# ANSWER EXTRACTION CHANGES CONCLUSIONS IN CONTROLLED MATHEMATICAL REASONING TESTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A mathematical answer score combines the response a model produces with the rule used to read that response. We separate these two components in a controlled study of six mathematical template families. The dataset contains 240 programmatically checked problems, each presented with and without a task-specific reasoning cue. Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct produce 960 responses under a common decoding protocol. A strict final-answer rule assigns correct scores to 217 of 480 Qwen responses and only 2 of 480 Llama responses. Yet the rule fails to extract an answer from 477 Llama responses, including texts that state a numerical conclusion in ordinary language. A post hoc, label-blind terminal-answer sensitivity recovers 242 additional Llama predictions, raising its correct count to 135 without changing any generated text. Under this sensitivity, the mean cue effect is $-0.4$ percentage points for Qwen and $-2.9$ points for Llama, with descriptive family-level intervals spanning zero. Individual family effects vary substantially. These results do not establish either a universal cue benefit or a model ranking in mathematical reasoning. They show why a benchmark must distinguish answer-format compliance, extractable correctness, and the effect of a reasoning intervention. We release paired prompts, exact labels, complete responses, and both scoring rules so that these quantities can be recomputed separately.

## 1 INTRODUCTION

Mathematical word problems offer apparently simple evaluation targets: a model produces a number, and an evaluator compares it with a reference. This arrangement underlies widely used benchmarks such as GSM8K and MATH (Cobbe et al., 2021; Hendrycks et al., 2021). In practice, a language model produces a text rather than a number. The evaluator must decide which expression is the final answer, whether its units agree with the question, and whether an incomplete or contradictory response has a usable conclusion. Each decision changes what an accuracy score measures.

This distinction matters when a study changes the wording of a question. A reasoning cue can affect both the calculation and the form of the final response. If the scoring rule accepts only one form, the measured intervention effect can combine a change in reasoning with a change in answer-format compliance. The same issue affects comparisons between models that respond differently to an identical formatting instruction. A common prompt is an important control, but it does not make the interpretation of every unextracted answer identical.

We examine this problem in six elementary mathematical families: equal-distance average speed, climbing with overnight regression, conditional dice probabilities, digit occurrences, successive percentage changes, and rectangle counting. Forty parameter instances are sampled for each family. Each instance receives a narrative prompt and a second prompt with one mathematical cue. Both request an exact integer or fraction after `Final:`. The cue supplies useful reasoning information; it is not a meaning-preserving paraphrase. Two open instruction-tuned models generate one greedy response per prompt.

The largest cross-model difference comes from answer extraction. The strict rule identifies only three final predictions among 480 Llama responses, although many texts state a conclusion without the requested marker. A conservative terminal-cue sensitivity increases the number of extracted predictions to 245. Its correct-answer count rises from 2 to 135. Qwen changes less, from 217 to 229 correct answers. These are changes to scoring, not improvements to either model. A low score under



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.

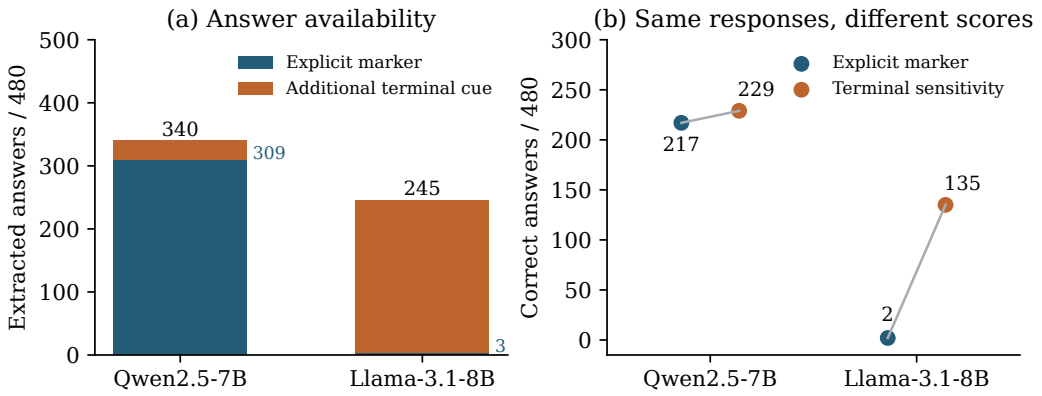


Figure 1: Measured scoring sensitivity on the same 480 responses per model. Left: answers extracted by the explicit-marker rule and the additional answers recovered by the terminal-cue sensitivity. Right: correct answers under the two rules. The sensitivity changes no response and uses no reference label during extraction. An extracted answer can still be mathematically incorrect.

a requested output contract is valid evidence about that contract; it is insufficient by itself to diagnose an absence of mathematical reasoning.

