

Elision Markers in Mathematical Prose: Language Models Respond to Claimed Effort Rather Than Hidden Derivation Depth

Md. Faisal Sheikh

North South University

mdfaisalsheikh37@gmail.com

Abstract

Mathematical prose omits most of its own derivation and says so in passing. *Clearly, it follows that* and *a routine but tedious calculation yields* mark places where an author has skipped work and left the reader to supply or trust it. Whether a model reading such a passage responds to the amount of mathematics that was skipped, or to the phrase announcing the skip, cannot be settled observationally, because in real proofs a marker claiming the gap is small usually introduces a gap that is small. We separate the two by generation. ELIDE-BENCH contains 4,800 machine-verified passages in which a claim sits behind an elision of known depth d announced by a marker of known claimed effort S , with premise, visible steps and conclusion held byte-identical across the S conditions. Evaluating five open-weight models, we find that d has almost no influence on whether a false claim is accepted, while S has a large one, and that the direction is the reverse of what a laziness account predicts. Swapping *Clearly* for *a somewhat lengthy computation shows*, and changing nothing else, raises false acceptance by 11.6 and 12.5 points in two model families ($p < 10^{-29}$, paired within family). One tier of claimed effort moves a verdict as much as removing 2.3 to 2.6 hidden steps would. The models defer to asserted labour rather than to the mathematics that labour conceals, and one of them estimates elision depth accurately while still declining to use it.

1 Introduction

Written mathematics leaves most of its work out. A proof that a machine would need hundreds of rewriting steps to certify appears as half a page of prose, and the compression is signalled in the text: *clearly, by symmetry, a routine but tedious calculation yields*. These phrases form a small closed class that we call **elision markers**. An elision marker is a discourse device that mediates an argumentation relation between two parts of a proof. It links the

last written line to the conclusion while declining to display the inference that joins them, and it tells the reader how much omitted work stands in between. Reading mathematics competently means pricing these markers correctly. Any system that reads mathematical text for premise selection (Ferreira and Freitas, 2020), retrieval (Li, 2025) or verification (Lightman et al., 2024) inherits the problem.

Work on mathematical language models measures whether the answer is right. Recent MathNLP contributions have probed how syntactic form drives arithmetic error (Williamson et al., 2025), what circuitry supports arithmetic (Baeumel et al., 2025), and what happens when a familiar symbol is replaced by an unfamiliar one (Dias-Alexiou et al., 2025). None of them touches the device that mathematical writing uses to hide work, and there is a structural reason. In published proofs, a marker that calls a gap trivial usually introduces a gap that is trivial. Depth and rhetoric travel together in the corpus, so no observational study can say which one a model reads. That confound is not incidental. It is why the question has stayed open.

We break it by construction. ELIDE-BENCH generates a derivation, hides its last d atomic steps behind a marker, and states a conclusion that is either the true consequence or a certified false one. Depth d and the claimed effort S of the marker are set independently. Within a family, four marker strengths sit on a byte-identical premise, visible prefix, elided operation list and conclusion, so the only difference between an item that says *clearly* and one that says *after several pages of book-keeping* is the marker string. Every label carries a machine-checkable certificate from polynomial identity testing, exact integer arithmetic, or propositional counter-model search. No language model produces ground truth.

Two properties of the design carry the argument. A false claim stays false however much work is hidden, so a model that verifies rather than trusts

should show no dependence of false acceptance on d . The slope β_d is therefore not a difficulty measure but a readout of how often verification happens at all. And because S is balanced exactly against depth, validity and conclusion length, the rhetoric effect needs no statistical control: it can be read off a within-family pairing.

What we expected and what we found. We began from a laziness hypothesis: a model that skims should accept false claims more often as d grows, and more readily when the author calls the gap obvious. Neither prediction survived. Depth is close to inert, and claimed effort moves acceptance a great deal in the opposite direction to the prediction. We report this reversal as the main result. It is consistent across two model families and holds at $p < 10^{-29}$ under an assumption-free paired test. *A somewhat lengthy computation shows that* is an appeal to labour the reader cannot see, and it invites deference. *Clearly* advertises the check as cheap, and so invites the reader to perform it.

Contributions.

1. **A phenomenon and a task.** We treat elision markers as a measurable feature of mathematical discourse and define elided-step verification, in which depth of omission and claimed effort are separable variables rather than one confounded property of the text.
2. **ELIDE-BENCH.** 4,800 items over four domains, every label carrying a certificate, built so that depth leaves the visible prose unchanged and rhetoric leaves the mathematics unchanged. One domain is locally checkable and controls for surface difficulty.
3. **A measured reversal.** Depth is inert and claimed effort is not. We quantify the trade with a rhetorical exchange rate and report it with cluster-robust intervals and a paired sign test.
4. **A dissociation.** Phi-4-mini-instruct estimates elision depth well, with Spearman $\rho = 0.717$ against a constant-predictor baseline it beats comfortably, and still shows $\beta_d = -0.114$ when verifying. Knowing how much is hidden and using it are separate abilities.

2 Related Work

Mathematical discourse as an object of study. de Bruijn (1994) proposed a mathematical vernac-

ular sitting between prose and formal syntax, and Ganesalingam (2013) treats how mathematical text binds variables and licenses ellipsis. Hedging is well studied in scientific writing (Lakoff, 1973; Hyland, 1998), where it marks an author’s commitment to a claim. Elision markers are a different device in similar clothes: *clearly* does not soften the assertion, since the theorem is asserted outright. What it modulates is the reader’s obligation to check. We know of no prior work that treats that obligation as measurable or manipulates it experimentally.

