989	0.983	0.947	22
14b	512	0.28/1.00	302/512	0.990	0.973	0.875	56
14b	1024	0.04/0.98	302/1024	0.997	0.999	0.954	8
14b	2048	0.02/0.75	302/2048	0.996	0.998	0.989	4

cap rate: share of outputs that reached the cap, base / post-trained. *med. tok.*: median generated tokens the generator recorded. *16f*, *deg.*, *lex.*: within-scale reading on the full sixteen features, on the degenerate block and on the lexical block. *paired* n: size of the cell that scores both classes on the same problems. The six-feature column of Section 3.3 is in the released table. Generated by `scripts/_p2_cap_sens_analyze.py`.

Table A4: What length matching selects. Keeping only the outputs that reached the cap selects the base branch's hardest problems: the n column is that pool, the level columns are the MATH500 difficulty prior of the pool against the whole set, and the last column is a one-sided Mann–Whitney test. The lower block gives the overlap of these pools between neighbouring scales, which is the population a transfer reading moves between.

scale	pool n	pool mean	pop. mean	Δ	one-sided p
0.6b	26/100	3.73	3.21	$+0.52$	0.0355
1.7b	33/100	3.67	3.21	$+0.46$	0.0442
4b	33/100	4.00	3.21	$+0.79$	0.0015
8b	47/100	3.70	3.21	$+0.49$	0.0181
14b	28/100	3.71	3.21	$+0.50$	0.0372
overlap of the pools between neighbouring scales					
pair	intersection	Jaccard			
0.6b $\leftrightarrow$ 1.7b	13/46	0.28			
1.7b $\leftrightarrow$ 4b	21/45	0.47			
4b $\leftrightarrow$ 8b	24/56	0.43			
8b $\leftrightarrow$ 14b	24/51	0.47			

F What length matching selects

Keeping only the outputs that reached the cap is the standard way to hold length fixed, and Table A4 measures what else it holds fixed. Because the branch determines which problems produce a long output — the base branch writes to the cap only where it fails or loops, the post-trained branch almost everywhere — the surviving base pool is the hardest part of the set: its mean MATH500 level is 0.46 to 0.79 points above the population in all five scales, with a one-sided Mann–Whitney p between 0.0015 and 0.044. The lower block gives the overlap of these pools between neighbouring scales: a transfer reading moves between populations that share only 28% to 47% of their members, which is the mechanism behind the asymmetry of rows 2 and 3 of Table 1. Pairing the questions removes the selection by construction, and it is the control that changes the cross-scale conclusion.

Table A5: External and control detectors under the same audit. Each row is a detector; the four column groups are the cells of the ladder: *none* is uncontrolled, *lev.* keeps only capped outputs, *pair* adds question-id pairing and *trunc.* truncates the prefix to 200 characters. Text-only methods lose the reading as the controls are added; the two readouts that use the model's own logits keep more of it, and the cap bit is undefined once every remaining output has reached the cap.

	$0.6b$				$1.7b$			
detector	none	lev.	pair	trunc.	none	lev.	pair	trunc.
16 features (ours)	0.924	0.802	0.719	0.746	0.962	0.920	0.879	0.747
TF-IDF + LR	0.952	0.887	0.825	0.691	0.944	0.896	0.756	0.788
Burrows' Delta	0.862	0.870	0.726	0.518	0.901	0.835	0.730	0.612
token entropy	0.528	0.867	0.802	0.722	0.711	0.793	0.777	0.600
token NLL	0.549	0.863	0.809	0.788	0.720	0.793	0.785	0.537
cap bit	0.870	—	—	—	0.835	—	—	—
length	0.870	—	—	—	0.835	—	—	—

Table A6: The same detectors in transfer: fit on one scale, scored on the other. The signed AUC is reported, so a value near or below 0.5 is a failed transfer; the grey-box readouts are labelled in the text.

	$0.6b \to 1.7b$				$1.7b \to 0.6b$			
detector	none	lev.	pair	trunc.	none	lev.	pair	trunc.
16 features (ours)	0.911	0.828	0.755	0.589	0.876	0.598	0.643	0.355
TF-IDF + LR	0.942	0.903	0.925	0.521	0.950	0.860	0.917	0.670
Burrows' Delta	0.870	0.875	0.817	0.664	0.912	0.843	0.899	0.682
token entropy	0.711	0.793	0.777	0.600	0.528	0.867	0.802	0.722
token NLL	0.280	0.207	0.215	0.537	0.451	0.137	0.191	0.212
cap bit	0.835	—	—	—	0.870	—	—	—
length	0.835	—	—	—	0.870	—	—	—

G External detectors under the same audit

The ladder of §4.2 is a property of the apparatus, not of our classifier, so it should move other detectors in the same way. Table A5 and Table A6 run three text-only detectors that a black-box lineage claim would plausibly use — the 16 features of §3.3, a TF-IDF classifier over word and character n-grams, and Burrows' Delta, the standard method of the authorship-attribution literature — together with two readouts that use the model's own logits (mean token entropy and mean token NLL) and two degenerate controls (the cap bit and the raw length). Every text-only detector is strong uncontrolled (0.86 to 0.96) and loses the reading as the controls are added: within a scale they fall to 0.52 to 0.79 in the paired and truncated cell, and in transfer to 0.36 to 0.68. The cap bit cannot be scored in those cells at all, because every surviving output has reached the cap. The two logits readouts split: mean token entropy keeps 0.60 to 0.72 in transfer, while mean token NLL falls to 0.21 to 0.54; neither is available in the black-box setting, and what the entropy reads is the repetition pathology directly rather than a calibrated text statistic.

