TELLING REASONING MODELS WHAT TO REPORT REMOVES THE COST OF CONFLICTING SOURCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented and agentic systems routinely place disagreeing sources in one context, and reasoning models pay for it. When two equally trusted passages disagree, six open model configurations spend 2.0–57× the completion tokens of matched agreeing passages across 23 Wikipedia contradictions and 32 constructed documents (≥176× in the one read cut short by repetition loops at our generation cap), and 89.2–99.8% of the tokens spent on conflicting items are reasoning. We show what removes this cost: a decision about what to report. One sentence naming the passage to answer from removes most of it, and it needs no authority. In a follow-up on 47 further contradictions, where the premium and its removal replicated in every model, an explicitly arbitrary directive was cheaper than an authoritative rule in all four models by point estimate (0.72–0.93×), and permission to give either value cost 1.3–2.8× the directive. Inside that permission frame, adding a command cuts cost by 19–60%, while an assertion of which passage is operative costs 1.4–3.0× the command, in every model and wording. A factorial experiment isolates the cause: telling the model what to report removes the cost in either grammatical form (a declarative statement of what to report costs 0.93–1.05× the command), while handing it a claim about which passage is operative instead of a decision costs 1.55–3.71×, in all three models and both wordings. Fixing the answer blindly trades accuracy for tokens, failing 66–97% of the pairs that other documents settle. An evidence-conditional stopping rule removes 46–88% of the cost of 150 unsettleable conflicts no model had seen, answers CONFLICT on 94–100% of them, keeps or improves accuracy on two of three models, and repays its extra checking once more than 4.1% (Qwen3.8-27B) to 12.9% (gpt-oss-20b) of conflicts are unsettleable. Response policy is a first-class lever on the inference cost of retrieval.

1 INTRODUCTION

Retrieval-augmented and agentic systems routinely place sources that disagree into one context, and the reasoning models that now serve them deliberate before they answer. Deliberation is paid for in tokens: on the anchor model, an unresolved conflict in a constructed note costs 57× the completion tokens of a matched agreeing note (Table 2).

What does the extra reasoning pay for? A model given two equally trusted passages may be deciding which source to trust; then naming the authoritative passage should end the deliberation, and an instruction that disclaims authority should not. Or it may be waiting to be told what to answer; then any sentence that fixes the answer should end it, whether or not it claims to be right, while permission to give either value, which removes the risk of error but leaves the choice open, should end it less. The two readings prescribe different instructions.

The answer is a decision. Authority is unnecessary: on 47 further contradictions an explicitly arbitrary directive was cheaper than an authoritative rule in all four models by point estimate. An operative-source claim is not sufficient: asserting that only passage k states the operative fact, without saying what to report, left 1.4–3.0× the cost of a command naming passage k. A factorial experiment shows that the decision works in either grammatical form ("the value from passage k is the one to report" costs 0.93–1.05× "give the value from passage k") and that the cost returns when the model is handed a claim about which passage is operative instead of a decision (1.55–3.71×). What reasoning models need is a decision about what to report. Our contributions:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

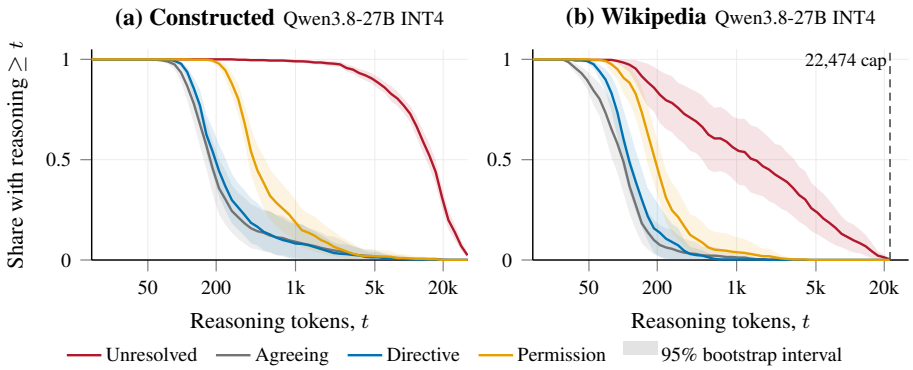


Figure 1: Conflicting passages keep the model reasoning; a directive brings it back to the agreeing baseline, and permission ends it later. The unresolved curve remains above all three fixed-response curves at longer reasoning lengths in both corpora. Shading shows document-bootstrap 95% intervals; the dashed line marks Wikipedia's 22,474-token cap.

Table 1: Each instruction fixes, licenses, or leaves open the response. Instructions tested (constructed wording).

Instruction	Exact sentence
unresolved	In these notes the appendix gives background only, and readers often skip it when they are short of time.
precedence clause	Precedence. If the two entries disagree, entry [k] controls.
stated rule	In these notes entry [k] states the operative rule, and that settles it for every question asked here.
arbitrary directive	For this exercise, answer from entry [k]; that is a convenience and says nothing about which entry is right.
permission	If the two entries disagree, you may answer from either entry, and either value counts as a full answer.
conflict licence	If the two entries disagree, you may answer `CONFLICT` instead of a value, and that counts as a full answer.

- **Measurement** An unresolved conflict multiplies completion tokens 2.0–57× across six open configurations and two corpora (≥176× in one cap-censored read), and the excess is reasoning (Section 3).
- **What removes the cost** One sentence that fixes the answer removes most of it without claiming authority; permission to give either value removes less (Sections 3 and 4).
- **The factorial** The form of the decision changes cost by at most 7% (0.93–1.05×); an operative-source claim attached to the command costs 0.96–1.27×; the same claim in place of the decision costs 1.55–3.71× (Section 6).
- **A stopping rule that transfers** Answering `CONFLICT` only when the other documents leave the answer open removes 46–88% of unsettleable cost on 150 new items, keeps or improves accuracy on two of three models, and pays for its checking once 4.1–12.9% of conflicts are unsettleable (Section 5).

Three studies use open-weight models from 7B to 120B parameters: instruction comparisons on constructed and natural contradictions (Study 1), a known-answer study that prices each response policy in accuracy as well as tokens (Study 2), and a follow-up on 47 further contradictions with known-answer contrasts, the factorial, and a transfer test of the stopping rule (Study 3). Every estimate is a ratio with a 95% interval, per model and wording; estimates whose intervals include 1, and opposite-direction results, stay visible in their tables.

2 Tasks, Models, and Measurements

Each prompt shows two entries stating different values of one constant, then asks for the value. Table 1 lists the instructions tested (constructed wording; natural-text variants say passages for entries). The directive fixes the choice while disclaiming authority; permission allows either value without naming one. **Models** Five open families (Qwen3.8-27B as anchor at INT4 plus its unquantised bf16 weights, gpt-oss-20b, Nemotron-3-Nano-30B, Olmo-3-32B-Think, K2-Horizon-7B) in Study 1; the anchor, gpt-oss-20b and Olmo-3-7B-Think in the accuracy study (Study 2); these three plus gpt-oss-120b on the 47 articles of Study 3, where gpt-oss-120b also gives a settled-versus-unsettled contrast

Table 2: Unresolved conflicts multiply cost on every read. Unresolved / agreeing cost ratios (A / B: unresolved costs k× agreeing), with family-wise simultaneous 95% intervals (12-read family). K2 constructed is cap-censored (≥).

Model	Corpus	Ratio [95% CI]	Simultaneous 95%
Qwen3.8-27B INT4	constructed	57.42 [42.08,78.37]	[35.83,92.03]
Qwen3.8-27B bf16	constructed	51.52 [39.20,67.72]	[34.03,78.00]
gpt-oss-20b	constructed	10.50 [8.34,13.21]	[7.41,14.88]
Nemotron-3-Nano-30B	constructed	5.58 [4.05,7.69]	[3.43,9.07]
Olmo-3-32B-Think	constructed	4.39 [3.59,5.36]	[3.23,5.95]
K2-Horizon-7B	constructed	≥176.20 [108.17,287.02]	[≥84.06,369.34]
Qwen3.8-27B INT4	Wikipedia	15.23 [7.51,30.87]	[5.12,45.25]
Qwen3.8-27B bf16	Wikipedia	10.03 [5.27,19.12]	[3.72,27.10]
gpt-oss-20b	Wikipedia	3.73 [2.70,5.14]	[2.27,6.11]
Nemotron-3-Nano-30B	Wikipedia	2.04 [1.57,2.66]	[1.36,3.07]
Olmo-3-32B-Think	Wikipedia	2.99 [2.04,4.38]	[1.66,5.38]
K2-Horizon-7B	Wikipedia	5.85 [3.63,9.41]	[2.81,12.17]

on the 50 worlds of its first generator (Table 11). The anchor ran the unresolved, rule, directive and permission instructions on these 47 articles (2 samples each) in an exploratory pilot; those rows enter no estimate reported here, and Study 3 re-measures the anchor on new samples. **Measures** Cost is completion tokens per call; each contrast is a ratio A / B (A costs k× B) with a 95% interval, no-difference value 1; accuracy contrasts are paired failure-rate differences in percentage points with intervals, no-difference value 0. The 95% interval is the primary display; Bonferroni simultaneous 95% intervals within each declared family (Appendix C) are the conservative check, and where a statement rests on the 95% interval alone the text says so. Every reported ratio, difference and share was computed by two independent implementations; point estimates agree exactly and bootstrap interval bounds agree to within 0.005 on the log scale. Decision-curve mean tokens and answer classes come from one implementation; simultaneous intervals and break-even shares follow by the formulas of Appendix C.

3 Conflict cost across corpora

An unresolved conflict multiplied completion cost on every read (Table 2, Fig. 2): 4.4× to 57× on 32 constructed documents in five configurations and at least 176× on K2-Horizon-7B, and 2.0× to 15× on 23 Wikipedia contradictions, every 95% and simultaneous interval above 1. The K2 constructed read is a lower bound: 223 of 512 rows hit the generation cap and 222 of the 223 were repetition loops (Appendix G). One sentence fixing which source to answer from removed most of the premium (Fig. 3). On constructed documents a precedence clause left 1.23–1.60× of agreeing controls in all six configurations and a stated rule 1.24–1.59× in five, with 2.38× [1.88, 3.00] on K2-Horizon-7B; the arbitrary directive left 0.99–1.54× in five and 20.91× [13.33, 32.78] on K2-Horizon-7B, where 32 of its 512 responses ran to the cap. On Wikipedia the directive left 1.02–1.59× in all six. The directive cost less than the stated rule by point estimate in 11 of the 12 reads (0.67–0.99×); K2-Horizon-7B's constructed notes are the exception. Permission removed less: it cost 1.25–3.66× the directive across the ten reads with a permission run, with the 95% interval excluding no-difference in nine (the exception is Nemotron on Wikipedia, 1.25 [0.99, 1.58]) and the simultaneous interval in eight (Table 8).

3.1 Where the cost sits

The reasoning segment carries the cost: it makes up 89.2–99.8% of completion tokens on conflicting items in every condition and model of Studies 1 and 2, and 92.6–99.7% on unresolved conflicts in Study 1 (Appendix E). Unresolved multipliers on reasoning tokens alone match or slightly exceed completion (anchor constructed 61.5× reasoning vs 57.4× completion; Wikipedia 16.6× vs 15.2×), while answer-token multipliers stay between 0.73× and 2.27× (Fig. 4).