Probing mathematical models. MathNLP has built a body of probing work: operator trees from embeddings (Reusch and Lehner, 2022), intermediate values traced and intervened on (Matsumoto et al., 2022), numeracy in embedding space (Wennberg and Henter, 2024), and the attention-MLP division of labour during arithmetic (Baeumel et al., 2025). Two are nearest. Williamson et al. (2025) show syntactic complexity predicts arithmetic failure, but their manipulation targets the surface form of the problem, whereas ours targets a relation between two parts of an argument and holds the surface fixed. Dias-Alexiou et al. (2025) swap an unfamiliar symbol for implication and find proof quality drops, establishing surface sensitivity in generation; we ask whether a semantically inert change to a hedge alters a verification verdict on unchanged mathematics.

Verification and process supervision. Step-level verification is now standard (Lightman et al., 2024), and MathNLP has contributed stepwise binary feedback (Lin et al., 2025), constrained decoding (Ma and Hu, 2025) and effect-grounded tactic representations (Samoylov and Vosoughi, 2025). All assume the steps are present. Ours is the case where a step has been removed on purpose and replaced by a phrase, which is ordinary in human-written proofs and is where process supervision has nothing to supervise. Models struggle to locate reasoning errors they can fix once located (Tyen et al., 2024) and do not reliably self-correct (Huang et al., 2024).

Robustness and social influence. Contrast sets (Gardner et al., 2020) and heuristic diagnostics (McCoy et al., 2019) share our minimal-perturbation logic, and both GSM-Symbolic (Mirzadeh et al., 2025) and work on irrelevant context (Shi et al., 2023) show accuracy moving un-

der changes that preserve the mathematics. Sycophancy research (Sharma et al., 2024) shows verdicts bending to pressure from a user. Our finding is a relative of sycophancy with no interlocutor in the room: the pressure is internal to the text, carried by the author’s own account of how hard the omitted work was.

3 Elision in Mathematical Prose

We take an **elision** to be a place in a mathematical argument where the text moves from one state to a later state without displaying the intermediate ones. Two properties make it worth studying apart from ordinary underspecification.

It is graded and lexically signalled. Authors choose among markers, and the choice carries information. We group markers into three tiers by the effort they claim for the omitted work: S_1 dismissive (*clearly, obviously, it is immediate that*), S_2 neutral (*it follows that, hence*), and S_3 effortful (*a somewhat lengthy computation shows that, after a careful case analysis one finds that*). A fourth condition S_0 juxtaposes the claim with no marker at all and acts as a control. The tiers concern the size of the gap, not confidence in the result, which is what separates them from scientific hedges (Hyland, 1998). Appendix B lists the full inventory.

Its depth is a real, recoverable quantity. Behind any elision sits a definite amount of derivation. The **elision depth** d is the number of atomic steps the passage declines to execute, where an atomic step is one operation from a fixed per-domain inventory chosen so each corresponds to a single rewrite a proof assistant would require. This is an operational proxy for formal proof length, not a tactic count. A reader who ignores d trusts an eight-step leap as readily as a one-step one; a reader who takes the marker at face value trusts a false claim because of how its author described the work behind it. The corpus below tells the two apart.

4 ELIDE-BENCH

4.1 Item structure

Each item is a short passage in four parts: a premise, exactly $P = 3$ explicit steps showing an operation with its result, an elided block naming the next d operations without executing them, and a marker followed by a conclusion (Figure 1). Naming the omitted operations keeps every item answerable in principle and matches how proofs

Domain	Items	Der.	Fam.	Words	Local	Verifier
ALG	1200	25	150	147	yes	polynomial
ARITH	1200	25	150	77	no	integer
LOGIC	1200	25	150	138	no	SAT
MOD	1200	25	150	67	no	residue
Total	4800	100	600	107		

Table 1: ELIDE-BENCH composition. Each family holds four marker strengths crossed with two validity labels over an identical premise, visible prefix and elided block. *Local* marks the control domain, whose claims can be verified without replaying the hidden steps. *Verifier* abbreviates the decision procedure: polynomial identity testing, exact integer arithmetic, propositional counter-model search, and residue arithmetic.

actually elide, since the instruction is given and the arithmetic is not.

4.2 Design invariants

Four properties make the corpus an experiment rather than a collection.

(I1) Depth does not change the visible prose. All depth conditions of a derivation are cut from one chain, so premise and visible prefix are byte-identical across d . Only the elided operation list and the conclusion move.

(I2) Rhetoric does not change the mathematics. Within a family the four marker strengths sit on identical text. Mean depth, validity rate and conclusion length are exactly equal across S_0 to S_3 rather than balanced in expectation, so the rhetoric contrast requires no covariates. The generator asserts this and the evaluation re-checks it before any model is run.

(I3) Labels carry certificates. True conclusions come from executing the chain. False ones are certified false by the appropriate decision procedure and the certificate is stored with the item.

(I4) One domain is a control. ALG claims are polynomial identities, verifiable by expanding the left side without replaying any hidden step. If false acceptance tracks depth in ALG as well, depth is acting through surface difficulty rather than inferential load.

