WHEN GRADING IS NOT MODE-INVARIANT: ANSWER-EXTRACTION AND NORMALIZATION ERRORS IN THINK/NO-THINK COMPARISONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Choosing whether to use an explicit thinking phase requires a reliable estimate of its reported (grader-measured) accuracy gain. Yet on identical MMLU outputs, a vendored Qwen multiple-choice extractor applied to full transcripts reports a think-minus-no-think gap of -15.7 points, while two other mainstream extractors report $+17.0$ to $+17.3$; the disagreement persists on the complete MMLU test set. Scoring only the answer segment flips the MMLU gap's sign; one GSM8K extractor's failure persists under the same change. We audit 200,328 Qwen3-8B candidates from five benchmark-derived item sets, using 595 stratified reference labels (LLM-prescreened, human-checked) and separate follow-up and transfer studies. Source-code interventions identify repeated answer-trigger phrases that, in affected outputs, truncate the transcript before final-answer extraction; on 240 Qwen candidates of a separately sampled MMLU follow-up, we evaluate the answer-segment rule whose population verdicts were recorded before that draw. Against the returned question-answer key, the repair removes 47 reference-grading errors and introduces four among the 51 sampled think verdicts that change (51 and zero against benchmark gold). The estimated population error reduction for the 2,560 think outputs has a conditional 95% interval $[20.1, 43.9]$ points; no-think verdicts are unchanged. This does not establish the semantic accuracy ranking. On GSM8K, a last-number extractor's reversal survives the same repair. An exact error decomposition relates mode-specific errors to the reported gap; conditional bounds specify when the evidence supports a comparison; agreement alone cannot certify validity. These results support checking repairs against reference labels and reporting the remaining uncertainty; inference concerns fixed corpora and reference conventions, without adjustment for historical search or undetected label errors.

1 INTRODUCTION

An evaluator choosing between a language model's *think* and *no-think* modes needs to know whether extra computation buys an accuracy gain (Yang et al., 2025; Gan et al., 2026; Sui et al., 2025), whether a reported gain survives grading error, and whether a repair reduces it. A change in answer style can change how a grader (the parser or judge mapping text to a verdict) measures correctness. On one fixed set of MMLU candidates, the vendored Qwen multiple-choice extractor `qwen_choice`, applied to full transcripts, reports a think-minus-no-think gap of -15.7 points, whereas other mainstream graders on the same outputs report $+17.0$ to $+17.3$ (Figure 1). The outputs are unchanged; the scoring rule differs — an *observed-score reversal* of the extraction contract, not the developer's complete pipeline or published scores.

Think/no-think holds model weights fixed while changing the output distribution — a setting for measuring differential error. In `qwen_choice`, repeated "choice is"/"answer is" phrases can truncate a trace before trigger/fallback parsing; in lm-eval flexible, a last-number rule can select a later numeric or punctuation-only match: branches shift with output mode, so we measure mode-conditional error.

Three common checks are insufficient: high pooled accuracy does not protect the direction of a small gap, repository provenance does not certify behavior on these outputs, and grader agreement does not

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.

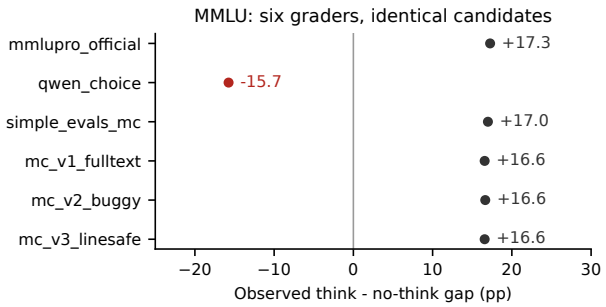

Figure 1: Census observed think-minus-no-think gaps for six graders on identical MMLU outputs: the negative `qwen_choice` gap against five positive alternatives.

certify validity (Proposition 1). Public evaluations have fed reasoning traces to graders and reported score discrepancies and suspected extraction failures (Hugging Face, 2025a; Padmanabhan & Goyal, 2026; EleutherAI, 2026) (Appendix D.21).

We audit 200,328 Qwen3-8B candidates from five benchmark-derived item sets (two modes, 1,473 fixed items). An eight-grader mainstream panel and a seven-grader in-house mechanism panel (plus one mechanism probe) score the same outputs; 995 labeling records support the primary model (595 feed the primary estimate, 400 the semantic extension), and 6,100 separate records support five transfer models (Section 3 and Appendix D.7); we ask whether error depends on mode and which repairs are supported by reference evidence. A separate MMLU follow-up tests answer-segment extraction on 240 sampled records: against the returned question key, it reduces reference-grading errors with conditional 95% interval $[20.1, 43.9]$ points over all 2,560 think outputs (Appendix D.3), while both follow-up accuracy-difference intervals still span zero (Table 6) — a useful repair and a resolved mode ranking are separate outcomes.

Contributions and scope. (1) *Fixed-output evidence*: standard graders give opposite observed think/no-think directions on identical outputs; mode-conditional FP/FN channels quantify the error behind the flips (Section 5). (2) *Mechanisms and evaluated repairs*: source-code rules and surface rewrites locate deterministic extraction failures, and a paired reference audit measures repaired and newly introduced errors (Appendices D.3 and D.14). (3) *Conditional comparison rules*: an error decomposition and reversal boundary identify what an observed gap can support; an agreement impossibility result and a bounded-overlap content bound separate information from assumptions (Sections 3.2, 3.4, 4.1 and 4.2 and Appendix B.2). The accompanying count-based intervals cover repairs and task gaps; a design-based certification rule abstains familywise on every cell, and a direct count inversion certifies per-cell directions where the counts allow (Tables 18 and 19). The primary audit covers one model and five fixed item sets; five transfer corpora, a new-generation experiment, a complete-test-set comparison without additional reference annotations (Appendix D.9), and a separate extraction follow-up are reported with their own designs (Section 5.4 and Appendices D.2 and D.5). Reference-label estimates are conditional on the 1,473 selected items and the benchmark-gold convention; follow-ups keep their own reference keys.

2 Related Work

Measurement error. Measurement invariance is a classical condition for comparing groups (Meredith, 1993; Millsap, 2012; Holland & Wainer, 1993): when it holds in a positively oriented channel, score gaps transmit accuracy gaps (attenuated, never reversed); when it fails, comparisons call for correction or conditional error bounds. Latent-class methods estimate judge misclassification without any gold set, under structural assumptions (Hui & Walter, 1980; Albert & Dodd, 2004; Lash et al., 2014); prediction-powered methods propagate machine-judge error against a labeled reference target (Angelopoulos et al., 2023; Saad-Falcon et al., 2023; Chen et al., 2026; Shen et al., 2026). Dorner et al. (Dorner et al., 2025) model judge rates as functions of the evaluated system; Chen et al. (Chen et al., 2026) compare several corrections. These frameworks can define groups by generation mode, but they do not study a paired think/no-think extractor setting. Stratified judge-error estimation

already exists (Fisch et al., 2024); we use the same differential-outcome-misclassification channel, not a new estimator.

