# SCORING THE SAME OUTPUTS: HOW ANSWER-EXTRACTOR CHOICE AND CONSISTENCY CHANGE BENCHMARK COMPARISONS

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

## ABSTRACT

Free-response benchmarks do not score model text directly: an answer extractor first decides which number counts as the answer, so two pipelines scoring the same outputs can report different gaps. We make that scoring step the object of study. For a frozen class of seven deterministic extraction rules, we compute every gap fixed outputs can yield when one shared rule scores every response, and how far the range expands when the rule may vary per question or per system—oracle stress tests of assignment consistency. A registered reanalysis of a disclosed development cohort (22 comparisons, four 1–2B families) and a preregistered held-out confirmation locate both robustness and fragility, and the pipeline falsifies its own development sign flip. On 193 harder ContextMATH pairs the audit separates two regimes: Qwen2.5's scenario effect has no unique value until the extraction rule is declared (a shared-rule range of $-13.0$ to $+12.4$ points at 7B, $-14.5$ to $+8.8$ at 32B—rule dependence, not a confidence interval), while four newer checkpoints from two families earn exact within-class robustness certificates: one identical negative gap under every rule and assignment. Bigger models alone do not remove scoring ambiguity. Declare one extractor before scoring, apply it to every system, and report the shared-rule range over a prespecified class.

## 1 INTRODUCTION

Suppose two papers report GSM8K accuracy for two small models. One number came from the authors' script, which accepts an answer only after an explicit final-answer marker; the other from a harness whose default takes the last number. Is the gap a property of the models or the scripts? Scored by different rules, the comparison cannot separate them. This is a measurement paper about that scoring step; it proposes neither a new model nor a new benchmark.

That scores move with the rule is well known (Biderman et al., 2024; Hochlehnert et al., 2025). Consider a completion whose last line is "So she has 20 left." (gold 20): a rule requiring a final-answer marker records no answer, scored incorrect; a rule that can also use an unmarked number on the final line records 20, scored correct. Both are reasonable. When a 1.5B-class Qwen model moves from its original protocol (three-shot concise, 768 tokens, temperature 0.8) to a corrected one (zero-shot chain-of-thought, 2048 tokens, temperature 0.7), 31.2% of its first completions carry an unmarked number on the final line and 22.6% end with the prompt's own placeholder line, "The answer is N.", which every rule in our class treats as no answer. Three fixed extractor styles outside our class, applied identically to both protocols, read that change as $+14.0$, $-8.9$ or $-25.5$ points (Section 4).

This paper is about a second, quieter dependence. A comparison also depends on where the rule was allowed to vary: one rule shared by every response; one rule per question, shared by both systems; or a rule chosen separately for each system and question. Until a paper says which rule it used and that the same rule scored every system, its reported gap is not fully pinned down. Within a frozen class of seven deterministic rules, we compute every gap fixed outputs can yield under each constraint. The shared-rule range is the primary, controlled comparison; the two looser ranges are oracle stress-test envelopes—what assignment freedom could produce once outcomes are known. When all three collapse to a point, the comparison is robust to every rule in the class; otherwise 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.

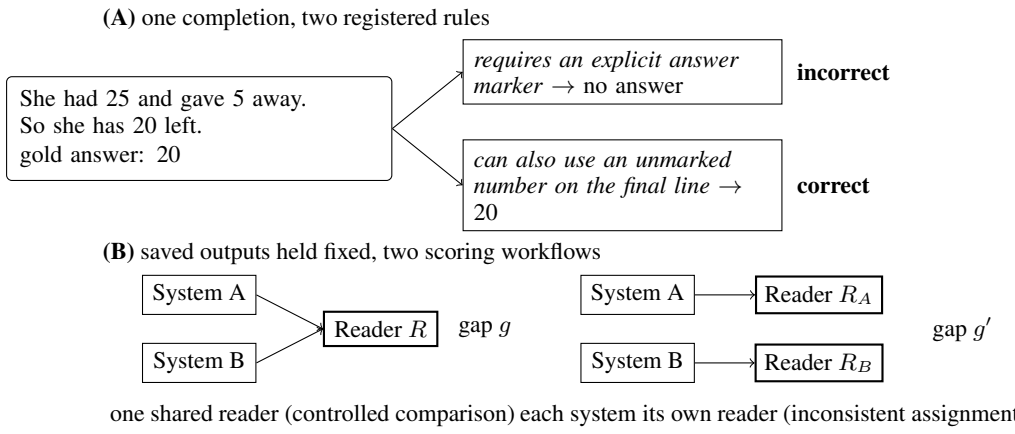



Figure 1: How can fixed responses yield different reported gaps? (A) One completion ending with an unmarked number on its final line is scored correct by a rule that can use such numbers and incorrect by a rule that requires an explicit answer marker; both rules belong to the seven-rule class studied here. (B) With the saved outputs held fixed, the reported gap between two systems changes when each is scored by its own reader ($R_A$, $R_B$) instead of one shared reader $R$—an inconsistent-assignment example. Scoring each system with its own reader is one point inside the free-assignment envelope reported in Section 4; the shared-reader gap is the controlled comparison. Synthetic responses.

range is the report. Section 5 asks whether any of this survives harder problems, larger models, and newer training recipes.

We make three contributions. **Method:** for a finite, frozen class of seven deterministic answer-extraction rules, the exact set of gaps that fixed outputs can yield under each of three assignment disciplines, with exact algorithms and a containment result (the shared-rule set lies inside the per-question set, which lies inside the free set; Section 2); the analysis plan was frozen and hash-pinned before the registered rescoring was executed, on disclosed development data that informed the design (Section 3). **Evidence:** across 22 comparisons of four 1–2B families on numeric grade-school benchmarks, sensitivity is heterogeneous and concentrated in Qwen, with a directly measured answer-placement shift; three fixed extractor styles outside the class, applied identically to both arms, can report the opposite observed direction—the last-number style on Qwen's prompt comparisons, the strict terminal style on Gemma's—so the class is one auditable slice of the scoring landscape (Section 4); and on 193 harder ContextMATH pairs the same audit separates regimes: Qwen2.5 is extractor-sensitive at 7B and at 32B, while four newer checkpoints from two families earn exact negative certificates under every rule and assignment discipline (Section 5). **Confirmation:** a separately preregistered GSM1k study resolves two greedy prompt comparisons as negative and reproduces the near-identical-output control, while the development sign flip does not recur (Section 4.1).

The 22-comparison study is a development-cohort study—those predictions predate the design and shaped it, and we say so wherever such numbers appear; the deliberately narrow confirmation and the Section 5 cells are the places this paper reports beyond that cohort (Section 7).

## 2 Evaluator-policy assignment coupling

This section formalizes the extraction rules of Section 1 as *evaluator policies* over a frozen observation grammar, and the three ways of assigning them as *assignment disciplines*.

### 2.1 Observations and the registered policy class

A completion is first mapped, by one frozen deterministic observation grammar, to a sequence of *candidate observations*: numeric spans with canonical rational values and lexical relation labels. The grammar (Section 3) distinguishes four mutually exclusive relations for a candidate: *terminal-*

*marked* (an explicit final-answer marker followed by the numeral, terminating its line), *marked* (marker-bound but not terminal), *terminal-unmarked* (an unmarked numeral terminating the completion's last content line), and *body-unmarked* (any other eligible numeral). Exclusion rules (quoted text, code spans, instruction echoes) and an abstention rule on same-relation ambiguity are part of the grammar, not of the policies.

**Definition 1** (Policy class)**.** *Order the four relations by evidential strength as the diamond poset with the four cover relations*

$$\textit{terminal-marked} \succ \textit{marked}, \qquad \textit{terminal-marked} \succ \textit{terminal-unmarked},$$
$$\textit{marked} \succ \textit{body-unmarked}, \qquad \textit{terminal-unmarked} \succ \textit{body-unmarked},$$

*where marked and terminal-unmarked are incomparable. A registered evaluator policy is a pair (acceptance set, priority order): an upward-closed subset of this poset together with a linear extension of the poset restricted to that subset. Given a completion's candidates, a policy restricts to candidates whose relation it accepts, selects the highest-priority* populated *relation under its linear extension, and returns the unique normalized value among that relation's candidates; if the values of the selected relation's candidates are not unique, the policy abstains.*

Appendix N (Figure 3) illustrates the diamond and a worked completion.

**Theorem 1** (Enumeration)**.** *The class contains exactly seven policies: five upward-closed acceptance sets, of which three admit one linear extension and two admit two.*

The seven rules vary only in which evidence levels they accept and whether marked or final-line evidence takes priority; all abstain when the selected level contains conflicting values. The enumeration and its decision vectors on a pinned 105-case adversarial battery are frozen by hash. Three historical extractor styles—flexible last-number extraction, marker-bound extraction, and strict terminal parsing—are proven *outside* the class by executed witnesses; the class is not a relabeling of existing extractors, and claims about the class never bound extractors outside it.

### 2.2 Assignment disciplines and attainable effect sets

Fix a comparison: two arms (baseline, candidate) over a shared item frame, a metric, and per-item scored decisions for every policy. An *assignment* maps responses to policies. Three disciplines are registered:

**Definition 2** (Attainable effect sets)**.** *Let $P$ be the registered policy class, and for policy $p$ let $\Delta(p)$ denote the weighted effect (candidate minus baseline) when every response in both arms is scored by $p$. Then $D_{\mathrm{G}} = \{\Delta(p) : p \in P\}$ is the* globally coupled *attainable set. $D_{\mathrm{I}}$ extends the assignment to vary per item while remaining shared across arms within each item; $D_{\mathrm{P}}$ lets the policy vary freely per arm–item pair. Each is an exact finite set of rationals, computed by an exact output-sensitive dynamic program over per-item decision vectors; in the registered equal-weight settings all attainable values live on a bounded rational grid, so the computation is polynomial there (Appendix C).*

**Theorem 2** (Containment)**.** $D_{\mathrm{G}} \subseteq D_{\mathrm{I}} \subseteq D_{\mathrm{P}}$*, with equality not holding in general.*

The containment is elementary; the paper's object is not the containment but the *measured geometry* of these sets for real systems under a registered class, and the discipline that treats them asymmetrically: $D_{\mathrm{G}}$ and its hull are the fair-comparison objects because a reader interpreting "the effect of the intervention" presumes one evaluator applied uniformly; $D_{\mathrm{I}}$ and $D_{\mathrm{P}}$ quantify the additional conclusion-space that weaker coupling disciplines would open, are reported as observed-finite-sample descriptive diagnostics only, and are never used to rescue or upgrade a primary claim.

### 2.3 Estimands and inference

Effects are exact weighted means: equal problem weights on GSM8K ($n = 1319$), and equal template weights on GSM-Symbolic (100 templates $\times$ 5 instances; the template is the independent unit). The primary metric is First@1; Majority@16 and observed-any-correct@16 over the persisted 16-draw bundles are registered as descriptive-only companions. Sampling uncertainty is separated from evaluator variation: for the single joint primary family of all 154 comparison$\times$policy components, a centered sup-norm bootstrap over independent unit resamples (shared resample vectors within a

frame, independent blockwise vectors across frames) yields one simultaneous band per component at joint confidence 0.95; the bands induce endpoint bands for the minimum and maximum of $D_{\mathrm{G}}$ through a single shared coverage event. Sign claims use exactly three calibrated states—positive, negative, or not resolved at simultaneous 95—and are conditional-superpopulation statements about the named frozen systems: the coverage target conditions on the persisted checkpoints, decode settings, and draws, and establishes no method-level generalization across checkpoints, seeds, or decoding randomness. Exact-set facts (hull widths, point-invariance, sign-sensitivity) are always reported as observed finite-sample facts, never as superpopulation claims.