Cue effects also depend on the mathematical family. Neither model has a resolved average benefit across the six selected families under the terminal sensitivity. Qwen improves on the dice family and worsens on digit counting. Llama's effects on percentage changes and rectangle counting have opposite signs. These observations argue for reporting intervention effects at the family level, together with extraction availability, instead of interpreting a pooled accuracy change as one general robustness property.

The paired prompts and two scoring rules let us trace changes in the estimated cue effect to changes in extracted answers. Exact programmatic labels make those comparisons reproducible. The six families form a controlled diagnostic rather than a representative sample of mathematical reasoning tasks.

## 2 Related Work

**Mathematical robustness and controlled inputs.** GSM-Plus evaluates mathematical perturbations, and GSM-Symbolic uses symbolic templates to vary problem instances (Li et al., 2024; Mirzadeh et al., 2025). MathCheck organizes mathematical evaluation around a checklist of behaviors (Zhou et al., 2025); CORE-MATH studies the transition from abstract problems to contextual formulations (Cao et al., 2026). Our work shares their interest in controlled variation but focuses on a different comparison: two extraction rules applied to identical responses from a paired cue intervention. Using six elementary families makes the reference calculations and cue interventions inspectable, at the cost of substantially narrower coverage than those benchmarks.

**Prompting and verification.** Chain-of-thought and zero-shot reasoning prompts change the way intermediate work is elicited (Wei et al., 2022; Kojima et al., 2022). Self-consistency aggregates sampled solutions (Wang et al., 2023); Tree of Thoughts and Self-Refine introduce further search or revision (Yao et al., 2023; Madaan et al., 2023). Process supervision and automated step-level supervision assess intermediate mathematical work (Lightman et al., 2024; Wang et al., 2024). We neither compare those methods nor grade the validity of each intermediate step. Our intervention consists of one fixed cue, and our primary outcome concerns the final numerical answer. Program-aided reasoning (Gao et al., 2023) provides another way to separate language generation from calculation; here, programs validate dataset labels rather than solve the model's answer at inference time.

**Measurement and uncertainty.** CheckList motivates behavioral tests that expose failures hidden by aggregate accuracy (Ribeiro et al., 2020). Work on semantic uncertainty emphasizes that surface forms and answer meanings need not coincide (Kuhn et al., 2023; Farquhar et al., 2024). Our exact-rational outcomes have a simpler equivalence relation, but the distinction between surface form and meaning remains essential. Statistical comparisons likewise depend on the sampling unit (Dror et al., 2018; Efron, 1979). We preserve each prompt pair during resampling and distinguish the six chosen families from their 240 parameter instances. The resulting intervals describe this finite study; they do not certify generalization to an unspecified population of mathematical tasks.

## 3 What a Final-Answer Score Measures

Let $x$ denote a problem, $y(x) \in \mathbb{Q}$ its reference answer, and $T$ the generated text. An extractor $E$ maps $T$ to a rational number or an abstention symbol $\perp$. Define

$$A_E = \Pr\{E(T) = y(X)\}, \tag{1}$$

$$C_E = \Pr\{E(T) \neq \perp\}, \qquad Q_E = \Pr\{E(T) = y(X) \mid E(T) \neq \perp\}. \tag{2}$$

When $C_E > 0$, $A_E = C_E Q_E$. The first quantity is the operational score when unextracted answers count as wrong. The second is extraction availability. The third is correctness conditional on extraction. None alone identifies the probability that a reader could infer a correct answer from the complete text.

A strict output contract can be appropriate for an application that requires machine-readable answers. In that setting, failure to emit the requested marker is a real failure. The problem arises when that same score is described as a pure measure of mathematical competence. Conversely, a permissive reader can incorrectly promote an intermediate number to a final answer. Increasing extraction availability is not automatically an improvement in validity.

