WHEN FORMATTING DETERMINES RANKINGS: SEPARATING SEMANTIC ACCURACY, COMPLIANCE, AND STRICT ACCURACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model evaluations can conflate correctness, format adherence, and parser recognition. Across five models, MATH-500 and a balanced 13-domain MMLU-Pro subset, we vary final serialization while preserving the reasoning instruction and report extracted-answer accuracy, terminal-grammar compliance, and strict interface accuracy. The original 24,750 generations comprise a 17,250-response main experiment and a paired token-budget study. Paired intervals and multiplicity correction identify 12 strict-accuracy and 5 compliance rank reversals, but no reliable extracted-answer accuracy reversal. Nine original strict reversals persist under a wrapper-tolerant policy. A targeted 2,000-response, same-session intervention removes angle brackets around the answer placeholder, eliminating the Qwen32–DeepSeek-V3.2 Final/JSON strict point-order reversal on both tested subsets. A further study yields 3,000 outputs on held-out questions: complete prospective MMLU-Pro results support the prespecified model-by-format template interaction, while MATH results after an additional recovery request constitute a protocol-deviation sensitivity analysis, not overall prospective confirmation. Template effects are heterogeneous, including substantial accuracy declines in some conditions, and do not establish a universal repair. These findings demonstrate sensitivity of specified model–protocol snapshots, not invariant model deficiencies or latent capability ordering. Evaluations should report extracted-answer accuracy, compliance, and strict accuracy separately, alongside prompts, extraction rules, and token budgets.

1 INTRODUCTION

Benchmark leaderboards compress a complex interaction into one number. A model receives a prompt, produces a reasoning trace, serializes a final answer, and is then judged by an evaluator that must locate and verify that answer. Failures at any stage are usually recorded as the same event: an incorrect response. This is convenient operationally, but it makes a leaderboard score ambiguous. A low score can mean that the model reasoned incorrectly, that it ignored a formatting instruction, or that the evaluator failed to recognize a substantively correct answer.

Long reasoning traces make token cost and truncation consequential (Fan et al., 2025); structured-output requests can also affect task performance (Lee et al., 2026). Dedicated extractors and verifiers improve answer recovery and equivalence judgment (Yu et al., 2025; Chen et al., 2025). Other work studies answers recoverable through continuations from intermediate reasoning prefixes (Hammoud et al., 2025; Wang & Zhu, 2026). These are different mechanisms from terminal-interface success.

We ask a deliberately narrow question: *when the reasoning request, questions, decoding settings, and verifier are fixed, how much can changing only the required final serialization alter scores and model rankings?* We evaluate five models spanning local 7B checkpoints and larger API-served systems on a mathematical benchmark (MATH-500) and a non-mathematical, multi-domain benchmark (MMLU-Pro). Every prompt permits step-by-step reasoning and differs only in its required final line: `Final Answer: <answer>`, `\boxed{<answer>}`, or `{"answer":"<answer>"}`.

The central methodological choice is to report three quantities rather than one. *Semantic answer accuracy* uses a declared, gold-independent parser and task verifier. *Format compliance* checks the

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.

final nonempty line against the requested contract. *Strict protocol accuracy* requires compliance and a correct answer extracted through that contract. Strict-protocol rankings reverse repeatedly after multiplicity correction, whereas no reliable semantic reversal is detected. This concerns reported scores; it does not establish equivalence or stability of underlying capability ordering.

Our contributions are:

- A controlled, paired evaluation across five models, two benchmarks, three formats, and two budgets (24,750 original generations), plus a targeted 2,000-generation same-session template intervention.
- Evidence that interface-conditioned model gaps can reverse reliably even when extracted-answer reversals are not detected, with reversal-identity tracking under a changed terminal grammar. The decomposition itself is not claimed as new.
- A heterogeneous template intervention, followed by 3,000 held-out-instance outputs: complete prospective MMLU evidence and explicitly post-deviation MATH recovery sensitivity. Neither supports a universal repair.
- A traceable audit workflow connecting frozen-output rescoring, paired ranking tests, and a separately generated template intervention. Released artifacts include an external-extractor replay, regression tests, raw records, and the original targeted author audit; no manual score overrides are used.

2 Related Work

Answer extraction and verification. xFinder demonstrates model-ranking differences between extraction frameworks and answer representations; xVerify targets efficient equivalence judgment (Yu et al., 2025; Chen et al., 2025). Intermediate-prefix continuation studies investigate answer recoverability, rather than only parsing a fixed final response (Hammoud et al., 2025; Wang & Zhu, 2026). Our contribution is not the existence of extractor-dependent rankings, but paired terminal-serialization interventions, multiplicity-corrected interface-ranking tests, and an explicit scoring audit.

Prompt and output-format sensitivity. MMLU-Pro was designed to be more difficult and less prompt-sensitive than MMLU (Wang et al., 2024). The Format Tax separates reasoning-quality scoring from format compliance and identifies mathematical extraction failures that inflate measured format losses (Lee et al., 2026). Thus neither separate reporting nor extraction-induced apparent degradation is new here. We focus on terminal-only serialization with an otherwise fixed reasoning request, explicit interface-conditioned model gaps, and paired multi-budget evidence. This operational comparison does not isolate reasoning from extraction effects or establish validity for external leaderboard scoring rules.

Reasoning benchmarks and test-time budget. MATH and MATH-500 are standard tests of competition-level mathematical problem solving (Hendrycks et al., 2021; Lightman et al., 2024); MMLU-Pro broadens evaluation across knowledge and reasoning domains (Wang et al., 2024). Test-time token use is itself an evaluation variable, especially for reasoning models (Fan et al., 2025). We therefore make the 4,096-token budget part of the main protocol and run a sample-matched 1,024-token ablation.

3 A Three-Way Decomposition of Evaluation

Let y_{mifb} be the response from model m on item i, requested format f, and budget b. Let g_i be the reference answer, E_{sem} a fixed marker-based extractor, E_f the extractor for the requested final

serialization, and V the task verifier. We define

$$A_{\text{sem}}(m, f, b) = \frac{1}{N} \sum_i V(E_{\text{sem}}(y_{mifb}), g_i), \tag{1}$$

$$C_{\text{fmt}}(m, f, b) = \frac{1}{N} \sum_i \mathbb{1}[y_{mifb} \text{ ends in a valid instance of } f], \tag{2}$$

$$A_{\text{strict}}(m, f, b) = \frac{1}{N} \sum_i \mathbb{1}[y_{mifb} \text{ complies with } f] V(E_f(y_{mifb}), g_i). \tag{3}$$

The term "semantic" is operational: it means correctness after the declared gold-independent extractor and verifier, not human proof validation or a direct measure of latent capability. Because E_{sem} and E_f can select different candidates, strict correctness is not defined as the product of semantic correctness and compliance. Their aggregate gap is descriptive, not a uniquely identified causal decomposition.

Semantic extraction. The extractor scans balanced `\boxed{...}`, `Final Answer:`, valid JSON `answer` fields, and explicit `<answer>...</answer>` tags, then selects the latest substantive candidate. On MMLU-Pro it also recognizes standalone letters, declarative answer patterns, and in-prose option mentions, or uniquely maps option text to the provided choices. This map never consults the gold label. These permissive patterns can capture rejected or tentative choices; Section 6.1 tests an alternative. On MATH-500, if no candidate is found, the last nonempty line is passed to Math-Verify (Kydlíček, 2026); it may be incomplete prose rather than an answer. Unparseable predictions receive no exact-string fallback. MMLU-Pro uses exact A–J letter equality.

Compliance and strict scoring. Compliance requires the final nonempty line to satisfy a declared terminal grammar, not byte-for-byte equality. The frozen grammar trims outer whitespace; FINALaccepts case and whitespace variations in its label; JSON is parsed as a single `answer`-key string object, accepting legal whitespace and standard decoder behavior. MMLU-Pro requires one unwrapped option letter, normalized to uppercase. The appendix gives accepted and rejected examples, including the decoder's duplicate-key behavior. Metasyntax such as `<F>` or `<answer>F</answer>` remains noncompliant. Strict accuracy measures correctness through this particular interface, not pure reasoning ability.