## 3 Registered design and execution firewall

Evaluator studies are unusually exposed to researcher degrees of freedom—the analyst controls both instrument and analysis—so the instrument, the analysis plan, and the execution machinery were frozen and hash-pinned before the registered rescoring opened any prediction. One boundary is stated precisely: the scored predictions predate the design (the disclosed development corpus, previously analyzed under unregistered evaluators), so what is registered is the reanalysis—class, estimands, gates, execution—not the corpus's first exposure. Three mechanisms carry the discipline (full detail in Appendix M):

- **A finite, frozen observation grammar.** A registered conservative ASCII-numeric evaluator with a deterministic input boundary; unsupported answer forms yield reported abstention, never a silent score; a pinned battery of 105 adversarial cases fixes all 735 policy decisions exactly.
- **Preregistration with a fail-closed authority.** The hash-pinned preregistration and a separate code-pinned authority file gate every scored-table code path; the component manifest binding every comparison was built and validated *output blind*—predictions entered only as byte streams for hashing.
- **Adversarial review before results.** Multi-round fresh-context review on hash-named artifacts with no rerolling and every adverse report preserved verbatim; gates only ever tightened; scoring and analysis were separately author-authorized against exact digests, executed once each.

**Cohort.** The scored predictions are a frozen development corpus: 31 evaluation runs of four instruction-tuned families (Gemma-3 1B, Qwen 1.5–2B-class, Llama 1B-class, OLMo 1B-class) under controlled prompt, budget, temperature, and protocol variations on GSM8K (full 1319-item test split) and GSM-Symbolic (500 instances of 100 templates), persisted with per-run configuration snapshots and byte-hashed prediction files. These predictions predate the design and shaped it; every result computed on them is a development-cohort result (Section 7).

**Held-out confirmation protocol.** A second preregistration froze the confirmation before any completion existed: the full 1,205-problem GSM1k test release (Zhang et al., 2024a) at a pinned revision (private-evaluation terms, no contamination-free claim), one pinned 1.5B-class Qwen instruct checkpoint, an eight-cell prompt/budget/decoding grid with partial-grid interpretation forbidden, the frozen policy class unchanged (disclosed as development-informed), the same clipping gate, and two prespecified replication tests whose failure must be reported as nonreplication, never repaired (Section 4.1; Appendix G).

## 4 Registered results

Appendix D (Table 9) reports every primary development-cohort comparison: the size and hull of the observed attainable set $D_{\mathrm{G}}$ under global assignment, and the registered simultaneous sign state. Every registered statistic in the tables and macro-backed prose of this section is generated mechanically, fail-closed, from the hash-pinned registered results file; the remaining literals are quoted from the frozen preregistration and recorded validity blocks. Three registered disciplines shape what may be claimed: exact-set facts are observed finite-sample facts about this cohort; sign claims exist only in the three calibrated simultaneous states; and a prespecified validity gate suppresses any comparison whose arms exceed the registered maximum clip fraction ($1/10$), without unit subsetting—

clipped completions score as abstention-then-incorrect, and the whole claim is ineligible rather than repaired.

**Assignment coupling is model-dependent.** Of the 22 primary comparisons, 20 keep a single-sign observed hull (13 entirely positive, 7 entirely negative), one is exactly null under every policy, and one contains both signs. 5 comparisons are exactly point-invariant—every one of the seven policies yields the identical rational effect: 4 of the sign-robust comparisons, the historical Gemma comparison being the cleanest instance ($D_{\mathrm{G}} = \{239/1319\}$), and the exactly-null comparison itself—while the remaining sign-robust comparisons keep their qualitative sign with nonzero hull width. Where the class is inert, evaluator-policy assignment coupling costs nothing to control—and controlling it converts an implicit robustness assumption into a measured fact.

**Where it is not inert, it is large, and it concentrates in one family.** Every wide attainable set in the cohort belongs to a Qwen comparison (Table 9). On the historical protocol comparison, Qwen's attainable set is $\{-275/1319,\ -265/1319,\ -43/1319,\ -33/1319\}$: the same frozen predictions support a measured effect anywhere from -20.8 to -2.5 percentage points—an order-of-magnitude spread in claimed effect magnitude decided entirely by which registered policy scores the comparison. A gold-blind post-registered replay of the first persisted completion makes the placement shift directly observable: among 1319 unclipped completions per historical Qwen arm, terminal-marked presence falls from 98.6% to 36.7%, terminal-unmarked rises from 4.0% to 31.2%, and a literal terminal metavariable rises from 0.1% to 22.6%; GSM-Symbolic repeats all three directions, with exact counts and denominators in Appendix D. These nonexclusive lexical rates support an answer-placement interpretation; they do not diagnose a latent computation or add an inferential claim. A post-registered marker-sensitivity diagnostic sharpens the concentration claim (Appendix H). Public code instantiates the reversal: three version-pinned public evaluators (lm-eval `flexible-extract`, Qwen's published last-number rule, Math-Verify) give positive deltas on 10 Qwen comparisons (4 held out) where all seven rules give negative deltas (Appendix P). The disagreement has a measurable structure. On the historical Qwen chain-of-thought arm, lm-eval `flexible-extract` credits 144 completions whose final line is the literal unfilled placeholder ("…was $26. The answer is N."). For this reader, differential credit to completions without an eligible terminal candidate contributes $+16.6$ points, decomposing its $+12.9$-point effect into a $-3.7$-point remainder; the remainder is negative in all 30 reader–comparison cases (Appendix P). These are two measurement contracts: one abstains unless the response identifies a unique answer at a prespecified evidence level; the other recovers an answer from the reasoning trace using positional or symbolic rules. A public reader may correctly recover an answer the first contract rejects; neither is universally right, and the comparison is undefined until one is declared.

**Exactly one primary First@1 comparison is observed sign-sensitive.** The comparison factorial_qwen_budget_sampled_cot has $D_{\mathrm{G}} = \{-14/1319,\ -12/1319,\ 8/1319,\ 10/1319\}$: the attainable set contains both negative and positive effects. Under globally coupled assignment—the fair discipline—the direction of this conclusion is decided by the evaluator policy, not by the systems being compared. This is the phenomenon the framework exists to expose, occurring within a registered class of only seven conservative policies, on the primary metric of a standard benchmark. The hull is small in magnitude (-1.1 to +0.8 percentage points), and its importance is qualitative conclusion-direction sensitivity, not effect size. As an observed finite-sample fact it is exact; at the registered simultaneous level the comparison's sign is not resolved, which is itself the correct report: a comparison whose sign flips with the evaluator should not resolve. The held-out confirmation tests exactly this observation—and falsifies it (Section 4.1). The descriptive sampled metrics add three further observed-sign-sensitive comparisons, every one Qwen (Appendix D).

**Simultaneous sign states and the validity gate.** Of the 17 claim-eligible comparisons, 6 resolve positive at simultaneous 95 (all historical protocol-correction effects that resolve, plus two symbolic transfers), 2 resolve negative (both Qwen prompt-format cells, whose entire attainable hulls are negative), and 9 are not resolved at the joint level. The remaining 5 comparisons—all Gemma factorial cells whose chain-of-thought arms clip up to $320/1319$ of completions against the registered $1/10$ maximum—are suppressed by the validity gate exactly as prespecified. The gate fired against our own model family and was applied unchanged.

Two post-registered cross-frame-dependence sensitivity constructions changed no registered resolution and are reported in full in Appendix L.

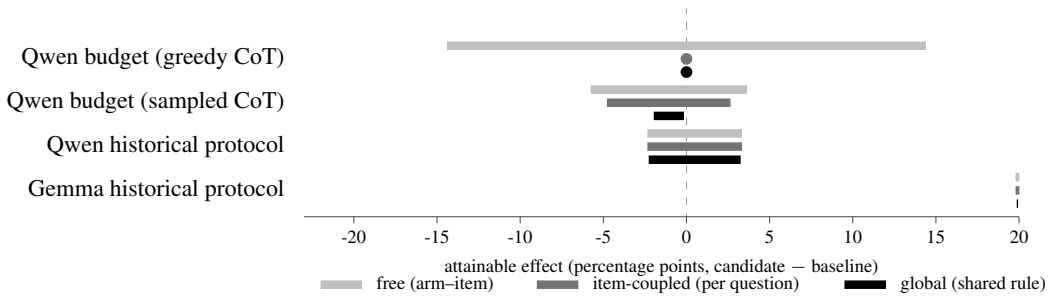



Figure 2: The observed assignment lattice $D_{\mathrm{G}} \subseteq D_{\mathrm{I}} \subseteq D_{\mathrm{P}}$ for four registered comparisons (development cohort; observed-finite-sample descriptive diagnostics). Each bar is an attainable-effect hull under one assignment discipline (legend in figure); dots mark singleton sets. Every coordinate is generated from the pinned results; hull endpoints, set sizes, and all 22 comparisons appear exactly in Appendix D.

**What assignment consistency buys: the $D_{\mathrm{G}} \subseteq D_{\mathrm{I}} \subseteq D_{\mathrm{P}}$ expansion.** The sign flip above is the policy-choice layer of a broader assignment-robustness problem: what happens when the *assignment discipline* itself is relaxed. Figure 2 shows the observed expansion for four registered comparisons (all sets are observed-finite-sample descriptive diagnostics; no primary claim uses the relaxations). The taxonomy is sharp. On the greedy budget comparison, global and item-coupled assignment agree exactly on zero ($D_{\mathrm{G}} = D_{\mathrm{I}} = \{0\}$), yet arm–item assignment spans $\pm 14.4$ percentage points across 381 attainable values: any conclusion in that range could be manufactured by scoring the two arms with different policies per item, while any consistent per-item discipline pins the comparison to exactly null. On the sampled budget comparison the middle layer already breaks: $D_{\mathrm{G}}$ spans about one point, but $D_{\mathrm{I}}$ reaches -13.5 to +13.2 and $D_{\mathrm{P}}$ -18.0 to +17.7 percentage points—item-wise policy variation alone can manufacture large effects in either direction. On the historical comparison the expansion is negligible: the damage is already done at the policy-choice layer inside $D_{\mathrm{G}}$. Which layer of the lattice a pipeline fixes determines which conclusions the same frozen predictions can support.

### 4.1 Held-out registered confirmation

Development-cohort evidence cannot confirm itself. A second preregistration therefore froze, before any completion was generated, an eight-cell confirmation grid on the full 1,205-problem test release of GSM1k (Zhang et al., 2024a)—held out, distribution-matched to GSM8K, used under private-evaluation terms with no contamination-free claim—for one pinned 1.5B-class Qwen instruct checkpoint: {few-shot concise, zero-shot CoT} $\times$ {768, 2048 tokens} $\times$ {greedy, sampled $N$=16}, giving 8 First@1 comparisons in one joint family of 56 components (radius 0.0423). The policy class carried over unchanged and is recorded as development-informed; the execution discipline is that of Section 3 (Appendix G).