### 3.1 Paired intervention effects

Write $z = 0$ for the narrative prompt and $z = 1$ for its cued counterpart. For a fixed extractor, the measured effect is

$$\Delta_E = \frac{1}{n} \sum_{i=1}^{n} \big[\mathbf{1}\{E(T_{i1}) = y_i\} - \mathbf{1}\{E(T_{i0}) = y_i\}\big]. \tag{3}$$

Changing $E$ can change this effect even when all $T_{iz}$ remain fixed. Let $E_0$ be the strict rule and $E_1$ an extension that preserves its valid predictions. If $r_z$ is the fraction of responses in condition $z$ newly extracted correctly by the extension, then

$$\Delta_{E_1} - \Delta_{E_0} = r_1 - r_0. \tag{4}$$

This accounting identity does not assume that newly recovered answers are representative of missing answers. It shows why a scoring change can alter the estimated intervention effect without altering generation. In our implementation a complete-number check also guards against malformed numeric prefixes; any change to an existing prediction is counted separately.

### 3.2 Reference independence and ambiguity

The extractor must not use $y$ to decide which substring is the answer. Otherwise the same response may receive different inferred predictions under different candidate labels. A useful reference-independence check holds the response fixed and varies the reference label: the predicted value should remain unchanged. Our extraction function enforces this separation at its interface by accepting text only. The final equality comparison necessarily uses the reference; selecting the predicted answer must not.

Reference independence is necessary but insufficient. An extractor can be label-blind and still misread an intermediate calculation, a negated conclusion, or a percentage. We therefore retain abstention, use a restricted numerical grammar, and report the terminal extension as a sensitivity analysis rather than a validated replacement for mathematical judgment. For example, a response can correctly state that a price is $98.79\%$ of its original value and then box $9879/10000$. The latter is a ratio, not the requested percentage. Selecting the earlier percentage because it agrees with the label would conceal a unit inconsistency rather than resolve it.

Table 1: The six families. Each has 40 paired instances. $a, b$ are speeds or daily climbing quantities as appropriate; $h$ denotes height, $s$ the number of die sides, $p$ a percentage, and $H, V$ the numbers of grid lines.

| Family | Exact target | Second computation route |
|---|---|---|
| Average speed | $2ab/(a+b)$ for equal distances | Total distance divided by integer travel times at a common distance |
| Climbing | $1 + \lceil (h-a)/(a-b) \rceil$ for $h > a > b > 0$ | Day-by-day ascent and overnight regression |
| Conditional dice | Favorable ordered outcomes divided by $2s-1$ | Enumerate all ordered rolls satisfying the condition |
| Digit occurrences | Count every occurrence in $1, \ldots, n$ | Positional counting with a separate leading-zero correction |
| Percentage change | $100 - p^2/100$ percent | Exact sequential price updates from an integer initial price |
| Rectangles | $\binom{H}{2}\binom{V}{2}$ | Enumerate pairs of horizontal and vertical lines |

### 3.3 What remains unidentified after answer recovery

Extraction sensitivity does not identify an ideal score for every response. A useful finite-sample calculation makes the remaining uncertainty explicit. Consider any completion of the evaluation that retains its $c$ correct and $w$ incorrect extracted answers but assigns each of the $m$ abstentions either status. There are $n = c + w + m$ responses.

**Proposition 1** (Abstention completion bounds)**.** *Under this fixed-decision convention, the completed accuracy lies in*

$$\left[\frac{c}{n}, \frac{c+m}{n}\right]. \tag{5}$$

*Both endpoints are attainable by assigning all abstentions incorrect or all correct. For two models $a, b$, the completed accuracy difference lies in*

$$\left[\frac{c_a}{n_a} - \frac{c_b + m_b}{n_b}, \frac{c_a + m_a}{n_a} - \frac{c_b}{n_b}\right]. \tag{6}$$

The proof is direct: if $u$ of the missing responses are assigned correct, the completed count is $c + u$ with $0 \le u \le m$. These are accounting bounds, not confidence intervals or a new statistical estimator. They require retaining the extracted decisions; if a semantic reader also overturns extracted answers, the convention no longer applies.