4 Prompt contrasts on 47 articles

The premium and its removal replicated on 47 Wikipedia contradictions: unresolved / agreeing 17.92× [10.53,30.50] on the anchor, 3.84× [2.99,4.93] on gpt-oss-20b, 2.86× [2.21,3.70] on Olmo-3-7B-Think and 3.08× [2.53,3.75] on gpt-oss-120b, with an arbitrary directive removing 81–99% of that log premium across models and wordings (Table 12). The directive was cheaper than the

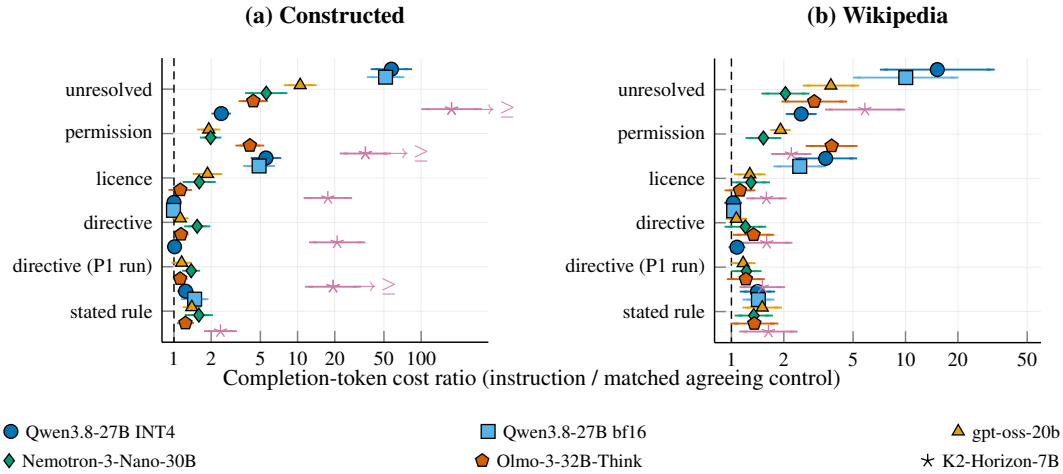


Figure 2: Unresolved costs are highest; settling categories cluster near 1×. Whiskers show 95% intervals; arrows with ≥ mark censored lower bounds.

(a) Constructed notes (32)

Qwen INT4	Qwen bf16	gpt-oss 20b	Nemo-tron	Olmo 32B	K2 7B	
57.4	51.5	10.5	5.58	4.39	≥176	unresolved
2.41	–	1.91	1.98	4.11	≥35.4	permission
5.55	4.88	1.87	1.60	1.13	17.6	licence
1.00	0.99	1.13	1.54	1.14	20.9	directive
1.01	–	1.15	1.38	1.12	≥19.4	directive (P1 run)
1.24	1.47	1.39	1.59	1.24	2.38	stated rule
1.28	1.23	1.23	1.60	1.25	1.39	precedence clause
1.07	1.07	1.13	1.11	1.05	1.18	rule + clause

(b) Wikipedia (23 articles)

Qwen INT4	Qwen bf16	gpt-oss 20b	Nemo-tron	Olmo 32B	K2 7B
15.2	10.0	3.73	2.04	2.99	5.85
2.52	–	1.91	1.53	3.76	2.21
3.48	2.47	1.28	1.30	1.12	1.59
1.02	1.03	1.07	1.20	1.34	1.59
1.08	–	1.17	1.22	1.21	1.51
1.41	1.43	1.50	1.34	1.35	1.63

no in-document clause on Wikipedia

Figure 3: Fixing the answer brings cost near agreeing controls; leaving the choice open costs more. Study 1 cost multipliers per instruction and model: conflicting over matched agreeing completion tokens, geometric mean over documents (intervals in Table 8). Darker cells cost more (log scale); ≥ marks a cap-censored lower bound; – not run (Qwen bf16 had no permission run). P1 run: the separate permission run, which regenerated the directive.

authoritative rule in all four models and both wordings by point estimate (0.72–0.93×); six of the eight 95% intervals and five of the eight simultaneous intervals exclude no-difference (Table 3). Permission cost 1.3–2.8× the directive, with every 95% and simultaneous interval above 1. Inside the permission frame, a command naming the passage cut cost by 19–60% relative to permission alone. Stating instead that the passage is operative did not substitute for the command: the assertion cost 1.4–3.0× the command in every model and wording, every interval above 1, while against permission it changed direction across models (permission / assertion 0.47–1.61×; Fig. 5, Table 3). On known-answer items the ordering replicated on the two models that followed the command (on 1.00 and 0.99 of unsettleable items): permission / command 2.37× [2.13,2.63] on the anchor and 1.48× [1.31,1.68] on gpt-oss-20b. Compliance moderates the command's effect: Olmo-3-7B-Think followed it on 0.68 of those items and on 0.86–0.88 of the 47 articles, and on known-answer items its permission cost 0.88× [0.78,1.00] the command (Tables 12c and 13).

Table 4 maps each candidate explanation to the contrast that tests it; its last three rows are the factorial of Section 6.

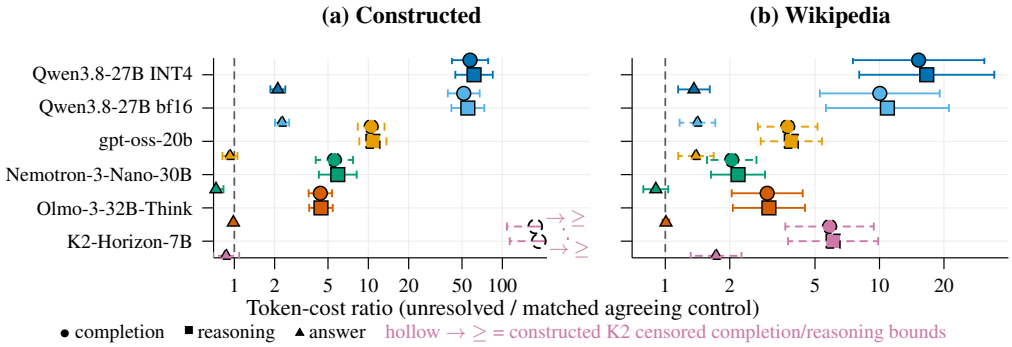


Figure 4: Unresolved costs are concentrated in completion and reasoning tokens; answer-token ratios are much smaller.

Table 3: A decision removes the premium; permission and an operative-source assertion leave more. Study 3 contrasts on the 47 articles in both wordings (a, b): ratio A / B of completion tokens (A costs k× B) with 95% interval; second line: Bonferroni simultaneous 95% interval within the contrast's family (32 permission-frame contrasts; 8 directive / stated-rule contrasts).

Contrast (A / B)	w.	Qwen3.8-27B	gpt-oss-20b	Olmo-3-7B-Think	gpt-oss-120b
permission / directive	a	1.90 [1.65,2.20] sim [1.50,2.42]	1.61 [1.44,1.81] sim [1.34,1.95]	2.76 [2.11,3.61] sim [1.76,4.33]	1.38 [1.28,1.49] sim [1.21,1.57]
	b	1.76 [1.47,2.10] sim [1.31,2.37]	1.48 [1.32,1.66] sim [1.22,1.79]	2.65 [2.02,3.46] sim [1.69,4.14]	1.35 [1.25,1.46] sim [1.19,1.54]
permission / command	a	1.35 [1.14,1.60] sim [1.02,1.79]	1.55 [1.42,1.70] sim [1.34,1.80]	2.40 [1.90,3.04] sim [1.62,3.56]	1.28 [1.21,1.34] sim [1.17,1.39]
	b	1.41 [1.20,1.66] sim [1.08,1.85]	1.30 [1.19,1.42] sim [1.12,1.50]	2.48 [1.92,3.22] sim [1.61,3.82]	1.23 [1.14,1.33] sim [1.08,1.40]
permission / assertion	a	0.97 [0.81,1.17] sim [0.71,1.32]	0.52 [0.47,0.58] sim [0.43,0.63]	1.61 [1.28,2.03] sim [1.09,2.36]	0.68 [0.61,0.75] sim [0.57,0.80]
	b	0.86 [0.72,1.03] sim [0.64,1.17]	0.47 [0.42,0.53] sim [0.39,0.57]	1.41 [1.12,1.77] sim [0.96,2.06]	0.73 [0.67,0.80] sim [0.62,0.85]
assertion / command	a	1.39 [1.22,1.59] sim [1.11,1.74]	2.98 [2.63,3.37] sim [2.42,3.66]	1.49 [1.29,1.73] sim [1.17,1.90]	1.89 [1.71,2.09] sim [1.60,2.23]
	b	1.63 [1.46,1.82] sim [1.36,1.97]	2.75 [2.48,3.04] sim [2.32,3.25]	1.77 [1.52,2.06] sim [1.37,2.28]	1.68 [1.56,1.81] sim [1.49,1.91]
directive / stated rule	a	0.82 [0.75,0.90] sim [0.72,0.93]	0.89 [0.83,0.95] sim [0.80,0.98]	0.84 [0.74,0.95] sim [0.70,1.00]	0.92 [0.85,1.00] sim [0.82,1.03]
	b	0.72 [0.64,0.82] sim [0.60,0.86]	0.77 [0.72,0.82] sim [0.70,0.84]	0.91 [0.80,1.03] sim [0.76,1.08]	0.93 [0.85,1.02] sim [0.82,1.06]

5 AN EVIDENCE-CONDITIONAL STOPPING RULE

Fixing the answer blindly trades accuracy for tokens. On known-answer worlds (Study 2) the arbitrary directive was the cheapest instruction on unsettleable items (0.07–0.09× the unresolved cost on the anchor, 0.19–0.21× on gpt-oss-20b) but failed 92.0%, 96.7% and 66.0% of settleable pairs on the anchor, gpt-oss-20b and Olmo-3-7B-Think, because it overrides evidence that settles the answer (Fig. 6). An evidence-conditional stopping rule (use the documents if they settle the question, otherwise answer `CONFLICT`) takes the saving only where the evidence leaves the answer open: on the anchor it removed 67–71% of unsettleable cost, answered `CONFLICT` on 148–150 of 150 unsettleable items, and failed 1 and 0 of 300 settleable pairs, against 1 with no instruction. Savings track execution: gpt-oss-20b answered `CONFLICT` on 114–121 of 150 and saved 36–39%. The rule's scope is a model that already tells settled from unsettled conflicts. Olmo-3-7B-Think fails 101 of 300 settleable pairs with no instruction and spends the same on unsettled as on settled conflicts (1.06× [0.97, 1.16], against 3.50–18.09× on the other Study 2 models; Table 11); under the rule it answered `CONFLICT` on 24–32 of 150 and saved nothing (1.13–1.15×). On 150 unsettleable items no model had seen, with the 300 settleable worlds of Study 2, the rule's transfer wording removed 88% of unsettleable cost on the anchor [87–90%], 46% on gpt-oss-20b [40–51%] and 59% on gpt-oss-120b [56–62%], answering `CONFLICT` on 100%, 94% and 95% of those items (Table 5). Settleable accuracy was unchanged on the anchor (0 of 300 pairs failed under either instruction) and improved on gpt-oss-120b (−13.7 points, from 23.7% to 10.0% of pairs failed); gpt-oss-20b lost 6.7 points [4.3, 9.0], and almost all of those added failures are abstentions: of its 53 failing settleable responses under the rule, 51 answered `CONFLICT` and 2 gave a wrong value, against 5 wrong values with no instruction. Checking costs tokens on settleable items, 1.35–1.36× no instruction in all three models, and the added tokens are reasoning tokens: reasoning tokens rose 1.36–1.39× while

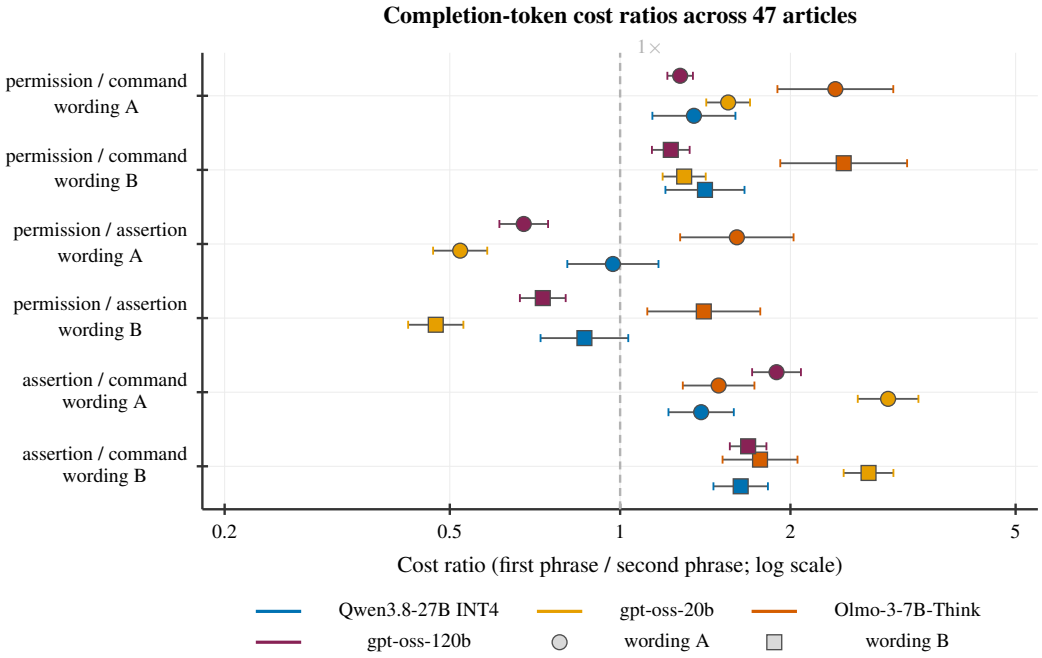


Figure 5: A command cuts cost inside the permission frame; an operative-source assertion costs 1.4–3.0× the command. Completion-token cost ratios on the 47 articles (first phrase / second phrase; 95% intervals; values in Table 3). Color identifies model; shape identifies wording A/B.