4 Experimental Design

4.1 Models and serving

We evaluate Qwen2.5-7B-Instruct (Yang et al., 2025), DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025a), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023; Mistral AI, 2024) locally with vLLM. Their immutable Hugging Face revisions are recorded in the supplement. We additionally query Qwen2.5-32B-Instruct and DeepSeek-V3.2 (DeepSeek-AI et al., 2025b) through SiliconFlow's OpenAI-compatible API. These endpoints lack exposed immutable backend revisions and represent provider- and date-specific snapshots.

Local prompts are rendered with each checkpoint's official chat template. Local generation uses vLLM 0.19.0, temperature 0, an 8,192-token context cap, and seed 20260823. API generation uses temperature 0 and top-$p = 1$. Greedy/API-zero-temperature decoding reduces stochastic variation but does not guarantee bitwise reproducibility across serving systems.

4.2 Benchmarks and evaluation matrix

The main experiment uses all 500 MATH-500 test items and 650 MMLU-Pro test questions: 50 seeded questions from each of 13 non-mathematics categories (biology, business, chemistry, computer science, economics, engineering, health, history, law, other, philosophy, physics, and psychology). MMLU-Pro offers up to ten choices (Wang et al., 2024); this does not itself protect against systematic extraction errors.

For each model, every item is evaluated under all three formats at 4,096 maximum new tokens, yielding $5 \times (500 + 650) \times 3 = 17{,}250$ main generations. The budget study draws 120 MATH-500

items and 10 items per MMLU-Pro category (130 total) from the benchmarks, with separate 1,024- and 4,096-token generations: $2 \times 5 \times (120 + 130) \times 3 = 7{,}500$. Its MATH sample is nested within the main 500; the MMLU sample is not nested within the main 650 (21 overlapping items). This does not affect the complete, unique 1,024/4,096 within-budget pairing.

4.3 Controlled output formats

All prompts begin: *"Solve the problem carefully. You may reason step by step."* MATH answers must contain only the final mathematical answer; MMLU-Pro answers must contain exactly one A–J letter without option text. The only experimental manipulation is the requested final line:

FINAL	`Final Answer: <answer>`
BOXED	`\boxed{<answer>}`
JSON	`{"answer":"<answer>"}`

The angle brackets are metasyntactic placeholders, not literal markup. Their occasional literal rendering by API models motivated the parser audit in Section 6.

4.4 Uncertainty and multiplicity

Comparisons are paired by question identifier. The frozen analysis reports unadjusted percentile 95% intervals from 3,000 paired item-bootstrap draws (seed 20260823). Exact two-sided McNemar tests are Holm-adjusted separately within dataset $\times$ metric families: 30 model-pair/format tests, 15 within-model format tests, and 15 budget tests, respectively. Reliability requires both opposing intervals to exclude zero and both adjusted $p < .05$; intervals themselves are not simultaneous. A supplementary MMLU analysis preserves each of the 13 domain sample sizes in every bootstrap draw, targeting an equally weighted domain mixture. MATH covers all 500 fixed items, so its bootstrap represents hypothetical item-population resampling stability, not uncertainty in the arithmetic score of those 500 items.

5 Results

5.1 Extracted-answer accuracy and strict protocol success differ

Figure 1 shows the main result. DeepSeek-V3.2 has the highest semantic answer accuracy for all six dataset–format combinations. Under strict scoring, however, it ranges from first to fourth depending on serialization. R1-Distill-7B is similarly stark on MATH-500: semantic accuracy is 0.776–0.800, but strict accuracy is 0.000–0.082 because its terminal line rarely follows the requested contract.

Table 1 reports semantic/compliance/strict triples. The gap is not confined to one family. Qwen2.5-32B's MMLU-Pro semantic accuracy is almost constant (0.672–0.686), while strict accuracy ranges from 0.403 under FINALto 0.686 under JSON. Conversely, Qwen2.5-7B is almost perfectly compliant on MMLU-Pro, so its semantic and strict scores are close.

5.2 Reliable rank reversals occur only in compliance-based metrics

Across model and format pairs, 21 reversals satisfy the two unadjusted interval conditions; 17 survive Holm correction: 5 compliance and 12 strict-accuracy reversals. There are zero *statistically reliable* semantic reversals, not zero point-estimate reversals: Qwen32 minus R1 on MATH is $+.008$ for BOXED, $+.024$ for FINAL, and $-.042$ for JSON, but neither inverse comparison satisfies the interval criterion. The 12 strict reversals concern six dataset–model-pair combinations, all involving an API endpoint; they are not 12 independent replications. Table 2 gives examples.

Under MMLU-Pro FINAL, DeepSeek-V3.2 has the highest extracted-answer accuracy at 0.832 but is fourth in strict accuracy at 0.246. Under JSONit leads at 0.823 and 0.778, respectively. Observed strict success is sensitive to the requested interface and resulting response; separately generated responses do not hold the answer fixed.

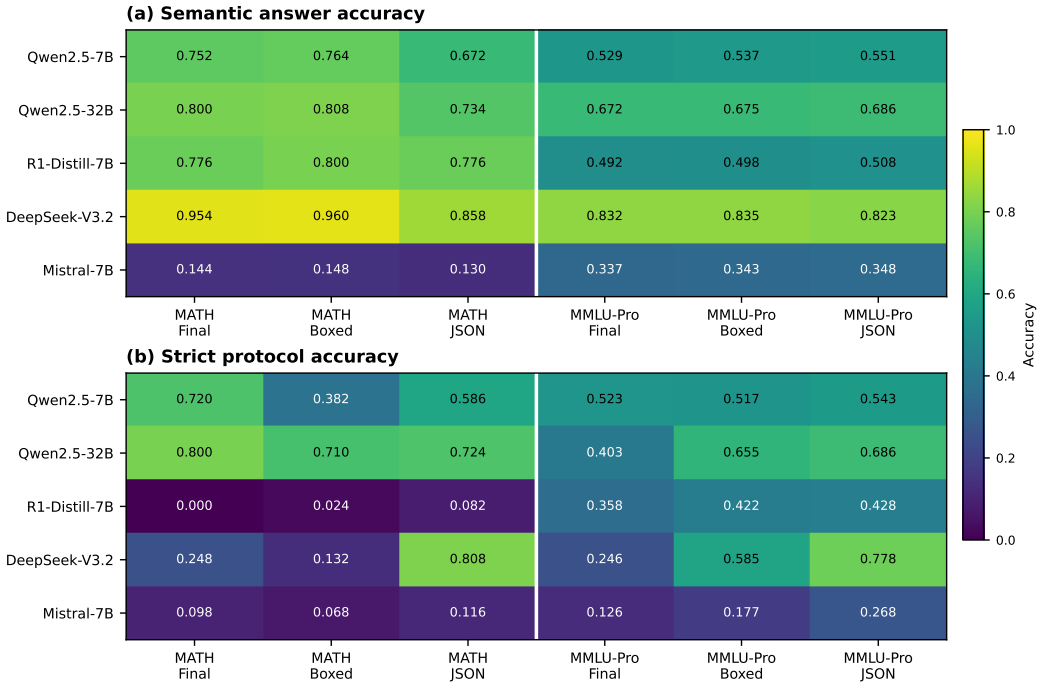

Figure 1: The same 4,096-token generations scored two ways. Semantic accuracy (top) shows fewer point-estimate ordering changes; strict protocol accuracy (bottom) combines answer correctness with declared terminal-format compliance and substantially changes rankings. Neither panel establishes capability-order equivalence. Each cell contains 500 MATH-500 or 650 MMLU-Pro examples.