H The twenty ordered pairs in full

§4.3 reports the ordered pairs as an average across the twenty and a count of how many meet each criterion. Tables A7 and A8 print the cells. The pairs are written train $\to$ test, and each row is one cell of the double-paired construction: the same problems on both sides, the same visible length, and the same bucket of outputs selected by whether they reached the cap. The n column is that cell's size, and it is not constant across rows, because the bucket and the pairing do not retain the same number of problems for every pair.

Two features of these tables are worth pointing at, because the averages in the body hide them. The only pair that survives correction lies in the full feature set, and the only pair whose interval lies entirely below zero on the other side lies in the length-free set, so the two feature groups do not agree about which direction the anomaly runs. And the pair that comes closest to meeting the first

Table A7: The twenty ordered scale pairs on the full sixteen-feature set, scored on one double-paired cell. Pairs are written train → test. Δ is the within-scale reading minus the cross-scale reading, so a positive Δ means the pair reads lower across scales than within. H1 asks whether the cross-scale interval lies entirely below 0.5; H2 whether the Δ interval lies entirely above zero. In the H2 column *reverse* marks the one row per table whose interval lies entirely below zero, that is, the pair that reads higher across scales than within. The last column is H2 after Holm correction over the pooled forty one-sided p-values.

pair	n	within	cross	Δ	Δ 95% CI	H1	H2	H2 Holm
0.6b → 1.7b	66	0.747	0.590	+0.159	[+0.009, +0.320]	no	yes	no
0.6b → 4b	66	0.737	0.579	+0.160	[-0.000, +0.321]	no	no	no
0.6b → 8b	94	0.661	0.592	+0.069	[-0.084, +0.219]	no	no	no
0.6b → 14b	56	0.865	0.716	+0.149	[+0.038, +0.273]	no	yes	no
1.7b → 0.6b	52	0.746	0.355	+0.390	[+0.151, +0.605]	no	yes	no
1.7b → 4b	66	0.737	0.800	-0.063	[-0.167, +0.031]	no	no	no
1.7b → 8b	94	0.661	0.690	-0.031	[-0.149, +0.085]	no	no	no
1.7b → 14b	56	0.865	0.843	+0.024	[-0.123, +0.162]	no	no	no
4b → 0.6b	52	0.746	0.364	+0.382	[+0.156, +0.588]	no	yes	yes
4b → 1.7b	66	0.747	0.752	-0.005	[-0.124, +0.120]	no	no	no
4b → 8b	94	0.661	0.776	-0.114	[-0.239, +0.012]	no	no	no
4b → 14b	56	0.865	0.765	+0.102	[-0.050, +0.243]	no	no	no
8b → 0.6b	52	0.746	0.484	+0.261	[+0.030, +0.482]	no	yes	no
8b → 1.7b	66	0.747	0.705	+0.039	[-0.066, +0.144]	no	no	no
8b → 4b	66	0.737	0.641	+0.095	[-0.032, +0.223]	no	no	no
8b → 14b	56	0.865	0.640	+0.223	[+0.058, +0.386]	no	yes	no
14b → 0.6b	52	0.746	0.677	+0.070	[-0.088, +0.224]	no	no	no
14b → 1.7b	66	0.747	0.665	+0.083	[-0.054, +0.226]	no	no	no
14b → 4b	66	0.737	0.805	-0.067	[-0.187, +0.054]	no	no	no
14b → 8b	94	0.661	0.633	+0.028	[-0.107, +0.163]	no	no	no

Counts on this feature group: H1 0 of 20; H2 before correction 6 of 20; H2 after Holm 1 of 20; Δ interval entirely below zero 0 of 20. The pair that survives Holm is 4b→0.6b.

criterion is the one whose cross-scale reading is lowest: the pair that transfers from the larger scale down to the smaller one, whose interval ends just above the line.

I THE DISGUISE ARM IN FULL

§4.5 reports a withdrawal and the control that forced it. Table A9 prints both scales and every control in that run, including the construction as it was first reported.

The order of the rows is the order of the argument, and each row rules out one explanation for the reading below it. The first is the reading as first published. The next two measure, without any attack at all, whether a text that has been perturbed can be told apart from one that has not, and they say that it can: this is what makes the first row a property of the construction rather than of the two branches. The symmetrised row then adds the same perturbation to both classes, and it is the number the paper reports. The three rows after it drop one family of features at a time from that cell, which is the question of whether the reading is carried by any single family; it is not, and neither is it carried by the lexical features alone. The last two rows take the construction in two further directions, one of which removes the length difference between the classes and one of which repeats the original one-sided error on a differently built decoy.

J THE CHAT-TEMPLATE ARM IN FULL

§4.6 reports the protocol comparison as a difference of cap rates in two directions. Table A10 prints every arm collected for it.