**Sign states.** Of the 8 comparisons, 0 resolve positive, 2 resolve negative, 6 are not resolved, and 0 are suppressed—every cell passed the clip gate (largest arm clip fraction $2/241$). Both resolved comparisons are greedy prompt contrasts (zero-shot CoT minus few-shot concise); they share one attainable hull, -22.0 to -8.3 percentage points, entirely negative under every policy—direction robust to assignment. The sampled prompt contrasts are observed-negative under every policy but unresolved.

**The assignment lattice replicates.** The greedy CoT budget contrast reproduces the development pattern exactly: $D_{\mathrm{G}} = D_{\mathrm{I}} = \{0\}$, while $D_{\mathrm{P}}$ spans $\pm 13.7$ percentage points across 331 attainable values. Consistent per-item discipline again pins the comparison to exactly null; arm–item assignment could manufacture double-digit effects either way.

**The sampled sign flip is falsified.** The development cohort's one sign-sensitive comparison does not generalize. On GSM1k the sampled CoT budget contrast is point-invariant at the global and item-coupled layers: $D_{\mathrm{G}} = D_{\mathrm{I}}$ contain the single value +0.17 percentage points—every registered policy

returns the identical effect—while $D_{\mathrm{P}}$ still spans -15.2 to +15.5 points across 371 values. Under the prespecified replication test this is a falsification, and we report it as one: the development sign flip is cohort-specific, not a property of the model family. Structure replicates; the most striking single observation does not.

## 5 Does the ambiguity persist on harder problems and larger models?

The development cohort and the confirmation are grade-school benchmarks and 1–2B models; we therefore repeated the audit on 193 ContextMATH pairs (Cao et al., 2026)—MATH-500 problems rewritten as real-world scenarios, restricted to the exact-rational level-4/5 stratum so both arms share identical gold answers—under one shared fixed protocol (zero-shot chain-of-thought ending "The answer is N.", greedy decoding, an 8,192-token budget), not tuned to any model. Each cell is one paired comparison scored by all seven rules under all three assignment disciplines. The analysis plan, class, and validity gate are those of Section 3; the first two cells were frozen before generation, and the remaining cells are post-registered extensions under the same frozen design, labeled as such in Table 1.

Table 1: ContextMATH regime map: one frozen 193-pair frame, one shared fixed protocol, seven prespecified rules; effects are scenario minus original, in percentage points. $D_{\mathrm{G}}$ is the shared-rule attainable set (Section 2); a singleton means the gap is identical under all seven rules—an exact robustness certificate. "Extension" rows are post-registered under the frozen design; the suppressed Llama row is reported, not omitted, and carries no population claim; the Gemma-3-4B row is a repaired run; per-cell clipping and marking counts are in Appendix J.

| Model (year) | Evidence | Shared-rule gaps $D_{\mathrm{G}}$ (pp) | Verdict |
|---|---|---|---|
| Qwen2.5-7B (2024) | registered; stress test | $[-13.0, +12.4]$ | sign-sensitive |
| Qwen3-4B-2507 (2025) | registered | $\{-10.9\}$ | identical, all 7 rules |
| Qwen2.5-32B (2024) | extension; targeted | $[-14.5, +8.8]$ | sign-sensitive |
| Qwen3.8-27B (2026) | extension | $\{-8.8\}$ | identical, all 7 rules |
| Gemma-4-E4B (2026) | extension | $\{-15.5\}$ | identical, all 7 rules |
| Gemma-3-4B (2025) | extension; repaired run | $\{-17.1\}$ | identical, all 7 rules |
| Llama-3.2-3B (2024) | extension; suppressed | ($-15.0$, descriptive) | clip gate exceeded |

**A selected stress test and a certificate.** For a 2024 model selected as a stress test (Qwen2.5-7B), the scenario effect ranges from $-13.0$ to $+12.4$ points across the seven shared rules: the sign of the reported effect is the extractor's choice. We had preregistered the opposite—a one-directional loss—and report that prediction as falsified. For a 2025 model (Qwen3-4B-2507), all seven rules and all three assignment disciplines give exactly $-10.9$ points: within the class, no extractor choice and no assignment freedom can change the conclusion—an exact robustness certificate.

**Sensitivity persists at 32B; four newer checkpoints are stable in these cells.** A targeted same-lineage follow-up (Qwen2.5-32B) tested whether the 7B sensitivity disappears at scale. It does not: the shared-rule set spans $-14.5$ to $+8.8$ points—against a preregistered expectation of a narrow one-signed effect, again reported as falsified. Under the same frozen design, Qwen3.8-27B, Gemma-4-E4B (2026) and Gemma-3-4B (2025) each give one identical negative gap under every rule and assignment ($-8.8$, $-15.5$ and $-17.1$ points). One further cell is reported in Table 1 as descriptive only: Llama-3.2-3B exceeded the prespecified clipping gate (20 of 193 scenario completions clipped). The Gemma-3-4B row is a repaired run; its first run, with a misconfigured stop set, clipped every completion and is retired as a technical failure (Appendix J).

**The mechanism is visible in the outputs.** Both sensitive Qwen2.5 cells show the same qualitative placement pattern: in the original arm, many completions end inside LaTeX display math (`\[\boxed{...}\]`), so rules disagree about whether a terminal answer exists (7B: 114 vs. 59 such completions across arms, terminal-marked 76 vs. 114; 32B: 73 vs. 20, terminal-marked 96 vs. 147). The certified models place a marked terminal answer almost always (187/182, 188/183,

191/189, and 183/176 of 193), with identical correctness scores across rules. Counts differ across cells; the pattern, not the magnitudes, repeats.

**What may and may not be concluded.** Bigger models do not automatically remove scoring ambiguity: on the same 193 pairs, Qwen2.5 remains extractor-sensitive at both 7B and 32B, while four newer checkpoints receive stable negative certificates under all seven prespecified rules. These seven cells do not estimate how often either regime occurs, and we make no causal claim about model age, scale, or training recipe; "certificate" is always relative to the frozen class. A post-registered LaTeX-aware reader outside the class, declared before rescoring the frozen predictions, selected negative observed effects for both sensitive Qwen2.5 cells and exactly reproduced the three certificates tested; public Math-Verify reproduces two exactly and the third within one item (Appendices O, P).

**For practitioners.**

1. Choose the answer reader before scoring.
2. Apply it identically to every system.
3. Report the shared-reader range over a prespecified plausible class.
4. Treat item-wise and per-system assignment ranges only as stress tests.
5. Report abstention and unsupported-answer rates.

## 6 Related work

**Evaluator sensitivity and evaluation reproducibility.** That scoring choices move benchmark numbers is well documented: harness and extraction differences change reported accuracies (Biderman et al., 2024; Hochlehnert et al., 2025), and prompt-format perturbations alone induce large swings (Sclar et al., 2023), and minute answer-selection changes reorder leaderboards (Alzahrani et al., 2024), and even nominally deterministic settings leave measurable run-to-run instability (Atil et al., 2024). This literature varies the evaluator and reports the induced spread; we instead fix a finite registered class, make the *assignment discipline* an explicit design axis, and compute the exact attainable conclusion sets each discipline induces.

**Evaluator-assignment effects on decisions.** Recent work establishes the broad phenomenon that *which evaluator is assigned* changes downstream decisions in the LLM-judge setting: Afrasyab (2026) show that same-provider judge assignment shifts clinical-domain model rankings. That line defines no policy class over a shared observation grammar, no assignment lattice, and no exact attainable-set algorithms; it concerns generative judge identity. It is the nearest occupied ground; our claim is the narrower formal and registered one: the source-preserving global/item/arm–item assignment-consistency lattice with exact set computation. Terminology: "evaluator coupling" names a different quantity (an evaluator-agent coupling measure $\gamma$) in Liu (2026); we therefore always write *evaluator-policy assignment coupling*.

**Answer extraction and scoring error.** A complementary line measures extraction itself: Molfese et al. (2025) document systematic scoring errors in benchmark answer matching; Wong et al. (2026) separate when a model decides an answer from when it lexically binds it; trained verifiers such as xVerify (Chen et al., 2025) and xFinder (Yu et al., 2025) and answer regeneration (Jo et al., 2025) replace rules with an additional model inference, and deterministic rule-plus-retrieval pipelines extend equivalence handling (Bernard et al., 2025). Our instrument is deliberately the opposite design point: deterministic, source-preserving policies over a frozen grammar—a policy can only return a normalized value associated with an eligible numeral span present in the completion, or abstain; it can never revise, infer, re-solve, or introduce a novel value—which is precisely what makes exact attainable sets and the assignment lattice well-defined objects. Multiple-choice conversion removes extraction fragility by changing the measured construct (Zhang et al., 2024b); generative judges recover recall at the price of stochastic, identity-dependent scoring. Ours is the end where the attainable-effect question has an exact answer—and the reporting recommendation travels to any deterministic pipeline.

**Multiverse analysis and partial identification.** Reporting a conclusion across all defensible specifications originates in psychology (Steegen et al., 2016; Simonsohn et al., 2020); evaluation multiverses expose score variability from model-design and evaluation choices (Simson et al., 2024), with recent work on distributional geometry, multiplicity control, and post-selection inference across

multiverses (Saraceno & Calcagnì, 2026; Rognon-Vael & Rossell, 2026; Vesely & Andreella, 2026). We claim none of those ingredients. Partial identification bounds a latent target under linking assumptions (Manski, 2003; Chen et al., 2026); our setting has no latent target—once a policy and assignment are fixed the effect is point-determined—so $D_{\mathrm{G}}$, $D_{\mathrm{I}}$, and $D_{\mathrm{P}}$ are exact attainable robustness sets over a declared class, not identified sets or confidence regions, and their containment records progressively weaker assignment-consistency constraints.

**Benchmarks.** We evaluate on GSM8K (Cobbe et al., 2021) and the template-perturbed GSM-Symbolic (Mirzadeh et al., 2024), and our historical comparisons revisit protocol-correction effects documented for small open models (Hochlehnert et al., 2025). Sampling-based metrics follow the self-consistency and pass@k lineage (Wang et al., 2023); the registered design treats them as descriptive companions to First@1.

## 7 Limitations and scope

**One development cohort, one narrow held-out confirmation.** Every number in the 22-comparison study is computed on predictions that predate this design and informed it; the preregistration firewalls the *analysis*, not the corpus's first exposure. The held-out confirmation (Section 4.1) is genuinely out-of-development—frozen before any completion existed—but deliberately narrow: one model family at one scale, one benchmark under private-evaluation terms, eight First@1 comparisons, a policy class the development cohort shaped (disclosed, not removed); beyond it, the seven Section 5 cells are the paper's only other out-of-development evidence, and they support no prevalence estimate.

**The class is conservative and registered, not universal.** The evaluator is an ASCII-numeric extraction class over a frozen grammar. Completions whose answers appear only in unsupported forms (prose-only numerals beyond the registered word grammar, LaTeX, non-ASCII scripts) yield measured, reported abstention, but the class makes no claim about evaluators outside it (the three historical extractor styles proven outside the class are bounded by nothing here); wider classes—semantic matchers, LLM judges—need their own registered treatment. The attainable sets are likewise relative to the frozen grammar's boundary choices; a post-registered diagnostic probed three marker edits (Appendix H)—the sole substantively exposed edit reshaped only Qwen sign categories—but this probes marker edits, not broader lexical or grammar robustness. The hull is one auditable slice, not an envelope of practice: pinned public readers fall outside it on the Qwen comparisons, largely by crediting answers that the abstain-unless-unique contract rejects (Appendix P).