Judges and extractors. LLM judges exhibit order, length, and self-preference biases (Gu et al., 2024; Li et al., 2024; Zheng et al., 2023; Wang et al., 2023; Dubois et al., 2024). Extractor–human disagreement varies with chain-of-thought and formatting (Molfese et al., 2025); judge and verifier benchmarks measure whether evaluators distinguish correct from incorrect answers (Tan et al., 2024; Li et al., 2025), a category-level audit decomposes math-verifier false negatives through certified-equivalent rewrites (Xin, 2026), and imperfect-judge calibration with statistical guarantees also exists (Feng et al., 2026). These works profile or calibrate the instrument across tasks and answer forms; our target is the instrument's mode-conditional error within one model and its propagation into a paired comparison. Public reports document full-transcript feeding and score discrepancies (Hugging Face, 2025a; Red Hat AI, 2025; EleutherAI, 2026; Hugging Face, 2025b; Padmanabhan & Goyal, 2026). Jo et al. (Jo et al., 2025) show that reasoning-model rankings depend on the answer extractor; verifier training and standardization are other repair paths (Chen et al., 2025; Rumiantsev et al., 2025). Those paths repair or replace the instrument; we instead estimate mode-conditional FP/FN rates of a fixed instrument against reference labels, propagate their asymmetry into a paired gap, and audit the source-code rules that create it.

Think/no-think comparisons. Hybrid-reasoning work motivates the accuracy–compute trade-off (Yang et al., 2025; Gan et al., 2026; Sui et al., 2025; Chitale et al., 2026). Benchmark-measurement science usually compares already-graded models (Truong & Koyejo, 2026); our target is the grader as an instrument across modes within one model. A recheck of the six close works listed above (Molfese et al., 2025; Jo et al., 2025; Chen et al., 2025; Rumiantsev et al., 2025; Dorner et al., 2025; Chen et al., 2026) found no paired extractor audit of this form: Molfese et al. (2025) audit extractor–human disagreement across models without mode-conditional labels within one model; Jo et al. (2025) vary answer localization across reasoning models, not a paired within-model mode contrast; Dorner et al. (2025) and Chen et al. (2026) estimate or correct judge error against labels, not conditioned on generation mode; and Chen et al. (2025) and Rumiantsev et al. (2025) repair or standardize the instrument rather than audit a fixed one. The recheck covered vendor documentation and recent papers and is a bounded search, not an exhaustive absence claim.

Position. The distinction is the error representation and target: mode-conditional FP/FN channels of one fixed grader, their propagation into a paired mode contrast, and a decision rule that resolves a direction (possibly against the observed scores), abstains, or reports incompatibility under an explicit error budget.

3 Study Design

We score a fixed corpus with two grader panels and sample candidates for reference labeling. For each cell, the estimated mode-specific error rates support one of three calls under supplied error caps: the direction survives the budget, the analysis abstains, or the supplied set is physically incompatible (Figure 5); the empirical cap table is a sensitivity device, not a coverage realization. A grader is the complete extraction–decision contract: parser rules, input segmentation, and normalization. The protocol and plain-language definitions are in Appendices A and C.

3.1 Corpus, modes, and graders

The corpus is one Qwen3-8B model on five benchmark-derived, difficulty-stratified item sets: GSM8K (201), MATH (312; table key `math_full`), MMLU (320), MMLU-Pro (320), and TriviaQA (320) (Cobbe et al., 2021; Hendrycks et al., 2021; 2020; Wang et al., 2024; Joshi et al., 2017), from middle pilot-accuracy bands of the same model; the 1,473-item estimands are conditional on this outcome-dependent selection, not benchmark-wide prevalence. All-wrong and near-ceiling items were excluded (they give no reference-correct support for FN and little reference-incorrect support for FP estimation; numeric band boundaries in Appendix C.2); composition checks are in Appendices C.2 and C.3 and the supplement (Section S.6). The pilot screen's contract and the corpus finish reasons are not in the frozen record; a token-cap proxy covers all ten task–mode cells (supplement Section S.6; only MATH no-think rows reach the cap). Each item has 128 independent no-think and 8 think

candidates (think outputs average about 4,200 tokens; configuration in Appendix C.3); a label-free complete-test check is in Appendix D.9.

The mainstream panel has eight version-pinned grading contracts derived from widely used evaluation code (upstream sources retrieved July 2026, archived byte-verbatim, executed with four disclosed adaptations; Appendix C.6). Applied to our fixed outputs, they do not recreate each vendor's end-to-end prompt (full evaluated set in Appendix C.6). The in-house mechanism panel defines the primary analysis: seven historical answer extractors (three multiple-choice, two shared-core math, and two exact-match, Table 4) plus an eighth math probe outside every multiplicity family. The seven-grader primary panel yields 487,696 verdicts on 200,328 candidates (the probe counted separately). The mainstream panel (assembled after the in-house audit results were known) has 13 applicable cells (three prompt-contract-degenerate) and 496,672 verdicts; degenerate cells are format mismatches, not silently treated as failures (Appendix C.6).

3.2 Estimands and decision rules

Let S be reference correctness and g a grader verdict. For mode $a \in \{\text{no-think}, \text{think}\}$, estimate $\mathrm{FP}_a = \mathbb{P}(g = 1 \mid S = 0, a)$ and $\mathrm{FN}_a = \mathbb{P}(g = 0 \mid S = 1, a)$, together with their think-minus-no-think gaps. The census observed gap is $D_{\mathrm{obs}} = A_{\mathrm{obs,think}} - A_{\mathrm{obs,no\text{-}think}}$, where $A_{\mathrm{obs},a} = \mathbb{P}(g = 1 \mid a)$, all candidate-level quantities under the fixed 128 no-think versus 8 think candidate protocol (candidate-matching checks in Appendix D.6); the targets S, C, and Y stay distinct throughout (C: committed-answer agreement with gold under the content rule; Y: semantic correctness), and unless a bridge to Y is stated, every gap D below is with respect to S. The mode contrast is bundled (template, cap, candidate count, and seed streams differ; protocol in the supplement, Section S.4); comparisons below are mode/contract contrasts, not isolated computation effects. We also report

$$A_{\mathrm{corr},a} = \frac{A_{\mathrm{obs},a} - \mathrm{FP}_a}{1 - \mathrm{FN}_a - \mathrm{FP}_a}, \qquad D_{\mathrm{corr}} = A_{\mathrm{corr,think}} - A_{\mathrm{corr,no\text{-}think}}; \tag{1}$$