Table A8: The twenty ordered scale pairs on the length-free four, scored on one double-paired cell. Pairs are written train → test. Δ is the within-scale reading minus the cross-scale reading, so a positive Δ means the pair reads lower across scales than within. H1 asks whether the cross-scale interval lies entirely below 0.5; H2 whether the Δ interval lies entirely above zero. In the H2 column *reverse* marks the one row per table whose interval lies entirely below zero, that is, the pair that reads higher across scales than within. The last column is H2 after Holm correction over the pooled forty one-sided p-values.

pair	n	within	cross	Δ	Δ 95% CI	H1	H2	H2 Holm
0.6b → 1.7b	66	0.696	0.485	+0.214	[+0.046, +0.391]	no	yes	no
0.6b → 4b	66	0.615	0.582	+0.033	[-0.106, +0.176]	no	no	no
0.6b → 8b	94	0.461	0.533	-0.071	[-0.221, +0.080]	no	no	no
0.6b → 14b	56	0.746	0.786	-0.041	[-0.168, +0.082]	no	no	no
1.7b → 0.6b	52	0.719	0.476	+0.241	[-0.036, +0.478]	no	no	no
1.7b → 4b	66	0.615	0.568	+0.045	[-0.147, +0.232]	no	no	no
1.7b → 8b	94	0.461	0.689	-0.230	[-0.360, -0.099]	no	reverse	no
1.7b → 14b	56	0.746	0.607	+0.136	[-0.048, +0.325]	no	no	no
4b → 0.6b	52	0.719	0.574	+0.143	[-0.054, +0.338]	no	no	no
4b → 1.7b	66	0.696	0.625	+0.070	[-0.073, +0.223]	no	no	no
4b → 8b	94	0.461	0.518	-0.057	[-0.222, +0.117]	no	no	no
4b → 14b	56	0.746	0.792	-0.046	[-0.117, +0.028]	no	no	no
8b → 0.6b	52	0.719	0.630	+0.087	[-0.106, +0.280]	no	no	no
8b → 1.7b	66	0.696	0.525	+0.171	[-0.010, +0.365]	no	no	no
8b → 4b	66	0.615	0.487	+0.127	[-0.113, +0.348]	no	no	no
8b → 14b	56	0.746	0.420	+0.323	[+0.099, +0.547]	no	yes	no
14b → 0.6b	52	0.719	0.587	+0.131	[-0.029, +0.294]	no	no	no
14b → 1.7b	66	0.696	0.586	+0.112	[-0.040, +0.264]	no	no	no
14b → 4b	66	0.615	0.644	-0.029	[-0.117, +0.052]	no	no	no
14b → 8b	94	0.461	0.550	-0.089	[-0.257, +0.076]	no	no	no

Counts on this feature group: H1 0 of 20; H2 before correction 2 of 20; H2 after Holm 0 of 20; Δ interval entirely below zero 1 of 20. No pair survives Holm on this group.

The table is worth reading down the cap-rate column rather than across it, because the comparison the body reports is a difference between two rows and the difference is not stable in either the format or the scale. Under the continuation format the instruction-tuned branch caps on a larger share of problems than its base at both scales. Under the chat template that ordering reverses at the smaller scale and all but vanishes at the larger one, so the direction of the effect is a setting of the instrument rather than a property of the branches. The last column records whether a chat template was requested and could not be applied, and it is zero in every collected arm, so no row here is a continuation row wearing a chat label. The criterion pre-registered for this arm asked a narrower question than the table can answer, and landed inside its undetermined band, which is why the arm is reported as a reading rather than as a result.

K THE DEGENERATE CLASSIFIER ON HUMANEVAL

§4.7 asks whether the degeneracy channel is a property of these two branches or of the mathematics collection. Table A11 gives the full reading: the IT branch caps on 100% of problems at $0.6b$ and 99% at $1.7b$, and the single bit of whether generation reached the cap separates the two branches with an AUC of 0.965 and 0.990, clearing the threshold we set before collecting.

The single bit of whether generation reached the cap separates the two branches on this task as well, against a threshold that was fixed before the collection and is not adjusted here. What the table also makes visible is why the paper reports the single-bit number rather than the full-length within-task reading beside it. On the instruction-tuned side the bit has almost no variance, because generation reaches the cap on essentially every problem, so a classifier built on it is close to a constant; that full-length reading is the same degeneracy seen from the other side, and it says only that the two branches can be told apart once the length of the output is available to the classifier. The single-bit

Table A9: The disguise arm in full, both scales, in the order of the argument: the construction as first reported, the two measurements of the leakage it contains, the symmetrised comparison, the three feature-drop variants of that comparison, and two further constructions that repeat the same error. Every number comes from one run of the leakage control or of the attack script.

setting	$0.6b$	$1.7b$	n	what it is
disguise, decoys on one class only (as first reported)	1.000	1.000	200	the reading as first reported
same, with the token unit aligned	0.999	1.000	200	the same construction, token unit corrected
can a classifier tell the text was decoyed, base side	0.997	1.000	200	the leakage measured directly
can a classifier tell the text was decoyed, IT side	1.000	1.000	200	the same measurement on the other branch
decoys on both classes	0.950	0.938	200	the decisive symmetrised comparison
symmetrised, length family dropped	0.935	0.949	200	does the length family account for it
symmetrised, repetition family dropped	0.937	0.919	200	does the repetition family account for it
symmetrised, lexical features only	0.832	0.913	200	do the remaining lexical features account for it
one-sided, length-matched to the cap	0.834	0.906	129	length no longer separates the classes
one-sided, decoy prefix on one class only	1.000	1.000	200	the same error in a second construction

The first row reproduces the construction as first reported: the decoys were added to the base class only, which makes *was this perturbed* a perfect stand-in for the label, and the reading is 1.000. The decoy-detector rows measure that stand-in directly and put it at or above 0.997 in all four cells, so the 1.000 is a leakage reading rather than an attack result. Adding the same decoys to both classes is the decisive comparison and reads 0.950 and 0.938. Dropping the length family, the repetition family, or all but the lexical features from that cell leaves the reading between 0.832 and 0.949, so no single family accounts for it. Removing the length difference between the classes on the one-sided construction drops it from 1.000 to 0.834 and 0.906, which is the same statement from the other side.

number says which part of the length carries the separation, and that is the part the criterion was written about.