Table 4: A decision about what to report is the only explanation consistent with every contrast. Mechanism map: candidate explanation and the contrast that tests it (A / B); observed ratios across models and wordings, and the reading.

Candidate; contrast (A / B)	Observed; reading
Authority; directive / stated rule	0.72–0.93, below 1 in all four models by point estimate (6 of 8 95% and 5 of 8 simultaneous intervals exclude 1; Table 3); removal needs no authority.
Stakes; permission / directive	1.35–2.76, every interval above 1 (Table 3); an open choice removes less.
Source-scoping; rule that uses both sources / directive (known-answer)	Qwen 0.78 [0.68,0.90] and 0.75 [0.66,0.85]; gpt-oss-20b 1.17 [1.04,1.33] and 1.03 [0.92,1.16] (Table 13); reading both sources costs at most 17% more than an arbitrary pick.
Operative-source claim; assertion / command	1.39–2.98, every interval above 1; the claim does not substitute for a decision.
Processing load; unresolved / agreeing	2.0–57 (≥176 censored); one sentence removes most of it, so the input alone does not set the cost.
Grammatical form of the decision; declarative decision / command	0.93–1.05 (Table 17); form changes cost by at most 7%.
Claim attached to the command; command with claim / command	0.96–1.27 (Table 17); a small addition.
Claim in place of the decision; claim alone / declarative decision	1.55–3.71, every interval above 1 (Fig. 8, Table 17); the cost driver.

answer tokens fell (0.39–0.96×). The overhead differs by generator (1.15–1.85×) and in absolute terms across models (+464 to +805 tokens per settleable item; Table 14), so the near-identical pooled ratios do not reflect a fixed checking cost. Because abstaining on unsettleable items is correct, the rule beats no instruction on accuracy and tokens together once more than 4.1% (Qwen3.8-27B), 10.2% (gpt-oss-120b) or 12.9% (gpt-oss-20b) of conflicts are unsettleable. These are the exact crossings of the two expected-cost lines (Fig. 7); on the 0.05 grid of the bootstrap bands the first cheaper point is 0.05 for Qwen3.8-27B and 0.15 for both gpt-oss models.

6 Reporting instructions versus operative-source claims

Study 3's assertion ("only passage k states the operative fact") differed from its command ("give the value from passage k") in two ways at once: it was declarative, and it replaced the decision about what to report with an operative-source claim. We separated the two on the same 47 articles, in both wordings, for the three models on which the assertion did not cost less than permission in Study 3.

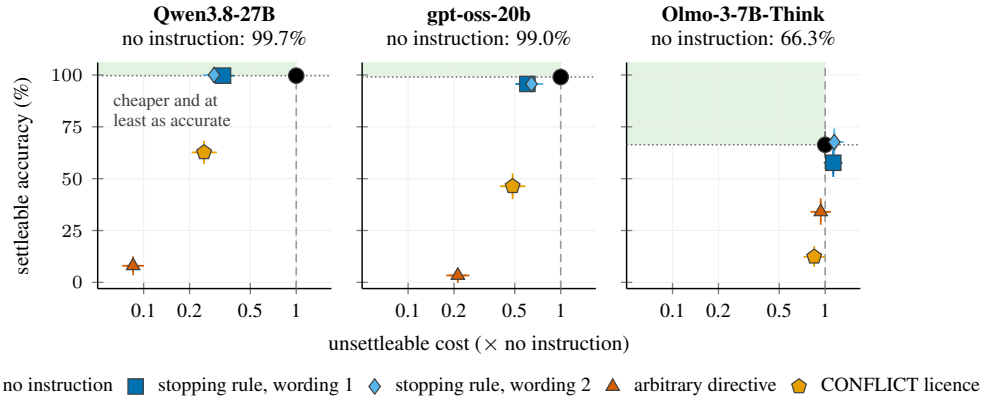


Figure 6: The stopping rule cuts unsettleable cost without the directive's accuracy loss. Per model (Study 2): cost on unsettleable items against accuracy on settleable pairs, with 95% intervals; shaded: cheaper than no instruction and at least as accurate. The directive is cheapest but fails 66–97% of settleable pairs. Olmo-3-7B-Think, which already fails a third of them with no instruction, answers `CONFLICT` on only 16–21% of unsettleable items under the rule.

Table 5: The stopping rule transfers to new unsettleable items. Stopping rule against no instruction on 150 unsettleable items new to every model and the 300 settleable worlds of Study 2 (cost ratios A / B: the rule costs k× no instruction; accuracy in points; `CONFLICT` shares in %), with 95% world-cluster bootstrap intervals; second line: simultaneous 95% over the 12 outcomes. Qwen settleable failures were 0 of 300 under both instructions (exact one-sided 97.5% upper bound 1.22 points).

Measure	Qwen3.8-27B	gpt-oss-20b	gpt-oss-120b
Settleable failure difference (points)	0.0 [0.0,0.0] sim [0.0,0.0]	6.7 [4.3,9.0] sim [3.3,10.1]	−13.7 [−17.5,−9.7] sim [−19.4,−7.9]
Unsettleable cost ratio	0.12 [0.10,0.13] sim [0.10,0.14]	0.54 [0.49,0.60] sim [0.47,0.63]	0.41 [0.38,0.44] sim [0.37,0.46]
Unsettleable `CONFLICT` share (%)	100 [100,100] sim [100.0,100.0]	94 [92,97] sim [90.7,98.0]	95 [93,97] sim [92.2,98.5]
Settleable cost ratio	1.36 [1.30,1.43] sim [1.27,1.46]	1.36 [1.31,1.41] sim [1.29,1.43]	1.35 [1.31,1.38] sim [1.30,1.40]

The cells are the command; a declarative decision ("the value from passage k is the one to report"); the command with the claim attached ("…, which is the only passage that states the operative fact"); and the claim alone, which is the Study 3 assertion (Table 16). Each ran inside the permission frame with 16 samples per article and cell on Qwen3.8-27B and 8 on gpt-oss-20b and -120b, and 97.3–100% of responses in every one of these cells gave passage k's value. Telling the model what to report removes the cost in either grammatical form: the declarative decision cost 0.93–1.05× the command. Attaching the claim to the command changed cost by 0.96–1.27×. Handing the model the claim instead of a decision is what costs: 1.55–3.71× the declarative decision in all three models and both wordings, every 95% and simultaneous interval above 1, and the excess is reasoning tokens, whose ratios lie within 4% of the completion ratios (Fig. 8, Table 17). Because the claim-alone cell has no decision, this design separates form from the claim for the command only; averages over its four cells are in Table 17. A follow-up run adds the missing cell, the declarative decision with the claim attached ("the value from passage k, which is the only passage that states the operative fact, is the one to report"). On both gpt-oss models the claim then costs little (1.10–1.14× the declarative decision, every upper bound at most 1.22×), form stops mattering (1.00–1.08× the claim with the command), and the bare claim costs 1.5–3.2× the same claim with the decision attached (across runs): the missing decision is the cost. Qwen3.8-27B deliberates over the claim itself, which costs 1.47–1.48× even beside a declarative decision (Table 17). Two independent implementations of this analysis agree on every estimate.

7 Discussion and Conclusion

Conflicting sources impose a large cost that sits in the reasoning trace, and a decision removes it. One sentence fixing the answer removes most of the cost without claiming authority, and leaving

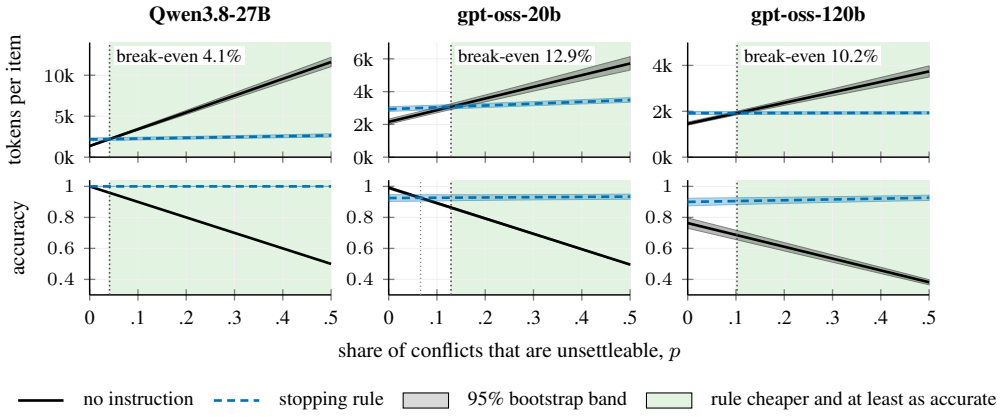


Figure 7: The stopping rule repays its extra checking once a small share of conflicts is unsettleable. Expected completion tokens (top) and accuracy (bottom) per item on the transfer test as the unsettleable share p varies. Dotted lines: exact break-even, 4.1% (Qwen3.8-27B), 12.9% (gpt-oss-20b) and 10.2% (gpt-oss-120b); on gpt-oss-20b the rule is at least as accurate from 6.6% (faint dotted line). Both curves are linear in p, so the order at $p = 0.5$ persists to $p = 1$.

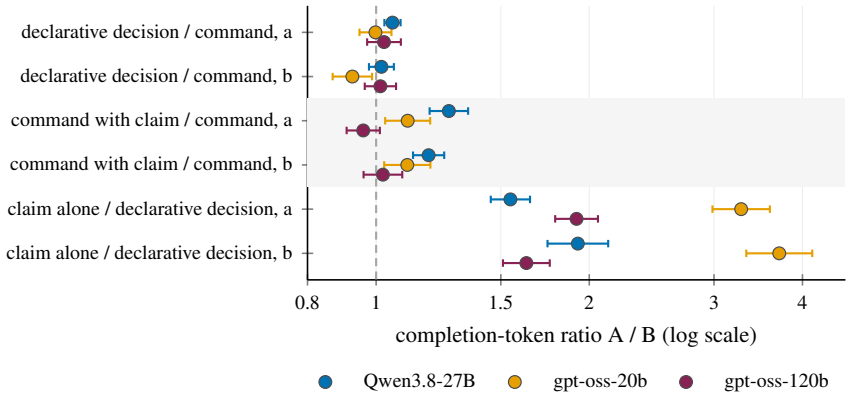


Figure 8: Telling the model what to report removes the cost in either grammatical form; a claim in place of the decision does not. Factorial contrasts per model and wording (a, b) on the 47 articles, with 95% intervals: the declarative decision costs 0.93–1.05× the command; the operative-source claim attached to the command 0.96–1.27× the command; the claim alone 1.55–3.71× the declarative decision, every interval above 1 (Table 17).

the choice open removes less. The decision works in either grammatical form, while a claim about which passage is operative, handed over in place of the decision, leaves 1.55–3.71× the cost of a declarative decision naming the same passage. Reasoning models do not need to be told which source is right, only what to report. Where other evidence can settle the answer, an evidence-conditional stopping rule takes the saving without the accuracy loss of a fixed answer: on 150 new unsettleable items it removed 46–88% of the cost and kept or improved accuracy on two of three models, with savings tracking how fully a model executes it. **Scope** Single-turn prompts on open-weight models from 7B to 120B parameters; the cost measure is the generated-token budget; accuracy items are synthetic known-answer worlds; the stopping rule presupposes a model that can already tell settled from unsettled conflicts. We expect the same rule to help wherever the input leaves the response undetermined, such as missing premises and ambiguous questions.