Algebraically the correction requires only $J_a \neq 0$; we further restrict reporting to positively oriented channels ($J_a = 1 - \mathrm{FN}_a - \mathrm{FP}_a > 0$), reporting $J_a \leq 0$ or near-zero cells as undefined or excluded (Table 15 last row; the three prompt-contract-degenerate cells are excluded as format mismatches, not algebraic undefinedness). We also report design-based D_{HT} and additive-residual D_{res} estimates (Appendix B.6), which target the reference-label gap; they coincide with the semantic gap whenever the two modes carry equal mean label bias, in particular when the reference labels equal truth (Appendix B.6). This per-mode slope condition is distinct from the average-slope condition $j > 0$ used in the joint set inversion (Theorem 2): $j > 0$ does not imply $J_a > 0$ for both modes. H1 is a point-estimate reporting screen: per error type, the estimated FN or FP mode difference exceeds two percentage points in absolute value with its interval excluding zero; H3 tests attenuation (H3a, across the panel); H2 asks whether observed sign and CI status agree. Historical H1/H3a multiplicity used 24 and 12 tests within their panel families (`math_v4_normfix` outside); we keep these as analysis labels only, not as confirmatory claims.

3.3 Labeling records

The primary model has 995 author-managed labeling records, not 995 independent human annotations. Round 1 (595 records) checks agreement with benchmark gold and feeds the primary estimate; Round 2 (400 sampled records, of which 393 are binary and 7 held) allows semantic and format equivalence and feeds only the semantic bridge (Appendices C.5 and D.16). Five transfer models contribute 6,100 records as separate per-model substudies. A later, separately sampled 400-record answer-extraction follow-up is reported in Appendix D.2; it does not replace either round. The supplement (Section S.4) summarizes the instrument, recoverability, and human review of every cohort.

Round-1 records were sampled by task, mode, and grader-agreement strata, oversampling disagreements, with Horvitz–Thompson weights. The prescreen judge saw only question, gold, and candidate text plus record index fields (explicit mode, accuracy, and verdict fields hidden), so it is metadata-blind but not necessarily mode-blind (candidate indices and thinking blocks remain visible). A second judge and mechanical re-extraction checked 388 rows (68 semantic and 320 mechanical) without disagreement; 132/132 targeted human adjudications agreed. A blind human relabeling of 194 round-1 records also found 0 disagreements, but its mixed carrier-plus-random design is not an

unweighted corpus sample. The records therefore probe label error without establishing independent external-annotator agreement.

3.4 INFERENCE

Every candidate is graded, so D_{obs} is a census quantity; label inference conditions on the fixed items, outputs, implementations, and benchmark-gold convention. The original 595 keys and strata are reconstructed from the recorded seed; within each original stratum, samples conditional on the grader-specific accepted/rejected counts are simple random without replacement (Appendix D.1). We invert hypergeometric tails for population reference-correct counts, propagate a joint count set through FN/FP ratios, and report per-cell 95% and simultaneous intervals (26 mainstream and 24 in-house comparisons); zero observed events retain nonzero intervals and zero-denominator-compatible rates remain undefined. Coverage is for the fixed reference convention and grader menu, not semantic truth, panel-search history, or future generations.

Variance estimates for the HT and additive residual estimators include finite-population corrections; normal and item-bootstrap intervals are diagnostic sensitivity analyses, not fixed-population coverage claims. The machine-augmented PPI analysis and old adaptive-transfer claims are audited separately and withdrawn where their target or sampling rule does not match the fixed population (Appendices D.7 and D.19). Synthetic calibration checks (20 design–scenario combinations, 5,000 simulations each) give joint count coverage 0.9696–1.0000 (minimum $4848/5000$), validating implementation behavior, not label accuracy or grader selection.

4 MEASUREMENT THEORY

4.1 WHEN DOES A SHARED GRADER PRESERVE A COMPARISON? AN ERROR DECOMPOSITION

Theorem 1 (Figure 2) identifies the channel quantities that decide a comparison, justifying the estimands of Section 3.2; Theorem 2 defines the audit's decision through joint-set inversion. The modes $a \in \{0, 1\}$ are any two text-generating modes scored by one grader: decoding regimes, prompt styles, pipelines, or leaderboard models; here $0 =$ no-think, $1 = \text{think}$, and every gap is think minus no-think $(1 - 0)$. A shared grader *implementation* does not imply a shared *measurement channel*: when the modes shift the surface distribution of answers, the grader's error process can shift with them. Let $A_a = \mathbb{P}(S{=}1 \mid a)$ and $D = A_1 - A_0$ be the reference accuracy and gap under the fixed target S, and $J_a = 1 - \text{FN}_a - \text{FP}_a$ the channel slope (the Youden index (Youden, 1950) of the verdict against the reference); a positive J_a means the verdict is positively oriented. For any quantity x, let $\Delta x = x_1 - x_0$ and $\bar{x} = (x_0 + x_1)/2$; we write $\pi = \bar{A}$ and $j = \bar{J}$ where they shorten formulas, and assume $A_1 = \pi + D/2$ and $A_0 = \pi - D/2$ both lie in $(0, 1)$. Proofs are in Appendices B and B.1; the ingredients are classical (Särndal et al., 1992; Lash et al., 2014).

An exact identity decomposes the observed gap:

$$D_{\text{obs}} \;=\; \bar{J}\, D \;-\; \bar{A}\, \Delta\text{FN} \;+\; (1 - \bar{A})\, \Delta\text{FP}. \tag{2}$$

The observed gap is the true gap *transmitted* through the average channel, plus FN and FP differentials weighted by average accuracy and error rate; the identity holds to machine precision on the audit's cells (Table 15; the check audits the algebra, not semantic truth). Three validity targets organize the analysis: *invariance* ($\text{FN}_0 = \text{FN}_1$ and $\text{FP}_0 = \text{FP}_1$, the equalized-odds condition (Hardt et al., 2016) with mode as the protected attribute), *gap-unbiasedness*, and *rank validity* (definitions in Appendix B.2).

Theorem 1 (Feasible channel differentials and the sharp reversal boundary)**.** *Write* $x = \Delta\text{FN}$, $y = \Delta\text{FP}$, $\pi = \bar{A}$, $j = \bar{J} \in (0, 1]$, $D = D \neq 0$, *and* $\ell_\pi(x, y) = \pi x - (1 - \pi)\, y$.