Table 1: Main 4,096-token results. Cells are $A_{\text{sem}}/C_{\text{fmt}}/A_{\text{strict}}$. Each MATH cell has $n = 500$ and each MMLU-Pro cell $n = 650$.

Dataset	Model	FINAL	BOXED	JSON
MATH	Qwen2.5-7B	.752/.934/.720	.764/.450/.382	.672/.792/.586
	Qwen2.5-32B	.800/.996/.800	.808/.844/.710	.734/.916/.724
	R1-Distill-7B	.776/.000/.000	.800/.026/.024	.776/.090/.082
	DeepSeek-V3.2	.954/.268/.248	.960/.134/.132	.858/.914/.808
	Mistral-7B	.144/.606/.098	.148/.406/.068	.130/.656/.116
MMLU-Pro	Qwen2.5-7B	.529/.992/.523	.537/.963/.517	.551/.985/.543
	Qwen2.5-32B	.672/.586/.403	.675/.965/.655	.686/.997/.686
	R1-Distill-7B	.492/.660/.358	.498/.749/.422	.508/.754/.428
	DeepSeek-V3.2	.832/.300/.246	.835/.700/.585	.823/.949/.778
	Mistral-7B	.337/.308/.126	.343/.448/.177	.348/.735/.268

5.3 APPARENT SEMANTIC FORMAT EFFECTS DEPEND ON EXTRACTION

Under the frozen evaluator, JSONyields lower MATH-500 semantic accuracy than both FINALand BOXEDfor Qwen2.5-7B, Qwen2.5-32B, and DeepSeek-V3.2. The six Holm-significant differences are 0.066–0.102 (Table 3). However, all six lose significance under a literal-mathematics JSON control (Section 6.1), with residual differences of only 0.002–0.014. Thus the apparent penalty is extraction-sensitive; these data do not establish that JSON impairs reasoning. We find no corrected semantic format effect on MMLU-Pro under the primary rule.

5.4 TOKEN BUDGET IS A MODEL-DEPENDENT CONFOUND

The paired ablation varies budget while matching items and requested format. On the 120-item MATH subset, increasing the budget from 1,024 to 4,096 tokens raises R1-Distill-7B semantic

Table 2: Representative reliable strict-accuracy rank reversals. Differences are model A minus model B with 95% paired-bootstrap intervals; both directions have Holm-adjusted $p < .05$.

Dataset	Pair (A–B)	Format 1: difference [CI]	Format 2: difference [CI]
MATH	Qwen32–V3.2	FINAL: $+.552$ [.500,.602]	JSON: $-.084$ [-.124,-.046]
MATH	Qwen7–V3.2	FINAL: $+.472$ [.414,.526]	JSON: $-.222$ [-.264,-.180]
MMLU-Pro	Qwen32–Qwen7	FINAL: $-.120$ [-.166,-.074]	JSON: $+.143$ [.105,.182]
MMLU-Pro	Qwen7–V3.2	FINAL: $+.277$ [.226,.328]	JSON: $-.235$ [-.277,-.192]
MMLU-Pro	R1-7B–V3.2	FINAL: $+.112$ [.063,.163]	JSON: $-.351$ [-.392,-.309]

Table 3: Within-model MATH-500 effects under the frozen evaluator ($n = 500$). All six are Holm-significant here but not under literal-math JSON extraction (Table 9). Positive values favor the non-JSON format.

Model	BOXED− JSON [95% CI]	FINAL− JSON [95% CI]
Qwen2.5-7B	.092 [.056,.130]	.080 [.048,.114]
Qwen2.5-32B	.074 [.040,.108]	.066 [.034,.098]
DeepSeek-V3.2	.102 [.074,.130]	.096 [.066,.130]

accuracy by 0.308–0.450 and DeepSeek-V3.2 by 0.183–0.217, while reducing their truncation rates by 0.208–0.600 (Table 4). All displayed effects remain significant after Holm correction. Effects for the non-reasoning 7B checkpoints are much smaller. These are total budget effects across separate generations: longer computation, different continuations, and completion of the final serialization are not independently isolated.

6 EVALUATOR AUDIT AND REPRODUCIBILITY

Post-hoc parser repair. During the pre-writing audit we found that some API responses rendered the prompt's metasyntactic brackets literally, for example `Final Answer: <F>`, `\boxed{<C>}`, or `<answer>G</answer>`. A second model-assisted pass found additional gold-independent surface forms—Markdown-emphasized labels, explicit `[ANS]` tags, bold option letters, and a heading/newline case—that the original semantic parser missed. Before final analysis we froze a repair that recognizes only these explicit forms while continuing to mark them noncompliant. Regression tests cover closed, unclosed, boxed, JSON, bare-tag, placeholder-only, Markdown, bracket-tag, and "latest substantive marker wins" cases. On the 17,250-row main experiment the complete repair changes 1,127 semantic predictions and 901 semantic correctness labels (mostly API outputs), but only 14 strict correctness labels. Crucially, the repair *removes* apparent semantic instability; it does not create the paper's strict-rank-reversal result.

Exhaustive flags and author verification. We scan all 24,750 rows after repair. There are 399 empty semantic extractions: 332 coincide with length truncation; the remainder include no explicit marker (8), unrecognized explicit-answer forms (54), literal placeholders (4), and one invalid multiple-choice string. Seven mathematical predictions are unparseable by Math-Verify, and 59 outputs reproduce part of the prompt. Empty or unparseable answers receive no manual rescue; prompt reproduction is an annotation, not an automatic failure. The supplement contains every flag plus a seeded 249-row stratified audit form covering 217 unique outputs. A model-assisted prescreen marked 125 issue rows for author verification; an author manually reviewed every marked row, confirmed the prescreened judgments, and found no case requiring a score override. The qualitative audit characterizes failure modes only: all reported scores remain outputs of the frozen evaluator.

6.1 OFFLINE SENSITIVITY CHECKS

We apply post-hoc controls to the same 17,250 main outputs, without new generations or manual score overrides. Each policy is fixed before its execution; the mathematics JSON control is additionally motivated by the new random audit. All policies and outcomes are retained, not selected by whether

Table 4: Paired MATH-500 budget effects ($n = 120$): 1,024 $\rightarrow$ 4,096 tokens. ΔS is semantic-accuracy change and ΔT is truncation-rate change; brackets are 95% CIs.

Model	Format	ΔS [CI]	ΔT [CI]
R1-Distill-7B	FINAL	+.308 [.233,.392]	-.450 [-.542,-.367]
	BOXED	+.450 [.358,.542]	-.600 [-.683,-.517]
	JSON	+.317 [.233,.408]	-.417 [-.508,-.333]
DeepSeek-V3.2	FINAL	+.183 [.117,.258]	-.208 [-.275,-.142]
	BOXED	+.217 [.142,.292]	-.258 [-.342,-.183]
	JSON	+.200 [.125,.283]	-.258 [-.342,-.183]

they preserve reversals. Each uses the same paired tests and correction families as the primary analysis; these are sensitivity analyses, not independent replication.

A wrapper-tolerant terminal policy removes only declared outer Markdown/math delimiters or a complete terminal code fence, retaining the requested serialization and forbidding trailing prose. It yields 7 compliance and 13 strict reversals, versus 5 and 12 originally. Of the 12 original strict reversals, 9 persist and 3 disappear; 4 new ones emerge. Semantic scores are held fixed. Table 10 distinguishes changes in reversal identity from total counts. Thus reversals are not limited to exact outer typography under this particular relaxation. An alternative explicit-marker MCQ policy removes broad in-prose option patterns, retains answer declarations, and adds unwrapped terminal letters. It is not a nested subset of the original parser and yields no reliable semantic reversal. Nevertheless, R1 MMLU semantic accuracy falls by 5.08, 5.69, and 5.85 percentage points for BOXED, FINAL, and JSONrespectively (Table 11); no reliable rank reversal does not imply stable absolute scores. Both extraction policies make mistakes: among 49 completed disagreement cases, model-assisted review preferred the alternative in 12, the original in 34, and neither in 3; 282 truncated disagreements remain unadjudicated beyond terminal-text triage.