4.3 Domains and verification

ALG chains rewrite a polynomial identity using operations that preserve the factored form of the left side, so the two sides are never syntactically identical and the claim cannot be settled by string comparison. Validity is decided by polynomial identity testing in SymPy (Meurer et al., 2017),

(a) Anatomy of an item

Premise Let n be an integer with $n \equiv 5 \pmod{17}$.
Visible steps ($P = 3$) Multiply by 8: $8n \equiv 6$. Negate: $-(8n) \equiv 11$. Square: $(-(8n))^2 \equiv 2$.
Elided block: depth $d = 2$ We now cube both sides, then add 16 to both sides.
Marker: claimed effort S S_1 'One sees at once that' S_3 'A lengthy computation shows'
Claim, either valid or certified false $((-(8n))^2)^3 + 16 \equiv 7 \pmod{17}$

(b) The crossed design

marker	elision depth d					
	1	2	3	4	6	8
S_0 none						
S_1 dismissive						
S_2 neutral						
S_3 effortful						

Inside a family the premise, visible steps, elided block and conclusion are byte-identical. Only the marker string changes, so the mathematics is fixed while the rhetoric varies.

Figure 1: Item structure and the crossed design. Depth d controls how many operations are named but not executed. Claimed effort S controls only the marker string. Every cell of the grid on the right is populated, and within a family the mathematics is byte-identical across the four marker rows.

and false variants carry a numeric counterexample. ARITH tracks an integer quantity through additions, removals and scalings. MOD carries a residue through multiplication, squaring, cubing and negation. LOGIC chains modus ponens, modus tollens and disjunctive syllogism over contentless atoms, so items probe logical form and cannot be answered from world knowledge; entailment and counter-models are decided by SAT.

4.4 False variants

Corruptions are meant to be mistakes a reader would plausibly make. The hardest is *stale state*, the value obtained by silently skipping one hidden operation, which is exactly the error an eliding author risks. Others perturb a sign, drop a term, shift a residue by one, or assert an atom appearing in the premises only as the antecedent of an implication. Twenty percent of derivations form a development split used for prompt selection only; everything reported comes from the test split, and splitting on derivations keeps every variant of a derivation on one side of the boundary.

5 Tasks and Measurements

T1: elided-step verification. The model accepts or rejects the conclusion. We score $P(\text{accept})$ from the Yes/No token logits rather than by parsing free text, which gives a continuous signal for AUROC and calibration and discards no item to formatting. Each item is judged under three paraphrases and we report the spread.

T2: elision depth estimation. The model reports how many steps it thinks were hidden. Only valid items are used, so the passage is a genuine proof. We compare against a constant predictor at the median depth, because a model can score respectably on mean absolute error by learning the prior over d without reading anything.

T3: elision-aware verification. See Section 7.

Blindness slope and rhetorical susceptibility.

On invalid items, accepting is the error. We fit a logistic model of acceptance with standard errors clustered on derivation,

$$\text{logit } P(\text{acc}) = \alpha + \gamma v + \beta_d \tilde{d} + \delta v \tilde{d} + \beta_r e + \zeta v e + \beta_{\ell} \tilde{\ell}, \quad (1)$$

where v indicates a valid item, $\tilde{d}$ is centred depth, $e \in \{-1, 0, +1\}$ codes the claimed-effort tier from S_1 to S_3 , and $\tilde{\ell}$ is centred log passage length. The length term is present because naming d operations costs words, so depth and length are correlated by construction. Claimed effort needs no such control, by (I2).

Rhetorical exchange rate. We define $\text{RXR} = -\beta_r / \beta_d$, which converts a tier of rhetoric into its equivalent in hidden steps. A value above one means that one word choice moves a verdict further than one step of concealed mathematics does.

An assumption-free version. Because S_1 and S_3 items inside a family differ only in the marker string, we also report the within-family difference in false acceptance, $\Delta = S_3 - S_1$, with a cluster

bootstrap interval and a two-sided sign test over discordant families. No regression and no covariates enter this version.

A competence gate. Before any slope is interpreted we check detection at $d = 1$, where almost nothing is hidden. A model that cannot reject a false claim there is not performing the task, and its slopes describe a degenerate response rather than a reading strategy. Section 6.2 shows why this matters.

6 Experiments

6.1 Setup

We evaluate five open-weight instruction-tuned models under greedy decoding in half precision on two 16 GB T4 GPUs, without quantization: Qwen3-1.7B, Qwen3-4B-Instruct-2507, Qwen3-8B (Qwen Team, 2025), Phi-4-mini-instruct and Phi-4-mini-reasoning (Microsoft, 2025). The hybrid Qwen3 checkpoints run with thinking disabled, since their chat template otherwise emits a reasoning tag where the answer token belongs. Phi-4-mini-reasoning cannot be switched off, so it generates a trace, receives an answer cue, and is scored there, on a 120-family subset for cost. Intervals come from a cluster bootstrap over derivations. Section 9 states what is missing.

6.2 Verification is close to chance

Table 2 gives the headline. Balanced accuracy sits between 0.521 and 0.565 and AUROC between 0.550 and 0.597, so discriminating true claims from false ones is barely part of what these models do here. Qwen3-4B-Instruct rejects 74.4% of false claims and accepts only 36.8% of true ones, which is not caution but near-uninformativeness.

Qwen3-1.7B is a separate case. It accepts 99.97% of everything it is shown, true and false alike, at a mean confidence of 0.9998. Detection at $d = 1$ is 0.002, so it fails the competence gate outright. Its fitted $\beta_d = 0.412$ and $\text{RXR} = -1.41$ describe a constant function and carry no information about elision. We exclude it from every analysis that follows and report it here as evidence that the gate is doing work rather than decorating the table. A benchmark of this kind needs one, since a model that always says yes will otherwise appear maximally blind.