8 Related Work

Knowledge conflicts in language models include gaps between context and parametric memory and disagreements within retrieved evidence (Xu et al., 2024; Hou et al., 2024; Wu et al., 2024). Our outcome is reasoning-token expenditure under controlled response instructions. Contradiction-induced overthinking is established: Zhang et al. (2026) build adversarial contradictions to inflate reason-

ing cost and test brevity, budget, and filtering defenses; Kumar et al. (2025) show cost amplification through decoy tasks; missing-premise work shows extended deliberation need not resolve an ill-posed problem (Fan et al., 2025; Chen et al., 2024). Abstention and clarification work studies when models should decline or ask: Kirichenko et al. (2025) evaluate when models should withhold answers across 20 datasets, and conflict-aware retrieval seeks reliable answers from inconsistent evidence (Wu et al., 2024; Zhang et al., 2024; Zhang & Choi, 2025). Three lines are closest. First, Zhang et al. (2026) first showed contradiction-induced inflation and defenses; we contribute controlled settlement, especially authority-disclaiming directives versus free choice and `CONFLICT` licences. Second, Hou et al. (2024) established equal-trust conflicting evidence; we contribute controlled cost measurement and intervention comparisons on 23 of its contradictions. Third, missing-premise work establishes that ill-posedness prolongs reasoning; we distinguish absent premises from incompatible supplied premises and show cost cuts without adding truth-relevant information.

AI USE STATEMENT

We used large language model assistants (Anthropic Claude, OpenAI Codex, Meta Muse Spark) throughout this project, under author direction. **Research ideation and execution:** the assistants helped formulate hypotheses and design the studies and follow-up experiments, implemented the task builders and generation runners, ran the experiments on cluster and rented GPUs, wrote the analysis code (including two independent implementations whose outputs we reconciled), and generated the figures and tables. **Retrieval and discovery:** the assistants searched for related work and verified every cited reference against its published source. **Writing:** the assistants drafted, edited, and polished the manuscript and appendix, and ran simulated reviews to find weaknesses before submission. The authors reviewed the AI-assisted experiments, code, analyses, figures, and prose and take responsibility for the final work and its claims. Every reported number is computed by script from the study outputs and source rows identified in the paper and appendix.

ETHICS STATEMENT

All natural text is public Wikipedia content. No human subjects were involved.

REPRODUCIBILITY STATEMENT

Prompts, task files, analysis code, and derived results are in the supplementary material; generated response rows will be released with the camera-ready version. Prompts, data construction, serving settings, and estimators are in Appendices A–C. Every reported ratio, difference and share was computed by two independent implementations; point estimates agree exactly and bootstrap interval bounds agree to within 0.005 on the log scale.

REFERENCES

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2+3=? on the overthinking of o1-like LLMs. *arXiv preprint arXiv:2412.21187*, 2024.

Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. Missing premise exacerbates overthinking: Are reasoning models losing critical thinking skill? *arXiv preprint arXiv:2504.06514*, 2025.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. Wikicontradict: A benchmark for evaluating LLMs on real-world knowledge conflicts from wikipedia. *arXiv preprint arXiv:2406.13805*, 2024.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. Abstentionbench: Reasoning LLMs fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*, 2025.

Abhinav Kumar, Jaechul Roh, Ali Naseh, Marzena Karpinska, Mohit Iyyer, Amir Houmansadr, and Gene Bagdasarian. Overthink: Slowdown attacks on reasoning LLMs. *arXiv preprint arXiv:2502.02542*, 2025.

Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. *arXiv preprint arXiv:2404.10198*, 2024.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. *arXiv preprint arXiv:2403.08319*, 2024.

Michael J. Q. Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with LMs. In *Findings of NAACL*, 2025.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proc. ACL*, 2024.

Xiaolei Zhang, Xiaojun Jia, Liquan Chen, and Songze Li. CODE: A contradiction-based deliberation extension framework for overthinking attacks on retrieval-augmented generation. *arXiv preprint arXiv:2601.13112*, 2026.

A PROMPTS AND EXAMPLE INPUTS

Every call reported here is a single turn with reasoning mode on. The system message is a fixed base text followed by one instruction sentence (all sentences verbatim in Table 6); the user message contains the documents and the question. The two inputs below are complete.

A.1 COMPLETE EXAMPLE INPUTS

Constructed notes Each note states one fictional rule twice, as Entry 1 and Entry 2, in its body, each entry followed by a worked example that applies it; in a conflicting item the entries differ in one constant only. The body ends with an "Edition." line, which the precedence clause replaces in the clause condition, and a background paragraph headed "Appendix" that contains neither entry and no constant and is identical in every condition of the note. The unresolved instruction's sentence refers to that paragraph and is the same in the conflicting item and its agreeing controls, so it cancels in every ratio to agreeing. The two agreeing controls give both entries the same constant (below: both "three hundred" or both "five hundred", with matching worked examples). Each of the note's eight inputs is a separate item, and which entry states which constant, and which entry a rule, clause or directive names, is balanced across the eight.

Constructed note D05, item 3: conflicting version, no clause, unresolved instruction

System:

````
You are a helpful assistant.  Answer using only the note and the table below.
Give the final line as `ANSWER: <value>`.  In these notes the appendix gives
background only, and readers often skip it when they are short of time.
````

User:

````
# Halden long years --- Reference Note (D05)
Source:  Halden Calendar Code HCC-6 (halden.example.org/specs/hcc-6)
This note is authoritative for questions about Halden long years.
Entry 1.  A Halden year is long exactly when divisible by six, except that century
years are long only when divisible by three hundred.
Worked example 1.  The year 900 is a long year.
Entry 2.  A Halden year is long exactly when divisible by six, except that century
years are long only when divisible by five hundred.
Worked example 2.  The year 900 is a common year.
Elaboration.  Halden years number twelve moons in fixed order, century marks fall on
years ending in two zeroes, and priests proclaim festival dates elsewhere.
Edition.  This note follows the printed index for entries.
Appendix (background detail).  Moon names cycle without change, season rites follow
the thaw, and tide tables sit outside this note.

```csv
qid,year
D05_Q1,1200
D05_Q2,2400
D05_Q3,3600
D05_Q4,4800
D05_Q5,600
D05_Q6,1800
D05_Q7,2100
D05_Q8,3300
```

Is 3600 a long year or a common year?  Answer long or common.
````

Wikipedia passages Each item gives two passages from one article and a question. The two agreeing controls replace the value in one passage by the other's (below: Passage 1 reads "8 December 1967", or Passage 2 reads "12 August 1967"). This article has the median unresolved ratio of the 23 for Qwen3.8-27B INT4, whose mean completion tokens, conflicting / agreeing over 16 repetitions, are 2,867 / 128 unresolved, 187 / 136 under the stated rule, 155 / 158 under the directive, 1,996 / 143 under the licence and 288 / 130 under permission.

Wikipedia item "Marina Baker": conflicting version, unresolved instruction

System: `You are a helpful assistant. Answer using only the two passages below. Give the final line as `ANSWER: <value>`. In these passages the wording follows the source article, and readers often skim it when they are short of time.`

User:

```
Passage 1.
Marina Augusta Baker was born on 12 August 1967 (age 56).

Passage 2.
Marina Baker was born on 8 December 1967.

Question: Marina Augusta Pepper also called Marina Baker is an English Liberal
Democrat local politician, journalist. When was Marina Baker born?
```

B DATA, MODELS AND SERVING

Constructed notes (Study 1) 32 notes, each a fictional reference rule stated twice with different constants (Appendix A.1). Each note has a conflicting and two agreeing versions, run without and with the precedence clause under the unresolved instruction and the stated rule, and without the clause under the directive and the licence: 32 notes $\times$ 18 conditions $\times$ 8 inputs $=$ 4,608 prompts per model, each run twice. A separate run adds permission and regenerates the directive on the notes without the clause (3,072 responses per model).

Wikipedia contradictions (Study 1) 23 contradictions from WikiContradict (Hou et al., 2024), one per article, in which each conflicting value appears verbatim in its passage and is absent from the other. Items were selected mechanically (distinct contexts, verbatim values, a clean value swap, no duplicates); two readers, blind to each other and to all model output, then checked every edited control, and an item was kept only if both kept it. Articles used during development are excluded. Each prompt (23 items $\times$ 3 versions $\times$ 4 instructions $=$ 276) is run 16 times.

Wikipedia contradictions (Study 3) 47 further contradictions, one per article and none among the 23, from a screen of 1,052 Wikipedia articles that carry an intra-article contradiction tag (WikiContradict articles and development articles excluded): 88 candidate pairs (at most one per article), 54 kept by two blind readers, 47 kept by a third reader. Passages, agreeing controls and the named passage k are built as in Study 1. The 11 instructions give 47 $\times$ 3 $\times$ 11 $=$ 1,551 prompts, run 4 times (Qwen3.8-27B INT4), 16 times (gpt-oss-20b) or 8 times (Olmo-3-7B-Think, gpt-oss-120b).

Known-answer worlds (Studies 2 and 3) Three generators build small document sets around one question: who held a role on a date, which rate applied on a day after a delayed activation, and how many members two cohorts share. Two documents answer the question directly with different values; the other documents determine which value is right (settleable), are relevant but do not determine it (unsettleable), or have the determining record removed (masked); in agreeing items both reports agree. Settleable items come in twin pairs that differ only in which value the evidence supports, so a policy that ignores the evidence fails one twin of every pair. Every world is checked by exhaustive enumeration of its admissible assignments (a unique answer that reverses between settleable twins; both values admissible when unsettleable) and by an independent text-only validator (500 of 500 items per generator); a classifier on word, character and layout features separates settleable from unsettleable items near chance (AUC 0.536, 0.528, 0.506 by generator). Study 2 uses 300 settleable pairs (100 worlds per generator) and 150 unsettleable, 150 masked and 150 agreeing items (50 per generator), one call per item and instruction; the Study 3 known-answer contrasts re-run their reference instructions on the same items in the same job. The transfer test uses 150 unsettleable items no model had seen and the 300 settleable worlds, with the unresolved instruction and ET-3 in the same run (one call per item for Qwen3.8-27B INT4, two for the gpt-oss models).

Models and serving Table 7. All models were served with vLLM 0.28.0 with top-k disabled; K2-Horizon-7B, which has no native vLLM implementation, ran through vLLM's Transformers backend.

Table 6: Every instruction sentence, verbatim. The system message is the base text followed by one sentence; k is the entry or passage named for that item (balanced across items). ET-3 is the stopping-rule wording of the transfer test; CR-Q and PD-Q-1 are the known-answer contrasts of Study 3.