Under terminal scoring, the bounds are $[47.7, 76.9]\%$ for Qwen and $[28.1, 77.1]\%$ for Llama. The corresponding Qwen-minus-Llama difference can range from $-29.4$ to $48.8$ percentage points. This width does not imply that the missing responses would all be correct in one model and wrong in the other. It shows which additional assumptions a format-independent ranking would need. Without inspecting those responses under a validated semantic procedure, the observed score gap does not determine that ranking. Reporting the operational score alongside availability is more informative than silently dropping missing cases or imputing their quality.

## 4 Controlled Problem Construction

### 4.1 Six families with exact labels

We sample 40 instances from each family using seed 20260921. All arithmetic uses exact integers and rational numbers. Two computation routes must agree on every reference answer. Table 1 gives the mathematical target and the second route. Agreement checks implementation consistency; it is not independent human validation of the natural-language prompt.

The sampling procedures reject duplicate parameter tuples. Each family also specifies a plausible shortcut, such as the arithmetic mean for equal-distance speed or the count of elementary cells for

Table 2: Parameter domains before rejection of duplicate tuples or instances whose designated shortcut is correct. The accepted 40 instances per family are fixed for both prompt conditions and models.

| Family | Sampling domain | Designated shortcut |
|---|---|---|
| Average speed | Distinct $a, b \in \{10, 15, \ldots, 120\}$ | Arithmetic mean $(a+b)/2$ |
| Climbing | $3 \le a \le 11$, $1 \le b < a$, $a+2 \le h \le 59$ | $\lceil h/(a-b) \rceil$ |
| Conditional dice | $s \in \{6, 8, 10, 12\}$, $1 \le k \le s$, $k+1 \le t \le k+s$ | Named-die probability $1/s$ |
| Digit occurrences | $80 \le n \le 799$, digit $0 \le d \le 9$ | Number of integers containing $d$ |
| Percentage change | 40 distinct integer percentages $1 \le p \le 79$ | Unchanged price, $100\%$ |
| Rectangles | Integers $3 \le H, V \le 15$ | Elementary cells $(H-1)(V-1)$ |

rectangle counting. Instances are retained only when the exact answer differs from that shortcut. This selection is useful for a diagnostic but changes the task distribution: the study is enriched for cases where the designated shortcut fails. It does not estimate failure rates on naturally occurring word problems.

The dice family explicitly conditions on at least one die showing a value, rather than observing a named die. The climbing family stops on reaching the wall's top before an overnight slide. Digit counting excludes leading zeros and counts repeated occurrences within one integer separately. Percentage questions ask for a percentage of the original price, not the corresponding unitless ratio. These distinctions determine the reference labels and remain fixed across each prompt pair.

### 4.2 Sampling domains and selection effects

Table 2 makes the numerical distribution explicit. Some nominally different parameter tuples induce related calculations. Swapping the two speeds leaves the equal-distance average unchanged, and swapping the grid dimensions leaves the rectangle count unchanged. Our exact-string uniqueness check does not turn these symmetries into distinct mathematical mechanisms. Accordingly, the 40 instances within a family are treated as a selected collection of parameterized tasks, not 40 independently discovered reasoning skills.

Rejecting cases where a shortcut coincides with the reference also affects interpretation. For digit occurrences, counting integers containing a digit is wrong only when at least one integer contributes multiple occurrences. The accepted subset therefore emphasizes precisely that distinction. An intervention effect on the subset estimates performance under this diagnostic selection, not the average effect over every number range. The same principle applies to the other designated shortcuts. We report all accepted instances and fixed ranges so that a reader can change the target population explicitly rather than infer one from the aggregate score.

### 4.3 Cue intervention

The two prompts in a pair share all numerical parameters and the requested output form. The second adds a fixed family-specific cue: use total travel time; stop before the final slide; enumerate the full conditional sample space; count each digit position; apply the second percentage to the changed price; or include rectangles spanning several cells. The cue can reveal part of the solution strategy. Thus an improvement would be evidence that this assistance helps under the tested conditions, not invariance to a harmless wording change. A decline would not by itself demonstrate deception or the absence of a reasoning procedure.

The same 240 instances appear in both models and both conditions. We keep the six families equally weighted. There is one generated response per condition, so the study measures checkpoint behavior under greedy decoding rather than variation over stochastic decoding seeds. Multiple instances within a family vary parameters but reuse the same mathematical and linguistic structure.