The API-channel comparison is an archive-provenance check, not an independent sensitivity experiment: all 6,900 API main records have empty `reasoning_content` and `raw_output` identical to nonempty `final_content`. Zero score changes therefore follow from identical inputs and provide no evidence about separate reasoning/final-channel behavior. Table 8 summarizes reversal counts. The MCQ and literal-JSON controls leave the strict scoring branch fixed by design; only the wrapper policy tests strict rankings under a changed terminal grammar.

The audit-triggered MATH control accepts a terminal, single-key JSON-like answer string. It retains standard JSON decoding if valid and free of control characters; otherwise it preserves the literal payload, including LaTeX backslashes. This is a nonstandard semantic sensitivity, not repaired JSON compliance. It changes 199 candidates and 121 correctness labels, leaving compliance and strict scores fixed. One symbolic verification times out; assigning that case either outcome changes only Mistral JSON accuracy by 0.002 and leaves the reported conclusions unchanged. Qwen7, Qwen32, and V3.2 JSON accuracy becomes 0.750, 0.798, and 0.948: all six previously significant gaps now have intervals spanning zero and Holm $p = 1$ (Table 9). We therefore do not interpret the frozen JSON deficit as evidence of reduced reasoning ability.

Additional descriptive audit. A correctness-independent hash sample selects three nonempty outputs per model–dataset–format cell (90 total). Terminal-text inspection exposes tentative MCQ candidates, unfinished MATH fallback text, and JSON/LaTeX escaping losses. A separate 17-case sample of changed strict outcomes contains only declared wrappers. These new checks are model-assisted, not covered by the earlier author sign-off, and do not estimate a population error rate or independently validate Math-Verify.

Design and integrity checks. Fixed-domain MMLU bootstrap retains all 17 primary reliable reversals. Existing matched-prompt MATH repeat-run intervals contain zero, not evidence of equivalence; the 21-item MMLU overlap is descriptive only (Appendix H). Archived prompts, revisions, raw responses, and preflight checks establish complete pairing and 17,250 unique main records, with 3,750 at each ablation budget.

6.2 PLACEHOLDER INTERVENTIONS AND HELD-OUT VALIDATION

The failure inventory suggested that the literal `<answer>` placeholder might induce terminal meta-syntax rejected by our grammar. Before new generation, we fixed a targeted follow-up using the existing 120 MATH and 130 MMLU items, Qwen32 and V3.2, FINAL/JSON, and original/revised prompts (2,000 outputs). The revision replaces every `<answer>` in instruction prose and examples with `answer`, leaving other text and scoring fixed. Both arms are newly queried in the same session in randomized eight-condition question blocks, rather than compared with historical API outputs.

Figure 2 separates this intervention from fixed-output rescoring. Originally Qwen32 leads V3.2 in Final but trails in JSON on both subsets; after revision it trails in both. The tested point-order reversal disappears, not necessarily all 12 original reliable reversals. V3.2 MATH Final strict accuracy rises from .267 to .842 while extracted accuracy stays .950. Conversely, Qwen32 MATH JSON strict accuracy falls from .700 to .392; its mean output length falls from 495 to 127 tokens, and literal-math extraction does not remove the decline. Appendix J retains every metric and paired interval. These observations neither identify hidden reasoning changes nor establish a universal repair.

A. Extraction sensitivity	**B. Interface sensitivity**	**C. Prompt sensitivity**
Same outputs; change answer recovery.	Same outputs; relax wrappers.	New outputs; remove angle brackets.
Six significant MATH JSON deficits → none under literal extraction.	9/12 original strict reversals retained; 3 lost, 4 new.	MATH strict: V3.2 Final .267 → .842; Qwen32 JSON .700 → .392. No universal fix.
Strict scores unchanged.	*Semantic scores unchanged.*	

Figure 2: Three different checks, not a causal decomposition. A–B reuse main outputs (500 MATH or 650 MMLU items per cell); C uses the same-session follow-up (120 MATH items per displayed cell). All three concern measured protocol outcomes, not isolated changes in latent reasoning. Full follow-up matrices remain in Appendix J.

Let g_{ft} be the Qwen32-minus-V3.2 strict accuracy gap for format f and template t. The fixed follow-up contrast $(g_{\text{Final,new}} - g_{\text{Final,old}}) - (g_{\text{JSON,new}} - g_{\text{JSON,old}})$ is $-.317$ on MATH (95% paired CI $[-.467, -.167]$) and $-.308$ on MMLU ($[-.438, -.177]$). The paired intervals are the principal uncertainty summaries. As an assumption-dependent Monte Carlo sensitivity, sign-flip tests, Holm-adjusted over these two contrasts, give $p = .000130$ and $.000120$. This test assumes question-level sign exchangeability; it is not an exact test of every composite interaction null. The audit-motivated, reused-question design supports sensitivity to this particular template revision, not intrinsic reasoning changes or generic prompt robustness.

Held-out-item validation. A prospectively frozen 3,000-output study uses 120 unused MATH test items outside MATH-500 and 130 unused MMLU items (13 domains), the same endpoints/formats, and three placeholders: `<answer>`, `answer`, and `[answer]`. The no-angle interaction on MMLU is $-.262$ (97.5% paired interval $[-.431, -.092]$, Bonferroni allocation across two primary estimands). MATH's original analysis is incomplete: one request exhausted its two-attempt allowance during an interruption. Its missing-outcome mean bounds are $[-.425, -.417]$, not a confidence interval. One explicitly authorized third request, made after inspecting the original results, gives a completed-data sensitivity of $-.425$ ($[-.583, -.258]$); this is not pristine prospective confirmation. Appendix M reports all cells, the cross-day interruption, and extraction controls. In particular, square-placeholder Qwen MMLU Final extracted accuracy is .177 under the frozen parser but .631 under a narrow terminal-bracket sensitivity. The no-angle Qwen mathematics JSON decline persists under literal extraction. Neither placeholder is a universal repair.

Under the prespecified two-dataset criterion, the original prospective study therefore does not establish overall confirmation: MMLU supports the interaction, while the recovered mathematics result remains a sensitivity analysis.

7 DISCUSSION

What reverses, and what does not. The central lesson is that robustness to a scoring change and robustness to a prompt change are different questions: relaxing wrappers retains nine original strict reversals, whereas a newly generated placeholder intervention removes the tested Final/JSON point-order reversal. Neither check substitutes for the other. This concerns specified model–protocol snapshots, not external leaderboard changes. Undetected extracted-answer reversals do not establish capability-order equivalence; neither metric independently identifies reasoning ability.

A practical reporting standard. Freeze outputs and report all metrics; vary extraction and terminal acceptance separately while tracking reversal identities; then test suspected prompt defects with newly generated, matched arms (Appendix K). Strict accuracy matters for machine-readable interfaces, but a permissive extractor or improved prompt cannot retroactively repair an invalid response.

External extractor replay. Replaying the pinned official MMLU-Pro local-script extractor over all 9,750 main MMLU outputs changes 142 correctness labels, but yields no cross-format model point-order reversal. Appendix L reports every cell, missing-answer handling, and paired intervals. This extends the extraction sensitivity check beyond our own policies, not the validation of strict rankings: the terminal contract is unchanged. Our custom prompts and subset also preclude interpreting this replay as official leaderboard performance.

Broader impact. Metric decomposition can make model comparisons more transparent and reduce incentives to optimize undocumented formatting quirks. Conversely, publishing a menu of protocols can enable selective reporting. We therefore recommend a preregistered reference protocol, complete protocol metadata, and multiplicity-corrected stability analysis rather than choosing the most favorable score after seeing results.