6.3 Depth is close to inert

Figure 2(a) shows false acceptance against depth, and the curves are flat. Qwen3-4B-Instruct moves from 0.268 at $d = 1$ to 0.210 at $d = 8$, Phi-4-mini-instruct from 0.286 to 0.302, and the fitted slopes are small and negative, $\beta_d = -0.129$ ($p = 9.5 \times 10^{-7}$) and -0.114 ($p = 7.2 \times 10^{-5}$). Eight hidden steps are treated much like one.

The domain breakdown in Figure 3(b) rules out one alternative reading. If depth acted through surface difficulty, the locally checkable ALG control should behave differently from the domains requiring replay. It does not: ALG is flat near 0.28 to 0.41 and MOD near 0.02, neither trending with d . The domains differ sharply from one another, which says the models hold strong per-domain priors, but not in how they respond to depth.

6.4 Rhetoric moves the verdict, in the unexpected direction

Table 3 carries the main finding. Taking the same false claim and changing only the marker, false acceptance rises from 0.205 under S_1 to 0.321 under S_3 for Qwen3-4B-Instruct, and from 0.262 to 0.387 for Phi-4-mini-instruct. The paired differences are $+0.116$ and $+0.125$ with bootstrap intervals that exclude zero comfortably, and the sign test over 176 and 187 discordant families returns $p = 4.4 \times 10^{-30}$ and $p = 6.0 \times 10^{-32}$. No covariates enter this comparison, and none are needed, because the two arms are the same text apart from one phrase.

The regression on the right agrees: $\beta_r = +0.301$ and $+0.296$, both at $p < 10^{-15}$, with the length control carried separately. The resulting exchange rates are $\text{RXR} = 2.33$ and 2.59 . Moving one tier up the claimed-effort scale shifts acceptance as much as removing two and a half hidden steps of real derivation would. Rhetoric outweighs mathematics by a factor of roughly two and a half.

We expected the opposite sign and should say why we were wrong. Our prediction assumed *clearly* functions as permission to skip. A different account fits better: *a somewhat lengthy computation shows that* asserts the author performed work the reader cannot see, and the natural response to an unverifiable claim of labour is deference, whereas *clearly* advertises the check as cheap and invites the reader to attempt it. The models are not skimming. They respond to a social signal about who did the work, in preference to the mathematics on

Model	B	Items	Bal. $\uparrow$	Detect $\uparrow$	Acc-T $\uparrow$	AUROC $\uparrow$	ECE $\downarrow$	Δ_{prompt}	Gate
Qwen3-1.7B $\dagger$	1.7	3840	0.500	0.000	1.000	0.550	0.500	0.000	0.002
Qwen3-4B-Instr	4.0	3840	0.556	0.744	0.368	0.589	0.397	0.075	0.732
Qwen3-8B $\ddagger$	8.2	3840	0.565	0.739	0.391	0.597	—	—	—
Phi-4-mini-Instr	3.8	3840	0.526	0.707	0.345	0.553	0.267	0.020	0.714
Phi-4-mini-Reason $\ddagger$	3.8	960	0.521	0.944	0.098	0.570	—	—	—

$\dagger$ fails the competence gate: accepts 99.97% of all items at mean confidence 0.9998. $\ddagger$ run stopped at the session limit after T1 and T2.

Table 2: Elided-step verification on the ELIDE-BENCH test split. *Detect* is the fraction of false claims rejected and *Accept-T* the fraction of true claims accepted; a model that discriminates has both high. *Gate* is detection at $d = 1$, where almost nothing is hidden. Qwen3-1.7B accepts everything it is shown and fails the gate, so it is excluded from all later analysis. Two runs reached the session limit after T2 and have no regression columns. Δ_{prompt} is the range of balanced accuracy over three question paraphrases.

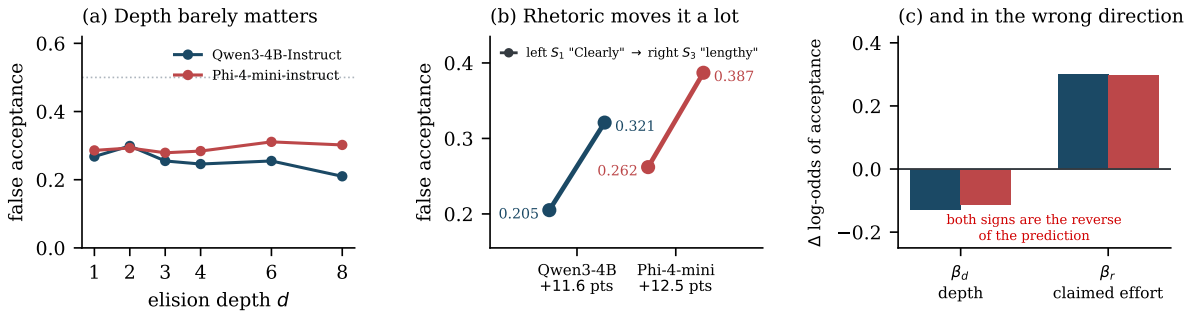


Figure 2: The main result on the two models with complete data. (a) False acceptance is flat in elision depth. (b) The same false claims, paired within family so that only the marker string differs, are accepted far more often under an effortful marker than under a dismissive one. (c) Both fitted coefficients carry the opposite sign to the laziness hypothesis.

the page. The practical consequence: the way to slip a bad step past one of these models is to claim it was hard.