Condition	Label	Sentence
Study 1, constructed notes Base: "You are a helpful assistant. Answer using only the note and the table below. Give the final line as `ANSWER: <value>`."		
unresolved	V0	In these notes the appendix gives background only, and readers often skip it when they are short of time.
stated rule	V1	In these notes entry k states the operative rule, and that settles it for every question asked here.
directive	D	For this exercise, answer from entry k; that is a convenience and says nothing about which entry is right.
permission	P	If the two entries disagree, you may answer from either entry, and either value counts as a full answer.
licence	L	If the two entries disagree, you may answer CONFLICT instead of a value, and that counts as a full answer.
In-note line: *no clause* "Edition. This note follows the printed index for entries."; *precedence clause* "Precedence. If the two entries disagree, entry k controls."		
Studies 1 and 3, Wikipedia passages (Study 1: unresolved, wording-a stated rule, directive and permission, licence; Study 3: all but the licence). Base: "You are a helpful assistant. Answer using only the two passages below. Give the final line as `ANSWER: <value>`."		
unresolved		In these passages the wording follows the source article, and readers often skim it when they are short of time.
stated rule	a	In these passages passage k states the operative fact, and that settles it for every question asked here.
	b	Passage k states the operative fact in these passages; for every question asked here, that settles it.
directive	a	For this exercise, answer from passage k; that is a convenience and says nothing about which passage is right.
	b	For this exercise, use passage k to answer; this is a convenience and says nothing about which passage is right.
permission	a	If the two passages disagree, you may answer from either passage, and either value counts as a full answer.
	b	When the two passages disagree, either passage may supply your answer, and either value counts as a full answer.
command	a	If the two passages disagree, give the value from passage k, and either value counts as a full answer.
	b	When the two passages disagree, passage k must supply your answer, and either value counts as a full answer.
assertion	a	If the two passages disagree, only passage k states the operative fact, and either value counts as a full answer.
	b	When the two passages disagree, the operative fact is only in passage k, and either value counts as a full answer.
licence (Study 1)		If the two passages disagree, you may answer CONFLICT instead of a value, and that counts as a full answer.
Studies 2 and 3, known-answer documents Base: "You are a helpful assistant. Answer using only the documents below. Give the final line as `ANSWER: <value>`."		
unresolved	U	The documents below were collected from different places and are shown in no particular order.
stopping rule	ET-1	If the documents disagree, check whether the other documents settle which answer is right; if they do not, answer CONFLICT.
	ET-2	When the documents conflict, give the answer the remaining documents establish; if they establish neither, answer CONFLICT.
	ET-3	If the documents disagree, check whether the other documents alone determine the answer without assumptions; if they do not, answer CONFLICT.
directive	DA-1	If the documents disagree, give the answer from the first-listed document that directly answers the question; this is a convenience and says nothing about which is right.
	DA-2	When the documents conflict, report the answer of the first-listed document that directly answers the question; the choice is arbitrary and does not mean that document is correct.
permission	P-1	If the documents disagree, you may give the answer from either document that directly answers the question; either counts as a full answer.
	P-2	When the documents conflict, you may report the answer of either document that directly answers the question; either one is an acceptable, complete answer.
licence	LC	If the documents disagree, you may answer CONFLICT instead of a value, and that counts as a full answer.
computed rule	CR-Q-1	If the documents disagree, give the smaller of the answers from the documents that directly answer the question; this is a convenience and says nothing about which is right.
	CR-Q-2	When the documents conflict, report the smaller of the answers of the documents that directly answer the question; the choice is arbitrary and does not mean that answer is correct.
command	PD-Q-1	If the documents disagree, give the answer from the first-listed document that directly answers the question; either counts as a full answer.

Table 7: Models, settings and generation caps (maximum completion tokens per call). Capped: responses at the cap among the 512 unresolved conflicting responses on constructed notes (Study 1).

Model	Weights (revision)	Reasoning	T / top-p	Cap	Capped	Studies
Qwen3.8-27B INT4	RedHatAI/Qwen3.8-27B-INT4 (2fb0debc)	thinking on	0.6 / 1.0	32,768[a] / 126,976	13	1, 2, 3, F
Qwen3.8-27B bf16	Qwen/Qwen3.8-27B (1d4bf0f2)	thinking on	0.6 / 1.0	61,440	0	1
gpt-oss-20b	openai/gpt-oss-20b (6cee5e81)	effort high	1.0 / 1.0	126,976	0	1, 2, 3, F
Nemotron-3-Nano-30B	nvidia/NVIDIA-Nemotron-3-Nano-30B-A3B-BF16 (bf77c317)	thinking on	1.0 / 1.0	61,440	0	1
Olmo-3-32B-Think	allenai/Olmo-3-32B-Think (f2edda15)	always on	0.6 / 0.95	28,672	9	1
K2-Horizon-7B	IFM/K2-Horizon-7B (eca6bb60)	effort high	1.0 / 0.95	61,440	223	1
Olmo-3-7B-Think	allenai/Olmo-3-7B-Think (d97e442d)	always on	0.6 / 0.95	61,440	–	2, 3
gpt-oss-120b	openai/gpt-oss-120b (b5c939de)	effort high	1.0 / 1.0	126,976	–	3, F

[a] On constructed notes; 126,976 on every other item. F: factorial experiment of Section 6 (Appendix F). No response reached the cap in Study 3 or the transfer test; in Study 2, two Olmo-3-7B-Think responses did and none of the other models'.

C MEASUREMENT AND STATISTICS

Cost and ratios The cost of a call is its number of completion tokens, reasoning and answer together (at least 1). For each unit (a constructed note, pooling its 8 inputs and repetitions; a Wikipedia article, pooling its repetitions) and instruction X, let $\bar{c}_X$ be the mean cost of the conflicting version and $\bar{g}_X$ the average of the mean costs of its two agreeing controls under the same instruction; the unit's log premium is $\ell_X = \ln(\bar{c}_X/\bar{g}_X)$. The ratio "$X$ / agreeing" is exp of the mean of ℓ_X over units (a geometric mean), and a contrast "A / B" is exp of the mean over units of $\ell_A - \ell_B$, paired within unit, so each instruction is measured against its own agreeing controls. Intervals are 95% t intervals on the log scale with $n-1$ degrees of freedom ($n = 32$ notes, 23 or 47 articles), exponentiated. The share of the unresolved log premium that X removes is $(\bar{\ell}_U - \bar{\ell}_X)/\bar{\ell}_U$.

Known-answer items The unit is a world, resampled with its twins, repetitions and instructions; every estimate is the equal-weight mean of the three generators' world means. A cost ratio "A / B" is exp of the world mean of $\ln(c_A/c_B)$ on the same item (settleable twins averaged). A settleable pair fails unless both twins are answered with the correct value (per world, averaged over repetitions). Intervals come from 10,000 resamples of worlds within generator (seed 3500): basic bootstrap for contrasts and differences, percentile for shares and per-instruction values; when no pair changes we give the exact one-sided 97.5% Clopper–Pearson bound.

Simultaneous intervals The 95% interval is the primary display. As the conservative check, each declared family gets a Bonferroni family-wise 95% interval with the same point estimate and standard error, computed from the unrounded estimates: quantile $t_{1-0.025/m}$ for the $m = 12$ Study 1 unresolved reads (df 31 or 22), the $m = 10$ Study 1 permission / directive reads, the $m = 32$ Study 3 permission-frame contrasts (four contrasts $\times$ four models $\times$ two wordings; df 46), the $m = 8$ Study 3 directive / stated-rule contrasts, and the $m = 18$ factorial contrasts of Figure 8 (df 46); and $z_{1-0.025/m}$ for the $m = 12$ transfer-test outcomes, whose standard error is the half-width of the 95% bootstrap interval (log scale for ratios) over $z_{0.975}$. The main text reports, for every pattern statement, how many intervals of each kind exclude no-difference whenever the two counts differ.

Answers and grading The answer is the last line of the form `ANSWER: value` after the model's reasoning block (for gpt-oss, in the final channel), after removing trailing end-of-sequence tokens and, in Study 3 and the transfer test, markdown emphasis. On known-answer items, a response with no such line after its reasoning block is graded from one inside it. Wikipedia answers are matched to the two passage values (either value, both, neither, `CONFLICT`, or none; `CONFLICT` when the first word is "conflict"). Known-answer answers are correct when the matched candidate equals the true value; `CONFLICT`, both values, other strings, no answer and a capped response are not correct. A response follows an instruction when it gives the value the instruction selects (the named passage's or document's value; for the computed rule, the smaller value; for permission, either value). A response at the generation cap keeps the cap as its cost.

Decision curve With a share p of conflicts unsettleable, an instruction's expected accuracy is $(1-p)(1-e) + p\,h$ and its expected cost $(1-p)\,t_E + p\,t_N$ (e: share of settleable pairs failed; h: `CONFLICT` share on unsettleable items; t_E, t_N: mean completion tokens per settleable and unsettleable item). Both are linear in p, so the break-even share at which the stopping rule (ET-3)

becomes cheaper than no instruction (U) is exact: $p^* = (t_E^{\mathrm{ET}} - t_E^{\mathrm{U}})/[(t_E^{\mathrm{ET}} - t_E^{\mathrm{U}}) + (t_N^{\mathrm{U}} - t_N^{\mathrm{ET}})]$, and the accuracy crossing follows in the same way from e and h (Table 14). The 95% world-bootstrap bands of Figure 7 are computed at steps of 0.05 in p.

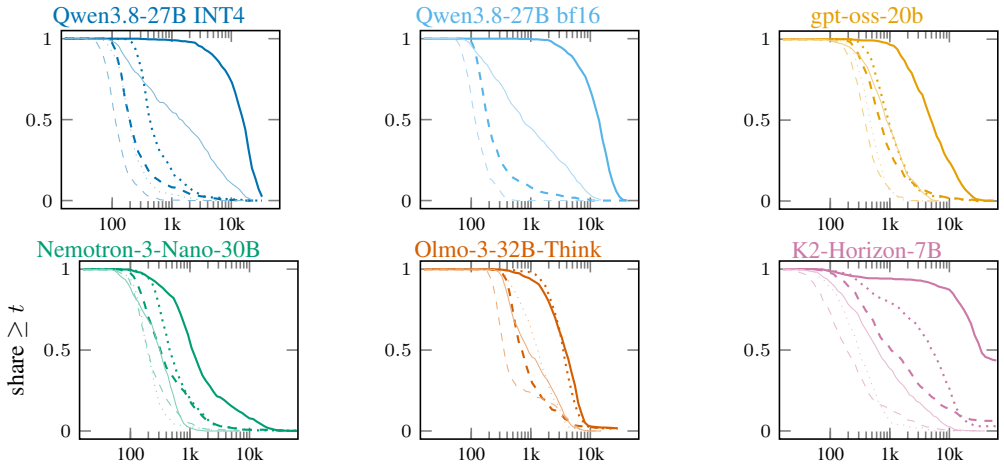
D RESULTS BY STUDY

Every estimate behind the main text, per model and wording, with its 95% interval. Ratios read "A / B" and exceed 1 when A costs more.

D.1 STUDY 1: THE COST OF AN UNRESOLVED CONFLICT AND WHAT REMOVES IT

Every unresolved / agreeing interval lies above 1, in all six configurations on both corpora (Table 8), and the conflicting version costs more than its agreeing controls on at least 20 of the 23 Wikipedia articles in every configuration (Table 9). On constructed notes the precedence clause, the stated rule and the directive leave 0.99–1.60 in five configurations; in K2-Horizon-7B they leave 1.39, 2.38 and 20.91 [13.33, 32.78], the last with 32 of 512 directive responses at the cap (Appendix G). Permission costs 1.25–3.66 times the directive in all ten configuration $\times$ corpus pairs, with the 95% interval excluding 1 in nine (the exception is Nemotron on Wikipedia, 1.25 [0.99, 1.58]) and the simultaneous interval over the ten in eight (Nemotron and K2-7B on Wikipedia include 1). The directive costs less than the stated rule by point estimate in 11 of the 12 reads; on K2-7B's constructed notes it costs 20.91 against 2.38. On Olmo-3-32B-Think permission leaves the premium in place: 4.11 on constructed notes against 4.39 unresolved, and 3.76 on Wikipedia against 2.99 unresolved (2.90 in the permission run's own measurement).

Robustness of the constructed-note premium (Qwen3.8-27B) Raising the INT4 model's generation cap from 32,768 to 126,976 tokens raises its unresolved ratio from 57.42 to 61.88 [46.95, 81.56], with 2 instead of 13 of 512 responses at the cap; the unquantized bf16 weights give 51.52 [39.20, 67.72] with none at the cap. The numeric question labels, whose parity is aligned with the counterbalancing of constants, pulled the INT4 model's conflicting answers toward constant a (document mean 0.824 [0.777, 0.871]); replacing them with letter codes brings that share to 0.540 [0.470, 0.610] and raises the unresolved ratio to 76.27 [55.73, 104.39] (72 of 512 at the cap, so also a lower bound).



Solid: unresolved; dashed: directive; dotted: permission. Dark lines: constructed notes; light lines: Wikipedia. Common log axis, 14 to 65,000 tokens, in every panel. Qwen bf16 had no permission run.

Figure 9: Distribution of reasoning length on conflicting items: share of responses with at least t reasoning tokens, per configuration (Study 1). K2-Horizon-7B's constructed unresolved curve ends at the generation cap with 223 of 512 responses (44%) still running.

D.2 STUDY 2: KNOWN-ANSWER ITEMS

Table 10 gives accuracy, cost and `CONFLICT` shares for every instruction and model, and Table 11 the cost of the same conflict when the other documents settle it and when they do not, including gpt-oss-120b's measurement from Study 3. Olmo-3-7B-Think fails 0.337 of settleable pairs with

Table 8: Study 1 cost ratios with 95% intervals (t intervals over documents or articles). Each row reads "A / B": the geometric-mean completion-token cost of A divided by that of B; values above 1 mean A costs more. Rows "X / agreeing" compare conflicting items under instruction X with their matched agreeing controls under the same instruction. Below each permission / directive row: Bonferroni simultaneous 95% interval over the ten permission-run reads.