8 LIMITATIONS

Our model set comprises three local 7B checkpoints and two unpinned API endpoints; all original primary strict reversals involve the latter. The first template follow-up reuses questions and is exploratory. The held-out study covers new instances, not new problem families or providers; its cross-day interruption and post-result third-attempt recovery limit confirmation, and its MATH distribution differs from MATH-500. Three placeholders do not establish cross-template invariance or isolate punctuation from instruction interpretation. Code, multilingual and open-ended tasks may differ. The external replay tests one extractor, not an official leaderboard. Our evidence concerns model–protocol snapshots.

Semantic accuracy is an operational parser-and-verifier metric. We intentionally use no gold-dependent rescue and no manual score overrides. Although an author manually verified every audit row flagged by the model-assisted prescreen, this targeted check is not a blinded, multi-annotator, population-random prevalence study. Mathematical verification inherits Math-Verify's parsing and equivalence limitations; MMLU-Pro inherits benchmark-label noise. The main matrix uses one generation per condition. Overlapping items from the separately generated 4,096-token budget run permit a limited run comparison, not estimation of general sampling-seed variance. Finally, the 1,024/4,096 ablation uses 250 items per model rather than the full test sets and consists of separate generations; API zero-temperature behavior need not be bitwise deterministic.

9 CONCLUSION

Formatting can determine protocol-specific rankings, not which model reasons best. Extraction and wrapper controls address different sensitivities. Template interventions show heterogeneous effects; held-out evidence provides prospective MMLU support but only post-deviation MATH recovery sensitivity. Report extracted-answer accuracy, compliance, strict accuracy, and truncation separately.

Reproducibility statement

The supplement contains the frozen prompts and prompt hashes, immutable revisions for local models and datasets, provider timestamps, generation metadata, row-level scored outputs, regression tests, preflight integrity checks, the full scorer-revision audit, bootstrap and exact-test outputs, and the author-verification form. Section 6 and the appendix describe the main reproducibility safeguards and all post-hoc evaluator changes.

AI use statement

Generative AI assisted experimental-design critique and refinement, generation/scoring/analysis implementation and debugging, qualitative prescreening, result interpretation, literature discovery, and manuscript drafting and editing. Evaluated models generated the studied response data; prompts and outputs are archived. The author verified the original prescreened cases and found no warranted score overrides; later model-assisted sensitivity audits are not covered by that sign-off. Checks include regression tests, preflight invariants, row-level audits, reproducible reanalysis, frozen-artifact comparisons, and citation checks against source publications. The author takes responsibility for the final content, claims, code, and artifacts.

References

Ding Chen, Qingchen Yu, Pengyuan Wang, Mengting Hu, Wentao Zhang, Zhengren Wang, Bo Tang, Feiyu Xiong, Xinchi Li, Chao Wang, Minchuan Yang, and Zhiyu Li. xverify: Efficient answer verifier for reasoning model evaluations. *arXiv preprint arXiv:2504.10481*, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948v1*, 2025a.

DeepSeek-AI, Aixin Liu, Aoxue Mei, et al. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025b. URL https://arxiv.org/abs/2512.02556.

Siqi Fan, Bowen Qin, Peng Han, Shuo Shang, Yequan Wang, and Aixin Sun. The price of a second thought: On the evaluation of reasoning efficiency in large language models. *arXiv preprint arXiv:2505.22017*, 2025.

Hasan Abed Al Kader Hammoud, Hani Itani, and Bernard Ghanem. Beyond the last answer: Your reasoning trace uncovers more than you think. *arXiv preprint arXiv:2504.20708*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Hynek Kydlíček. Math-verify: Math verification library, 2026. URL https://github.com/huggingface/Math-Verify. Version 0.9.0, released January 10, 2026.

Ivan Yee Lee, Loris D'Antoni, and Taylor Berg-Kirkpatrick. The format tax. *arXiv preprint arXiv:2604.03616*, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Mistral AI. Mistral-7B-Instruct-v0.3 model card, 2024. URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. Accessed September 25, 2026; experimental checkpoint revision recorded in the supplement.

Hanyang Wang and Mingxuan Zhu. The detection–extraction gap: Models know the answer before they can say it. *arXiv preprint arXiv:2604.06613*, 2026.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115v2*, 2025.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu Li, Feiyu Xiong, Bo Tang, and Ding Chen. xfinder: Large language models as automated evaluators for reliable evaluation. In *International Conference on Learning Representations*, 2025.

A Exact Prompt Templates

The full MATH instruction is:

> Solve the problem carefully. You may reason step by step. For `<answer>`, use only the final mathematical answer. End with exactly this final line and no following text:

The full MMLU-Pro instruction is:

> Solve the problem carefully. You may reason step by step. For `<answer>`, use exactly one option letter from A to J. Do not include the option text inside `<answer>`. End with exactly this final line and no following text:

The sentences in each instruction are joined with single spaces. A newline then introduces exactly one of `Final Answer: <answer>`, `\boxed{<answer>}`, or `{"answer":"<answer>"}`. Two newlines follow, then `Problem:`, a newline, and the benchmark question (including rendered choices for MMLU-Pro). No demonstrations or format-specific reasoning instructions are used. Local chat-template rendering follows Section 4; archived prompt hashes identify exact instances.

Frozen terminal grammar examples. For MMLU-Pro FINAL, both `Final Answer: D` and `finalanswer:d` pass; `Final Answer: <D>` and a following prose line fail. JSON whitespace is allowed, as is a lowercase single-letter value. Python's decoder also accepts duplicate `answer` keys using the last value: `{"answer":"A","answer":"D"}` passes and selects D. This inherited behavior is disclosed, not recommended as a strict JSON schema. Additional distinct keys fail. The scoring code and regression tests define these normalization boundaries; frozen scores are not silently changed to impose a stricter schema.

B Full Experimental Inventory

Table 5: Original generation inventory (24,750 responses). The separate 2,000-response template follow-up is detailed in Appendix J.

Run	Items/model	Formats	Rows
Local main, 4,096	1,150	3	10,350
API main, 4,096	1,150	3	6,900
Local paired, 4,096	250	3	2,250
API paired, 4,096	250	3	1,500
Local paired, 1,024	250	3	2,250
API paired, 1,024	250	3	1,500
Total			24,750

Local software versions are Python 3.12.13, PyTorch 2.10.0+cu129, Transformers 4.57.6, Datasets 4.8.4, and vLLM 0.19.0 on Linux. The engine context length is 8,192 and request batch size is 16. API requests use the OpenAI Python client 2.31.0 with maximum concurrency 8. The supplement records immutable local model and dataset hashes and provider timestamps.

C Rank-Reversal Accounting

A reversal row compares one model pair under two formats. The full CSV includes the paired differences, percentile intervals, discordant counts, exact McNemar p-values, and Holm-adjusted values for every model pair and format.

D Scorer Revision Audit

The repair is deliberately asymmetric between semantics and compliance. It recognizes the substantive content of a provider-rendered placeholder without pretending that the model met the exact requested

Table 6: Number of format-pair rank reversals with nonzero paired-bootstrap intervals, before and after Holm correction.

Dataset	Metric	Unadjusted	Holm
MATH-500	Semantic accuracy	0	0
MATH-500	Format compliance	4	4
MATH-500	Strict accuracy	4	4
MMLU-Pro	Semantic accuracy	0	0
MMLU-Pro	Format compliance	5	1
MMLU-Pro	Strict accuracy	8	8

Table 7: Changes introduced by the frozen prompt-metasyntax repair. "Prediction" is the extracted answer string; "correct" is semantic correctness.

Run	Rows	Prediction	Correct	Strict correct
Local main 4,096	10,350	78	31	0
API main 4,096	6,900	1,049	870	14
Local paired 4,096	2,250	17	3	0
API paired 4,096	1,500	209	170	3
Local paired 1,024	2,250	19	6	0
API paired 1,024	1,500	179	141	4

syntax. No reference answer is read during extraction, and all aggregate tables are regenerated from the repaired row-level files.

E Offline Control Details

Table 8: Reliable reversal counts under offline policies. MCQ and literal-JSON policies leave strict scores fixed by design; API final-only uses identical inputs. Only the wrapper row changes the terminal grammar. Policies are not independent replications.