**Coverage remains sparse.** The development cohort is four 1–2B instruction-tuned families on grade-school arithmetic word problems; Section 5 extends the audit to 193 harder ContextMATH pairs and checkpoints up to 32B, where both regimes recur. That within-class sensitivity appears only in one lineage across our cells is an observation about those cells, not a family claim.

**Statistical scope.** The simultaneous-95 sign states are conditional-superpopulation statements about the exact frozen checkpoints, decode settings, and persisted draws named in the component manifest; different checkpoints or fresh randomness are outside the registered inference. GSM-Symbolic templates derive from GSM8K items, so the frames share provenance; two post-registered dependence-sensitivity constructions weakened no registered resolution (Appendix L). The sampled metrics are statistics of the persisted 16-draw bundles under a stated-not-verified exchangeability assumption—one reason they are descriptive-only.

**Suppression is loss, not neutrality.** The clipping gate suppressed 5 development-cohort comparisons outright; re-running them at adequate budgets remains future work.

## 8 Conclusion

A reported gap depends on the extraction rule and on whether every response shared it. Within seven declared rules we computed, exactly, every gap the same outputs can produce under three assignment disciplines; held out, the development sign flip was falsified. On harder problems the audit returned both verdicts: sign-flipping sensitivity for Qwen2.5 at two sizes, identical negative gaps under every rule for four newer models. Declare the rule, apply it symmetrically, report the range.

## REPRODUCIBILITY STATEMENT

The complete apparatus is deterministic and hash-pinned: the preregistration configuration, evaluator implementation, 105-case decision battery, output-blind component manifest, pre-output freeze record, fail-closed authority file, scored unit table, and registered results each carry canonical SHA-256 digests, as do the held-out confirmation's own preregistration, component manifest, score table, and results, and each registered analysis is reproducible from the frozen predictions by a single pinned command. To respect double-blind review, the *anonymized* supplementary material contains the source code, the preregistration, scrubbed summaries of the full adversarial review chronology, and blinded artifact identifiers; it contains no repository URLs, usernames, local paths, issue or pull-request numbers, or searchable version-control hashes. The complete hash-linked record—including every adverse review verbatim and the full digest chain—is preserved privately and will be released upon de-anonymization. The development cohort uses the public GSM8K test split and the GSM-Symbolic release at pinned revisions, and the held-out confirmation uses the GSM1k test release at a pinned revision; GSM-Symbolic is used for evaluation only under its license (CC-BY-NC-ND-4.0), GSM1k under private-evaluation terms (its dataset card omits an explicit license), and no benchmark data is redistributed. Evaluator validation rests on the frozen 105-case adversarial battery, the structural test suites, and the executed fresh-context review ladder, whose reviewing model families overlap the drafting families (disclosed as AI review, not human validation).

## AI USE STATEMENT

Generative AI assisted with study design, formal reasoning, implementation, analysis, literature review, and writing. The author reviewed all AI-assisted work and takes full responsibility for the submission.

## REFERENCES

Koyar Afrasyab. Evaluating medical AI under missing information: Same-provider judges and human raters change apparent safety. *arXiv preprint arXiv:2607.18828*, 2026.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M. Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.744.

Berk Atil, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. Non-determinism of "deterministic" LLM settings. *arXiv preprint arXiv:2408.04667*, 2024.

Raymond Bernard, Shaina Raza, Subhabrata Das, and Rahul Murugan. EQUATOR: A deterministic framework for evaluating LLM reasoning with open-ended questions. *arXiv preprint arXiv:2501.00257*, 2025.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, Francois Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

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. *arXiv preprint arXiv:2601.23048*, 2026.

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.

Xiaohong Chen, Ashesh Rambachan, and Elie Tamer. Partial identification from LLM prompts. *arXiv preprint arXiv:2606.15031*, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. *arXiv preprint arXiv:2504.07086*, 2025.

Hwiyeol Jo, Joosung Lee, Jaehone Lee, Sang-Woo Lee, Joonsuk Park, and Kang Min Yoo. Finding answers in thought matters: Revisiting evaluation on large language models with reasoning. *arXiv preprint arXiv:2510.14773*, 2025.

Zewen Liu. Mapping the evaluation frontier: An empirical survey of the bias-reliability tradeoff across eleven evaluator-agent conditions. *arXiv preprint arXiv:2607.00304*, 2026.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

Francesco Maria Molfese, Luca Moroni, Luca Gioffré, Alessandro Scirè, Simone Conia, and Roberto Navigli. Right answer, wrong score: Uncovering the inconsistencies of LLM evaluation in multiple-choice question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18477–18494, 2025.

Paul Rognon-Vael and David Rossell. E-values for multiplicity control in multiverse analysis. *arXiv preprint arXiv:2607.17596*, 2026.

Giovanni Saraceno and Antonio Calcagnì. Towards a geometric characterization of multiverse analysis. *arXiv preprint arXiv:2607.11345*, 2026.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2023.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4:1208–1214, 2020.

Jan Simson, Florian Pfisterer, and Christoph Kern. One model many scores: Using multiverse analysis to prevent fairness hacking and evaluate the influence of model design decisions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1305–1320, 2024. doi: 10.1145/3630106.3658974.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016.

Anna Vesely and Angela Andreella. Post-selection inference for multiverse analysis in mixed-effects models (PIMAX). *arXiv preprint arXiv:2607.03225*, 2026.

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.

Hugh Mee Wong, Rick Nouwen, and Albert Gatt. When models decide and when they bind: A two-stage computation for multiple-choice question answering. *arXiv preprint arXiv:2601.03914*, 2026.

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.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024a.

Ziyin Zhang, Zhaokun Jiang, Lizhen Xu, Hongkun Hao, and Rui Wang. Multiple-choice questions are efficient and robust LLM evaluators. *arXiv preprint arXiv:2405.11966*, 2024b.

## A THE SEVEN REGISTERED POLICIES

**Enumeration.** The relation poset is the diamond of Section 2: terminal-marked (TM) covers marked (M) and terminal-unmarked (TU); M and TU each cover body-unmarked (BU); M and TU are incomparable. Its upward-closed subsets are exactly $\{\mathrm{TM}\}$, $\{\mathrm{TM}, \mathrm{M}\}$, $\{\mathrm{TM}, \mathrm{TU}\}$, $\{\mathrm{TM}, \mathrm{M}, \mathrm{TU}\}$, and the full set (any upward-closed set containing M, TU, or BU must contain TM; containing BU forces both M and TU). A policy adds a linear extension of the induced subposet: the three chain-like sets admit exactly one extension each, while $\{\mathrm{TM}, \mathrm{M}, \mathrm{TU}\}$ and the full set each admit exactly two (the only freedom is the order of the incomparable pair M, TU; TM is always first and BU always last). This gives $1 + 1 + 1 + 2 + 2 = 7$ policies, proving Theorem 1.

**Identifiers.** The frozen enumeration emits, in order, policies $P_1$–$P_7$ with self-describing identifiers of the form `accept[·]__priority[·]`:

| | Acceptance set | Priority order |
|---|---|---|
| $P_1$ | TM, M, TU, BU | TM > M > TU > BU |
| $P_2$ | TM, M, TU, BU | TM > TU > M > BU |
| $P_3$ | TM, M, TU | TM > M > TU |
| $P_4$ | TM, M, TU | TM > TU > M |
| $P_5$ | TM, M | TM > M |
| $P_6$ | TM, TU | TM > TU |
| $P_7$ | TM | TM |

**Why these axioms.** The class rests on two normative commitments. Upward closure is evidence-strength monotonicity: the relation poset orders numeral placements by how strongly the completion binds them as its answer (marker presence, terminal position), and a policy that accepted a weaker placement while rejecting a stronger one would be selectively crediting worse evidence—the classic degree of freedom behind extraction cherry-picking. Priority orders that extend the poset make disagreement resolution respect the same ordering, so no policy can prefer weaker evidence while stronger evidence is populated. Together with determinism, source preservation, and mandatory abstention on empty acceptance, these axioms admit exactly the seven policies of Theorem 1. Deliberately excluded are non-monotone acceptance sets (marker-only-unless-terminal and similar), value rewriting beyond the declared lexical forms, cross-line candidate joins, and content-dependent policy switching—the last being precisely the coupling failure the framework measures rather than permits.

**Frozen decisions.** The enumeration, its identifier list, and all 735 policy decisions on the pinned 105-case adversarial battery are frozen by digest (digests in the post-review record). Three historical extractor styles—flexible last-number extraction, marker-bound extraction, and strict terminal parsing—are proven outside the class by executed witness completions on which each historical extractor returns a value no class member can return (or fails to abstain where every member must); the class therefore bounds only its own members.

## B THE FROZEN OBSERVATION GRAMMAR

**Input boundary.** Before any lexical rule runs, a deterministic normalization maps the completion to the grammar's input domain: Unicode NFC composition; CRLF and the declared line-separator repertoire (CR, VT, FF, FS, GS, RS, NEL, U+2028, U+2029) to LF; and the declared Unicode space repertoire (NBSP, U+1680, U+2000–200A, U+202F, U+205F, U+3000) to ASCII space. All observation offsets refer to normalized text; the raw input is separately bound by a total, version-tagged hash defined for every Python string, including unpaired surrogates.

**Lexical classes.** Recognition is ASCII-only by construction: digits are `[0-9]`; sign words and number words are explicit ASCII case-variant spellings (no case-folding primitive appears in any decision path); final-answer markers are a frozen list matched case-insensitively within ASCII; fractions and thousands-grouped numerals follow explicit finite rules; sign and currency prefixes bind within one line through horizontal whitespace only. Every numeric candidate is discovered within one LF-delimited line and has a newline-free span. Every codepoint outside the declared repertoire has exactly one inert separator role: it never joins tokens, never changes a value, and never crosses a decision boundary. Structural tests forbid implicit Unicode whitespace, word, digit, case, or line-boundary primitives in any decision path, and property tests assert the two global invariants (no candidate contains a newline; undeclared codepoints never change decision vectors).

**Relations and abstention.** Each candidate receives exactly one relation label (TM, M, TU, BU as in Section 2); exclusion rules remove candidates inside quoted spans, code spans, and instruction echoes. A policy selects the highest-priority populated accepted relation and returns its unique normalized value; if that relation's candidates disagree in value, the policy abstains. Unsupported answer forms yield no candidate and hence abstention, which is reported as a diagnostic. The battery pins all 105 cases $\times$ 7 policies = 735 decisions exactly, including candidate spans and line indices.

## C EXACT SET COMPUTATION AND COMPLEXITY

Fix a comparison with items $1, \ldots, n$, exact rational weights $w_i$ summing to one, and per-policy per-item metric outcomes $b_p(i), c_p(i) \in \{0, 1\}$ for baseline and candidate.

**Computation.** $D_{\mathrm{G}} = \{\sum_i w_i(c_p(i) - b_p(i)) : p \in P\}$ is computed by direct evaluation ($|P| = 7$ weighted sums). $D_{\mathrm{I}}$ is the Minkowski sum of the per-item sets $B_i = \{w_i(c_p(i) - b_p(i)) : p \in P\}$, computed by the dynamic program $S_0 = \{0\}$, $S_i = \{s + b : s \in S_{i-1},\, b \in B_i\}$, with $D_{\mathrm{I}} = S_n$; $D_{\mathrm{P}}$ is identical with $B_i = \{w_i(c_p(i) - b_q(i)) : p, q \in P\}$.