A / B	Qwen INT4	Qwen bf16	gpt-oss-20b	Nemotron	Olmo-32B	K2-7B
Constructed notes (32)						
unresolved / agreeing	57.42 [42.08,78.37]	51.52 [39.20,67.72]	10.50 [8.34,13.21]	5.58 [4.05,7.69]	4.39 [3.59,5.36]	$\geq$176.20 [108.17,287.02]
precedence clause / agreeing	1.28 [1.20,1.36]	1.23 [1.17,1.30]	1.23 [1.15,1.32]	1.60 [1.38,1.86]	1.25 [1.13,1.39]	1.39 [1.24,1.55]
stated rule / agreeing	1.24 [1.14,1.35]	1.47 [1.22,1.77]	1.39 [1.27,1.53]	1.59 [1.31,1.92]	1.24 [1.14,1.35]	2.38 [1.88,3.00]
clause and rule / agreeing	1.07 [1.01,1.13]	1.07 [0.99,1.15]	1.13 [1.06,1.19]	1.11 [0.98,1.26]	1.05 [0.98,1.13]	1.18 [1.08,1.30]
directive / agreeing	1.00 [0.94,1.06]	0.99 [0.93,1.05]	1.13 [1.04,1.22]	1.54 [1.30,1.83]	1.14 [1.07,1.22]	20.91 [13.33,32.78]
licence / agreeing	5.55 [4.50,6.85]	4.88 [3.91,6.11]	1.87 [1.53,2.28]	1.60 [1.27,2.02]	1.13 [0.97,1.30]	17.58 [12.05,25.66]
permission / agreeing†	2.41 [2.17,2.69]	n/a	1.91 [1.66,2.20]	1.98 [1.75,2.25]	4.11 [3.39,4.99]	$\geq$35.40 [23.66,52.96]
permission / directive†	2.39 [2.14,2.66]	n/a	1.66 [1.43,1.93]	1.44 [1.26,1.64]	3.66 [3.02,4.44]	1.83 [1.29,2.58]
simultaneous (10 reads)	sim [2.03,2.80]	n/a	sim [1.33,2.07]	sim [1.18,1.74]	sim [2.76,4.87]	sim [1.09,3.05]
Wikipedia contradictions (23)						
unresolved / agreeing	15.23 [7.51,30.87]	10.03 [5.27,19.12]	3.73 [2.70,5.14]	2.04 [1.57,2.66]	2.99 [2.04,4.38]	5.85 [3.63,9.41]
stated rule / agreeing	1.41 [1.18,1.69]	1.43 [1.22,1.67]	1.50 [1.22,1.85]	1.34 [1.10,1.64]	1.35 [1.04,1.76]	1.63 [1.17,2.28]
directive / agreeing	1.02 [0.96,1.09]	1.03 [0.98,1.08]	1.07 [0.98,1.17]	1.20 [0.96,1.50]	1.34 [1.07,1.67]	1.59 [1.20,2.13]
licence / agreeing	3.48 [2.42,5.00]	2.47 [1.84,3.31]	1.28 [1.09,1.49]	1.30 [1.06,1.58]	1.12 [0.96,1.30]	1.59 [1.28,1.97]
permission / agreeing†	2.52 [2.16,2.93]	n/a	1.91 [1.76,2.08]	1.53 [1.27,1.84]	3.76 [2.82,5.03]	2.21 [1.78,2.73]
permission / directive†	2.34 [1.96,2.79]	n/a	1.64 [1.40,1.91]	1.25 [0.99,1.58]	3.11 [2.05,4.73]	1.47 [1.03,2.09]
simultaneous (10 reads)	sim [1.79,3.06]	n/a	sim [1.30,2.07]	sim [0.88,1.78]	sim [1.66,5.84]	sim [0.86,2.50]

†Separate permission run on the same items, with the directive regenerated alongside; permission / directive is the paired within-run contrast (the directive's own ratio in that run is the permission row divided by it). Qwen bf16 had no permission run; Wikipedia items have no in-document clause. K2-7B's constructed unresolved value is a lower bound: 223 of 512 of those responses reached the 61,440-token generation cap (Appendix G). The permission run re-measured the unresolved condition on Wikipedia: Qwen INT4 15.65×, gpt-oss-20b 4.10×, Nemotron 2.32×, Olmo-32B 2.90×, K2-7B 5.56×.

no instruction (101 of 300; Qwen3.8-27B INT4 0.003, gpt-oss-20b 0.010), and the stopping rule's saving on unsettleable items follows its `CONFLICT` share in every model.

D.3 Study 3: 47 further contradictions and known-answer contrasts

Table 3 gives the completion-token contrasts with 95% and simultaneous intervals; Table 12 adds the unresolved rows, every contrast on reasoning tokens, and how often each selecting instruction was followed. Every reasoning-token interval lies on the same side of 1 as its completion-token interval. The permission / assertion ratio differs in direction across models: below 1 on gpt-oss-20b and gpt-oss-120b, where the assertion costs more than permission; above 1 on Olmo-3-7B-Think, with both 95% intervals and the wording-a simultaneous interval excluding 1; and with every interval containing 1 on Qwen3.8-27B INT4. On known-answer items (Table 13) the computed rule, which uses both sources, costs 0.78 [0.68, 0.90] and 0.75 [0.66, 0.85] times the directive on Qwen3.8-27B INT4 but 1.17 [1.04, 1.33] and 1.03 [0.92, 1.16] on gpt-oss-20b; on Olmo-3-7B-Think, which followed the command on 0.68 of unsettleable items, permission costs 0.88 [0.78, 1.00] times the command.

D.4 The stopping rule on 150 new unsettleable items

The transfer test pairs 150 unsettleable items that no model had seen with the 300 settleable worlds of Study 2. Table 5 gives its four measures; Table 14 adds where the settleable overhead sits (in reasoning tokens; by generator), the settleable failure rates under each instruction, the mean tokens

Table 9: Unresolved cost ratio (conflicting / agreeing, geometric mean over 16 repetitions) for each of the 23 Wikipedia contradictions of Study 1. Italic: the conflicting version costs less than its agreeing controls. The geometric mean of each column is the corresponding Wikipedia entry of Table 8.

Article	Qwen INT4	Qwen bf16	gpt-oss-20b	Nemotron	Olmo-32B	K2-7B
1961 All-Ireland Senior Camogie Championship final	11.96	3.94	5.46	1.88	2.03	6.71
All Pakistan Federation of United Trade Unions	37.25	14.26	3.00	1.82	10.71	2.11
Angermünde	23.02	11.26	4.09	1.85	6.00	4.69
Ayapana triplinervis	130.98	85.65	7.97	3.67	7.30	7.56
Battle of Martín García	1.35	*0.85*	1.36	1.07	1.52	5.22
Battle of Monterrey	56.23	40.55	8.24	1.37	3.72	4.72
Beat Hotel	49.88	34.61	5.47	4.54	11.26	19.49
Blasphemy law in the United States	15.95	8.46	3.56	2.40	1.10	2.29
Blaster (flamethrower)	1.27	2.01	2.63	1.25	*0.54*	3.82
Bōsōzoku	5.21	4.60	1.40	1.27	2.68	1.03
Chartreuse (liqueur)	27.79	19.56	3.18	4.63	1.50	7.23
Convenience store	*0.95*	1.00	1.13	*0.71*	*0.93*	*0.82*
David Brady	122.12	74.06	6.42	2.20	8.11	23.34
Ger Browne	1.99	2.00	2.22	2.08	2.13	3.51
John Henry Seadlund	151.82	99.87	7.80	2.71	3.08	22.18
List of association football stadiums by country	130.13	77.67	9.28	4.95	8.50	30.06
Mari Fitzduff	2.82	2.61	2.71	1.55	1.37	2.21
Marina Baker	22.33	6.17	5.13	3.11	3.90	22.73
Mary Astor	44.93	19.91	7.27	5.19	7.31	18.15
Muhammad III of Alamut	7.03	6.98	7.70	1.69	2.70	8.10
October 2010 North American storm complex	66.99	48.43	9.40	4.36	7.37	26.12
Pierre Albert-Birot	3.49	1.74	1.07	*0.82*	*0.93*	1.49
Shaikh Asiri Lahiji	5.01	4.99	1.12	*0.76*	1.97	2.16
Geometric mean	15.23	10.03	3.73	2.04	2.99	5.85

Table 10: Study 2, known-answer items, per instruction (w. = wording). Settleable: share of the 300 settleable pairs failed (a pair fails unless both twins are answered with the correct value) and share of settleable responses that are CONFLICT. Unsettleable and masked (150 items each): cost relative to U (geometric mean of the per-item ratio; U = 1) and CONFLICT share. 95% world-cluster bootstrap intervals (10,000 resamples within generator). Settleable items were not run under DA-2, P-1 and P-2.

	Settleable		Unsettleable		Masked	
Instruction	pairs failed	CONFLICT	cost / U	CONFLICT	cost / U	CONFLICT
Qwen3.8-27B INT4						
unresolved (U)	0.003 [0.000,0.010]	0.000	1	0.000	1	0.000
stopping rule, w. 1 (ET-1)	0.003 [0.000,0.010]	0.000	0.331 [0.294,0.374]	0.987	0.124 [0.109,0.143]	1.000
stopping rule, w. 2 (ET-2)	0.000 [0.000,0.000]	0.000	0.289 [0.257,0.325]	1.000	0.105 [0.090,0.124]	1.000
directive, w. 1 (DA-1)	0.920 [0.890,0.950]	0.000	0.085 [0.076,0.096]	0.000	0.030 [0.027,0.034]	0.000
directive, w. 2 (DA-2)	n/a	n/a	0.070 [0.062,0.079]	0.000	0.027 [0.024,0.030]	0.000
permission, w. 1 (P-1)	n/a	n/a	0.185 [0.164,0.209]	0.000	0.051 [0.046,0.058]	0.000
permission, w. 2 (P-2)	n/a	n/a	0.159 [0.140,0.181]	0.000	0.050 [0.044,0.056]	0.000
licence (LC)	0.373 [0.333,0.413]	0.243	0.249 [0.217,0.287]	0.993	0.063 [0.055,0.073]	1.000
gpt-oss-20b						
unresolved (U)	0.010 [0.000,0.023]	0.000	1	0.000	1	0.000
stopping rule, w. 1 (ET-1)	0.043 [0.023,0.067]	0.018	0.605 [0.530,0.693]	0.807	0.190 [0.167,0.216]	1.000
stopping rule, w. 2 (ET-2)	0.043 [0.023,0.067]	0.017	0.643 [0.563,0.734]	0.760	0.227 [0.200,0.257]	1.000
directive, w. 1 (DA-1)	0.967 [0.947,0.987]	0.000	0.212 [0.186,0.241]	0.000	0.095 [0.085,0.106]	0.000
directive, w. 2 (DA-2)	n/a	n/a	0.191 [0.168,0.218]	0.000	0.093 [0.083,0.105]	0.000
permission, w. 1 (P-1)	n/a	n/a	0.341 [0.298,0.389]	0.000	0.129 [0.113,0.146]	0.000
permission, w. 2 (P-2)	n/a	n/a	0.304 [0.266,0.347]	0.000	0.110 [0.097,0.124]	0.000
licence (LC)	0.537 [0.490,0.583]	0.357	0.484 [0.419,0.559]	0.820	0.116 [0.101,0.133]	0.993
Olmo-3-7B-Think						
unresolved (U)	0.337 [0.287,0.387]	0.000	1	0.000	1	0.000
stopping rule, w. 1 (ET-1)	0.423 [0.367,0.477]	0.115	1.131 [1.027,1.240]	0.213	0.637 [0.555,0.726]	0.760
stopping rule, w. 2 (ET-2)	0.323 [0.273,0.373]	0.052	1.148 [1.046,1.265]	0.160	0.626 [0.552,0.707]	0.733
directive, w. 1 (DA-1)	0.660 [0.610,0.707]	0.000	0.935 [0.842,1.042]	0.000	0.454 [0.397,0.517]	0.000
directive, w. 2 (DA-2)	n/a	n/a	1.046 [0.944,1.161]	0.000	0.455 [0.396,0.520]	0.000
permission, w. 1 (P-1)	n/a	n/a	0.919 [0.834,1.012]	0.000	0.914 [0.823,1.011]	0.000
permission, w. 2 (P-2)	n/a	n/a	0.883 [0.802,0.972]	0.000	0.878 [0.796,0.966]	0.000
licence (LC)	0.877 [0.840,0.910]	0.452	0.848 [0.757,0.947]	0.380	0.374 [0.326,0.426]	0.793

Table 11: Cost of the same conflict when the other documents settle it and when they do not, under the unresolved instruction (ratios of geometric-mean completion tokens, 95% world-cluster bootstrap intervals). Qwen, gpt-oss-20b and Olmo-7B: first 50 worlds of each generator (150 worlds; settleable twins averaged). gpt-oss-120b: the 50 worlds of the first generator.