(i) Feasible region. *A pair* (x, y) *of channel differentials is physically realizable by two channels with average slope* j *if and only if*

$$(x, y) \;\in\; F_j \;=\; \{(x, y)\colon |x| \le 1,\; |y| \le 1,\; |x| + |y| \le 2(1 - j)\}. \tag{3}$$

(ii) Sign criterion. $D_{\text{obs}} = jD - \ell_\pi(x, y)$*; with* $s = \text{sign}(D) \in \{\pm 1\}$*, the comparison is* preserved *iff* $s\, \ell_\pi(x, y) < j|D|$, tied *iff equality holds, and* reversed *iff* $s\, \ell_\pi(x, y) > j|D|$.

(iii) Sharp safety boundary. *Let $K \subseteq F_j$ be a nonempty, compact, origin-symmetric uncertainty set ($K = -K$) for (x, y) with support function $h_K(w) = \max_{z \in K} w^\top z$, and put $w_\pi = (\pi, -(1-\pi))$. The sign is preserved for every channel in K iff $h_K(w_\pi) < j|D|$, and this condition is sharp: whenever the worst corner is physically realizable, equality is attained by legal channel probabilities. For the axis rectangle $K_\square = \{|x| \le u_f,\ |y| \le u_g\} \cap F_j$ the boundary evaluates in closed form (Appendix B.1), reducing to $B_\pi = \pi u_f + (1-\pi)\, u_g$ when the rectangle is itself feasible ($0 \le u_f, u_g \le 1$ and $u_f + u_g \le 2(1-j)$).*

Consequences and the FP ≈ 0 flip reduction: Appendix B.3.

Theorem 2 (Conditional joint-set inversion and the complete abstention rule)**.** *Fix the observed gap $d = D_{\text{obs}}$ on the fixed corpus and collect the channel quantities in $\eta = (\pi, j, x, y)$ as in Theorem 1. The true gap implied by Equation* (2) *is $\phi_d(\eta) = (d + \pi x - (1-\pi)\, y)/j$; define the physically compatible parameter set $\Theta_d = \{\eta\colon 0 \le \pi \le 1,\ 0 < j \le 1,\ (x, y) \in F_j,\ |\phi_d(\eta)| \le 2\min(\pi, 1-\pi)\}$, where the last constraint keeps both mode accuracies in $[0, 1]$ (the closure of the open set the conditional channel rates are defined on; boundary classes carry no conditional rates, and the completeness statement concerns the interior). If $\mathcal{C}_\alpha$ covers η jointly with probability at least $1 - \alpha$, then the image $\mathcal{I}_\alpha(d) = \phi_d(\mathcal{C}_\alpha \cap \Theta_d)$ covers D with probability at least $1 - \alpha$. The decision rule reports* incompatible *when $\mathcal{I}_\alpha(d) = \emptyset$, because the confidence set then contradicts the physical constraints. Otherwise it calls $+$ when $\mathcal{I}_\alpha(d) \subseteq (0, \infty)$, calls $-$ when $\mathcal{I}_\alpha(d) \subseteq (-\infty, 0)$, and abstains. Its directional error probability is at most α; the abstention is* complete *relative to this gap-only compatible set: within $\mathcal{C}_\alpha \cap \Theta_d$, it occurs exactly when a zero true gap is physically compatible or two compatible worlds imply opposite orderings; the two observed mode margins, not used here, could tighten the set further (Appendix B.4). Under the rectangular relaxation $\pi x - (1-\pi) y \in [-B, B]$ with $B := B_\pi$ (Theorem 1(iii)), the rule reduces to $|D_{\text{obs}}| > B$ (necessary and sufficient within the relaxation; the exact physical set uses the closed-form boundary of Theorem 1(iii)).*

Instantiations and the recommended workflow. The heuristic instantiation on every cell of both panels ($\bar{A}$ from the task's proxy grader, uncertainty not propagated) is a conditional sensitivity analysis, not an empirical realization of the coverage guarantee; the historical cap tables are superseded (claim-scope ledger in the supplement, Section S.1). The design-based realization on the count events of Appendix D.1, with $\bar{A}$'s uncertainty propagated through the same events, abstains on all 25 cells (Table 18): the heuristic passes are cap-dependent, not design-valid certificates. This realization decides on a conservative outer envelope of the exact joint image (the coordinate-wise box of Table 18, without the physical-accuracy constraint), so its abstentions do not inherit the completeness reading of Theorem 2: the exact image could certify cells on which the envelope abstains. A direct inversion of the same count events into the reference-accuracy gap, with no parameter relaxation, certifies a positive gap on 5 of the 25 cells at the per-cell level and abstains everywhere at the familywise level (Table 19; the five cells correspond to two distinct task-level reference gaps): the per-cell abstentions of the box realization are a property of the relaxation, not of the counts, and the direct inversion is the recommended per-cell decision rule. Near-degenerate channels ($j \approx 0$) weaken the inversion through $1/j$ (Table 15 last row; Appendices B.1, B.3 and B.4).

The same coverage event can support a required accuracy gain, an equivalence tolerance, or a fixed accuracy–cost trade-off (Corollary 2). An interval spanning zero alone supports neither equivalence nor a token-saving policy.

Proposition 1 (Agreement cannot certify invariance)**.** *No verdict-only statistic can give a non-trivial (strictly below 1) upper bound on a fixed grader's mode-conditional error $e_{ak} = \mathbb{P}(g_k \ne S \mid a)$ without assumptions that exclude shared defects: this includes agreement rates, Cohen's κ, majority vote, and sign survival across a grader panel. Fix any joint verdict distribution and any target grader k: in one world $S = g_k$ everywhere ($e_{ak} = 0$ in both modes); in another $S = 1 - g_k$ ($e_{ak} = 1$). Both worlds induce the same verdict distribution, so no verdict-only statistic separates them, and any upper bound on e_{ak} valid in both is trivial. A mode-conditional assignment shows invariance equally unidentifiable whenever $0 < \mathbb{P}(g_k{=}1 \mid a) < 1$ in both modes: $S = g_k$ on no-think and $S = 1 - g_k$ on think candidates gives $\text{FN}_0 = \text{FP}_0 = 0$ but $\text{FN}_1 = \text{FP}_1 = 1$, while the first world above is invariant.*

A shared mode-specific FN defect is the practically relevant instance of the second world (instance in Appendix D.12): reference labels, not peers, certify graders.