Policy	Semantic	Compliance	Strict
Frozen primary	0	5	12
Wrapper-tolerant terminal	0	7	13
Alternative MCQ markers	0	5	12
API final content only	0	5	12
Literal-math JSON (both timeout bounds)	0	5	12

Table 9: MATH semantic gaps after literal-math JSON extraction. All six Holm-adjusted p-values are 1. The single unresolved Mistral case does not affect these rows.

Model	Boxed− JSON [95% CI]	Final− JSON [95% CI]
Qwen2.5-7B	.014 [-.016,.046]	.002 [-.022,.028]
Qwen2.5-32B	.010 [-.016,.036]	.002 [-.026,.028]
DeepSeek-V3.2	.012 [-.004,.030]	.006 [-.014,.028]

The first three controls were fixed before execution. The mathematics control was added after the random audit exposed JSON escaping losses. Before completing it, implementation review corrected handling of valid Unicode escapes and literal matrix separators; an unbounded symbolic-verification attempt was interrupted without aggregate results. The final rule uses a 10-second independent-process verification cap and publishes both assignments for the one unresolved item. All policy outputs are retained. The sensitivity audit packet contains 530 distinct changed predictions and overlaps the 90-case sample; newly added mathematical differences are not claimed to have received exhaustive qualitative review.

The original v3 MATH records were copied from v2 while later parser edits targeted MCQ. Replaying the current extractor over all 24,750 outputs produces 39 MATH string differences (22 main), with no change to semantic correctness, compliance, format prediction, or strict correctness. Changed candidates were reverified; identical candidates reuse frozen correctness. Six compact historical score inputs now make the original scorer-change analysis self-contained.

F Reversal Identity and Score Sensitivity

Table 10: Strict-reversal identity under the wrapper-tolerant rule. Identity includes dataset, model pair, and format pair; rows are not independent replications.

Status	Count
Original reversals retained	9
Original reversals no longer reliable	3
Additional reversals under relaxation	4
Original / relaxed total	12 / 13
Distinct dataset–model pairs, original / relaxed	6 / 7
Purely local strict reversals, original / relaxed	0 / 1

The three lost relationships are BOXED–JSONcomparisons of Qwen32 versus V3.2 on MATH and MMLU, and Qwen7 versus V3.2 on MATH. The four new relationships are BOXED–FINALcomparisons for these same three dataset–model pairs, plus R1 versus Mistral on MATH. The supplement lists every retained, lost, and new identity with its paired differences.

Table 11: R1 MMLU semantic accuracy under the alternative explicit-marker extractor ($n = 650$). Correctness remains an operational parser metric, not human adjudication.

Format	Frozen	Alternative	Change (points)
BOXED	.49846	.44769	-5.08
FINAL	.49231	.43538	-5.69
JSON	.50769	.44923	-5.85

G Descriptive Strict-Failure Inventory

Table 12: Disjoint outcome categories across the 17,250 main generations. Noncompliant outputs are assigned the first applicable category in the displayed order.

Category	Outputs
Strict success	7,114
Compliant, but terminal answer incorrect or unverified	4,111
Noncompliant and length-truncated	925
Noncompliant with recognized outer wrapper	856
Noncompliant with terminal angle-tag metasyntax	1,157
Other terminal-grammar failure	3,087

The categories partition all outputs, not just semantically correct ones. Wrapper recognition uses the declared relaxed extractor; metasyntax matches terminal `<answer>` tags or single-letter angle tags. "Other" includes wrong serialization and unrecognized surfaces. Priority resolves overlaps, so these are not estimates of causal contributions. Model–dataset–format counts and the frozen semantic-correct count within each category are released to connect these patterns to specific comparisons.

H Sampling-Design and Existing Repeat-Run Checks

For MMLU, each of 3,000 draws samples paired differences separately within each of 13 domains, retaining 50 items per domain for the main matrix or 10 for budget comparisons, then averages the 13

domain means equally. MATH intervals, point estimates, exact McNemar tests, and each analysis's Holm families remain unchanged. This tests design sensitivity, not a new confirmatory dataset. All primary reliable reversal identities are retained. The supplement provides 2,535 domain-specific contrast rows and all revised intervals for model, format, and budget analyses.

Table 13: Joint interval-exclusion and Holm decisions under the frozen versus fixed-domain bootstrap. Counts are individual contrasts, not independent reversal events.

Analysis	Contrasts	Frozen pass	Fixed-domain pass
Model comparisons	180	161	161
Within-model format	90	54	54
Budget (including truncation)	120	27	27

The existing repeated 4,096-token runs are aligned by model, dataset, format, question identifier, and exact prompt hash; all matched hashes agree. There are 30 matched cells and 90 metric comparisons across both benchmarks. MATH has 120 overlapping items per cell. The 45 MATH intervals for semantic accuracy, compliance, and strict accuracy all contain zero; the full paired differences and intervals are released. This does not bound general seed or provider-version variance and cannot establish equivalence. Raw text differs even for local zero-temperature runs: 77–80 of 120 MATH responses for Qwen7, 104–115 for R1, and 74–84 for Mistral, versus 112–119 for API cells. Raw differences do not imply corresponding score differences or prove a provider revision change.

MMLU's main and budget samples overlap on only 21 identifiers, drawn from 8 domains with 10 biology items. These intersection-only scores must not be interpreted as a second balanced benchmark or compared to full-set scores as a causal run effect. The two budget levels themselves remain completely paired on their 130-item MMLU sample.

I Recommended Leaderboard Schema

We recommend that each submitted result include the following machine-readable fields: model and immutable revision when available; provider and query date; benchmark and data revision; prompt hash; decoding parameters; maximum and observed output tokens; semantic extractor and verifier versions; semantic accuracy; terminal-grammar compliance; strict accuracy; truncation rate; and a link to raw or privacy-preserving auditable outputs. A leaderboard may sort by strict accuracy when interface adherence is part of the task, but should expose the decomposition next to the sorted value.

J Targeted placeholder follow-up

Post-generation behavior and extraction audit. The adverse Qwen32 MATH JSON result coincides with shorter visible responses: mean generated tokens fall from 495.12 to 126.53, single-nonempty-line outputs increase from 0/120 to 78/120, and outputs below 50 tokens from 0/120 to 99/120. These are descriptive observations, not evidence about hidden reasoning. Examples include reference answers 16 and 3 receiving bare JSON answers 12 and 0; an escaping case with reference $11\sqrt{2}$ is also missed by standard JSON extraction. Applying the previously fixed literal-math JSON rule to both arms gives Qwen32 extracted-answer accuracy .800 (original) versus .433 (revised), and V3.2 .967 versus .950. Thus that particular extraction repair does not eliminate the Qwen32 decline. The control changes 43 candidates, with no unresolved verification; strict scores remain unchanged by design. All-cell length summaries and changed candidates are archived. This post-generation audit is model-assisted, not new human adjudication.

The follow-up reuses the archived paired-budget subset: 120 MATH questions and 10 questions in each of 13 non-mathematics MMLU domains. The selected questions are not chosen by follow-up outcomes. Two models, two formats and two templates yield 2,000 responses. All 1,000 original-template prompts match their archived prompt hashes; question text and gold labels match in all 2,000 cells. The intervention replaces every literal `<answer>` with `answer`, including both explanatory prose and serialization examples. This is one specific wording intervention, not an isolated test of punctuation alone. The JSON example still contains a placeholder. No gold-dependent extraction or manual score override is introduced.