**Correctness.** By induction on $i$: $S_i$ is exactly the set of weighted partial sums attainable by choosing one policy (respectively one policy pair) per item among the first $i$ items; item choices are unconstrained across items in $D_{\mathrm{I}}$ and $D_{\mathrm{P}}$, so the product structure factorizes into the Minkowski sum. For $D_{\mathrm{G}}$ the shared-policy constraint leaves exactly $|P|$ assignments. Containment (Theorem 2): a global assignment is the constant item-coupled assignment, and an item-coupled assignment is the diagonal arm–item assignment ($q_i = p_i$), giving $D_{\mathrm{G}} \subseteq D_{\mathrm{I}} \subseteq D_{\mathrm{P}}$ pointwise.

**When is item-coupled assignment fair?** The fairness claim for $D_{\mathrm{G}}$ is not that item-wise variation is inherently illegitimate; it is that *unconstrained* item-wise choice is an untracked degree of freedom. An item-wise assignment fixed ex ante—chosen output-blind before any completion is scored and applied identically to both arms—is simply a different single evaluator: a composite policy outside the registered class but inside the framework, whose comparison has a point effect, not a set. $D_{\mathrm{I}}$ is the envelope of all such composites, so its width measures exactly how much any post hoc item-wise choice could matter—which is why the registered design admits it only as a diagnostic. The recommendation is therefore not to forbid pre-committed item-wise designs but to require that any such assignment be frozen, output-blind, and arm-symmetric, and then reported as its own declared evaluator.

**Complexity.** The dynamic program is output-sensitive: each step costs $O(|S_{i-1}| \cdot |B_i|)$, and $|B_i| \leq |P|^2$. In the registered equal-weight settings every attainable value lies on a fixed rational grid—with $w_i = 1/n$ and binary metrics, partial sums lie in $\{k/n : |k| \leq n\}$, so $|S_i| \leq 2n + 1$ and the

total cost is $O(n^2|P|^2)$; the template-equal-weight symbolic frame is bounded the same way with denominator $T \cdot n_t$. For arbitrary rational weights the attainable-set size can grow exponentially (the problem contains subset-sum-style structure), which is why the manuscript claims polynomial time only for the registered equal-weight settings. The implementation additionally guards the state count and fails closed rather than approximating.

## D FULL REGISTERED TABLES

Bands for any primary component are its point estimate plus/minus the applicable simultaneous radius (registered joint 0.078; sensitivity radii in Section 4). Per-policy point estimates for all 22×7 primary components follow. All tables report percentage points for readability; every underlying value is an exact rational in the hash-pinned results record, and no rounding enters any registered computation.

Table 2: Complete assignment lattice for all 22 primary comparisons (percentage points; observed finite-sample descriptive diagnostics).

| Comparison | $D_{\mathrm{G}}$ (pp) | $\lvert D_{\mathrm{G}}\rvert$ | $D_{\mathrm{I}}$ (pp) | $\lvert D_{\mathrm{I}}\rvert$ | $D_{\mathrm{P}}$ (pp) | $\lvert D_{\mathrm{P}}\rvert$ |
|---|---|---|---|---|---|---|
| factorial_gemma_budget_greedy_cot | $+2.5$ | 1 | $+2.5$ | 1 | $[+2.4, +2.7]$ | 5 |
| factorial_gemma_budget_sampled_cot | $[+3.2, +3.3]$ | 2 | $[+3.0, +3.4]$ | 6 | $[+3.0, +3.4]$ | 6 |
| factorial_gemma_prompt_greedy_2048 | $[+13.9, +14.0]$ | 2 | $[+13.8, +14.2]$ | 6 | $[+13.8, +14.2]$ | 6 |
| factorial_gemma_prompt_greedy_768 | $[+11.5, +11.6]$ | 2 | $[+11.4, +11.8]$ | 6 | $[+11.4, +11.8]$ | 6 |
| factorial_gemma_prompt_sampled_2048 | $+17.7$ | 1 | $[+17.6, +17.9]$ | 5 | $[+17.6, +17.9]$ | 5 |
| factorial_gemma_prompt_sampled_768 | $[+15.1, +15.2]$ | 2 | $[+14.9, +15.3]$ | 6 | $[+14.9, +15.3]$ | 6 |
| factorial_gemma_temperature_sampled_fs | $-0.2$ | 1 | $[-0.3, -0.2]$ | 3 | $[-0.4, -0.1]$ | 5 |
| factorial_qwen_budget_greedy_cot | $+0.0$ | 1 | $+0.0$ | 1 | $[-14.4, +14.4]$ | 381 |
| factorial_qwen_budget_sampled_cot | $[-1.1, +0.8]$ | 4 | $[-13.5, +13.2]$ | 353 | $[-18.0, +17.7]$ | 473 |
| factorial_qwen_prompt_greedy_2048 | $[-25.1, -10.8]$ | 4 | $[-25.2, -10.7]$ | 192 | $[-25.2, -10.7]$ | 192 |
| factorial_qwen_prompt_greedy_768 | $[-25.1, -10.8]$ | 4 | $[-25.2, -10.7]$ | 192 | $[-25.2, -10.7]$ | 192 |
| factorial_qwen_prompt_sampled_2048 | $[-24.4, -5.6]$ | 4 | $[-24.4, -5.6]$ | 249 | $[-24.4, -5.6]$ | 249 |
| factorial_qwen_prompt_sampled_768 | $[-23.4, -6.4]$ | 4 | $[-23.4, -6.4]$ | 225 | $[-23.4, -6.4]$ | 225 |
| factorial_qwen_temperature_sampled_fs | $[+3.1, +3.6]$ | 2 | $[+3.1, +3.6]$ | 7 | $[+3.1, +3.6]$ | 7 |
| gsm8k_historical_gemma | $+18.1$ | 1 | $[+18.0, +18.3]$ | 5 | $[+18.0, +18.3]$ | 5 |
| gsm8k_historical_llama | $[+11.8, +14.9]$ | 2 | $[+11.8, +14.9]$ | 42 | $[+11.8, +14.9]$ | 42 |
| gsm8k_historical_olmo | $[+9.3, +9.9]$ | 5 | $[+9.2, +9.9]$ | 10 | $[+9.2, +9.9]$ | 10 |
| gsm8k_historical_qwen | $[-20.8, -2.5]$ | 4 | $[-21.2, -2.2]$ | 251 | $[-21.3, -2.0]$ | 255 |
| symbolic_matched_gemma | $[+16.8, +17.4]$ | 4 | $[+16.8, +17.4]$ | 4 | $[+16.8, +17.4]$ | 4 |
| symbolic_matched_llama | $[+7.6, +11.4]$ | 2 | $[+7.6, +11.4]$ | 20 | $[+7.6, +11.4]$ | 20 |
| symbolic_matched_olmo | $[+8.8, +9.6]$ | 3 | $[+8.6, +9.8]$ | 7 | $[+8.6, +9.8]$ | 7 |
| symbolic_matched_qwen | $[-18.4, -6.8]$ | 4 | $[-18.4, -6.8]$ | 59 | $[-18.4, -6.8]$ | 59 |

Table 3: Per-policy First@1 point estimates (percentage points, candidate minus baseline) for every primary comparison, in the frozen policy order $P_1$–$P_7$ of Appendix A.

| Comparison | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | $P_6$ | $P_7$ |
|---|---|---|---|---|---|---|---|
| factorial_gemma_budget_greedy_cot | +2.50 | +2.50 | +2.50 | +2.50 | +2.50 | +2.50 | +2.50 |
| factorial_gemma_budget_sampled_cot | +3.18 | +3.18 | +3.18 | +3.18 | +3.26 | +3.18 | +3.26 |
| factorial_gemma_prompt_greedy_2048 | +13.95 | +13.95 | +13.95 | +13.95 | +14.03 | +13.95 | +14.03 |
| factorial_gemma_prompt_greedy_768 | +11.52 | +11.52 | +11.52 | +11.52 | +11.60 | +11.52 | +11.60 |
| factorial_gemma_prompt_sampled_2048 | +17.74 | +17.74 | +17.74 | +17.74 | +17.74 | +17.74 | +17.74 |
| factorial_gemma_prompt_sampled_768 | +15.16 | +15.16 | +15.16 | +15.16 | +15.09 | +15.16 | +15.09 |
| factorial_gemma_temperature_sampled_fs | -0.23 | -0.23 | -0.23 | -0.23 | -0.23 | -0.23 | -0.23 |
| factorial_qwen_budget_greedy_cot | +0.00 | +0.00 | +0.00 | +0.00 | +0.00 | +0.00 | +0.00 |
| factorial_qwen_budget_sampled_cot | +0.76 | +0.76 | +0.76 | +0.76 | -0.91 | +0.61 | -1.06 |
| factorial_qwen_prompt_greedy_2048 | -10.77 | -10.77 | -10.77 | -10.77 | -24.56 | -11.30 | -25.09 |
| factorial_qwen_prompt_greedy_768 | -10.77 | -10.77 | -10.77 | -10.77 | -24.56 | -11.30 | -25.09 |
| factorial_qwen_prompt_sampled_2048 | -5.61 | -5.61 | -5.61 | -5.61 | -23.65 | -6.37 | -24.41 |
| factorial_qwen_prompt_sampled_768 | -6.37 | -6.37 | -6.37 | -6.37 | -22.74 | -6.97 | -23.35 |
| factorial_qwen_temperature_sampled_fs | +3.11 | +3.11 | +3.11 | +3.11 | +3.56 | +3.11 | +3.56 |
| gsm8k_historical_gemma | +18.12 | +18.12 | +18.12 | +18.12 | +18.12 | +18.12 | +18.12 |
| gsm8k_historical_llama | +14.86 | +14.86 | +14.86 | +14.86 | +11.75 | +14.86 | +11.75 |
| gsm8k_historical_olmo | +9.86 | +9.78 | +9.86 | +9.78 | +9.63 | +9.55 | +9.33 |
| gsm8k_historical_qwen | -2.50 | -2.50 | -2.50 | -2.50 | -20.09 | -3.26 | -20.85 |
| symbolic_matched_gemma | +17.40 | +17.40 | +17.40 | +17.40 | +17.00 | +17.20 | +16.80 |
| symbolic_matched_llama | +11.40 | +11.40 | +11.40 | +11.40 | +7.60 | +11.40 | +7.60 |
| symbolic_matched_olmo | +9.60 | +9.60 | +9.60 | +9.60 | +9.00 | +9.60 | +8.80 |
| symbolic_matched_qwen | -6.80 | -6.80 | -6.80 | -6.80 | -18.20 | -7.00 | -18.40 |

The hash-pinned diagnostic record contains the same exact counts for all 22 primary comparisons, plus parse-status and same-relation ambiguity partitions. It reads no gold answer, computes no correctness value or interval, and leaves every registered artifact and claim state untouched.