4.2 WHERE DO ERROR BOUNDS COME FROM WITHOUT LABELS? A BOUNDED-OVERLAP ERROR BOUND

Without reference labels, agreement alone does not certify validity. Source-code checks can supply additional information: a row-level safety assignment certifies agreement with a stated content rule, while bounded overlap between unsafe states limits undetected error. The exact identity, its pointwise-sharp constant, and code-verification conditions are in Theorem 3 and Appendix B.5. These are bounds relative to a rule that extracts a final numeric commitment and compares it to gold; they do not establish semantic truth. A direct baseline computes that content rule on every decidable row; on this corpus its all-completions interval $[+10.6, +14.2]$pp nearly coincides with the multi-grader safe-anchor interval $[+10.1, +14.3]$pp, so here the analysis's added value is grader attribution rather than a tighter comparison (Appendix D.22). A separate reference audit must bound disagreement between that rule and the intended correctness standard (Corollary 3).

Instantiation and the semantic bridge. On the fixed GSM8K corpus (27,336 candidates), the instantiation uses a code-audited panel of four graders; one member is the in-house `math_v1`, used only as a row-level safety anchor, and its known-defect rows are treated as hazard (Appendix B.5). One-sided content-level bounds (Table 17) are positive label-free for `qwen_math` and `math_verify` ($[+10.1, +14.3]$pp), with lm-eval flexible abstaining under that one-sided rule (the direct interval of Proposition 4, positive for all four and never wider, is the appropriate content-level comparison bound, Appendix D.22; Theorem 3 remains the per-grader attribution form when row-level information is reduced to aggregate overlap counts). The in-house grader's own bound is in Appendix B.7.

5 RESULTS

5.1 OBSERVED SIGN REVERSALS IN MAINSTREAM GRADERS

On identical MMLU candidates, `qwen_choice` reports a think-minus-no-think gap of -15.7 points where two other mainstream extractors report $+17.0$ to $+17.3$ — an observed-score reversal and a census fact (Figure 1). The 595-record reference substudy estimates a $+45.6$-point benchmark-reference FN gap for this cell (per-cell and simultaneous intervals in Table 1): compatible with a mode-concentrated failure pattern under the stated reference convention, without establishing a family-adjusted nonzero difference (Appendix D.19 does not resolve it).

Separately sampled extraction follow-up. A later 400-record extraction follow-up — a new sample, distinct from the 400-record semantic extension of Section 3.3 — targets `qwen_choice` on MMLU (240 Qwen3-8B candidates; 80 later think candidates each for DeepSeek and MiMo). The returned final-answer extractions give Qwen FN-gap intervals of $[+31.7, +72.9]$ points (benchmark gold) and $[+11.6, +65.2]$ (returned key), each with its own 95% simultaneous count guarantee over eight fixed strata ($386/400$ labels determined; submitter-confirmed human answering without seeing gold or candidates; Appendix D.2). The intervals condition on determined labels being correct and do not establish a semantic ranking; treating four flagged question answers as unresolved gives $[+10.3, +66.2]$ points.

The second observed-score reversal is on GSM8K: lm-eval flexible gives -13.2 points where `qwen_math` and `math_verify` give $+11.8$ and $+11.9$; no FN events are observed in the reference sample for these two graders, but the new design intervals (conditional on the fixed post-discovery grader menu) do not collapse to zero. For MMLU, the direct HT reference-correctness gap is $+30.6$ points (95% interval $[-9.1, +63.4]$); the additive residual point estimate is $+15.2$. Neither identifies a semantic-truth reversal; the unstable multiplicative inversion is confined to the estimator comparison in the supplement (Section S.7).

Propagating one 95% joint count set across all five task accuracy differences resolves none of them within two points (Appendix D.4 and Table 8), a statement about this audit's precision, not benchmark saturation.

Complete-test-set comparison. On the complete MMLU and GSM8K test sets (15,361 items, 368,664 candidates), the observed gaps under the two failing extractors are -21.0 and -22.0 points; the positive alternatives give $+2.3$ to $+4.9$ (the degenerate strict contract excluded) (Appendix D.9).

Table 1: Fixed-menu analysis; intervals do not adjust for earlier grader search. Mainstream headline results, in percentage points. FN rates are HT point estimates from the original 595 records, each against that arm's reference-correct support ($\mathrm{FN} = R/(R+T)$, Appendix D.1); observed accuracies are census values. Intervals use the actual sampling design (per-cell 95%; simultaneous over the 13 mainstream and 12 in-house cells, jointly at least 95%). D_{obs} is a census quantity; gaps from unrounded accuracies (displayed margins may differ by 0.1). Zero FN events do not give zero-width intervals; displayed 0.0 FN values are zero-event cells, not small rates rounded down. Supports and all cells: Appendix D.1; count inputs in the supplement (Section S.6).

Task	Grader	$\mathrm{FN}_{\text{no-think}}/\mathrm{FN}_{\texttt{think}}$	FN gap	Per-cell CI	Simultaneous CI	$\mathrm{Acc}_{\text{no-think}}/\mathrm{Acc}_{\texttt{think}}$	D_{obs}
mmlu	`qwen_choice`	5.3/50.9	+45.6	[+11.8, +77.0]	[−5.8, +81.8]	48.1/32.3	−15.7
gsm8k	lm-eval flexible	58.6/75.0	+16.4	[+3.3, +42.0]	[−4.6, +50.9]	26.8/13.6	−13.2
gsm8k	`qwen_math`	0.0/0.0	+0.0	[−12.2, +10.2]	[−18.3, +13.1]	76.0/87.8	+11.8
gsm8k	`math_verify`	0.0/0.0	+0.0	[−12.3, +10.2]	[−18.4, +13.1]	75.9/87.8	+11.9
mmlu_pro	`mmlupro_official`	13.1/8.7	−4.4	[−35.3, +17.1]	[−45.9, +23.0]	49.1/72.1	+23.0

Matched-candidate, length-stop-removed, and paper-subset comparisons preserve these directions (Appendix D.6; count tables in the supplement, Section S.6): selection changes magnitude, not sign.

5.2 Parser interventions and their costs

The vendored `qwen_choice` applies, in source order: (1) count the case-sensitive phrases "choice is" and "answer is"; (2) if either occurs more than once, an ICL heuristic truncates the input to its first paragraph; (3) search the retained text for a trigger, choosing the first option letter after the last trigger; (4) without a trigger, take the last standalone option letter (Appendix D.20): truncation precedes the trigger/fallback decision. In the primary MMLU think corpus, 79.5% of rows are both truncated and sent to fallback, the no-think share 0.64% (Appendix D.6); "no trigger" does not mean the original trace lacked triggers.