L TWO MORE PROTOCOLS, TWO MORE FAMILIES

Three further controls test whether what we have described is a property of Qwen3 base/IT pairs or a property of this measurement.

The protocol arm is in §3.1 and we do not repeat it: under a chat template the cap-rate difference between the branches runs the other way, and the statement we are willing to make is a scope limit rather than a result.

The family arm is Table A12 and Table A13, whose cap rates Figure 10 collects side by side. Mistral-7B, which has an independent tokenizer, shows the same reversal in the cap channel as the chat arm: its base caps on 62% of problems and its Instruct variant on 23%, a difference of -39 points, against +65 to +71 on Qwen3. The decomposition points the same way in both of its cells, with the lexical block above the degenerate one, and the criterion-3 cell reproduces the published composition and reading before it decomposes, so the decomposition runs on the same instrument as the number it decomposes.

OLMo-2-1B is the stronger version of the same test, because four post-training recipes — sft, dpo, instruct and rlvr — were run on one shared base, so nothing in the comparison varies except the recipe. The base caps at 53%; every recipe caps less, at 30%, 34%, 35% and 31%, and in all four the confidence interval of the difference lies entirely below zero. The decomposition agrees with

Table A10: Every arm collected for the protocol comparison: cap rate, median length, the tenth and ninetieth percentiles of length, the fraction of empty outputs, and the number of prompts for which a chat template was requested and could not be applied. That last column is zero in every collected arm: no row here is a continuation row in disguise.

model	arm	n	cap rate	median	p_{10}	p_{90}	empty	template fallback
0.6b (base)	continuation	100	0.29	228	68	512	0.00	0
0.6b (base)	chat, no thinking	100	0.98	512	512	512	0.00	0
0.6b	continuation	100	1.00	512	512	512	0.00	0
0.6b	chat, no thinking	100	0.37	354	151	512	0.00	0
0.6b	chat, thinking	100	1.00	512	512	512	0.00	0
1.7b (base)	continuation	100	0.35	405	169	512	0.00	0
1.7b (base)	chat, no thinking	100	0.43	486	31	512	0.00	0
1.7b	continuation	100	1.00	512	512	512	0.00	0
1.7b	chat, no thinking	100	0.41	422	189	512	0.00	0
1.7b	chat, thinking	100	1.00	512	512	512	0.00	0

Under the continuation format the instruction-tuned branch caps more than its base at both scales, by 71 and 65 points. Under the chat template the same comparison gives base 0.98 and 0.43 against IT 0.37 and 0.41, which is -61 and -2 points: the difference changes sign at $0.6b$ and closes at $1.7b$. The base branch was never post-trained to answer a chat template, so its chat outputs are out of distribution for it; the template itself was available for every arm collected here, which the last column records. The criterion pre-registered for this arm asked a different question - whether the IT branch still caps under chat - and allowed 20–80% as an undetermined band; both scales land inside that band, so the arm is recorded as undetermined rather than read in either direction.

Table A11: HumanEval, which is code generation rather than competition mathematics. Base and IT cap rates with Wilson 95% intervals, the cap-rate difference in points, the single-bit separator with its true and false positive rates, and the full-length within-task reading.

scale	base caps	95% CI	IT caps	95% CI	Δcap (pt)
0.6b	7/100 (7.00%)	[3.43, 13.75]	100/100 (100.00%)	[96.30, 100.00]	+93
1.7b	1/100 (1.00%)	[0.18, 5.45]	99/100 (99.00%)	[94.55, 99.82]	+98

scale	single-bit AUC	TPR	FPR	within AUC	95% CI	n (within)
0.6b	0.9650	1.0000	0.0700	0.964	[0.933, 0.990]	200
1.7b	0.9900	0.9900	0.0100	1.000	[0.999, 1.000]	200

The IT side reaches the cap on every problem at $0.6b$ and on 99 of 100 at $1.7b$, so the single-bit indicator has almost no variance on that side and its AUC is close to a constant score rather than a learned one. The full-length within-task reading at $1.7b$ is 1.000, which is the same degeneracy seen from the other side, and is why the body reports the single-bit number instead: it says what carries the separation, whereas 1.000 says only that the two classes can be told apart once length is available to the classifier.

Mistral rather than with Qwen3: the degenerate block falls below the lexical block in all four arms in the controlled cell, and at full length already in three of the four.

So the correction to our own claim is specific. The degenerate channel carries the reading for Qwen3 base/IT pairs under the continuation protocol, and in the two other families we can test, with five post-training recipes between them, it does not.

M THE CARDINALITY-MATCHED NULL

The claim that the degenerate block carries the strong face is a claim about a specific set of seven features, and any seven of the sixteen might do as well; this section calibrates it against all 11,440 seven-feature subsets, scored on the same cells. Table A14 prints each block's gap, the null it is compared against, and the share of subsets that come at least as close to the full set.

Table A12: Mistral-7B under the continuation protocol, block decomposition. In *both* cells the degenerate block scores below the lexical block, so what separation remains in this family is not carried by the block that carries Qwen3's.

cell	n	ALL	DEG
full length (crit. 2)	200	0.914 [0.876, 0.948]	0.654 [0.574, 0.728]
capqid + L=200 (crit. 3)	82	0.887 [0.813, 0.949]	0.707 [0.587, 0.820]

cell	n	LEX	REP
full length (crit. 2)	200	0.788 [0.719, 0.852]	0.798 [0.727, 0.861]
capqid + L=200 (crit. 3)	82	0.891 [0.817, 0.956]	0.849 [0.762, 0.921]

The criterion-3 cell reproduces the published composition ($n = 82$, base 62 / IT 20) and the published ALL $=$ 0.887, so the decomposition runs on the same instrument as the number it decomposes. The 95% CIs of DEG and LEX overlap in both cells as well.