A / B	Qwen3.8-27B INT4	gpt-oss-20b	Olmo-3-7B-Think	gpt-oss-120b
settleable / agreeing	2.08 [1.95,2.23]	1.94 [1.78,2.11]	2.56 [2.36,2.79]	n/a
unsettleable / agreeing	37.70 [34.31,41.19]	6.77 [6.02,7.61]	2.72 [2.43,3.03]	n/a
masked / agreeing	33.98 [31.54,36.59]	7.27 [6.51,8.12]	3.69 [3.32,4.12]	n/a
unsettleable / settleable	18.09 [16.40,19.88]	3.50 [3.07,3.96]	1.06 [0.97,1.16]	3.81 [3.28,4.43]
masked / settleable	16.31 [15.08,17.63]	3.75 [3.32,4.25]	1.44 [1.31,1.59]	6.44 [5.48,7.66]

Answers under U on the same unsettleable and masked items (value a / value b / both / other, of 150): Qwen3.8-27B INT4 62/50/4/34 and 41/33/0/76; gpt-oss-20b 68/80/1/1 and 60/68/13/9; Olmo-3-7B-Think 84/63/0/3 and 75/74/0/1. Settleable twins answered correctly under U: Qwen3.8-27B INT4 300/300; gpt-oss-20b 297/300; Olmo-3-7B-Think 244/300.

Table 12: Study 3 on the 47 articles, per model and wording (a, b; Table 6): ratio A / B with 95% t intervals over the 47 articles (df $= 46$). (a) Unresolved conflicts, completion tokens. (b) Every contrast of Table 3 computed on reasoning tokens alone. (c) Share of conflicting responses giving passage k's value. No Study 3 response reached the generation cap.

A / B	w.	Qwen3.8-27B	gpt-oss-20b	Olmo-3-7B-Think	gpt-oss-120b
(a) unresolved conflicts, completion tokens					
unresolved / agreeing		17.92 [10.53,30.50]	3.84 [2.99,4.93]	2.86 [2.21,3.70]	3.08 [2.53,3.75]
unresolved / directive	a	17.17 [10.10,29.19]	3.39 [2.59,4.45]	2.63 [1.99,3.46]	2.89 [2.34,3.57]
	b	16.18 [9.25,28.29]	3.55 [2.67,4.73]	2.35 [1.74,3.18]	2.84 [2.28,3.53]
(b) reasoning tokens					
unresolved / agreeing		19.58 [11.33,33.85]	4.02 [3.10,5.20]	2.93 [2.26,3.80]	3.22 [2.62,3.95]
unresolved / directive	a	18.72 [10.86,32.26]	3.53 [2.67,4.67]	2.68 [2.03,3.54]	3.02 [2.42,3.75]
	b	17.54 [9.87,31.20]	3.71 [2.76,4.97]	2.39 [1.76,3.25]	2.95 [2.35,3.70]
directive / stated rule	a	0.80 [0.73,0.89]	0.88 [0.82,0.95]	0.83 [0.73,0.94]	0.91 [0.84,0.99]
	b	0.70 [0.61,0.81]	0.75 [0.71,0.81]	0.90 [0.79,1.02]	0.92 [0.84,1.01]
permission / directive	a	2.03 [1.74,2.37]	1.66 [1.47,1.86]	2.82 [2.14,3.72]	1.42 [1.31,1.55]
	b	1.86 [1.54,2.25]	1.51 [1.34,1.70]	2.69 [2.04,3.54]	1.39 [1.28,1.51]
permission / command	a	1.39 [1.16,1.66]	1.59 [1.45,1.74]	2.45 [1.93,3.12]	1.30 [1.23,1.38]
	b	1.46 [1.24,1.73]	1.32 [1.20,1.44]	2.51 [1.93,3.28]	1.26 [1.16,1.36]
permission / assertion	a	0.99 [0.81,1.20]	0.52 [0.46,0.58]	1.63 [1.29,2.06]	0.67 [0.60,0.74]
	b	0.87 [0.73,1.05]	0.47 [0.42,0.52]	1.42 [1.12,1.79]	0.73 [0.66,0.80]
assertion / command	a	1.41 [1.23,1.62]	3.06 [2.70,3.48]	1.51 [1.30,1.75]	1.95 [1.76,2.16]
	b	1.67 [1.49,1.88]	2.82 [2.54,3.13]	1.78 [1.52,2.08]	1.73 [1.60,1.87]
(c) share of conflicting responses giving passage k's value					
stated rule	a / b	0.99 / 0.99	0.99 / 0.98	0.87 / 0.87	0.99 / 0.93
directive	a / b	0.99 / 1.00	1.00 / 0.99	0.92 / 0.93	1.00 / 1.00
command	a / b	0.99 / 0.99	0.99 / 1.00	0.86 / 0.88	1.00 / 1.00
assertion	a / b	0.96 / 0.98	0.98 / 0.98	0.85 / 0.85	0.99 / 1.00

Table 13: Study 3 contrasts on known-answer items (Study 2 worlds; one call per item): ratio A / B of costs with 95% world-cluster bootstrap intervals. Computed-rule contrasts use the two numeric generators (100 worlds); permission / command uses all three (150). Share of unsettleable responses giving the value the instruction selects (for permission, either value), Qwen / gpt-oss-20b / Olmo: computed rule w. 1 1.00/1.00/0.73; computed rule w. 2 1.00/0.99/0.81; directive w. 1 1.00/0.99/0.70; directive w. 2 1.00/1.00/0.71; command 1.00/0.99/0.68; permission (either value) 1.00/1.00/0.98.

Items	A / B	Qwen3.8-27B INT4	gpt-oss-20b	Olmo-3-7B-Think
unsettleable	computed rule / directive, w. 1	0.78 [0.68,0.90]	1.17 [1.04,1.33]	0.92 [0.81,1.05]
unsettleable	computed rule / directive, w. 2	0.75 [0.66,0.85]	1.03 [0.92,1.16]	0.94 [0.81,1.10]
unsettleable	permission / command	2.37 [2.13,2.63]	1.48 [1.31,1.68]	0.88 [0.78,1.00]
masked	computed rule / directive, w. 1	0.85 [0.77,0.93]	0.82 [0.73,0.93]	0.92 [0.80,1.05]
masked	computed rule / directive, w. 2	0.93 [0.83,1.04]	0.87 [0.77,0.98]	0.84 [0.73,0.97]

behind the decision curves of Figure 7, the exact break-even shares and the answers behind gpt-oss-20b's accuracy change.

E Reasoning and answer tokens

Every response row stores the serving runner's split of its completion tokens into the reasoning segment (the thinking block; for gpt-oss models, the analysis channel) and the final answer, and no row lacks either count. We recomputed every Study 1, permission-run and Study 2 estimate twice, with reasoning tokens alone and with answer tokens alone in place of completion tokens, using

Table 14: Transfer test beyond Table 5: the stopping rule (ET-3) against no instruction (U) on the 150 new unsettleable items and the 300 settleable worlds of Study 2. Ratios ET-3 / U of geometric-mean tokens with 95% world-cluster bootstrap intervals (10,000 resamples within generator); by-generator values are point estimates; mean tokens per item are the arithmetic means behind Figure 7.

Measure	Qwen3.8-27B	gpt-oss-20b	gpt-oss-120b
settleable items			
pairs failed, U / ET-3 (%)	0.0 / 0.0	0.8 / 7.5	23.7 / 10.0
cost, completion tokens	1.36 [1.30,1.43]	1.36 [1.31,1.41]	1.35 [1.31,1.38]
cost, reasoning tokens	1.39 [1.32,1.46]	1.36 [1.31,1.42]	1.39 [1.35,1.43]
cost, answer tokens	0.85 [0.78,0.92]	0.96 [0.94,0.98]	0.39 [0.35,0.42]
cost by generator (G1 / G2 / G3)	1.15 / 1.19 / 1.85	1.45 / 1.20 / 1.45	1.54 / 1.22 / 1.29
mean tokens per item, U / ET-3	1,361 / 2,166	2,156 / 2,931	1,456 / 1,920
unsettleable items			
cost, completion tokens	0.12 [0.10,0.13]	0.54 [0.49,0.60]	0.41 [0.38,0.44]
cost, reasoning tokens	0.12 [0.10,0.13]	0.54 [0.49,0.60]	0.41 [0.38,0.44]
cost, answer tokens	0.36 [0.30,0.44]	0.80 [0.76,0.84]	0.39 [0.34,0.46]
mean tokens per item, U / ET-3	21,862 / 3,139	9,278 / 4,043	6,021 / 1,941
mixtures with unsettleable share p			
break-even p^* (exact crossing)	0.041	0.129	0.102
first cheaper point on the 0.05 grid	0.05	0.15	0.15
ET-3 at least as accurate for	every p	$p \geq 0.066$	every p
ET-3 cheaper and at least as accurate for	$p > 0.041$	$p > 0.129$	$p > 0.102$

Qwen3.8-27B failed no settleable pair under either instruction; the exact one-sided 97.5% upper bound on its failure increase is 1.22 points. gpt-oss-20b's failing settleable responses: under ET-3, 51 `CONFLICT` answers and 2 wrong values; under U, 5, all wrong values. The break-even p^* solves $(1-p)\,t_E^{\mathrm{ET}} + p\,t_N^{\mathrm{ET}} = (1-p)\,t_E^{\mathrm{U}} + p\,t_N^{\mathrm{U}}$ with the mean tokens above (Appendix C).

the same code, units and intervals; Study 3 carries the reasoning-token version of every contrast (Table 12b). On conflicting items, reasoning is 93.4–99.7% of completion tokens on constructed notes, 89.2–98.4% on Wikipedia, 91.5–99.7% in the permission run, 98.6–99.7% in Qwen3.8-27B INT4's rerun at the higher cap and 96.5–99.8% on known-answer items (range of the mean share over conditions and models); on unresolved conflicting items of Study 1 it is 92.6–99.7%. Table 15 gives the unresolved ratios on each count. Over every Study 1 and permission-run instruction effect, the log effect on reasoning tokens differs from the one on completion tokens by at most 0.087, while the unresolved ratio on answer tokens stays between 0.73 and 2.27.

Table 15: Unresolved cost ratio (conflicting / agreeing) computed on completion tokens, on reasoning tokens alone and on answer tokens alone, Study 1 (95% t intervals). K2-7B's constructed completion and reasoning values are lower bounds (generation cap).

Model	completion	reasoning	answer
Constructed notes			
Qwen3.8-27B INT4	57.42 [42.08,78.37]	61.49 [44.58,84.81]	2.11 [1.86,2.40]
Qwen3.8-27B bf16	51.52 [39.20,67.72]	55.26 [41.68,73.28]	2.27 [2.01,2.56]
gpt-oss-20b	10.50 [8.34,13.21]	10.82 [8.56,13.66]	0.93 [0.81,1.06]
Nemotron-3-Nano-30B	5.58 [4.05,7.69]	5.91 [4.27,8.18]	0.73 [0.64,0.83]
Olmo-3-32B-Think	4.39 [3.59,5.36]	4.43 [3.62,5.43]	0.98 [0.95,1.01]
K2-Horizon-7B	$\geq$176.20 [108.17,287.02]	$\geq$187.01 [113.55,308.00]	0.87 [0.69,1.09]
Wikipedia			
Qwen3.8-27B INT4	15.23 [7.51,30.87]	16.62 [8.04,34.36]	1.36 [1.15,1.61]
Qwen3.8-27B bf16	10.03 [5.27,19.12]	10.89 [5.62,21.09]	1.41 [1.17,1.71]
gpt-oss-20b	3.73 [2.70,5.14]	3.88 [2.79,5.39]	1.39 [1.15,1.68]
Nemotron-3-Nano-30B	2.04 [1.57,2.66]	2.19 [1.63,2.92]	0.90 [0.79,1.03]
Olmo-3-32B-Think	2.99 [2.04,4.38]	3.05 [2.07,4.49]	1.01 [0.97,1.04]
K2-Horizon-7B	5.85 [3.63,9.41]	6.07 [3.74,9.86]	1.73 [1.31,2.27]

F FACTORIAL EXPERIMENT: DECISION FORM AND OPERATIVE-SOURCE CLAIM

The command and the assertion of Study 3 differ in two ways at once: grammatical form (imperative or declarative), and whether the sentence states a decision about what to report or the claim that the named passage is the only one stating the operative fact. This experiment crosses form with the claim inside the permission frame (Table 16).