Table 4: Descriptive sampled-metric groups (Majority@16 and observed-any-correct@16 over the persisted 16-draw bundles): observed $D_{\mathrm{G}}$ hulls and finite-sample sign states. These are registered descriptive companions, never primary claims.

| Comparison | Metric | $D_{\mathrm{G}}$ (pp) | $\lvert D_{\mathrm{G}} \rvert$ | Observed state |
|---|---|---|---|---|
| factorial_gemma_budget_sampled_cot | Maj@16 | $[+6.7, +6.9]$ | 2 | positive hull |
| factorial_gemma_budget_sampled_cot | Any@16 | $[+5.5, +5.5]$ | 2 | positive hull |
| factorial_gemma_prompt_sampled_2048 | Maj@16 | $[+20.5, +20.5]$ | 2 | positive hull |
| factorial_gemma_prompt_sampled_2048 | Any@16 | $[+23.4, +23.5]$ | 2 | positive hull |
| factorial_gemma_prompt_sampled_768 | Maj@16 | $[+14.3, +14.5]$ | 2 | positive hull |
| factorial_gemma_prompt_sampled_768 | Any@16 | $+18.5$ | 1 | positive hull |
| factorial_gemma_temperature_sampled_fs | Maj@16 | $[-0.5, -0.4]$ | 2 | negative hull |
| factorial_gemma_temperature_sampled_fs | Any@16 | $-4.1$ | 1 | negative hull |
| factorial_qwen_budget_sampled_cot | Maj@16 | $[-1.8, +1.3]$ | 4 | **sign-sensitive** |
| factorial_qwen_budget_sampled_cot | Any@16 | $[-2.0, -0.2]$ | 4 | negative hull |
| factorial_qwen_prompt_sampled_2048 | Maj@16 | $[-52.9, -11.8]$ | 4 | negative hull |
| factorial_qwen_prompt_sampled_2048 | Any@16 | $[-1.5, +4.1]$ | 4 | **sign-sensitive** |
| factorial_qwen_prompt_sampled_768 | Maj@16 | $[-51.4, -12.6]$ | 4 | negative hull |
| factorial_qwen_prompt_sampled_768 | Any@16 | $[+0.5, +4.2]$ | 4 | positive hull |
| factorial_qwen_temperature_sampled_fs | Maj@16 | $[-1.6, -1.4]$ | 4 | negative hull |
| factorial_qwen_temperature_sampled_fs | Any@16 | $[-0.8, -0.8]$ | 2 | negative hull |
| gsm8k_historical_gemma | Maj@16 | $[+20.7, +20.8]$ | 3 | positive hull |
| gsm8k_historical_gemma | Any@16 | $[+19.9, +19.9]$ | 2 | positive hull |
| gsm8k_historical_llama | Maj@16 | $[+13.7, +18.5]$ | 2 | positive hull |
| gsm8k_historical_llama | Any@16 | $[+10.3, +11.0]$ | 2 | positive hull |
| gsm8k_historical_olmo | Maj@16 | $[+7.6, +8.3]$ | 4 | positive hull |
| gsm8k_historical_olmo | Any@16 | $[+2.7, +2.9]$ | 3 | positive hull |
| gsm8k_historical_qwen | Maj@16 | $[-54.4, -13.3]$ | 4 | negative hull |
| gsm8k_historical_qwen | Any@16 | $[-2.3, +3.3]$ | 4 | **sign-sensitive** |
| symbolic_matched_gemma | Maj@16 | $[+19.6, +20.0]$ | 3 | positive hull |
| symbolic_matched_gemma | Any@16 | $[+23.4, +23.6]$ | 2 | positive hull |
| symbolic_matched_llama | Maj@16 | $[+9.6, +16.8]$ | 2 | positive hull |
| symbolic_matched_llama | Any@16 | $[+9.8, +15.2]$ | 2 | positive hull |
| symbolic_matched_olmo | Maj@16 | $[+6.4, +7.2]$ | 4 | positive hull |
| symbolic_matched_olmo | Any@16 | $[+6.2, +6.6]$ | 2 | positive hull |
| symbolic_matched_qwen | Maj@16 | $[-44.6, -10.8]$ | 4 | negative hull |
| symbolic_matched_qwen | Any@16 | $[+1.0, +7.2]$ | 3 | positive hull |

Table 5: Gold-blind post-registered First@1 placement diagnostic for the eight matched protocol comparisons. Each relation cell is baseline → candidate presence (%) among unclipped first completions; $n_B/n_C$ gives those exact denominators. Relations are nonexclusive. "N" is a literal terminal metavariable; Clip uses all first completions. The evidence is descriptive and from the disclosed development cohort.

| Comparison | $n_B/n_C$ | Terminal-marked | Marked | Terminal-unmarked | Body-unmarked | Literal N | Clip |
|---|---|---|---|---|---|---|---|
| GSM8K Gemma | 1272/1232 | 97.6→95.5 | 0.4→19.8 | 0.9→1.1 | 98.9→99.7 | 1.8→1.2 | 3.6→6.6 |
| GSM8K Llama | 1318/1314 | 97.0→88.8 | 0.0→0.1 | 2.0→9.9 | 99.9→100.0 | 2.0→1.4 | 0.1→0.4 |
| GSM8K OLMo | 1228/1307 | 99.8→96.7 | 0.8→5.2 | 11.5→6.9 | 88.8→97.2 | 0.2→0.2 | 6.9→0.9 |
| GSM8K Qwen | 1319/1319 | 98.6→36.7 | 0.3→1.8 | 4.0→31.2 | 98.6→95.1 | 0.1→22.6 | 0.0→0.0 |
| Symbolic Gemma | 493/471 | 98.2→95.1 | 2.4→22.3 | 1.4→1.1 | 97.8→99.6 | 1.4→1.9 | 1.4→5.8 |
| Symbolic Llama | 499/497 | 96.4→80.3 | 0.0→0.2 | 1.6→17.7 | 100.0→100.0 | 2.8→1.8 | 0.2→0.6 |
| Symbolic OLMo | 481/496 | 97.3→92.5 | 4.4→4.4 | 18.1→7.9 | 83.4→96.6 | 0.0→2.8 | 3.8→0.8 |
| Symbolic Qwen | 500/500 | 99.2→36.6 | 0.0→1.2 | 2.2→24.8 | 100.0→96.4 | 0.0→24.6 | 0.0→0.0 |

# E ADVERSARIAL REVIEW CHRONOLOGY

Every design surface was reviewed in fresh-context sessions, cross-assigned so a reviewer examines material it did not draft, under a no-rerolling rule with every adverse report preserved verbatim by digest. Gates were only ever tightened. The blinded chronology (products, digests, and raw reports appear in the post-review record):

| Stage | Rounds | Adverse findings | Outcome |
|---|---|---|---|
| Initial design (two role reviews) | 2 | 42 | all accepted; fix wave |
| Targeted design re-reviews | 8 | 4 fatal-class, 23 major, 4 minor | 7 reject, 1 pass |
| Cross-family and confirmation | 5 | 16 major, 8 minor | 4 reject, then consolidation |
| Finite-grammar closure and gate | 2 | 1 structural | pass; design frozen |
| Registered-run builder | 3 | 5 major, 1 minor | 2 reject, then pass |
| Authority/scorer execution gate | 1 | 0 (7 notes) | pass; single execution |
| Dependence sensitivity | 2 | 2 major, 3 other | 1 reject, then pass |
| Held-out confirmation gates | 2 | 2 launch defects | closed, then pass |

Two stop-rule events are part of this record: a novelty search triggered the prespecified stop rule on two adjacent works (both cited and distinguished in Section 6) and was adjudicated as continue-with-narrowed-claim; and the prespecified clipping validity gate suppressed five comparisons involving the drafting authors' own model family, a consequence accepted without repair (Section 4).

## F Development-cohort provenance

The scored corpus comprises 31 frozen evaluation runs of four public instruction-tuned checkpoints (Gemma-3 1B, and 1–2B-class Qwen, Llama, and OLMo instruct models, each at a pinned revision) over the full GSM8K test split (1319 problems) and the GSM-Symbolic release (100 templates $\times$ 5 instances, evaluation-only under its license). Each run persists its configuration snapshot, per-item predictions (byte-hashed), and metrics. Decode settings per arm (prompt variant, token budget, temperature, nucleus threshold, draw count, and the effective batch/seed scheme) are bound into per-comparison conditioning digests. The 22 registered comparisons span three families: four historical old-versus-corrected protocol comparisons, four symbolic-matched transfers, and fourteen prompt/budget/temperature factorial cells; every comparison covers its full frame with no unit subsetting. All predictions predate the registered design and were previously analyzed under different, unregistered evaluators; they are disclosed as development data throughout.

## G Held-out GSM1k confirmation

The confirmation cohort is the full 1,205-problem test release of GSM1k (Zhang et al., 2024a) at a pinned upstream revision, bound by raw file digest, canonical local digest, manifest, and audit before any generation; it is used under private-evaluation terms (the dataset card omits an explicit license), no record is redistributed, and no contamination-free claim is made. The generation grid is eight cells of one pinned Qwen2.5-1.5B-Instruct checkpoint—{few-shot concise (three pinned exemplars), zero-shot CoT} $\times$ {768, 2048 tokens} $\times$ {greedy, sampled $N{=}16$}—frozen with the prompt builders, decode controls, and seed before launch; partial-grid interpretation and post-access prompt or policy tuning are forbidden by the frozen protocol. The seven-policy class, its identifier digests, and the 105-case battery carry over from the development design unchanged and are disclosed as development-informed. Generation ran once, uninterrupted, on a single authorized device; the registered analysis (a paired-problem bootstrap with 20,000 draws over one 56-component First@1 family, the same $1/10$ clipping gate, and prespecified replication tests) executed once behind an exact-digest authority. Every registered cell is reported; Table 6 gives the complete exact sets.

## H Grammar-variant sensitivity

A post-registered, CPU-only diagnostic rescored every persisted primary arm through the frozen seven-policy class under three frozen marker edits—one per variant, each a minimal perturbation of the registered marker list—and recomputed every observed $D_{\mathrm{G}}$ exactly; the unpatched baseline must reproduce the registered exact sets or the run fails (it reproduced all 22). Exposure is reported with every claim: a variant that edits an unexposed channel cannot test robustness. Deleting the plain 'answer' marker—the one substantively exposed channel—changed 7 previously non-sign-sensitive Qwen comparisons to observed sign-sensitive (historical Qwen moves from its registered all-negative hull to [-3.0, +4.9] percentage points), left the sampled-budget comparison sign-sensitive, and left the greedy-budget comparison exactly null; every non-Qwen comparison retained

Table 6: Held-out GSM1k confirmation: complete assignment lattice and simultaneous claim states for all 8 First@1 comparisons (percentage points, candidate minus baseline; prompt contrasts are zero-shot CoT minus few-shot concise, budget contrasts 2048 minus 768 tokens). Sign states are calibrated to the joint 56-component band (radius 0.0423).

| Comparison | $D_{\mathrm{G}}$ (pp) | $\lvert D_{\mathrm{G}}\rvert$ | $D_{\mathrm{I}}$ (pp) | $\lvert D_{\mathrm{I}}\rvert$ | $D_{\mathrm{P}}$ (pp) | $\lvert D_{\mathrm{P}}\rvert$ | Claim state |
|---|---|---|---|---|---|---|---|
| Prompt (greedy, 768) | $[-22.0, -8.3]$ | 4 | $[-22.0, -8.3]$ | 166 | $[-22.0, -8.3]$ | 166 | negative |
| Prompt (greedy, 2048) | $[-22.0, -8.3]$ | 4 | $[-22.0, -8.3]$ | 166 | $[-22.0, -8.3]$ | 166 | negative |
| Prompt (sampled, 768) | $[-16.0, -0.9]$ | 4 | $[-16.3, -0.7]$ | 189 | $[-16.3, -0.7]$ | 189 | not resolved |
| Prompt (sampled, 2048) | $[-15.9, -0.7]$ | 4 | $[-16.1, -0.5]$ | 189 | $[-16.1, -0.5]$ | 189 | not resolved |
| Budget (greedy FS) | $+0.0$ | 1 | $+0.0$ | 1 | $+0.0$ | 1 | not resolved |
| Budget (greedy CoT) | $+0.0$ | 1 | $+0.0$ | 1 | $[-13.7, +13.7]$ | 331 | not resolved |
| Budget (sampled FS) | $+0.0$ | 1 | $+0.0$ | 1 | $[-0.2, +0.2]$ | 7 | not resolved |
| Budget (sampled CoT) | $+0.2$ | 1 | $+0.2$ | 1 | $[-15.2, +15.5]$ | 371 | not resolved |

its observed sign category, although hull endpoints moved in 20 of 22 comparisons (maximum endpoint shift 17.9 percentage points). The other two edits touch an absent and a nearly-unexposed channel (0 and 1 eligible candidates) and leave every set bit-identical—expected by exposure, not robustness evidence. This diagnostic probes three marker edits, not broader lexical or grammar robustness; all results are observed finite-sample facts with no inferential content, and no registered claim changes.