On the fixed primary MMLU outputs (Table 20), stripping the thinking segment changes D_{obs} to +16.2 points; disabling ICL truncation gives +14.1, and combining the changes gives +16.0 (not additive). Stripping moves 982 think verdicts from wrong to right and 165 from right to wrong against benchmark gold (20.0% of previously accepted think rows); accepted full-text verdicts can be accidental matches, so transitions are not automatically semantic improvements or harms (Appendices D.6 and D.14).

Does segmentation reduce grading error? On the 240 Qwen candidates of the later extraction follow-up, all 51 changed sampled think verdicts repair errors relative to benchmark gold and the extracted final answers; relative to the returned question key, 47 repair and four introduce errors. Design-based error-reduction intervals over all 2,560 think outputs are [31.4, 44.8] and [20.1, 43.9] points (Table 7). All 40,960 no-think verdicts are unchanged; the effect is conditional on the reference labels and settles neither answers nor the semantic mode ranking (Appendix D.3).

For lm-eval flexible on GSM8K, stripping leaves every verdict unchanged. A frozen post-hoc rule takes the last complete boxed answer in the answer segment, else the last explicit Answer line, else the segment itself; passing that text to flexible moves the primary gap to +22.4 points and the complete-set gap to +4.95, and 80 primary no-think verdicts from right to wrong (1.16% of previously accepted rows). The paired construction of Appendix D.3 audits it on the same 595 labels: all 39 sampled changes repair a full-rule error and none introduces one (Appendix D.20), while its reference FN gap straddles zero — a reference-evaluated parser intervention, not a validated generation-time repair.

New decoding and format experiment. New-generation experiments (4,160 outputs across temperatures and formats) retain the failure patterns and one failed prompt-contract repair (Appendices D.5 and D.20 and Table 9).

The in-house GSM8K graders illustrate the opposite error direction. They share an implementation core and agree on every candidate, yet both reject an estimated 37.9% of reference-correct no-think answers and none of the audited think answers (FN gap −37.9 points, simultaneous interval [−56.6, −0.8], Table 5). One concrete normalization failure is `\boxed{\$9.00}` scored

wrong against gold 9: agreement between shared-core implementations does not establish validity (Proposition 1 and Figure 7).

The content-level bounded-overlap calculation gives $[+10.1, +14.3]$ points for qwen_math/math_verify, while lm-eval flexible abstains under that one-sided rule (Table 17); keeping source-audited directional errors in a simpler bound gives that same positive range for all four graders on this corpus (Proposition 4 and Appendix D.22). Five semantic mismatches and seven held rows in the 400-record bridge study prevent promoting these content bounds to unconditional semantic certificates. Across the 12 applicable in-house task–grader cells, the historical item-cluster intervals retained positive observed signs; that panel-wide survival is a safeguard result, not a measurement-invariance certificate (Table 14).

5.3 What standard safeguards can and cannot establish

First, *provenance*: the MMLU flip occurs under the vendored Qwen multiple-choice extractor on these fixed outputs — a test of its extraction contract, not the developer's complete pipeline; provenance alone does not establish mode invariance. The MMLU-Pro official chain's two error channels nearly cancel relative to the proxy, leaving its observed gap close to the proxy gap (Table 15 and Appendix C.6). Second, *conclusion survival*: cross-grader comparison can detect a problem (qwen_choice); within the in-house panel, signs agree while magnitudes differ, so sign consistency establishes neither invariance nor magnitude. Third, *agreement*: by Proposition 1, no verdict-only statistic certifies mode-specific invariance (instance in Appendix D.12).

5.4 Transfer to five further models

We applied the graders to five further model corpora on the common domain of no-think 0–127 and think 0–7 (GLM-4.5-Air, Qwen3.6-35B-A3B, gpt-oss-20b, DeepSeek-V3.2, MiMo-V2-Flash; gpt-oss low/high reasoning effort; Appendix D.7; per-corpus generation protocol in the supplement, Section S.4). Observed reversals replicate across all five models, and mode-concentrated FN point patterns extend beyond the primary model: on GLM TriviaQA, two exact-match rules report -40.7 versus $+4.5$ points. The current intervals support less: Qwen3.6's simultaneous TriviaQA interval is the only one there excluding zero ($[+8.1, +100.0]$), and GLM's per-cell interval also excludes zero (Table 11). The earlier four-signal claim based on adaptive-sample bootstrap inference is withdrawn.

6 Discussion

Limitations and use. The primary reference audit covers one model and five fixed item subsets (c14–c16 near-ceiling bands excluded); the transfer, new-generation, and follow-up studies keep their own designs. Error rates depend on model, outputs, and grading contract (transfer results include opposite error directions), so source-model rates need extra evidence before reuse, and no universal score-gap threshold is established. The label-dependent audit did not run complete test sets; a label-free complete-set check reproduces the flip direction and fallback at larger magnitude (Appendix D.9). Reference labels were LLM-prescreened (an earlier review saw the verdicts); a blind relabeling of 194 records found no disagreement (0/194, Appendix D.18); transfer inference uses each model's fixed initial sample. The audited subsets and failing contracts are English (Hendrycks et al., 2020; Cobbe et al., 2021; Hendrycks et al., 2021; Wang et al., 2024; Joshi et al., 2017); contamination and multilingual outputs are untested. Semantic-level guarantees remain a sensitivity region (Appendices D.2 and D.5).

Use. For mode evaluation, report an observed gain with a reference-based interval, and check repairs against the same labels, counting removed and introduced errors. With a row-level code-audited safety assignment, the direct safe-anchor interval of Proposition 4 is the recommended content-level bound and Theorem 3 its per-grader attribution form; the MMLU audit supports segmentation under its conventions, while GSM8K shows segmentation alone is insufficient. Specify an accuracy tolerance or cost trade-off before interval-based mode choice (Corollary 2). The defects we name are verifiable in public upstream code; held-out or repaired graders are recommended for high-stakes comparisons; cross-model resolution and adaptive thinking remain unvalidated; paired reference checks evaluate repairs while leaving the mode ranking unresolved, so report the two separately.