Table A13: OLMo-2-1B, a second family with four post-training recipes on one shared base, under the continuation protocol. Δcap is the arm's cap rate minus the base's, in points, printed once per arm; the block columns are scored separately in the full-length cell and in the criterion-3 cell (question-id pairing with the base's capped ids, $L = 200$), whose size is the n column.

arm	Δcap (pt)	cell	n	ALL	DEG	LEX	REP	DEG<LEX
sft	-23 [-36, -10]	full length	200	0.798	0.756	0.753	0.736	no
sft		capqid + L=200	78	0.708	0.644	0.666	0.672	yes
dpo	-19 [-32, -6]	full length	200	0.868	0.743	0.817	0.710	yes
dpo		capqid + L=200	76	0.922	0.406	0.941	0.839	yes
instruct	-18 [-32, -4]	full length	200	0.940	0.847	0.881	0.652	yes
instruct		capqid + L=200	74	0.901	0.587	0.889	0.818	yes
rlvr	-22 [-35, -9]	full length	200	0.882	0.750	0.826	0.679	yes
rlvr		capqid + L=200	75	0.921	0.437	0.918	0.847	yes

The base caps at 53/100; every arm caps below it, so the cap-rate difference is negative for all four recipes with a CI entirely below zero - the opposite of the $+65$ to $+71$ points measured on Qwen3. At full length DEG already sits below LEX in three of the four arms, and in the fourth `sft` the two differ by 0.003; under the criterion-3 cell it falls below LEX in all four, and in `dpo` and `rlvr` its AUC is below chance. This is the second family to disagree with Qwen3's decomposition, and the only one where four recipes vary independently while the base weights are held fixed.

N Discussion in full

The apparatus decides the reading. Four times in this paper the same classifier, over the same feature set, on the same two branches, gives a different answer, and the only thing that changes is a setting of the instrument.

Held at full length and left unpaired, transfer across scales reads 0.885 and 0.892, and that is the number our earlier version reported as evidence of a scale-general lineage signal; held to the problems both branches answered and to a fixed visible length, the same transfer reads 0.487 and 0.458. Under a chat template the cap-rate difference between the branches changes sign at $0.6b$ and disappears at $1.7b$, and in two further families with five post-training recipes between them, the block that carries the Qwen3 reading carries nothing.

None of these is the correction of an arithmetic mistake. Each is the same arithmetic performed under a different setting, and that is the point we want to leave behind. A number in this literature is not a property of a model pair. It is a property of a model pair together with an apparatus, and the apparatus is usually the part that goes unreported.

Why the two faces are one. The title asks why a lineage fingerprint can look both robust and fragile. The answer our readings support is that neither property belongs to the lineage signal as such. Both belong to one channel: whether the model terminates, and how much text it emits before it does.

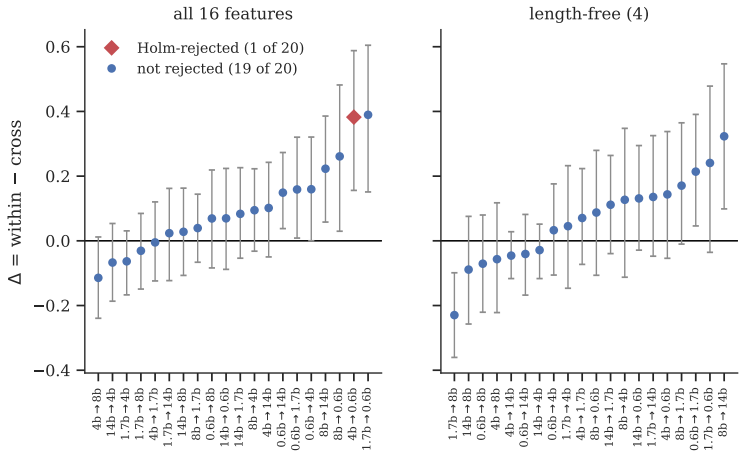

Figure 7: The twenty ordered pairs of Table A15, plotted as the paired difference between the within-scale and the cross-scale reading, with 95% bootstrap intervals on that difference. Six pairs clear the line before correction and one survives Holm, drawn in red. Almost every interval covers zero: on this basis transfer is not uniformly below within-scale, and it is not reliably above chance either.

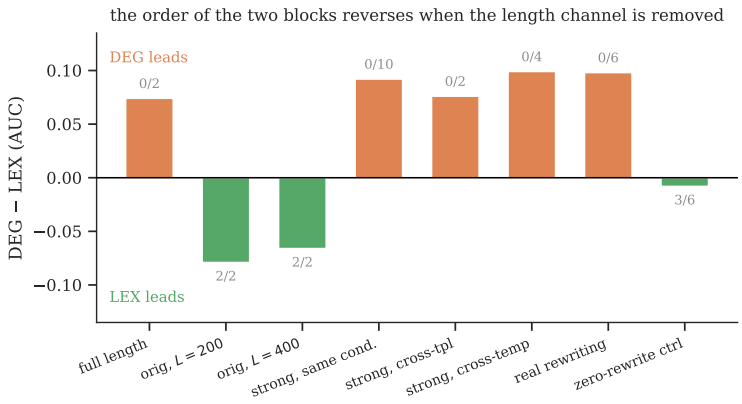

Figure 8: The block decomposition of Table A16 as a difference: each bar is the mean AUC of the degenerate block minus the mean of the lexical block over the cells of that condition, so a bar above zero means the degenerate block leads, and the labels count the cells where it scores below. Length and the two terminal markers lead at full length and in every attack that leaves the output long, and yield as soon as the visible text is truncated.