# 5 GENERATION AND SCORING PROTOCOL

## 5.1 MODELS AND RESPONSE COLLECTION

We evaluate Qwen2.5-7B-Instruct (Qwen et al., 2024) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). Each model uses its own chat template with one user message and no demonstration. Decoding is greedy, with a ceiling of 512 new tokens; model weights use bfloat16. The dataset, prompt pairings, and generation settings are identical across the 480 responses requested from each checkpoint. Complete text and generated token IDs are retained. The study contains 960 responses in total.

The token ceiling is part of the evaluated contract. Reaching it can prevent a final answer from appearing, so it should not be interpreted as a proof that a longer response would remain wrong. We report this endpoint separately: 99 Qwen responses and 137 Llama responses reach the ceiling. Neither extending a response nor changing a decoding setting is part of the scoring sensitivity. All comparisons between extraction rules use the exact same saved text.

## 5.2 EXPLICIT-MARKER SCORING

The original extractor looks for a numerical answer following `Final:`, or a boxed integer, decimal, or fraction. The selected expression is converted to an exact rational value, and missing or invalid predictions count as wrong. A complete-number grammar sensitivity rejects unsupported expressions such as a numeric prefix of scientific notation or an unfinished calculation. This grammar check does not use reference labels.

An explicit-marker rule rewards more than numerical correctness: a model must also place its answer where the rule expects it. That requirement is shared in the prompts but can have different compliance rates across checkpoints. We therefore report the number of extracted answers beside the correct count. We do not replace missing answers with a guessed value, use the designated shortcut to locate a substring, or search the response for a number equal to the label.

## 5.3 TERMINAL-CUE SENSITIVITY

After observing the high missing-answer rate, we add one conservative, post hoc extraction sensitivity. It first retains a valid explicit prediction. If an explicit marker is present but malformed, the rule abstains rather than searching earlier reasoning for a convenient answer. Otherwise it examines only the last nonempty line, requiring a recognizable conclusion cue and a single numerical expression. It rejects alternative answers, questions, approximate qualifiers, unsupported operations, and invalid fractions. Accepted decimals and fractions are normalized exactly. Units are not inferred from the reference answer.

The terminal rule is applied to every response from both models. Its purpose is to test whether conclusions depend heavily on a narrow output format, not to maximize benchmark accuracy. The rule cannot read arbitrary mathematics. It can miss conclusions spread across lines, refuse an answer sentence containing several quantities, and accept a sentence whose local syntax hides a deeper contradiction. We retain the original scores, expose every changed prediction, and make no claim of independent human validation. Section 6 treats its results as sensitivity estimates.

## 5.4 AGGREGATION AND UNCERTAINTY

For each family we report the paired difference between cued and narrative correctness. The overall point estimate averages the six family means. Descriptive hierarchical intervals use 20,000 draws, resampling six families with replacement and then 40 paired instances within each selected family. Pairing preserves the shared problem and reference answer. This differs from treating all 240 templates as independent task types.

Only six families are available, and they were selected by design. A hierarchical bootstrap cannot create a representative task population from that selection. Its intervals summarize sensitivity to the observed family composition and parameter instances. They are not simultaneous intervals for every family, nor a guarantee about all mathematical word problems. We report all family effects rather than selecting positive ones. No threshold or prompt cue is tuned to maximize a final aggregate score.

# 6 RESULTS

## 6.1 FORMAT COMPLIANCE DOMINATES THE STRICT CROSS-MODEL GAP

Figure 1 separates extraction availability from correctness. The explicit-marker rule extracts 309 of 480 Qwen responses and 3 of 480 Llama responses. Its correct counts are 217 and 2, respectively. The terminal-cue sensitivity extracts 340 Qwen responses and 245 Llama responses, with correct counts of 229 and 135. In percentage terms, Llama's measured score changes from 0.4% to 28.1%, while Qwen's changes from 45.2% to 47.7%.

The additional extracted responses are not uniformly correct. Of the 242 extra Llama predictions, 133 are correct; of the 31 extra Qwen predictions, 12 are correct. These counts distinguish additional extracted answers from additional correct answers. Recovery alone does not establish arithmetic correctness. The strict scores measure compliance with the requested interface as well as correctness, so their cross-model gap is not a pure comparison of mathematical competence.