REPRODUCIBILITY STATEMENT

The local research archive retains all analysis code, the analysis protocol with its content hashes, sampling manifests, the vendored mainstream graders (byte-verbatim, SHA-256-verified at import), and the adjudicated reference labels. It also retains the dated 400-row semantic-extension manifest, its seven held-row decisions, the five-row, four-item C–Y mismatch audit, the machine-augmented label set with judge provenance, the blind-relabeling packs and verdicts, and the PPI design and analysis outputs (Appendices D.18 and D.19). The accompanying supplement provides the claim-scope ledger, analysis provenance and audit trail, and analysis artifact index. It also contains the inputs needed to recompute the headline intervals of the original 25-cell design and the paired repair analysis: the repair audit's per-stratum counts, the complete stratum-level count inputs for all 25 cells, the current FP bounds, the follow-up strata, the generation protocol for every corpus, the annotation judge configuration, and the labeling-instrument matrix (supplement Sections S.4 and S.6); the 125 transfer cells recompute from `transfer_design.json`. Public hosting and licensing for the full code and data bundle have not been finalized. The audit design is described in Sections 3 and 3.3; grader provenance and execution adaptations are recorded in Appendix C.6. Primary candidates were generated with vLLM 0.11.0 at temperature 0.6 and top-p 0.95, with at most 8,192 output tokens per no-think candidate and 32,768 per think candidate; the primary corpus totals 178.7M output tokens on 200,328 candidates (no-think mean 684.8, think mean 4,209.0 tokens; recomputed row-by-row in `miaudit_token_accounting_2026-09-24.*`; per-shard wall-clock fields were not recorded). The new 4,160-output temperature and format experiment is reported in Appendix D.5, including its failed prompt intervention. The separately sampled 400-record extraction follow-up retains the returned question keys, transcription corrections, and unresolved labels separately from earlier audits (Appendix D.2). Its Qwen subset is reused for the paired repair analysis (Appendix D.3); the original labels also yield the five-task joint accuracy intervals in Appendix D.4. Both are retrospective reanalyses, without additional generations or annotations beyond the already described follow-up and original audits. The current fixed-population analyses use exact count inversion, with separate scripts for the original stratified sample, G1/G2 design audit, and fixed-initial-sample transfer analysis. The supplement records sample-key replays, interval validation, and the withdrawal of the old PPI and adaptive-transfer confidence claims. Historical bootstrap procedures use $B = 2{,}000$ replicates, with seeds 20260725, 20260727, 20260728, 20260809, and 20260821 for the primary, mainstream, two-phase, multi-model, and dated semantic-extension sensitivity analyses respectively.

AI USE STATEMENT

In this work, we used generative AI tools to assist with theoretical model development and proof checking, to implement analysis code, and to draft and edit parts of the paper. These assistant tools were commercial LLM coding and review agents used between July and September 2026; they did not produce reference labels outside the labeling instruments documented as research objects in Appendix C.4, and all reported numbers were recomputed from the frozen artifacts by deterministic scripts. LLM-based grading, pre-labeling, and targeted per-row inspection appear in this paper as objects and instruments of the research itself and are described as such in the main text and appendix. Deterministic recomputation and targeted reviews were used to check calculations and claims. These checks do not establish reference-label correctness; the paper reports the remaining limitations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Paul S. Albert and Lori E. Dodd. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60(2):427–435, June 2004. ISSN 1541-0420. doi: 10.1111/j.0006-341x.2004.00187.x.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference, 2023. URL https://arxiv.org/abs/2301.09633.

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, 2025. URL https://arxiv.org/abs/2504.10481.

Yiqun T. Chen, Sizhu Lu, Sijia Li, Moran Guo, and Shengyi Li. Efficient inference for noisy LLM-as-a-judge evaluation. In *International Conference on Machine Learning (ICML)*, 2026.

Rajiv Shailesh Chitale, Rahul Madhavan, Taneesh Gupta, Deepanway Ghosal, and Aravindan Raghuveer. Test-time scaling via error localization, 2026. URL https://arxiv.org/abs/2607.21453.

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, 2021. URL https://arxiv.org/abs/2110.14168.

Florian E. Dorner, Vivian Y. Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. In *International Conference on Learning Representations (ICLR)*, 2025.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *First Conference on Language Modeling (COLM)*, 2024.

EleutherAI. lm-evaluation-harness issue #3576: thinking-mode scores collapse against no-think on GSM8K flexible extraction. https://github.com/EleutherAI/lm-evaluation-harness/issues/3576, 2026. Accessed 2026-09-21.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of LLMs with imperfect judges. In *International Conference on Learning Representations (ICLR)*, 2026.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for hybrid language model evaluation, 2024. URL https://arxiv.org/abs/2406.04291.

Siyuan Gan, Jiaheng Liu, Boyan Wang, Tianpei Yang, Runqing Miao, Yuyao Zhang, Fanyu Meng, Junlan Feng, Linjian Meng, Jing Huo, and Yang Gao. Thinking-based non-thinking: Solving the reward hacking problem in training hybrid reasoning models via reinforcement learning, 2026. URL https://arxiv.org/abs/2601.04805.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 2024. URL https://doi.org/10.5281/zenodo.12608602.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on LLM-as-a-judge, 2024. URL https://arxiv.org/abs/2411.15594.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, 2016.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2020. URL https://arxiv.org/abs/2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Paul W. Holland and Howard Wainer (eds.). *Differential Item Functioning*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, December 1952. ISSN 1537-274X. doi: 10.1080/01621459.1952.10483446.

Hugging Face. Open LLM Leaderboard v2: evaluation of DeepSeek-R1-Distill-Qwen-32B (discussion). https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B/discussions/17, 2025a. Accessed 2026-09-21.

Hugging Face. lighteval issue #1094: Qwen3 thinking-mode extraction failures on MMLU-Redux. https://github.com/huggingface/lighteval/issues/1094, 2025b. Accessed 2026-09-21.

Hugging Face. Math-Verify: A robust mathematical expression evaluation tool. https://github.com/huggingface/Math-Verify, 2025c. Accessed 2026-09-21.

Hugging Face. Open R1: A fully open reproduction of DeepSeek-R1. https://github.com/huggingface/open-r1, 2025d. Accessed 2026-09-21.

Siu L. Hui and Stephen D. Walter. Estimating the error rates of diagnostic tests. *Biometrics*, 36(1): 167–171, March 1980. ISSN 0006-341X. doi: 10.2307/2530508.

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, 2025. URL https://arxiv.org/abs/2510.14773.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1147.

Leslie Kish. *Survey Sampling*. Wiley, 1965.

Evelyn M. Kitagawa. Components of a difference between two rates. *Journal of the American Statistical Association*, 50(272):1168–1194, 1955.

Timothy L. Lash, Matthew P. Fox, Richard F. MacLehose, George Maldonado, Lawrence C. McCandless, and Sander Greenland. Good practices for quantitative bias analysis. *International Journal of Environmental Research and Public Health*, 11(10):10129–10141, 2014.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge, 2024. URL https://arxiv.org/abs/2411.16594.