The channel explains the robust face, because it is where the near-ceiling reading mostly comes from — at full length the block built from length and the terminal markers scores 0.913 against the full set's 0.944. It explains the fragile face, because the ordinary perturbations leave exactly that channel alone: they preserve how much text is emitted and whether generation stops, and where a rewriting arm did drive the reading down, it did so by compressing the output to about a third of its length. And it explains why transfer collapses under control: pairing the questions and levelling the length removes the two things the channel was encoding, and what is left is a problem-difficulty statistic that does not carry across scales.

The correction runs in both directions and we state it in both. We do not claim the channel is an artifact that invalidates the readings. Within a scale, at a fixed length, a residual lexical signal separates the branches at 0.719 and 0.696, and that is a real reading of a real difference. What we claim is narrower and harder to avoid: a reading taken without controlling this channel is measuring the channel, and the number it returns does not tell you how much of it was lineage.

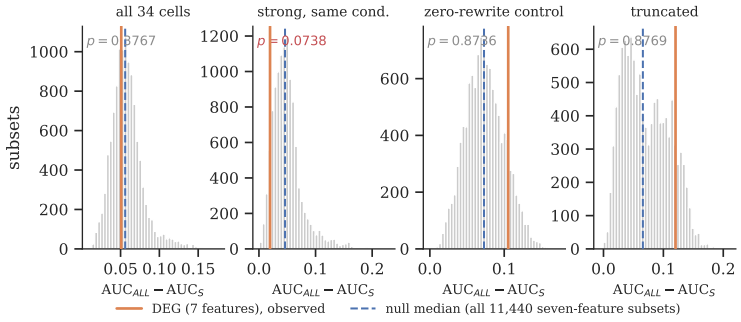

Figure 9: The cardinality-matched null of Table A14, in the appendix. Each panel is the histogram of the gap between the full set's AUC and a subset's AUC over all 11,440 seven-feature subsets, on one group of cells; the solid line is the degenerate block's own value and the dashed line is the null median. The degenerate block does not come closer to the full set than 7.4% of subsets do on the strong-attack cells, and it sits further from it than a typical subset on the truncated and zero-rewrite cells. This figure calibrates the decomposition and does not replace it.

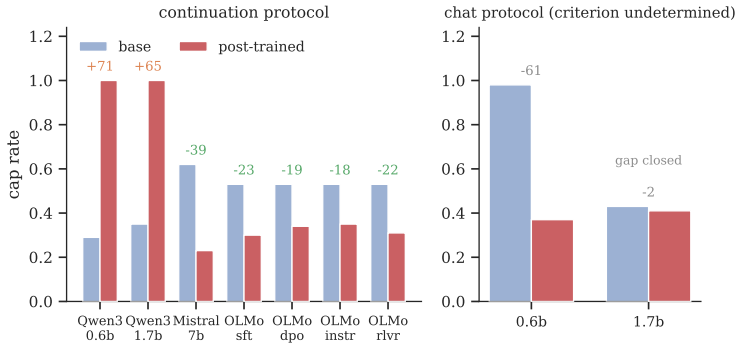

Figure 10: Cap rates for the base branch and its post-trained counterpart, under the continuation protocol (left) and under a chat template (right). Under continuation the post-trained branch caps more than its base at both Qwen3 scales, by 71 and 65 points, and less in every other family we collected: 39 points in Mistral and 18 to 23 in the four OLMo recipes that share one base. Under the chat template the comparison reverses at 0.6*b* and closes to 2 points at 1.7*b*; the criterion pre-registered for that arm landed in its undetermined band, so the right panel is reported as a reading and not as a result.

What we would ask a lineage claim to report. If a reading is a property of a pair together with an apparatus, then an apparatus that is not reported is not reproducible, and the fault is not in the reader. The controls in §3.4 are what we would ask to be reported with any black-box lineage claim, and the ladder in Table 1 is what happens when they are.

Four things carry most of the weight. The termination and cap rate for both branches, under the exact format the outputs were collected in, because that number predicts the reading more reliably than any property of the models. Whether the two classes were scored on the same problems, because length matching on its own leaves the base branch represented only by the problems it failed. The visible length at which the reading was taken, and whether it was taken at more than one. And whether the reading survives on features that carry neither length nor repetition — if it does not, the claim is about generation behaviour, which is worth making, but it is a different claim.

Our recommendation is drawn from the ladder and the decomposition, which we ran under the criteria of §3.5. The attack arms support it only within their own scope, and we hold them to it: the cross-template and cross-temperature cells are unpaired by question and are read as "the reading does not decrease", which is weaker than robustness, and the rewriting arm is read against its own control. We tested adversaries that work on the surface of the text. We did not test an adversary that targets the termination channel directly, and we would expect one to succeed.