The terminal sensitivity does not eliminate the Qwen–Llama gap in extractable correctness. Nor does it establish that its remaining gap is an unbiased competence estimate. Nearly half of Llama responses still have no extracted answer under the conservative rule. Among the reasons are lengthy reasoning without a terminal cue, more than one numerical expression in a concluding line, and responses reaching the output ceiling. Their possible correctness is not observed by either rule.

## 6.2 CONDITIONING ON EXTRACTION CHANGES THE EVALUATED SUBSET

One tempting alternative is to discard every unextracted response and report correctness only among the remaining ones. For the strict rule, this produces $217/309 = 70.2\%$ for Qwen and $2/3 = 66.7\%$ for Llama. The two percentages summarize very different selected subsets: the Llama estimate uses only three of the 480 requested outputs. Under terminal scoring, the corresponding conditional scores are $229/340 = 67.4\%$ and $135/245 = 55.1\%$. The denominators and the selected responses both change with the extractor.

The terminal rule increases each model's unconditional correct count while decreasing its correctness conditional on extraction. Newly recovered responses include a lower proportion of correct conclusions than the originally extracted set, accounting for the decline in conditional accuracy. The identity $A_E = C_E Q_E$ accounts for both movements. Reporting only $Q_E$ would obscure the improvement in answer availability; reporting only $A_E$ would obscure how much of the score is driven by availability.

A fair comparison cannot generally be obtained by assuming these omissions are random. Formatting, reasoning length, mathematical family, and whether a conclusion is reached can all influence extraction. Restricting both models to the intersection of items with extracted answers would answer a further conditional question, on a subset selected jointly by their behavior. It would not recover accuracy on the original 240 paired instances. We therefore keep the full denominator for the operational score and provide extraction counts as a separate endpoint. The completion bounds describe how much remains unknown under a fixed-decision convention rather than imputing the missing cases from the extracted ones.

## 6.3 CUE EFFECTS DIFFER BY MATHEMATICAL FAMILY

Figure 2 compares the cue effect under both extraction rules. Under terminal-cue scoring, Qwen improves by 20 points on conditional dice and 10 on average speed, while changes on digit counting, percentages, rectangles, and climbing are $-15$, $-2.5$, $-7.5$, and $-7.5$ points. Llama improves by 42.5 points on percentage changes and declines by 45 points on rectangle counting; its other family changes range from $-7.5$ to zero. These contrasts do not support a uniform response to the extra cue.

The equal-family mean terminal effect is $-0.4$ points for Qwen, with a descriptive 95% interval of $[-11.3, 11.3]$, and $-2.9$ points for Llama, with interval $[-23.8, 18.8]$. The original-rule means are $+1.25$ and $-0.83$ points. The sensitivity changes the average estimates but does not reveal a resolved general benefit. The near-zero averages combine effects of opposite signs. Reporting the six contrasts shows which mathematical structures and answer conventions drive that cancellation.

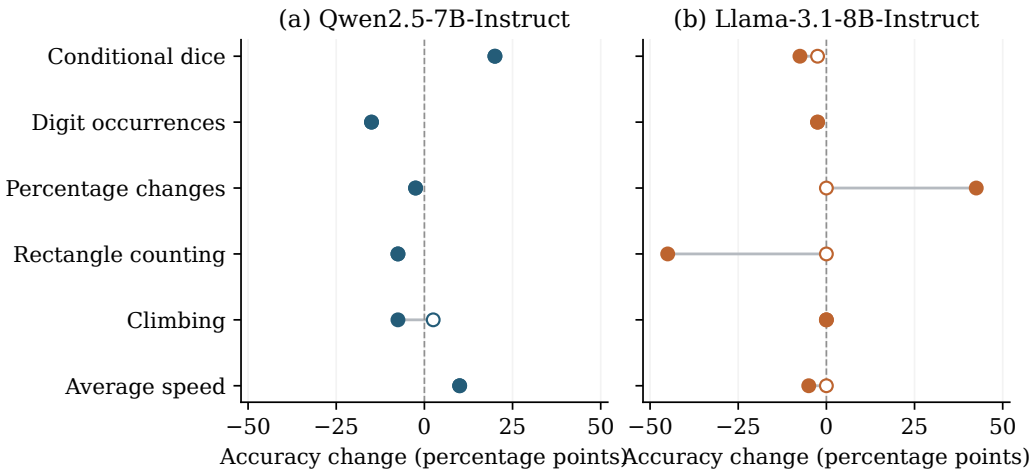


Figure 2: Scaffold-minus-narrative accuracy changes, in percentage points, for all six families. Each family contains 40 paired instances. Open markers use the original explicit rule; filled markers use the terminal-cue sensitivity. These are descriptive family estimates, not six independent demonstrations of a general cue effect.