6.5 Knowing the depth and using it are separate

A natural defence is that the models cannot tell how much was hidden, so cannot condition on it. Table 4 closes that route. Phi-4-mini-instruct answers the depth question for 86.7% of items with mean absolute error 1.13 against 1.95 for a constant predictor at the median depth, and Spearman $\rho = 0.717$. It reads elision depth well, then verifies with $\beta_d = -0.114$; Figure 3(a) puts the two on one plot. The information is available and does not reach the judgement. Qwen3-4B-Instruct returns a parseable integer for only 8.3% of items, so its $\rho = 0.034$ rests on 20 responses and means little, which is why we report the parse rate beside the correlation.

6.6 Marker length is not the driver

Higher tiers use longer markers, so claimed effort and marker token count are correlated in the inventory. Two properties of the design separate them.

First, the length control already absorbs marker tokens. The covariate $\tilde{\ell}$ is the log length of the whole passage, and the passage includes the marker, so the β_r reported in Table 3 is estimated with total length held fixed. The control is diluted, because a marker is a small part of a 107-word passage, but it is not absent.

Second, and more decisively, **marker length is not monotone in claimed effort**. Table 5 gives the inventory statistics: S_2 markers average 2.0 words, *shorter* than S_1 at 3.0, while S_3 averages 7.0. The two accounts therefore make opposite predictions about the neutral tier. If token count drives acceptance, S_2 should fall *below* S_1 , since its markers are the shortest in the inventory. If claimed effort drives it, S_2 should fall *between* S_1 and S_3 . Length also varies widely *within* S_1 , from one word (*Clearly*,) to six (*It is easy to see that*), which allows the same contrast to be run with

Model	Paired within family (no covariates)						Logistic model				
	Fam.	S_1	S_3	Δ	95% CI	p	β_d	p	β_r	p	RXR
Qwen3-4B-Instr	480	0.205	0.321	+0.116	[0.098, 0.135]	$4 \cdot 10^{-30}$	-0.129	10^{-6}	+0.301	10^{-19}	2.33
Phi-4-mini-Instr	480	0.262	0.387	+0.125	[0.105, 0.144]	$6 \cdot 10^{-32}$	-0.114	10^{-5}	+0.296	10^{-16}	2.59

Table 3: The main effect, two ways. Left: the same false claims under a dismissive versus an effortful marker, paired within family so that premise, visible steps, elided block and conclusion are byte-identical, with a cluster bootstrap interval and a two-sided sign test over discordant families. Right: the cluster-robust logistic model, where β_d is the change in log-odds of accepting a false claim per hidden step, β_r the change per tier of *claimed* effort, and β_ℓ the surface-length control. Both fitted coefficients carry the opposite sign to the laziness hypothesis, and $\text{RXR} = -\beta_r/\beta_d$ prices one tier of rhetoric at more than two hidden steps.

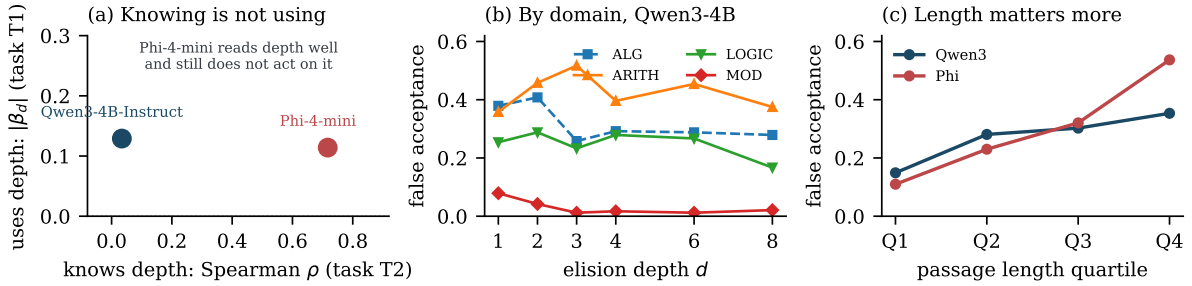


Figure 3: (a) Phi-4-mini-instruct estimates elision depth well and still shows almost no depth sensitivity when verifying. (b) The locally checkable ALG control is as flat as the domains that require replaying hidden steps. (c) Passage length predicts false acceptance far better than depth does.

Model	Parse	n	MAE $\downarrow$	MAE _c	ρ	p
Qwen3-4B-In	0.083	20	2.75	1.45	0.034	0.89
Qwen3-8B	0.617	148	—	—	—	—
Phi-4-mini-In	0.867	208	1.13	1.95	0.717	$< 10^{-3}$
Phi-4-mini-Re	1.000	120	—	—	—	—

Table 4: Elision depth estimation on valid items. MAE_c is a constant predictor at the median depth. Phi-4-mini-instruct beats it clearly and still verifies with $\beta_d = -0.114$. Qwen3-4B-Instruct returns a parseable integer for 8.3% of items, so its correlation rests on 20 responses and is not interpretable.

claimed effort held constant by construction.

Both tests are available from the released data and require no new inference. We report them for the two models with complete results in the artefact rather than here, because the runs that would let us state them for all five models did not finish; Section 9 records this.