Design Items: the conflicting versions of the 47 Study 3 items with the same named passage k; every contrast is between cells on the same item. Cells: the command, the declarative decision, the command with the claim and the claim alone, in each of two wordings, plus the permission sentence

Table 16: Factorial cells: X in "If [a] / When [b] the two passages disagree, X, and either value counts as a full answer." The declarative cell with the claim states no decision; it is the Study 3 assertion.

w.	form	no claim	operative-source claim
a	imperative	give the value from passage k (command)	give the value from passage k, which is the only passage that states the operative fact (command with claim)
	declarative	the value from passage k is the one to report (declarative decision)	only passage k states the operative fact (claim alone)
b	imperative	take the value from passage k (command)	take the value from passage k, which is the only passage that states the operative fact (command with claim)
	declarative	the value from passage k is the one to report (declarative decision)	the operative fact is only in passage k (claim alone)
a, b	declarative	n/a	the value from passage k, which is the only passage that states the operative fact, is the one to report (declarative decision with claim; follow-up run)

of each wording as reference (235 prompts per wording). The command of wording a is the Study 3 command, and the claim-alone cell is the Study 3 assertion in each wording. Models: Qwen3.8-27B INT4 (16 repetitions, as two seed blocks of 8), gpt-oss-20b and gpt-oss-120b (8 each), served as in Study 3; these are the three models on which the assertion did not cost less than permission in Study 3 (Table 3). Seed = base + repetition $\times$ number of prompts + prompt index, with base 4200 (4300 for Qwen's second block).

Estimands The outcome is completion tokens, with reasoning tokens reported alongside. For article i and cell x, y_{ix} is the log of the mean completion tokens over repetitions. With I_0 the command, A_0 the declarative decision, I_1 the command with the claim and A_1 the claim alone, the contrasts between cells, paired over articles, are $A_0 - I_0$ (form of the decision), $I_1 - I_0$ (claim attached to the command), $A_1 - A_0$ (claim in place of the decision) and $A_1 - I_1$, each reported as a ratio $\exp(\cdot)$ with a 95% t interval over the 47 articles (df $= 46$), and for the first three with a simultaneous interval over their 18 model $\times$ wording estimates. The averages over the four cells, $\frac{1}{2}[(A_0 - I_0) + (A_1 - I_1)]$ (declarative against imperative), $\frac{1}{2}[(I_1 - I_0) + (A_1 - A_0)]$ (with against without the claim) and their interaction $(A_1 - I_1) - (A_0 - I_0)$, are reported as well; because A_1 carries no decision, each average mixes form or claim with the loss of the decision. For every cell we also report the share of responses that give passage k's value.

Results Every decision and claim cell determines the response: 97.3–100% of rows give passage k's value (752 rows per cell on Qwen3.8-27B, 376 on each gpt-oss model; no capped rows). Against permission, the three cells that state a decision cost less (command 0.66–0.91, declarative decision 0.67–0.93, command with claim 0.79–0.96) and the claim alone costs more (1.23–2.87) across models and wordings. Between cells, the declarative decision costs 0.93–1.05 times the command; the claim attached to the command 0.96–1.27 times the command; the claim in place of the decision 1.55–3.71 times the declarative decision, every 95% and simultaneous interval above 1; and the claim alone 1.29–3.10 times the command with the claim. The averages over the four cells are 1.17–1.72 (declarative against imperative), 1.29–2.03 (with against without the claim) and 1.22–3.35 (interaction). Reasoning-token ratios track completion within 4% on the claim-in-place-of-decision contrast in all six model–wording cells, so the excess is reasoning tokens. Two independent implementations agree on all 48 contrasts of the design to 9×10^{-16}. Full estimates in Table 17.

G K2-HORIZON-7B AT THE GENERATION CAP

On constructed notes, 223 of K2-Horizon-7B's 512 unresolved conflicting responses reached the 61,440-token generation cap (the model accepts 524,288 tokens and was served at a 65,536-token context). We call a response a repetition loop when at least half of its last 4,000 characters is covered by strings of 50 or more characters that recur within its last 20,000 characters. 222 of the 223 capped responses are loops. Capped loops occur in 43.4% [36.3, 50.2] of unresolved responses, 0.6% under the precedence clause and under the stated rule, 0.2% under both, 6.3% [3.9, 9.0] under the directive and 0.8% under the licence, and in the permission run in 4.1% [2.3, 6.3] under the directive and 2.7% [1.6, 4.1] under permission (95% document-cluster bootstrap intervals). Sustained looping starts at no fixed length (median 19,712 tokens; 10th–90th percentile 11,584–30,144), and K2-Horizon-7B reached the cap on no Wikipedia item. Among responses without a detected loop, K2's ratios

Table 17: Factorial experiment (Section 6), completion tokens: ratio A / B with 95% t interval over the 47 articles; for the three contrasts of Figure 8 the second line is the Bonferroni simultaneous 95% interval over those 18 contrasts. Cells (Table 16): command; declarative decision; command with the operative-source claim; the claim alone (the Study 3 assertion). The averages over the four cells are exponentiated means of two log contrasts; because the claim-alone cell carries no decision, they combine form, claim and the loss of the decision. Compliance: share of responses giving passage k's value (rows).

A / B	Qwen3.8-27B	gpt-oss-20b	gpt-oss-120b
Wording a ("If the two passages disagree, X, and either value counts as a full answer.")			
cells against permission			
command / permission	0.75 [0.66,0.86]	0.75 [0.69,0.81]	0.91 [0.85,0.98]
declarative decision / permission	0.80 [0.69,0.91]	0.75 [0.69,0.81]	0.93 [0.88,0.99]
command with claim / permission	0.96 [0.80,1.15]	0.83 [0.74,0.93]	0.87 [0.82,0.92]
claim alone / permission	1.23 [1.06,1.43]	2.44 [2.17,2.75]	1.79 [1.65,1.94]
contrasts between cells			
declarative decision / command	1.05 [1.03,1.08] sim [1.01,1.10]	1.00 [0.95,1.05] sim [0.92,1.08]	1.03 [0.97,1.08] sim [0.94,1.12]
command with claim / command	1.27 [1.19,1.35] sim [1.15,1.40]	1.11 [1.03,1.19] sim [0.99,1.24]	0.96 [0.91,1.01] sim [0.88,1.04]
claim alone / declarative decision	1.55 [1.45,1.65] sim [1.40,1.71]	3.28 [2.99,3.60] sim [2.83,3.79]	1.92 [1.79,2.06] sim [1.72,2.14]
claim alone / command with claim	1.29 [1.18,1.40]	2.95 [2.68,3.25]	2.05 [1.92,2.20]
averages over the four cells			
declarative / imperative, averaged over the claim	1.17 [1.11,1.22]	1.72 [1.62,1.82]	1.45 [1.38,1.52]
with / without the claim, averaged over form	1.40 [1.34,1.47]	1.91 [1.78,2.04]	1.36 [1.29,1.42]
interaction (difference of the two form contrasts)	1.22 [1.12,1.33]	2.96 [2.69,3.26]	2.00 [1.84,2.17]
compliance: share giving passage k's value (rows)			
command	0.996 (752)	0.997 (376)	1.000 (376)
declarative decision	0.996 (752)	0.989 (376)	0.995 (376)
command with claim	0.980 (752)	0.995 (376)	1.000 (376)
claim alone	0.973 (752)	0.981 (376)	1.000 (376)
Wording b ("When the two passages disagree, X, and either value counts as a full answer.")			
cells against permission			
command / permission	0.66 [0.56,0.79]	0.84 [0.77,0.91]	0.87 [0.82,0.92]
declarative decision / permission	0.67 [0.58,0.78]	0.77 [0.71,0.84]	0.88 [0.84,0.93]
command with claim / permission	0.79 [0.67,0.92]	0.92 [0.83,1.03]	0.89 [0.84,0.94]
claim alone / permission	1.30 [1.09,1.54]	2.87 [2.59,3.18]	1.44 [1.34,1.54]
contrasts between cells			
declarative decision / command	1.02 [0.98,1.06] sim [0.95,1.08]	0.93 [0.87,0.99] sim [0.84,1.02]	1.01 [0.96,1.07] sim [0.94,1.10]
command with claim / command	1.19 [1.13,1.25] sim [1.10,1.28]	1.11 [1.03,1.19] sim [0.98,1.25]	1.02 [0.96,1.09] sim [0.93,1.13]
claim alone / declarative decision	1.93 [1.75,2.13] sim [1.65,2.25]	3.71 [3.33,4.13] sim [3.13,4.39]	1.63 [1.51,1.76] sim [1.45,1.84]
claim alone / command with claim	1.65 [1.52,1.79]	3.10 [2.76,3.49]	1.62 [1.51,1.73]
averages over the four cells			
declarative / imperative, averaged over the claim	1.30 [1.23,1.36]	1.69 [1.58,1.82]	1.28 [1.23,1.34]
with / without the claim, averaged over form	1.51 [1.42,1.62]	2.03 [1.89,2.17]	1.29 [1.22,1.36]
interaction (difference of the two form contrasts)	1.62 [1.49,1.77]	3.35 [2.96,3.80]	1.59 [1.46,1.74]
compliance: share giving passage k's value (rows)			
command	0.992 (752)	0.992 (376)	0.997 (376)
declarative decision	0.991 (752)	0.995 (376)	1.000 (376)
command with claim	0.989 (752)	0.997 (376)	1.000 (376)
claim alone	0.977 (752)	0.979 (376)	1.000 (376)
Follow-up (claim with a declarative decision), same 47 articles, completion tokens: A0C against A0 and against I1, ratio with 95% t interval. Bare-claim row is across runs (P3 claim alone over A0C). Compliance: share giving passage k's value (rows).			
claim with decision / declarative decision, wording a	1.48 [1.37,1.59]	1.13 [1.06,1.21]	1.10 [1.03,1.18]
claim with decision / declarative decision, wording b	1.47 [1.32,1.65]	1.13 [1.05,1.22]	1.14 [1.09,1.20]
claim with decision / command with claim, wording a	1.18 [1.12,1.26]	1.08 [1.02,1.14]	1.06 [1.00,1.12]
claim with decision / command with claim, wording b	1.30 [1.21,1.40]	1.00 [0.93,1.07]	1.07 [1.02,1.13]
bare claim / claim with decision, wording a, across runs	1.06 [1.00,1.12]	2.79 [2.58,3.03]	1.79 [1.66,1.94]
bare claim / claim with decision, wording b, across runs	1.26 [1.14,1.39]	3.21 [2.85,3.61]	1.49 [1.37,1.61]
compliance: claim with decision, wording a	0.984 (752)	0.997 (376)	1.000 (376)
compliance: claim with decision, wording b	0.993 (752)	0.987 (376)	0.997 (376)

of mean tokens, conflicting over agreeing, keep the same order: 31.2 unresolved, 5.1 directive, 1.9 stated rule, 1.2 precedence clause (descriptive, since this subset is selected on an outcome the instruction affects). Capped loops are rare elsewhere: 3 of 512 unresolved constructed responses for Olmo-3-32B-Think, none for Qwen bf16, gpt-oss-20b and Nemotron, 2 of 512 for Qwen INT4 in its rerun at the higher cap, and one response in total on Wikipedia. Every estimate keeps capped responses at the cap, so K2's constructed unresolved ratio, $\geq$176.20, is a lower bound on a cost that repetition drives.