Table 7: Marker-edit sensitivity of the observed $D_{\mathrm{G}}$ sets (development cohort; post-registered descriptive diagnostic). "Exposure" counts eligible candidates of the edited marker kind in unclipped first completions; "output changes" counts (arm, unit, policy) evaluator outputs—status and selected value—on first completions that differ from the unpatched baseline.

| Variant | Exposure | Output changes | Hulls moved | States changed | Max shift (pp) |
|---|---|---|---|---|---|
| drop plain 'answer' marker | 27,876 | 54,516 | 20 | 7 | 17.9 |
| drop #### marker | 0 | 0 | 0 | 0 | 0.0 |
| add 'result' marker | 1 | 0 | 0 | 0 | 0.0 |

# I PRACTITIONER CHECKLIST

Adopting the reporting discipline does not require this paper's full apparatus. For a free-response comparison scored by deterministic extraction:

1. *Freeze the evaluator before scoring.* Fix the extraction policies (or a small declared class) and the observation grammar before any completion is scored, and version them with the results.
2. *Declare the assignment discipline.* State whether one policy scores everything (global), varies by item while shared across arms (item-coupled), or varies per arm–item pair. Use global coupling for primary claims; if an item-wise design is scientifically necessary, fix it ex ante, output-blind, and identically across arms, and report it as its own declared evaluator (Appendix C).
3. *Report the attainable hull, not only a point.* With a finite policy set, $D_{\mathrm{G}}$ costs one scoring pass per policy; report its hull beside the headline delta and flag any comparison whose hull contains zero or both signs.
4. *Treat relaxations as diagnostics.* Compute $D_{\mathrm{I}}/D_{\mathrm{P}}$ only to quantify what weaker discipline could manufacture, never to rescue a conclusion.
5. *Report abstention.* Unsupported answer forms should abstain visibly rather than score silently; abstention rates are part of the result.
6. *Disclose all of the above* with the benchmark version and decode settings, so the comparison is reproducible at the scoring layer.

## J CONTEXTMATH AUDIT DETAILS

Table 8: Per-cell clipping and answer-marking counts for the ContextMATH regime map (original/scenario arm, out of 193). "Ends with marked answer" counts completions whose final answer carries the declared marker.

| Model | Clipped | Ends with marked answer | $D_{\mathrm{G}}$ (pp) |
|---|---|---|---|
| Qwen2.5-7B | 0/1 | 76/114 | $[-13.0, +12.4]$ |
| Qwen3-4B-2507 | 5/8 | 187/182 | $\{-10.9\}$ |
| Qwen2.5-32B | 1/0 | 96/147 | $[-14.5, +8.8]$ |
| Qwen3.8-27B | 4/9 | 188/183 | $\{-8.8\}$ |
| Gemma-4-E4B | 0/0 | 191/189 | $\{-15.5\}$ |
| Gemma-3-4B (repaired run) | 7/12 | 183/176 | $\{-17.1\}$ |
| Llama-3.2-3B | 17/20 | — | $(-15.0$, descriptive) |
| Gemma-3-4B | all/all | — | — |

The Section 5 frame is the exact-rational level-4/5 stratum of the ContextMATH (Cao et al., 2026) MATH-500 scenario pairs at a pinned revision: 193 ordered pairs, identical gold answers in both arms, one shared zero-shot chain-of-thought protocol ending "The answer is N." (greedy, 8,192-token budget, one seed), the same seven-rule class, observation grammar, clipping convention, and one-in-ten clipping gate as the registered design, and one sup-norm bootstrap band per cell. The first two cells were frozen before generation; the extension cells are post-registered under the same frozen design and are labeled in Table 1, which carries the per-cell attainable sets, clip counts, and terminal-marking counts.

Two preregistered predictions failed and are reported as such: the Qwen2.5-7B cell predicted a one-directional loss and returned a sign-sensitive set; the Qwen2.5-32B cell predicted a point certificate or a narrow one-signed effect and returned a sign-sensitive set spanning $-14.5$ to $+8.8$ points. For that 32B cell the assignment relaxations widen the shared-rule set of 4 values (width 23.3 points) to 62 values (width 31.6) under per-question choice and 98 values (width 50.3) under free arm–item assignment—the containment ordering of Section 2 at working scale. The Qwen3-4B cell completed under a disclosed pre-outcome operational amendment that removed only its GPU-time ceiling; the amendment predates any outcome exposure and changed no other frozen field. The one suppressed cell is a clipping-gate exceedance (Llama-3.2-3B, scenario arm 20 of 193 clipped; descriptive point effect $-15.0$ points). The Gemma-3-4B result comes from a repaired run: a first run of the same cell with an empty stop set clipped every completion and is retired as a technical failure, not a model result. Generation engines are declared per cell (local bf16 transformers harness for the 4–7B cells; a pinned vLLM engine for the 32B and 27B cells); prompts, decoding, and stop handling are identical across cells.

## K PRIMARY COMPARISON TABLE

## L CROSS-FRAME-DEPENDENCE SENSITIVITY

**Cross-frame-dependence sensitivity (post-registered).** The joint primary family resamples the two frames independently, while the symbolic templates derive from GSM8K source problems (Section 7). Two post-registered sensitivity constructions checked whether this matters; neither lost any registered resolution. A framewise 97.5% Bonferroni construction—valid without a cross-frame independence assumption whenever both framewise bands attain nominal coverage—changed the eligible classification from 6/2/9 to 6/4/7, resolving 2 additional Qwen comparisons as negative without overturning any registered resolution. A complementary source-linked cluster-ratio diagnostic—one frozen resampling construction in which each symbolic template enters with its pinned source problem's draw multiplicity—reproduced 6/2/9 exactly, with radius 0.0768 versus 0.0780 registered. The 5 suppressed comparisons remain suppressed under every construction. Relative to the two checked constructions, the registered classification is conservative: no registered resolution is overturned, and the only differences are additional resolutions.

Table 9: All 22 registered primary (First@1) comparisons. $|D_{\mathrm{G}}|$ is the number of distinct attainable effects under globally coupled evaluator-policy assignment across the seven registered policies; the hull gives the extreme attainable effects in percentage points (candidate minus baseline; the underlying values are exact rationals, preserved in the hash-pinned record). Sign states are calibrated to the joint simultaneous band (154 components, sup-norm radius 0.078 at confidence 0.95). Development cohort.

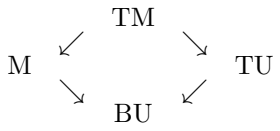
| Comparison | Frame | $\lvert D_{\mathrm{G}}\rvert$ | $D_{\mathrm{G}}$ hull | Registered sign state |
|---|---|---|---|---|
| Historical: Gemma | GSM8K | 1 | $+18.1$ | positive |
| Historical: Llama | GSM8K | 2 | $[+11.8, +14.9]$ | positive |
| Historical: OLMo | GSM8K | 5 | $[+9.3, +9.9]$ | positive |
| Historical: Qwen | GSM8K | 4 | $[-20.8, -2.5]$ | unresolved |
| Symbolic: Gemma | GSM-Sym. | 4 | $[+16.8, +17.4]$ | positive |
| Symbolic: Llama | GSM-Sym. | 2 | $[+7.6, +11.4]$ | unresolved |
| Symbolic: OLMo | GSM-Sym. | 3 | $[+8.8, +9.6]$ | positive |
| Symbolic: Qwen | GSM-Sym. | 4 | $[-18.4, -6.8]$ | unresolved |
| Gemma prompt (greedy, 768) | GSM8K | 2 | $[+11.5, +11.6]$ | suppressed |
| Gemma prompt (greedy, 2048) | GSM8K | 2 | $[+13.9, +14.0]$ | suppressed |
| Gemma prompt (sampled, 768) | GSM8K | 2 | $[+15.1, +15.2]$ | suppressed |
| Gemma prompt (sampled, 2048) | GSM8K | 1 | $+17.7$ | positive |
| Gemma budget (greedy CoT) | GSM8K | 1 | $+2.5$ | suppressed |
| Gemma budget (sampled CoT) | GSM8K | 2 | $[+3.2, +3.3]$ | suppressed |
| Gemma temperature (sampled FS) | GSM8K | 1 | $-0.2$ | unresolved |
| Qwen prompt (greedy, 768) | GSM8K | 4 | $[-25.1, -10.8]$ | negative |
| Qwen prompt (greedy, 2048) | GSM8K | 4 | $[-25.1, -10.8]$ | negative |
| Qwen prompt (sampled, 768) | GSM8K | 4 | $[-23.4, -6.4]$ | unresolved |
| Qwen prompt (sampled, 2048) | GSM8K | 4 | $[-24.4, -5.6]$ | unresolved |
| Qwen budget (greedy CoT) | GSM8K | 1 | $+0.0$ | unresolved |
| Qwen budget (sampled CoT) | GSM8K | 4 | $[-1.1, +0.8]$ | unresolved |
| Qwen temperature (sampled FS) | GSM8K | 2 | $[+3.1, +3.6]$ | unresolved |

## M REGISTERED-DESIGN DETAIL

**A finite, frozen observation grammar.** The evaluator is a *registered conservative ASCII-numeric evaluator*, not a universal parser. A deterministic input boundary normalizes completions (Unicode NFC; all declared line separators to LF; declared Unicode spaces to ASCII space) before any lexical rule runs; recognition is then ASCII-only, every numeric candidate is discovered within one LF-delimited line with a newline-free span, every undeclared codepoint has exactly one inert separator role, and raw-input hashing is total over every Python string. Unsupported answer forms yield no candidate and hence reported abstention rather than a silent score. Structural tests forbid implicit Unicode primitives in any decision path; a pinned battery of 105 adversarial cases fixes all 735 policy decisions exactly.