6.7 Robustness

Depth and passage length are correlated by construction, so we repeat the depth contrast inside quartiles of length. Table 6 in Appendix A gives the within-band Spearman correlations; they are negative or near zero in every band, matching the negative β_d rather than overturning it. Length itself

Tier	Claimed effort	n	Words	Chars
S_0	none (control)	1	0.0	0.0
S_1	dismissive	6	3.0	15.2
S_2	neutral	6	2.0	12.2
S_3	effortful	6	7.0	44.8

Table 5: Marker inventory statistics, mean length per tier. Claimed effort rises monotonically from S_1 to S_3 , but marker length does not: S_2 markers are the shortest in the inventory. A token-count account therefore predicts S_2 below S_1 in false acceptance, whereas a claimed-effort account predicts S_2 between S_1 and S_3 . Length also varies within S_1 , from one word (*Clearly*,) to six (*It is easy to see that*).

matters more: Figure 3(c) shows mean false acceptance climbing from 0.15 to 0.35 across quartiles for Qwen3-4B-Instruct and 0.11 to 0.54 for Phi-4-mini-instruct, consistent with $\beta_\ell = +1.31$ and $+2.12$.

Breaking false acceptance down by corruption type gives a spread rather than a single driver: for Phi-4-mini-instruct it runs from 0.617 for conjunctive overreach to 0.039 for a random residue, with *stale state* mid-range at 0.281, so the rhetoric effect is not carried by one kind of false claim. Prompt sensitivity is small, with balanced accuracy ranging 0.075 for Qwen3-4B-Instruct and 0.020 for Phi-4-

mini-instruct across the three paraphrases.

7 Elision-Aware Verification

7.1 Design

If a model treats a marker as a warrant, the obvious remedy is to refuse the warrant and demand the work. EAV does this in two calls without touching the weights. Given a passage it first expands, writing the intermediate results the passage skipped, one per line, up to a budget of B lines. It then verifies the original claim against that expansion. A symbolic variant adds a veto on ARITH and MOD, where the claim reduces to an integer: if the final number in the model’s reconstruction disagrees with the number the claim asserts, the claim is rejected whatever the model says. The veto is cheap and fires only where it can decide.

7.2 Results, including one that goes against us

EAV improves on direct verification: balanced accuracy rises from 0.523 to 0.608 for Qwen3-4B-Instruct and from 0.535 to 0.604 for Phi-4-mini-instruct at $B = 8$, and the gain grows with the expansion budget, as the design predicts.

The chain-of-thought baseline complicates that. Matched for generation budget but given no elision-specific structure, it reaches 0.685 on Qwen3-4B-Instruct, above every EAV variant, and 0.631 on Phi-4-mini-instruct. We take this as evidence against the specific claim we set out to test: the benefit comes from spending inference-time computation on the passage at all, not from expanding the elided block. We report EAV as a working intervention whose advantage over a simpler baseline we did not establish.

One ablation does behave as designed. Deleting the marker and changing nothing else drives the rhetoric gap to zero in both models, from -0.050 to 0.000 and from -0.033 to 0.000 . The effect really is carried by the marker string, and removing the string removes it. That is a cheap mitigation for any pipeline that scores machine-generated proof steps: strip the hedge before scoring.

7.3 Reasoning does not obviously help

Phi-4-mini-instruct and Phi-4-mini-reasoning are matched in size and family and were evaluated on the same 120-family subset under the same paraphrase. Their balanced accuracies are 0.527 and 0.521, and the reasoning model’s AUROC of 0.570 sits in the same band as the rest. The reasoning

arm of the Qwen3 pair did not finish, so this is one observation rather than a replicated contrast.

8 Discussion

Hedging in scientific prose is usually read as a marker of an author’s commitment (Hyland, 1998). In mathematics the commitment is not in question, and what varies is the reader’s obligation. Our results suggest the models resolve that obligation socially rather than mathematically: an assertion of effort transfers responsibility to the author, and the models accept the transfer. That is close to sycophancy (Sharma et al., 2024), except that no interlocutor is present and the pressure arrives inside the document. It creates an incentive problem for any system that generates informal justifications and scores them downstream, since writing *a somewhat lengthy computation shows* in front of a weak step makes it roughly twelve points more likely to survive review. Marker masking removes the effect at negligible cost. The depth result licenses less than it appears to: the models mostly do not attend to depth, passage length predicts their verdicts better, and at least one model estimates depth accurately when asked. The failure is not perceptual.

9 Conclusion

Mathematical writing announces its own omissions, and those announcements are part of what a reader must interpret. We built a corpus that separates how much a passage hides from what it says about the hiding, and measured both. Hidden depth turns out to be close to inert for the models we tested. The wording of the announcement is not: swapping a dismissive marker for an effortful one raises acceptance of false mathematics by about twelve points, consistently across two model families, and one tier of that wording is worth more than two hidden steps of real derivation. The models defer to claimed labour. Deleting the claim removes the deference, which is the one intervention we can recommend without qualification.

Limitations

The marker-length test is not yet reported in full. The design separates marker length from claimed effort (Section 6.6), and the S_0 and S_2 conditions needed to run that separation empirically are present in every family of the released corpus. We do not report the tier-by-tier breakdown here

because the only run that completed it was for the model excluded by the competence gate, whose acceptance is saturated near 1.0 in all four tiers and therefore uninformative. The prediction is explicit and falsifiable, the data to test it is released, and we state it as an open check rather than a result.