Table 3: Response-level counts by whether generation reaches the 512-token ceiling. "Missing" refers to no extracted numerical answer, not necessarily no mathematical conclusion in the text. Terminal scoring is the post hoc sensitivity.

| Model | At ceiling | Responses | Strict missing | Terminal missing | Terminal correct |
|---|---|---|---|---|---|
| Qwen | No | 381 | 73 | 42 | 228 |
| Qwen | Yes | 99 | 98 | 98 | 1 |
| Llama | No | 343 | 340 | 100 | 135 |
| Llama | Yes | 137 | 137 | 135 | 0 |

The Llama percentage result illustrates why formatting cannot be assumed to cancel in a paired contrast. Many percentage conclusions are written as ordinary sentences. Recovering those conclusions changes the difference between conditions, not only the model's absolute score. Conversely, the rectangle family includes responses that do not end in one extractable count, especially under the cue. Neither observation alone distinguishes mathematical difficulty, instruction following, or response organization as the cause.

### 6.4 Errors within apparently correct reasoning

Exact scoring of the final answer can disagree with the apparent quality of an intermediate calculation. In one Qwen percentage response, the text correctly derives a final price of $0.9879P$ and states $98.79\%$, then boxes $9879/10000$ as its answer. The question requests a percentage, whose exact numerical value is $9879/100$. Under the declared endpoint the boxed ratio is wrong. Another response makes the multiplication error $1.11 \times 0.89 = 0.9979$ before boxing the resulting ratio. These two responses share a wrong final score but have different failure mechanisms.

This example motivates retaining complete text and distinguishing units from extraction syntax. It does not justify changing a label to match the model's preferred representation. A future semantic grader could assess whether the text contains enough information to infer a correct result, but that would define a different outcome. It would require its own validation, especially for responses with contradictory conclusions.

### 6.5 MISSING ANSWERS ARE NOT ONLY A LENGTH-LIMIT EFFECT

Table 3 separates the extraction issue from the generation ceiling. Among Llama's 343 responses ending below the ceiling, 340 have no strict extracted answer. The terminal extension recovers 240 predictions in this stratum, including all 133 additional correct answers. Thus the strict missing rate cannot be attributed solely to a response ending at its maximum permitted length. Below-ceiling text often has an ordinary-language conclusion that is outside the requested output format.

The pattern is different at the ceiling: almost all responses remain unextracted under the terminal rule, and its only correct answer in that stratum is from Qwen. This is an association with observed response length, not an estimate of the effect of allowing more tokens. Difficult problems can generate longer responses, and repetitive or incomplete reasoning can reach the ceiling without a conclusion. The experiment does not distinguish these mechanisms by changing the budget. Stratification is useful for diagnosing the score, but selecting only below-ceiling responses would change the evaluated problem mix and invalidate a direct comparison with the original all-response accuracy.

## 7 IMPLICATIONS AND LIMITATIONS

**Operational accuracy and mathematical competence.** A strict answer contract answers an operational question: does the model deliver a correct value in the requested form? A broader terminal reader asks a different question about conclusions expressed in text. Both can be useful. A study should name its target, report extraction availability, and preserve abstentions. Calling every unextracted answer a mathematical error overstates what was measured; treating every natural-language number as a valid final answer creates the opposite error.

**Post hoc extraction.** The terminal rule was introduced after seeing output-format failures. Its recovered accuracy is not a new held-out performance result. The paired response corpus makes the source of the change inspectable, but does not erase the post hoc choice. We therefore retain the original endpoint and use the terminal rule to bound the interpretation of a strict comparison. The observed sensitivity is enough to reject a format-independent reading of that comparison, even though it does not identify the ideal semantic score.

**Task and intervention coverage.** A mathematical cue adds information, unlike a purely stylistic rewrite. Its effects can depend on the checkpoint, the exact wording, the numerical instance, and the permitted output length. We study two instruction-tuned checkpoints, six handcrafted families, one cue per family, and one greedy output per condition. We do not evaluate stronger reasoning checkpoints, trained verifiers, multiple cue paraphrases, or repeated stochastic generations. The exact-label procedures check arithmetic consistency, not human judgments about ambiguity or task representativeness.