**Preregistration with a fail-closed authority.** The preregistration (policy axioms, estimands, weighting, multiplicity family, validity gates, claim-language guard) is canonicalized and hash-pinned in code. A separate code-pinned authority file gates every scored-table code path behind three independent authorization bits; until an author-approved commit flips them, all public analysis APIs reject any caller-supplied freeze record. The component manifest binding every comparison—system identities, checkpoint revisions, decode-control digests, prediction-file hashes, ordered unit frames—was built and validated *output blind*: predictions entered only as byte streams for hashing, never parsed, and the builder hard-fails on any provenance contradiction.

**Adversarial review before results.** The design was frozen only after a multi-round fresh-context review protocol: every review ran in a fresh context on hash-named artifacts, cross-assigned so a reviewer examines material it did not draft; every adverse report was preserved verbatim under SHA-256 with no rerolling; gates were only ever tightened. The evaluator grammar survived six




**Synthetic worked example.**

```
A: We compute 20.
   The answer is N.
B: We compute 20.
   The answer is 20.
```

In A, the eligible numeral is body-unmarked and the last line is a literal metavariable, so only policies accepting the bottom relation can select 20. In B, 20 also occupies the top relation, so every registered policy selects it. The policies neither infer that N means 20 nor generate a replacement.

Figure 3: The frozen relation diamond and a completion-only illustration. Abbreviations take the relation-name initials. The example is explanatory, not an empirical observation.

review rounds, the registered-run machinery a further four. Scoring the frozen predictions and executing the registered analysis were separately author-authorized against the exact artifact digests, and neither implementer nor reviewers computed or inspected any comparison aggregate before the final gate passed. Scrubbed summaries of the review chronology with blinded artifact identifiers ship in the anonymized supplementary material; the complete hash-linked record, including every adverse review verbatim, is released upon de-anonymization.

## N THE RELATION DIAMOND

## O POST-REGISTERED READER DIAGNOSTICS

After the registered analyses closed, one additional deterministic reader outside the seven-rule class was declared and then applied once, symmetrically, to the frozen ContextMATH predictions (every prediction file re-hashed against its output-blind manifest, and every manifest and results artifact authenticated against its authority chain, before parsing). The *LaTeX-aware terminal reader* takes the completion's terminal segment (joining a trailing display-math block across lines), reads the last `\boxed{...}` there, falls back to the declared marker line, normalizes simple LaTeX (fractions, spacing commands, text wrappers), and abstains otherwise. Directional predictions were recorded before execution; one failed and is reported below.

Table 10: The LaTeX-aware reader on the frozen ContextMATH predictions (correct counts per arm out of 193; observed point effects, descriptive only). It selected negative observed effects for both sensitive Qwen2.5 cells and exactly reproduced the three registered point certificates tested (the repaired Gemma-3-4B run was not included); on the 32B cell its effect lies one problem (0.52 points) below the seven-rule hull.

| Cell | Orig. correct | Scen. correct | Effect (pp) | vs. seven-rule $D_{\mathrm{G}}$ |
|---|---|---|---|---|
| Qwen2.5-7B | 118 | 97 | $-10.9$ | inside hull |
| Qwen3-4B-2507 | 176 | 155 | $-10.9$ | equals the singleton |
| Qwen2.5-32B | 148 | 119 | $-15.0$ | 0.52 pp below hull |
| Qwen3.8-27B | 187 | 170 | $-8.8$ | equals the singleton |
| Gemma-4-E4B | 177 | 147 | $-15.5$ | equals the singleton |
| Llama-3.2-3B | 64 | 35 | $-15.0$ | equals the descriptive point |

Prediction scorecard: the certificate cells and the 32B directional prediction were confirmed; the 7B prediction (that the reader would land at or below the hull's negative end) failed—the reader landed inside the hull on the negative side—and is reported as failed. These are observed point effects of one declared reader: they do not inferentially resolve either sensitive comparison, and the certificates' exact reproduction is evidence under this one richer determinism, not general certificate generalization. A precommitted, arm-symmetric item-wise composite policy (a hash-assigned implementation check whose item-coupled-set membership holds by construction) ships in the supplementary material only.

## P UPSTREAM EVALUATOR CROSSWALK

After the registered analyses and the reader diagnostics closed, four public evaluator implementations were pinned by release or commit and applied once, symmetrically, to the same frozen outputs (every prediction file re-hashed against its manifest; every manifest and results artifact authenticated against its authority chain): the lm-evaluation-harness GSM8K `strict-match` and `flexible-extract` filters (release v0.4.12; regex, first and last match respectively, with the harness's exact-match normalization), the published Qwen GSM8K rule (last digit sequence in the completion, integer equality with the reference), and Math-Verify (v0.9.0; LaTeX and expression extraction with symbolic equivalence, library defaults). The pinned identifiers are recorded in the supplement. Effects are candidate minus baseline under the registered clipping convention.

`strict-match` looks for a `####` marker that our protocols never emit: it fails to extract on at least 96.9% of completions in every arm, so its zero deltas are coverage failures, not effect estimates, and it is excluded from the table. The three extracting readers agree on the binary correctness decision for 97.1% of development and GSM1k completions and 86.0% of ContextMATH completions.

Table 11: Public evaluators on the Qwen comparisons whose registered seven-rule hull is entirely negative (effects in percentage points). Every one of the 10 comparisons (6 development, 4 held-out GSM1k) is positive under all three extracting public readers. Math-Verify's deltas align with the number-recovering readers on these outputs; Table 12 decomposes the difference.

| Comparison | Seven-rule hull | flexible-extract | Qwen rule | Math-Verify |
|---|---|---|---|---|
| Qwen prompt (greedy, 2048) | $[-25.1, -10.8]$ | $+13.3$ | $+12.1$ | $+11.7$ |
| Qwen prompt (greedy, 768) | $[-25.1, -10.8]$ | $+13.3$ | $+12.1$ | $+11.7$ |
| Qwen prompt (sampled, 2048) | $[-24.4, -5.6]$ | $+9.8$ | $+8.9$ | $+9.4$ |
| Qwen prompt (sampled, 768) | $[-23.4, -6.4]$ | $+11.1$ | $+9.9$ | $+10.5$ |
| Historical: Qwen | $[-20.8, -2.5]$ | $+12.9$ | $+12.4$ | $+12.4$ |
| Symbolic: Qwen | $[-18.4, -6.8]$ | $+12.0$ | $+11.2$ | $+9.8$ |
| GSM1k prompt (greedy, 2048) | $[-22.0, -8.3]$ | $+15.1$ | $+14.0$ | $+12.9$ |
| GSM1k prompt (greedy, 768) | $[-22.0, -8.3]$ | $+15.0$ | $+13.9$ | $+12.8$ |
| GSM1k prompt (sampled, 2048) | $[-15.9, -0.7]$ | $+14.9$ | $+14.2$ | $+12.9$ |
| GSM1k prompt (sampled, 768) | $[-16.0, -0.9]$ | $+14.7$ | $+14.0$ | $+12.6$ |

On the non-Qwen development comparisons the public readers land inside or within a few points of the registered hulls; the greedy-budget control is exactly zero under all four readers in the development cohort, and on GSM1k the greedy few-shot budget contrast is exactly zero while the greedy chain-of-thought contrast is one item ($+0.08$ points) under the three extracting readers. On ContextMATH, Math-Verify—the on-label public evaluator for exact-rational answers—gives $-10.9$ and $-14.5$ points on the two sensitive Qwen2.5 cells (the negative side of the hull; the 32B value is the hull's endpoint), $-10.9$ and $-15.5$ on the Qwen3-4B and Gemma-4-E4B certificates (exactly the registered singletons), and $-9.3$ on Qwen3.8-27B (one item from its singleton). The integer-only readers are off-label on that frame and are reported in the supplement only. The public readers therefore instantiate one side of each sensitive comparison rather than resolving it, and the pinned Qwen script is a published official implementation, not a verified reconstruction of any specific model card's pipeline. Two ex-ante predictions failed and are reported in the supplement: Math-Verify was not within two points of the registered hull on the Qwen prompt comparisons, and the extracting public readers agreed with one another on the registered sign-sensitive comparisons.

**Credit decomposition.** To locate the disagreement, every first completion was classified gold-blind by the frozen observation grammar into: ends with the literal unfilled placeholder; has an eligible terminal numeric candidate (a number on its final line, marked or unmarked); carries numbers only inside the reasoning; or has no numeric candidate. Each public reader's effect was then split into the part carried by credits to completions with no eligible terminal candidate and the remainder (Table 12). On every reversed comparison and for all three readers, the remainder is negative—the sign of the seven-rule class—so the reversal is carried entirely by credit to completions that do not deliver a terminal answer under the grammar: 17–26% of chain-of-thought completions on these comparisons receive such credit, against at most 1.0% on the non-Qwen comparisons and the ContextMATH certificate cells tested. This is a structural category, not a judgment that the completion

presented no answer: inspected cases include completions that state the result in prose immediately before an unfilled placeholder (“. . . the original price was $26. The answer is N.”), which the recovery readers credit and the seven rules do not. The two readings are different measurement contracts—one abstains unless the response identifies a unique answer at a prespecified evidence level, the other recovers an answer from the reasoning trace using positional or symbolic rules—and the measured intervention effect changes sign between them. The framework declares, compares, and audits such contracts; it does not crown one as universally correct. On the Qwen historical chain-of-thought arm the class credits 562 completions and Math-Verify 760, with 557 in common; of Math-Verify’s 203 class-uncredited completions, 142 end with the placeholder, 56 carry the number only in the reasoning, and 5 have an eligible terminal candidate that the class read differently.

Table 12: Decomposition of the public readers’ effects on the reversed Qwen comparisons (percentage points). “No-terminal credit” is the differential credit to completions with no eligible terminal numeric candidate under the frozen grammar; “remainder” is the reader’s effect with that credit removed. Math-Verify columns give its effect and remainder.

| Comparison | flexible | no-terminal credit | remainder | Math-Verify | remainder |
|---|---|---|---|---|---|
| Qwen prompt (greedy, 2048) | $+13.3$ | $+24.6$ | $-11.3$ | $+11.7$ | $-10.3$ |
| Qwen prompt (greedy, 768) | $+13.3$ | $+24.6$ | $-11.3$ | $+11.7$ | $-10.3$ |
| Qwen prompt (sampled, 2048) | $+9.8$ | $+16.7$ | $-6.9$ | $+9.4$ | $-5.6$ |
| Qwen prompt (sampled, 768) | $+11.1$ | $+17.9$ | $-6.7$ | $+10.5$ | $-5.9$ |
| Historical: Qwen | $+12.9$ | $+16.6$ | $-3.7$ | $+12.4$ | $-2.5$ |
| Symbolic: Qwen | $+12.0$ | $+18.2$ | $-6.2$ | $+9.8$ | $-6.8$ |
| GSM1k prompt (greedy, 2048) | $+15.1$ | $+23.7$ | $-8.5$ | $+12.9$ | $-8.3$ |
| GSM1k prompt (greedy, 768) | $+15.0$ | $+23.6$ | $-8.5$ | $+12.8$ | $-8.3$ |
| GSM1k prompt (sampled, 2048) | $+14.9$ | $+16.3$ | $-1.4$ | $+12.9$ | $-1.3$ |
| GSM1k prompt (sampled, 768) | $+14.7$ | $+16.3$ | $-1.6$ | $+12.6$ | $-1.5$ |