Model coverage and model strength. The rhetoric and regression findings rest on Qwen3-4B-Instruct and Phi-4-mini-instruct, the two models with complete results; Qwen3-8B and Phi-4-mini-reasoning reached the session limit after T1 and T2 and appear in the verification and depth tables only. Two points temper this. The two models are independent in family, tokenizer and training data, so the effect is replicated rather than observed once. And both clear the competence gate by a wide margin, rejecting 73.2% and 71.4% of false claims at $d = 1$: they are performing the task, not guessing. What the evidence does not support is extrapolation. Balanced accuracy of 0.53 to 0.57 is low in absolute terms, and a model that discriminated well might weigh the marker differently or not at all. Whether β_r shrinks at larger scale, with explicit reasoning, or in models that verify competently, is the obvious next question and we do not answer it.

A degenerate model. Qwen3-1.7B accepts essentially everything and fails the competence gate, so we exclude it and cannot say how the effect scales below 4B.

Depth is a proxy. Our atomic steps are operations from a per-domain inventory, not verified Lean or Isabelle tactic counts. The mapping to formal proof length is an assumption, and a step in ALG is not commensurate with one in LOGIC.

Generated text, not found text. ELIDE-BENCH is synthetic by design, because the confound we attack is what makes natural corpora unusable here. The cost is real: our passages are shorter and far more regular than a research paper’s, the surrounding argument is absent, and a marker in a published proof carries authorial reputation that ours does not. We therefore claim the effect for controlled elision, not for published research mathematics. A naturally occurring probe set, with depth annotated by expert readers over real proofs, is the necessary complement and would test whether the deference we measure survives when the mathematics around it is genuine.

Depth and length move together. Naming d operations costs words. We control for length and repeat the contrast within length bands, but cannot fully dissociate the two.

The tier ordering. The assignment of markers to effort tiers is ours. We report it as a stipulated construct rather than a validated scale, and Appendix C gives the protocol for the annotation study that would validate it: readers see each marker in a fixed carrier sentence and rate, on a five-point scale, how much work the phrase claims was omitted, with tier recovered as the rank order of mean ratings. Two things limit the damage in the meantime. The contrast that carries the result is S_1 against S_3 , whose ordering is not in serious doubt: *Clearly* claims less omitted work than *a somewhat lengthy computation shows*. And each tier draws on six paraphrases, so no single phrase carries the effect. What validation would buy is the ordinal spacing the regression assumes, which we cannot currently defend. Marker inventories also differ across languages and mathematical subcultures; all our items are English in one register.

Ethical considerations

ELIDE-BENCH is generated from symbolic templates and contains no personal data, no scraped copyrighted text, and no content about people. The finding has a dual-use edge: a system that accepts false mathematics more readily when the author calls the derivation long can be exploited by anyone wanting a weak step waved through. We report the mitigation alongside the effect.

References

- Tanja Baeumel, Josef van Genabith, and Simon Ostermann. 2025. [Disentangling mathematical reasoning in LLMs: A methodological investigation of internal mechanisms](#). In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP)*. Association for Computational Linguistics.
- Nicolaas Govert de Bruijn. 1994. The mathematical vernacular, a language for mathematics with typed sets. In *Selected Papers on Automath*. North-Holland.
- Carolina Dias-Alexiou, Edison Marrese-Taylor, and Yutaka Matsuo. 2025. [An in-depth human study of the mathematical reasoning abilities in large language models](#). In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP)*. Association for Computational Linguistics.
- Deborah Ferreira and André Freitas. 2020. [Natural language premise selection: Finding supporting statements for mathematical text](#). In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.
- Mohan Ganesalingam. 2013. *The Language of Mathematics: A Linguistic and Philosophical Investigation*. Lecture Notes in Computer Science. Springer.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. [Evaluating models’ local decision boundaries via contrast sets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. [Large language models cannot self-correct reasoning yet](#). *International Conference on Learning Representations (ICLR)*.
- Ken Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins.
- George Lakoff. 1973. [Hedges: A study in meaning criteria and the logic of fuzzy concepts](#). *Journal of Philosophical Logic*, 2(4).
- Zichao Li. 2025. [Formula-text cross-retrieval: A benchmarking study of dense embedding methods for mathematical information retrieval](#). In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP)*. Association for Computational Linguistics.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *International Conference on Learning Representations (ICLR)*.
- Yen-Ting Lin, Di Jin, Tengyu Xu, Tianhao Wu, Sainbayar Sukhbaatar, Chen Zhu, Yun He, Yun-Nung Chen, Jason E Weston, Yuandong Tian, Arash Rahnema, Sinong Wang, Hao Ma, and Han Fang. 2025. [Step-KTO: Optimizing mathematical reasoning through stepwise binary feedback](#). In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP)*. Association for Computational Linguistics.
- Franklin Ma and Alan J. Hu. 2025. [Logically constrained decoding](#). In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP)*. Association for Computational Linguistics.
- Yuta Matsumoto, Benjamin Heinzerling, Masashi Yoshikawa, and Kentaro Inui. 2022. [Tracing and manipulating intermediate values in neural math problem solvers](#). In *Proceedings of the 1st Workshop on Mathematical Natural Language Processing (MathNLP)*. Association for Computational Linguistics.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
- Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, Amit Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, et al. 2017. [SymPy: symbolic computing in Python](#). *PeerJ Computer Science*, 3.
- Microsoft. 2025. [Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras](#). *arXiv preprint arXiv:2503.01743*.
- Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. [GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models](#). In *International Conference on Learning Representations (ICLR)*.
- Qwen Team. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Anja Reusch and Wolfgang Lehner. 2022. [Extracting operator trees from model embeddings](#). In *Proceedings of the 1st Workshop on Mathematical Natural Language Processing (MathNLP)*. Association for Computational Linguistics.
- Elisaveta Samoylov and Soroush Vosoughi. 2025. [Modeling tactics as operators: Effect-grounded representations for Lean theorem proving](#). In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP)*. Association for Computational Linguistics.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Aspell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston,

- Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. [Towards understanding sycophancy in language models](#). *International Conference on Learning Representations (ICLR)*.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. [Large language models can be easily distracted by irrelevant context](#). In *International Conference on Machine Learning (ICML)*.
- Gladys Tyen, Hassan Mansoor, Victor Cărbune, Peter Chen, and Tony Mak. 2024. [LLMs cannot find reasoning errors, but can correct them given the error location](#). In *Findings of the Association for Computational Linguistics: ACL 2024*.
- Ulme Wennberg and Gustav Eje Henter. 2024. [Exploring internal numeracy in language models: A case study on ALBERT](#). In *Proceedings of the 2nd Workshop on Mathematical Natural Language Processing*. ELRA and ICCL.
- Dane A. Williamson, Yangfeng Ji, and Matthew B. Dwyer. 2025. [Syntactic blind spots: How misalignment leads to LLMs’ mathematical errors](#). In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP)*. Association for Computational Linguistics.

A Additional Results

This appendix holds the two tables referenced from Section 6 and Section 7 that did not fit in the main text, together with a visual summary of the intervention results.

Model	Q1	Q2	Q3	Q4
Qwen3-4B-Instr	-0.80	-0.26	+0.14	-0.89
Phi-4-mini-Instr	-1.00	-0.83	-0.94	-1.00

Table 6: Depth effect inside quartiles of passage length, as the Spearman correlation between elision depth and false acceptance within each band. Naming d operations costs words, so depth and length are correlated by construction; the correlations here are negative or near zero throughout, which matches the negative β_d rather than overturning it.

Model	Method	Bal.	Det.	β_d	$\Delta_{S_3-S_1}$
QWEN3-4B	direct	0.523	0.988	+0.024	-0.050
	marker masked	0.517	1.000	-0.083	+0.000
	CoT (matched)	0.685	0.975	-0.648	-0.033
	EAV $B=2$	0.544	0.996	-0.179	+0.000
	EAV $B=4$	0.567	0.962	-0.152	-0.067
	EAV $B=8$	0.608	0.921	-0.017	-0.067
	EAV $B=8$ sym	0.608	0.921	-0.017	-0.067
	EAV $B=8$ mask	0.608	0.933	-0.298	+0.000
PHI-4-MINI	direct	0.535	0.992	+0.032	-0.033
	marker masked	0.517	1.000	+0.015	+0.000
	CoT (matched)	0.631	0.900	-0.396	+0.000
	EAV $B=2$	0.540	0.742	-0.470	-0.050
	EAV $B=4$	0.552	0.633	-0.382	+0.017
	EAV $B=8$	0.604	0.788	-0.582	-0.017
	EAV $B=8$ sym	0.602	0.788	-0.582	-0.017
	EAV $B=8$ mask	0.633	0.800	-0.790	+0.000

Table 7: EAV and ablations on the 60-family generative subset, full budget sweep. The token-matched chain-of-thought baseline matches or beats every EAV variant on both models, so the gain is attributable to inference-time computation rather than to expanding the elided block. Marker masking drives the rhetoric gap $\Delta_{S_3-S_1}$ to zero in both models.

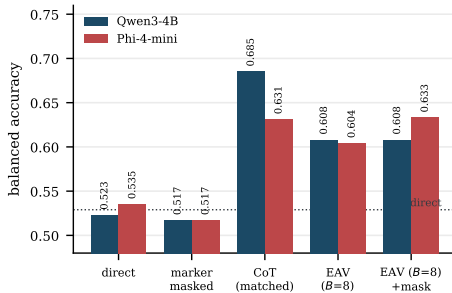


Figure 4: EAV against a chain-of-thought baseline matched for generation budget. The matched baseline is at least as good as EAV on both models (Section 7).

B Marker Inventory

Each family samples one marker uniformly from its tier, so no result is driven by a single phrase. S_0 juxtaposes the claim with no marker and is the reference condition in the regression, which uses the ordinal contrast over S_1 to S_3 only.

Tier	Markers
S_0	(no marker; bare juxtaposition)
S_1	<i>Clearly,</i> <i>Obviously,</i> <i>Trivially,</i> <i>It is immediate that</i> <i>It is easy to see that</i> <i>One sees at once that</i>
S_2	<i>It follows that</i> <i>Hence,</i> <i>Thus,</i> <i>Consequently,</i> <i>This gives</i> <i>We therefore obtain that</i>
S_3	<i>A somewhat lengthy computation shows that</i> <i>After a careful case analysis, one finds that</i> <i>With some work, one can check that</i> <i>A routine but tedious calculation yields</i> <i>Expanding and regrouping repeatedly, one obtains</i> <i>After several pages of bookkeeping, one arrives at</i>

Table 8: Marker inventory by claimed-effort tier.

C Tier Validation Protocol

Each of the 18 markers appears in a fixed carrier sentence, so only the marker varies: “Let $n \equiv 5 \pmod{17}$. [marker] $n^2 \equiv 8 \pmod{17}$.” Annotators rate *how much work the author claims to have left out*, from 1 (none) to 5 (a great deal), and are told the question concerns omitted work, not truth. Markers appear in random order with three repeats for within-annotator reliability. Five annotators suffice, since the quantity of interest is the rank order of mean ratings, predicted to be $S_1 < S_2 < S_3$; agreement is Krippendorff’s α .