Question order and the eight conditions within each question are shuffled using seed 20260913. Each question's eight requests are dispatched together, using at most eight workers, before proceeding to the next question. Both arms use temperature 0, top-$p = 1$, and a maximum of 4,096 output tokens; no seed is sent to the provider, consistent with the original API experiment. Provider revisions remain unpinned. New requests ran on September 13, 2026 and finished at 09:29 UTC. There were 2,014 attempts: 14 failed requests, each successfully recovered within the two-attempt limit. No question or condition was replaced. All 2,000 final records are unique and complete. All separate reasoning fields are empty. The two one-token connectivity probes are not experimental records. Successful-response token usage implies an estimated CNY 5.085 at the checked uncached rates; this is not an invoice and excludes unknown charges for failed calls. Conservative reservations including every failed attempt total CNY 34.223, below the CNY 90 runner stop line and CNY 100 authorization.

Table 14: Every follow-up condition. Entries are semantic/compliance/strict accuracy. Old and new are newly generated in the same session. Each row has 120 MATH or 130 MMLU items per template.

Dataset	Model	Format	Original	No-angle
MATH	Qwen32	Final	.808/1.000/.808	.725/1.000/.725
MATH	Qwen32	JSON	.725/.908/.700	.392/.942/.392
MATH	V3.2	Final	.950/.267/.267	.950/.875/.842
MATH	V3.2	JSON	.867/.925/.817	.875/.950/.850
MMLU	Qwen32	Final	.677/.738/.500	.708/1.000/.708
MMLU	Qwen32	JSON	.692/.992/.692	.685/.969/.677
MMLU	V3.2	Final	.831/.323/.254	.785/1.000/.785
MMLU	V3.2	JSON	.792/.969/.777	.777/1.000/.777

Table 15: New-minus-original strict accuracy changes with pointwise 95% paired intervals. These are descriptive secondary contrasts, not multiplicity-corrected discoveries.

Dataset	Model	Final	JSON
MATH	Qwen32	$-.083$ [$-.150, -.017$]	$-.308$ [$-.400, -.217$]
MATH	V3.2	.575 [.475, .667]	.033 [$-.008$, .083]
MMLU	Qwen32	.208 [.123, .285]	$-.015$ [$-.062$, .031]
MMLU	V3.2	.531 [.446, .615]	.000 [$-.046$, .046]

The pre-generation specification identifies two primary strict template-by-model-by-format contrasts, one per dataset, with the sign convention in Section 6.2. Confidence intervals use 10,000 paired resamples; MMLU resamples within each fixed domain and averages the 13 domain means equally. Two-sided sign-randomization tests use 100,000 draws, plus-one correction, and Holm adjustment over these two tests. Sign exchangeability is an assumption, not guaranteed by a zero mean interaction. The protocol was fixed before this follow-up's outcomes, but its motivation and questions come from the earlier study; it is not an independent confirmatory replication. Full cell intervals, secondary paired changes and raw records are released. We do not select analyses according to reversal retention or treat non-significance as equivalence.

There are 11 length-truncated responses: MATH Qwen32 JSON old/new 2/1, V3.2 Final 2/1, V3.2 JSON 2/1, and MMLU Qwen32 JSON 1/1. Other cells have zero. Frozen scoring is applied to all records, with independent-process mathematical verification capped at 15 seconds; no follow-up case remains unresolved. This does not validate Math-Verify against human mathematical truth. The prior author sign-off covers the original audit only, not these new outputs. The adverse Qwen32 MATH JSON result and the absence of new-template point-order reversals are both reported; neither is extrapolated to unseen templates or model families.

K Contribution Boundaries and an Executable Audit Workflow

What is added beyond the nearest work? xFinder already demonstrates extractor-dependent ranking changes, while The Format Tax already separates answer correctness from compliance and diagnoses mathematical extraction artifacts (Yu et al., 2025; Lee et al., 2026). Our contribution is not priority for any of these observations. It is the combination of controlled terminal-serialization comparisons, paired and multiplicity-adjusted interface-ranking evidence, explicit tracking of which relationships survive a changed grammar, and a same-session intervention that tests a concrete prompt artifact. These are bounded empirical findings, not a new universal evaluator or a guarantee of ranking stability.

Table 16: Audit steps, worked evidence, and permissible decisions. Each step changes a different component; none licenses selecting the most favorable score.

Step	Worked evidence	Decision boundary
Freeze and separate	V3.2 MMLU Final: extracted accuracy .832, compliance .300, strict accuracy .246 (Table 1).	Report answer recovery and interface success separately; do not infer latent capability from their gap.
Replay the scorer	Literal-math JSON removes six detected extracted-accuracy deficits; relaxed wrappers retain 9 of 12 original strict reversals.	Name the changed rule and affected ranking identities. Do not overwrite the reference protocol.
Intervene on prompts	Same-session placeholder revision removes the tested point-order reversal, with heterogeneous model effects.	The original finding depends on this protocol. Neither universal repair nor elimination of all original reliable reversals follows.

Minimal execution and review record. Before scoring, preserve question IDs, raw-output hashes, prompts, decoding settings, and model/provider versions. Declare a reference extraction rule and terminal contract independently. Reuse exactly the same outputs for scorer comparisons; join by question within model and format, report missing extraction, and keep complete score matrices. Fix bootstrap targets and multiplicity families before inspecting sensitivity outcomes. Sample disagreements without conditioning on whether the new rule improves accuracy, retaining unresolved cases rather than manually overriding scores. Only an intervention on generation requires new model responses. If a prospective claim needs a different task, provider, or prompt family, this audit cannot supply that generalization. The released scripts and manifests instantiate these steps for this study; they are not a validated general-purpose evaluation standard.

L External MMLU-Pro Extractor Replay

Scope and provenance. We apply the answer extractor from the official MMLU-Pro repository's `evaluate_from_local.py` (Wang et al., 2024), pinned to commit `f418b116db00b065c2aea046518d8fcf74d39872`. The supplement includes the original Apache-2.0 source and license, source hashes, a pre-scoring scope specification, and offline replay and verification scripts. Only the three unchanged extraction-function syntax trees are executed; upstream model loading and generation are not run. Ten unit tests cover priority, case sensitivity, wrappers, and missing extraction. An independently implemented extraction check reproduces all 9,750 predictions. This is a post-hoc sensitivity analysis on our zero-shot, terminal-format prompts and balanced subset, not a reproduction of the official few-shot protocol, API-script variant, full benchmark, or leaderboard scores.

Scoring and missing answers. The extractor first searches for a case-sensitive "answer is" pattern, then an answer-colon pattern, then the last standalone uppercase A–J letter. It can therefore select an earlier answer or an incidental letter. Primary replay accuracy counts missing predictions as wrong. The upstream aggregation instead guesses randomly among options when extraction fails; we separately report its exact expected score and missing-answer bounds rather than selecting a random realization. Across 9,750 responses, 38 have no extracted answer. The maximum expected-guessing uplift is 0.40 percentage points (Mistral Boxed). Even allowing every missing answer to be correct

leaves all model point orders unchanged. Frozen semantic, compliance, and strict fields are retained, not replaced.

Table 17: External-extractor accuracy on every main MMLU cell ($n = 650$ each), with missing extraction scored as wrong. Missing counts use Final/Boxed/JSON order. Full paired score changes, expected-guessing scores, and bounds are archived.

Model	Final	Boxed	JSON	Missing counts
Qwen7	.5262	.5369	.5508	0/0/0
Qwen32	.6723	.6754	.6862	3/1/0
R1-Distill-7B	.4800	.4846	.4892	0/0/0
V3.2	.8338	.8354	.8231	1/0/0
Mistral-7B	.3400	.3385	.3400	4/25/4

Paired results and interpretation. There are 142 changed correctness labels. R1's external-minus-frozen differences are -1.23, -1.38, and -1.85 percentage points for Final, Boxed, and JSON; corresponding pointwise 95% paired intervals are [-3.08,0.62], [-3.23,0.46], and [-3.69,-0.15]. Other cell changes range from -0.77 to +0.31 points. These secondary intervals are not multiplicity-adjusted discoveries. All intervals use 3,000 paired resamples within each of 13 fixed 50-item domains. For ranking sensitivity we test all 30 model-pair/format contrasts with exact McNemar tests, Holm-adjusted as one separate family, and require opposing nonzero paired intervals plus both adjusted $p < .05$. There are no point-order reversals and consequently no reliable reversals: every format orders V3.2, Qwen32, Qwen7, R1, and Mistral from highest to lowest. This is consistent with the absence of reliable extracted-answer reversals in the original analysis, not evidence of equivalence or an independent validation of strict-interface reversals. No new model responses are generated.