**Separate measurement validity from model improvement.** No weights change in this study. Recovering a correct prediction from a saved response improves neither the generated reasoning nor its interface compliance. Likewise, an aggregate cue effect near zero does not prove that cues are useless or that the two prompt forms are equivalent. The measurements distinguish output-format behavior, extractable correctness, and cue effects under each scoring rule; claims about broader competence require a validated semantic evaluation.

## 8 CONCLUSION

Controlled mathematical prompts do not by themselves ensure a controlled measurement of reasoning. In a 960-response paired study, a narrow answer-format rule leaves almost all Llama responses unextracted; a label-blind terminal sensitivity changes its correct count from 2 to 135 on the same text. Family-specific cue effects remain heterogeneous, and neither model shows a resolved average benefit across the six families under that sensitivity. The evidence supports reporting extraction availability, numerical correctness, and intervention effects separately. It does not support a general model ranking or a universal benefit from mathematical scaffolding.

## Reproducibility Statement

The supplementary material contains the exact problem generator, both calculation routes for every label, all paired prompts, response text and token counts, generation settings, scoring code, and scripts for the reported tables and figures. The model weights are obtained separately from their public releases. Recomputed scoring results are deterministic apart from the stated seeded bootstrap.

## Ethics Statement

The study uses programmatically generated mathematical problems and public model checkpoints. It contains no personal data or human-subject experiment. Scores are specific to the declared response and evaluation contracts and should not be interpreted as measurements of human cognition or as general rankings of model reliability.

## AI Assistance

OpenAI Codex was used to aid and polish manuscript writing and to assist with literature retrieval and discovery, including finding related work. The authors are responsible for the final manuscript and citations.

## References

Bowen Cao, Dongdong Zhang, Yixia Li, Junpeng Liu, Shijue Huang, Chufan Shi, Hongyuan Lu, Yaokang Wu, Guanhua Chen, Wai Lam, and Furu Wei. From abstract to contextual: What llms still cannot do in mathematics. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 24794–24816, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/2a6fc6774932e7835349089b0287ed2c-Paper-Conference.pdf.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Bradley Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. URL https://doi.org/10.1214/aos/1176344552.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10764–10799. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23f.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

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. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2961–2984, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.163. URL https://aclanthology.org/2024.acl-long.163/.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison 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. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 94743–94765, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ec2e7a896f8250986b3907f57621ce94-Paper-Conference.pdf.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek Wong, Xiaowei Huang, Qiufeng Wang, and Kaizhu Huang. Is your model really a good math reasoner? evaluating mathematical reasoning with checklist. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 34238–34281, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/54d2d38a56a74387d5916ee40e462295-Paper-Conference.pdf.

## A Exact Reference Computations

For equal-distance speed, choosing distance $d = \mathrm{lcm}(a, b)$ makes both travel times integers and yields $2d/(d/a + d/b) = 2ab/(a + b)$. For climbing with $h > a > b > 0$, the number of incomplete days is $\lceil (h - a)/(a - b) \rceil$; the final day has no slide. For the dice event, the conditional sample space contains $2s - 1$ ordered pairs. The favorable count is one when both dice show the specified value and two otherwise, under the sampled target range. The digit oracle enumerates decimal strings and checks them against positional counting, with leading zeros excluded. Percentage labels apply the two updates exactly to an initial price of 10,000. Rectangle labels enumerate pairs of horizontal and vertical lines and compare with the binomial-coefficient formula.

## B Scoring Sensitivity Details

The extraction function accepts text only. Reference answers and designated shortcuts enter only after extraction, when predictions are compared to labels. The terminal extension is deliberately conservative: it keeps explicit answers, refuses malformed explicit payloads, examines the last nonempty line, and requires a conclusion cue with one exact numerical value. Invalid denominators, unresolved alternatives, approximate qualifiers, and unsupported expressions cause abstention. The package records original and sensitivity predictions for all responses, including unchanged and incorrect predictions. Parser tests include scientific-notation prefixes, incomplete arithmetic, contradictory alternatives, and final lines that retract an earlier conclusion.