Table A14: Reading 1 — "the strong face is carried by DEG" — under a cardinality-matched null: all 11,440 seven-feature subsets of the 16, scored on the same cells. Gap is $\mathrm{AUC}_{\mathrm{ALL}} - \mathrm{AUC}_S$ averaged within the block, so a smaller gap means closer to the full set.

block	cells	DEG gap	null mean	null med.	≤DEG	p
all 34 cells	34	0.0509	0.0583	0.0559	4309	0.3767
full length (2)	2	0.0304	0.0477	0.0425	3427	0.2996
strong, same cond. (10)	10	0.0196	0.0503	0.0461	843	**0.0738**
strong, all (16)	16	0.0184	0.0485	0.0428	1205	0.1054
real rewriting (6)	6	0.0433	0.0633	0.0618	1865	0.1631
zero-rewrite (6)	6	0.1054	0.0742	0.0728	9994	0.8736
truncated (4)	4	0.1203	0.0714	0.0658	10032	0.8769

No block reaches $p < 0.05$; the closest is 0.0738, which does not survive a $\times 7$ correction. On the zero-rewrite and truncated blocks the direction reverses: DEG sits *further* from the full set than a typical seven-feature subset. This table calibrates Reading 1 and does not replace it; the main evidence stays the flip in Table A16.

Table A15: The five-scale extension: 20 ordered pairs. H1: the cross-scale CI lies entirely below 0.5. H2: the paired Δ CI lies entirely above 0. Holm is applied to the pooled 40 one-sided p-values.

features	within mean	cross mean	cross range	H1	H2 uncorr.	H2 Holm
ALL (16)	0.751 (5)	0.654 (20)	[0.355, 0.843]	**0/20**	6/20	**1/20**
LEX (4)	0.648 (5)	0.583 (20)	[0.420, 0.792]	**0/20**	2/20	**0/20**

The nearest miss on H1 is $1.7b \to 0.6b$ on ALL, upper bound 0.5120 — 0.012 above the line. Exactly one pair survives Holm, and it lies in ALL; none does in the length-free group. On the opposite side, one pair is significantly *higher* across scales than within (LEX, $1.7b \to 8b$); a second, $4b \to 8b$ on ALL, reaches one-sided $p = 0.0405$ but its paired-Δ CI still covers zero, so it does not meet the criterion this paper uses.

The conclusion, as it stood before the body was compressed:

A lineage fingerprint is usually reported as one number. We have tried to show what that number is made of. Across five scales, twenty ordered pairs, three families and a second task, the readings move further under changes to the instrument than under attacks on the text, and the part of the signal that survives every control is smaller than the part that does not.

The apparatus is not a nuisance to be averaged away: the same pair of models separates at 0.941 within a scale and at chance across scales, and the difference is not a property of the models but of which problems each branch was asked and how much text each of them wrote.

What survives is worth having, and worth stating carefully. Within a scale, at a fixed length, on the same problems, a residual signal separates the branches, and no surface attack we ran removes it. That is a real measurement of a real difference, and it is the claim we are prepared to defend. What we are not prepared to defend, and would ask others not to state without the controls that make it checkable, is that this signal carries across scale, across protocol, or across families.

O LIMITATIONS IN FULL

One collection environment. Every arm in a given pass was collected under one version of the inference stack, and those versions moved across passes. Arms collected in different passes are therefore not strictly comparable, which is why the two families of §4.6 were each collected in a single pass and are compared inside themselves. We did not re-collect the full five-scale set under a single environment, and doing so would be the first thing we would add.

Three families, five recipes. The family-level claim rests on Qwen3, Mistral and OLMo, with one post-training recipe in the second of those and four in the third. That is enough to show that the Qwen3 decomposition does not generalise, and not enough to say what decides whether a family follows it.

Table A16: The 16 features partitioned into three mutually exclusive, exhaustive blocks — DEG (length and density plus the terminal marker, 7), LEX (lexical residual, 4), REP (self-repetition, 5) — each scored alone. Values are block means over the cells listed; the last column counts cells where the degenerate block scores below the lexical block.

condition block	cells	ALL	DEG	LEX	REP	DEG<LEX
original output, full length	2	0.944	0.913	0.839	0.749	0/2
original output, $L = 200$	2	0.747	0.640	0.719	0.559	2/2
original output, $L = 400$	2	0.764	0.631	0.697	0.592	2/2
strong attack, same condition	10	0.936	0.916	0.824	0.705	0/10
strong attack, cross-template	2	0.892	0.884	0.808	0.666	0/2
strong attack, cross-temperature	4	0.947	0.927	0.828	0.706	0/4
real rewriting	6	0.741	0.698	0.600	0.568	0/6
zero-rewrite control	6	0.784	0.679	0.687	0.619	3/6

At full length DEG leads; once the length channel is removed by truncation the order flips. Excluded from every block are the two `xfer_full` cells, which have no within-scale ALL because transfer is not a within-scale quantity; the remaining 34 cells are partitioned exhaustively.

Table A17: The attack surface. No attack produces a discernible decrease in the reading. The right-hand column gives why each one fails to: none of them touches the termination and length channel that carries it.

attack	AUC	n	reading
disguise, one-sided (as first reported)	1.000	200	withdrawn: the control is one-sided
disguise, symmetrised, $0.6b$	0.950	200	correct reading
disguise, symmetrised, $1.7b$	0.938	200	correct reading
decoy detector ("was this perturbed?")	0.997	200	leakage upper bound
template perturbation, cross-template	0.908 / 0.876	200	no decrease
temperature $0.0 \rightarrow 0.7$	0.958 / 0.962	200	no decrease
answer rewriting, transfer	0.774 / 0.754	200	$\lvert\Delta\rvert \leq 0.013$ in 4 of 6 cells
zero-rewrite control (c')	0.787 / 0.762	200	collapses to 0.342 in one cell
cross-lingual, output pinned	0.945 / 0.960	100	no decrease