Disagreement inspection. A correctness-independent hash ordering selects up to two disagreements per cell (28 total); terminal excerpts were inspected by a model-assisted audit, not newly signed off by an author. For Qwen7 Final question `mmlupro_10440`, the external rule selects E from an earlier "Answer: EDCB" although the terminal answer is F. Conversely, for V3.2 Final `mmlupro_5678`, the external rule obtains the gold letter J from an option list even though the terminal line contains only the placeholder and the prose favors another option. Thus a higher score need not indicate better answer recovery; agreement on aggregate model order does not certify individual extraction decisions. No manual score overrides are made.

M Prospective Held-out-item Validation and Recovery Sensitivity

Design and provenance. Before generation on September 16, 2026, we froze the protocol, scorer, task list and analysis code with SHA256 hashes. This is a locally timestamped prospective protocol, not public preregistration or independent-team replication. Uniform seeded sampling selects 120 original MATH test instances from the remaining pool outside all previously used MATH-500 items; MMLU-Pro contributes 10 unused instances from each of the same 13 non-mathematics domains. The MATH source is `EleutherAI/hendrycks_math` revision `21a5633873b6a120296cce3e2df9d5550074f4a3`; MMLU retains its original fixed revision. All 500 historical MATH-500 problem texts match the source's 5,000 test problems. Across 46 available historical files, we exclude 500 unique mathematics questions and 760 MMLU stems by comparable identifier and Unicode/whitespace-normalized text. None of the 250 selected instances overlaps this inventory. Four high-similarity mathematics pairs were reviewed before generation and retained as distinct instances; some share problem families. This does not establish disjoint reasoning families or absence from model training. Reference solutions remain local and are never sent to the models. A second implementation confirms all selected gold extractions and MMLU choices against the pinned sources; this is not independent human adjudication of benchmark labels.

The factorial design is 250×2 endpoints $\times 2$ formats $\times 3$ templates, or 3,000 planned outputs. The original/no-angle conditions reproduce the previous intervention; the third substitutes every literal `<answer>` with `[answer]`. Other prompt text, 4,096-token budget, temperature zero and scoring remain fixed. Question blocks and all 12 conditions within each block are shuffled with

seed 20260916 and queried with concurrency eight. No historical output serves as a control arm. Generation was interrupted September 16 and resumed September 17; most conditions remain locally interleaved within question blocks, but the study is not a single uninterrupted API session. Backend revisions cannot be pinned.

Frozen inference and explicit deviation. The two primary estimands are the dataset-specific no-angle versus original interactions defined in Section 6.2. The frozen plan requires complete primary cells, uses 10,000 paired bootstrap draws, and preserves the 13 MMLU domain sample sizes. It reports pointwise 95% and individual 97.5% intervals, with Bonferroni allocation across two primary estimands. Bootstrap coverage is nominal/asymptotic, not an exact finite-sample guarantee. Support for the prespecified negative interaction requires the corresponding 97.5% upper endpoint below zero; overall prospective support requires both datasets. Approximate historical-variance planning gives 80%-power effect magnitudes .234 and .218 at the chosen sample sizes; these are not guaranteed powers, especially under the changed MATH distribution. Square-template contrasts are secondary descriptive analyses, not additional confirmatory tests.

The original execution produced 2,999 responses: one MATH/V3.2/Final/original cell timed out, then its second reserved attempt had an unknown outcome when execution stopped. The frozen two-attempt limit was respected in the resumed run. Its missing-value mean bounds are $[-.425, -.4167]$; no complete-case primary confirmation was declared. After the incomplete results were inspected, a third request for that exact cell was explicitly authorized on September 18. It succeeded with an unchanged prompt and settings. This is a protocol deviation and later-session sensitivity, not a retroactive alteration of the original analysis. Original responses, attempt ledger, scores and analysis remain byte-identical; the supplemental response and completed-data analysis are separately archived. Including recovery gives 3,003 reserved attempts, 3,000 successful outputs, two recorded timeouts and one interrupted request with unknown outcome. There are 11 length-truncated outputs, no unresolved mathematical-verification timeouts, and no nonempty separate reasoning channels. Successful-token cost is approximately CNY6.63, not a provider invoice; failed-attempt charges are unknown. Conservative maximum-token reservations total CNY57.91, below the authorized CNY100 ceiling. No additional models, sample replacements or outcome-dependent expansion were used.

Table 18: Held-out no-angle interaction. MATH supplemented analysis is explicitly post-deviation; its original prospective analysis is incomplete. Intervals are individual 97.5% paired bootstrap intervals with allocation across two primary estimands, not proof of universal robustness.

Dataset	Analysis status	Estimate	Interval
MATH	Missing-outcome mean bounds	$[-.425, -.417]$	Not computed
MATH	Supplemented sensitivity	$-.425$	$[-.583, -.258]$
MMLU	Complete prospective cells	$-.262$	$[-.431, -.092]$

Table 19: All held-out cells, including the supplemental mathematics cell. Each entry is extracted-answer/compliance/strict accuracy. There are 120 MATH or 130 MMLU items per cell. Frozen scores remain unchanged by the post-hoc extraction checks below. O: original; N: no-angle; S: square placeholder.

Dataset	Model	Format	O	N	S
MATH	Qwen32	Final	.842/1.000/.842	.750/1.000/.750	.842/1.000/.842
	Qwen32	JSON	.742/.942/.733	.425/.958/.425	.767/.942/.758
	V3.2	Final	.958/.250/.225	.950/.908/.892	.958/.892/.867
	V3.2	JSON	.875/.950/.842	.900/.967/.867	.858/.933/.817
MMLU	Qwen32	Final	.646/.585/.431	.662/.992/.662	.177/.131/.108
	Qwen32	JSON	.654/.985/.654	.638/.992/.638	.646/.992/.646
	V3.2	Final	.792/.285/.231	.808/.977/.785	.708/.800/.677
	V3.2	JSON	.800/.931/.738	.792/.992/.785	.792/1.000/.792

Secondary conditions and extraction audit. Square versus original interactions are $-.692$ on MATH (pointwise 95% interval $[-.792, -.583]$) and $-.708$ on MMLU ($[-.846, -.569]$); these are

descriptive, and the MATH value uses the recovery cell. A negative interaction does not imply universally improved accuracy. Under no-angle prompts, Qwen mathematics JSON strict accuracy decreases from .733 to .425. The pre-existing literal-mathematics JSON control, applied uniformly to all 720 mathematics JSON outputs, changes 66 predictions and 55 correctness labels without timeout. Its extracted accuracies for original/no-angle/square are .825/.483/.825 for Qwen and .967/.967/.958 for V3.2. Thus the Qwen decline survives this extraction policy; it is not evidence identifying hidden reasoning changes.

The square-template MMLU Final collapse requires a different interpretation. Inspection found explicit terminal answers such as `Final Answer: [A]` that the frozen semantic parser missed or replaced with an earlier in-prose candidate. We fixed a narrow, gold-independent diagnostic before executing it on all 1,560 MMLU outputs: only an exact terminal Final Answer label followed by one bracketed A–J letter replaces the semantic prediction; otherwise it remains unchanged. This changes 111 predictions and 70 correctness labels. Square/Final extracted accuracy becomes .631 rather than .177 for Qwen, and .792 rather than .708 for V3.2; other condition means are unchanged. This post-hoc sensitivity is not independently validated truth, and it never changes strict or compliance scores. It demonstrates why frozen extraction penalties cannot be read directly as reasoning losses. Full per-cell means, paired intervals, changed rows, prompt hashes and independent arithmetic checks are archived.