The as-first-reported disguise row is reproduced only to be withdrawn: its control was applied to one side only, which is a leakage construction rather than an attack. The decoy-detector row measures that leakage directly. The cross-template and cross-temperature cells are unpaired by question id and are read as "the reading does not decrease", a weaker statement than robustness. Rewriting is read against its own zero-rewrite control, which applies the same transformation to already-identical text; the control collapses in one cell because style transfer shortens the output to a median of 0.32 of its length, so the load-bearing statement there is that the *control* collapses, not that the attack fails. The cross-lingual row is the output-pinned arm, in which the prompt language changes but the answer language is fixed; without that pinning the two changes cannot be told apart and the arm is not readable as an attack. The output-side Chinese fraction in the translated arm is 0.13 at $0.6b$ and 0.15 at $1.7b$ ($n = 200$ each), against 0.00 in the untranslated arm, so translating the prompt does not turn the answers Chinese.

The chat arm is undetermined. The criterion we pre-registered for the chat template asked whether the IT branch still caps, and the measured rate fell inside the open band, so the arm is recorded as undetermined. Two further limits apply to that arm: the base branch was never chat-post-trained, so its chat-template outputs are out of distribution, and only the nothink arms have a base counterpart, so the think arms are not paired.

Two tasks. MATH500 and HumanEval. The second was chosen to leave mathematics, and it does, but its separation is carried by a cap indicator with almost no variance on the IT side, so it confirms that the classifier travels and not much more. We have no third task and no non-mathematical task with a graded rather than a binary separation.

Table A18: Family profile across the five Qwen3 scales under the unified protocol. *think effect* is the thinking pass accuracy minus the no-think pass accuracy on identical weights. n_{SC} is the number of problems on which the self-consistency columns are computed.

scale	n	len. ratio	think eff.	SC maj.	SC cons.	n_{SC}	overthink
0.6b	200	7.99	+0.065	0.62	0.81	190	2.62
1.7b	200	7.74	+0.045	0.66	1.00	161	1.98
4b	200	7.02	-0.010	0.66	1.00	161	2.06
8b	200	7.56	-0.020	0.64	0.88	152	1.94
14b	100	6.77	-0.080	0.72	1.00	87	1.50

The think effect is positive at the two smallest scales and negative from $4b$ on. n_{SC} is the post-backfill caliber; earlier drafts quoted $39/39/39/47$, which is the same measurement under an older caliber, not a second one.

Table A19: Within-scale discrimination on the double-paired cell (cap bucket + question-id pairing + prefix truncation L), which removes length and problem difficulty together. Ranges are 95% bootstrap CIs; n is the cell size at $L = 200$.

scale	features	n	$L = 200$	$L = 400$	$L = 800$
0.6b	ALL (16)	52	0.746 [0.601, 0.869]	0.754 [0.606, 0.881]	0.754 [0.608, 0.883]
	LEX (4)	52	0.719 [0.564, 0.848]	0.697 [0.534, 0.830]	0.619 [0.455, 0.774]
1.7b	ALL (16)	66	0.747 [0.618, 0.865]	0.834 [0.726, 0.923]	0.858 [0.757, 0.945]
	LEX (4)	66	0.696 [0.563, 0.820]	0.658 [0.524, 0.793]	0.758 [0.625, 0.883]
4b	ALL (16)	66	0.737 [0.602, 0.863]	0.901 [0.813, 0.972]	0.937 [0.867, 0.988]
	LEX (4)	66	0.615 [0.469, 0.754]	0.852 [0.747, 0.938]	0.860 [0.758, 0.943]
8b	ALL (16)	94	0.661 [0.545, 0.768]	0.647 [0.528, 0.761]	0.718 [0.611, 0.819]
	LEX (4)	94	0.461 [0.343, 0.582]	0.481 [0.366, 0.604]	0.630 [0.506, 0.748]
14b	ALL (16)	56	0.865 [0.749, 0.958]	0.898 [0.805, 0.967]	0.966 [0.918, 0.994]
	LEX (4)	56	0.746 [0.608, 0.872]	0.883 [0.783, 0.960]	0.953 [0.893, 0.992]

The reading is not stable in L. At $4b$ the length-free column moves from 0.615 to 0.852 between $L = 200$ and $L = 400$, a shift of 0.237 — larger than the headline within-scale effect. L is an instrument setting with no principled value; nothing in the protocol fixes it.

Surface features, and a reconstruction. The classifier reads 16 statistics of the visible text. Our comparison against a representation-level method is a reconstruction of a published pipeline on 40 problems, and it preserves the ordering of readings without preserving their thresholds; the two are not on one scale. We also did not test an adversary that reasons about the mechanism, as opposed to the surface, of the channel we identify.

Instrument settings that we chose. The truncation length L has no principled value, and §4.1 and Table A19 show the reading moving with it, so it is a setting and not a physical constant. The cap bucket used by rows 2 and 3 of Table 1 conditions on the base branch having failed the problem at full length, which means the paired sample is not a random one; we pair on it because it is the only way to hold length fixed while keeping the two classes on the same problems, and we report the sample size per cell so that the cost of doing so is visible. The scaler in row 1 of Table 1 is fitted differently from every other row, which is why that row is labelled 17-caliber and must not be compared against the rows beneath it. The generation cap is the one setting we did vary: re-collecting every scale at 1024 and 2048 (Appendix E) leaves the channel in place at four of the five scales and weakens it, without dissolving it, at $14b$.