Xuzhao Li, Xuchen Li, Shiyu Hu, Yongzhen Guo, and Wentao Zhang. VerifyBench: A systematic benchmark for evaluating reasoning verifiers across domains, 2025. URL https://arxiv.org/abs/2507.09884.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003. doi: 10.1007/978-0-387-22796-9.

William Meredith. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543, December 1993. ISSN 1860-0980. doi: 10.1007/bf02294825.

Roger E. Millsap. *Statistical Approaches to Measurement Invariance*. Routledge, March 2012. ISBN 9780203821961. doi: 10.4324/9780203821961.

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*, 2025.

OpenAI. simple-evals. https://github.com/openai/simple-evals, 2023. Accessed 2026-09-21.

Shankar Padmanabhan and Tanya Goyal. Updating parametric knowledge with context distillation retains post-training capabilities, 2026. URL https://arxiv.org/abs/2602.16093.

Red Hat AI. DeepSeek-R1-Distill-Qwen-7B-quantized.w4a16 model card. https://huggingface.co/RedHatAI/DeepSeek-R1-Distill-Qwen-7B-quantized.w4a16, 2025. Accessed 2026-09-21.

Pavel Rumiantsev, Soumyasundar Pal, Yingxue Zhang, and Mark Coates. FEval-TTC: Fair evaluation protocol for test-time compute, 2025. URL https://arxiv.org/abs/2511.01203.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems, 2023. URL https://arxiv.org/abs/2311.09476.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer New York, 1992. doi: 10.1007/978-1-4612-4378-6.

Minghe Shen, Ananth Balashankar, Adam Fisch, David Madras, and Miguel Rodrigues. Robust LLM performance certification via constrained maximum likelihood estimation, 2026. URL https://arxiv.org/abs/2604.03257.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. Stop overthinking: A survey on efficient reasoning for large language models, 2025. URL https://arxiv.org/abs/2503.16419.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges, 2024. URL https://arxiv.org/abs/2410.12784.

Sang T. Truong and Sanmi Koyejo. *AI Measurement Science*. Stanford University, 2026. URL https://aimslab.stanford.edu/textbook/.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators, 2023. URL https://arxiv.org/abs/2305.17926.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL https://arxiv.org/abs/2406.01574.

Esther Xin. Where the verifier fails: A category-level audit of reward signals in RLVR, 2026. URL https://arxiv.org/abs/2609.01354.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement, 2024. URL https://arxiv.org/abs/2409.12122.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

William John Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*, 2023.

Guide to the appendix. Appendix A explains the statistical terms used throughout the paper. Appendix B proves the main-text results. Appendix C summarizes the audit protocol. Appendix D reports the extended results behind the main-text claims. The anonymized supplement contains the verbatim instruments and per-stratum count tables (Sections S.4 and S.6), extended theory (Section S.5), meta-audit details (Section S.7), and the claim-scope ledger, analysis provenance and audit trail, and analysis artifact index (Sections S.1–S.3). Display precision: effect sizes are shown to one decimal, fine-grained sensitivities and small-sample ratios to two, and simulation diagnostics and q-values at the precision they require; exact content-bound arithmetic checks keep three decimals and content-bound component masses are shown as fractions to four, so the displayed identities recompute; exact synthetic arithmetic, code-level census diagnostics, and planning constants keep the decimals they require; all computation uses unrounded values.

A Statistical Terms and Notation

Table 2 gives plain-language definitions; Table 3 then maps the similarly named gap quantities onto what each one answers and what it requires.

Table 2: Plain-language glossary of the statistical terms used in this paper, with each term's role here.

Term	Plain meaning and role in this paper
Mode (no-think / think)	The two generation modes compared for one model. The primary corpus uses 128 no-think and 8 think candidates per item; follow-up candidate counts are specified separately.
Grader	A parser or judge mapping free-form model text to a binary verdict (correct / incorrect). The *mainstream panel* is eight graders vendored from widely used evaluation code; the *in-house mechanism panel* is seven answer-extraction graders from our own evaluation history (results in Appendix D.12).
Misclassification channel (FP_a, FN_a)	Per-mode rates at which the grader accepts truly incorrect candidates (FP) and rejects truly correct ones (FN). The paper's central quantities: measurement invariance fails when these differ across modes; comparative validity is then assessed through the error decomposition, not presumed lost.
Channel slope J (Youden index)	$1 - \mathrm{FN} - \mathrm{FP}$: how strongly the verdict tracks truth; a near-zero J means a nearly uninformative grader.
Equalized odds	$g \perp a \mid S$: the verdict is independent of the mode given truth. Borrowed from fairness, here the comparative-measurement validity condition (invariance).
D_{obs} / D / D_{corr}	Observed think−no-think accuracy gap / the reference-accuracy gap under the fixed target S / the gap after removing each mode's estimated grading errors (Equation (1)).
D_{HT} / D_{res}	Two design-based reference-gap estimates from the reference substudy: audit labels alone (Horvitz–Thompson), and full-corpus verdicts plus audit residuals (Equation (8)).
Horvitz–Thompson (HT) weights	Inverse within-stratum inclusion probabilities. HT totals and means are design-unbiased for the specified finite-population reference labels; semantic interpretation additionally requires valid labels.
Kish effective size (Kish, 1965) n_{eff}	$(\sum w)^2 / \sum w^2$: how many equally weighted rows would have the same weight concentration. This summary does not establish design-valid binomial confidence bounds.
Item-cluster paired bootstrap	Resamples whole items with all their candidates, preserving within-item and cross-mode covariance; a historical item-composition sensitivity ($B = 2{,}000$), not current inference.
Two-phase bootstrap	Additionally re-draws the audited rows within strata; a historical nested sensitivity, not a validated sampling-design model.
Benjamini–Hochberg (BH) FDR / q-value	The BH-adjusted analogue of a p-value. FDR control requires valid tests and appropriate dependence conditions; the paper's historical q-values are not current guarantees.
Manski identified set (Manski, 2003)	The range of reference-accuracy gaps compatible with the observed data given assumed per-mode error-rate caps. A conditional sensitivity device ("given these caps"), not a confidence set.
SRSWOR	Simple random sampling without replacement, applied separately within each original reference-substudy stratum.
Content-level vs. semantic truth	C: whether the committed answer is numerically equal to gold (code-decidable). Y: reference-labeled semantic correctness. Content-level bounds hold for C unless lifted across the bridge δ.
Earlier labeling records	995 for the primary model (595 round-1 design-weighted + 400 semantic extension; 988 binary, 7 held) and 6,100 across five per-model transfer substudies. The primary estimate uses only the 595; the later extraction follow-up remains separate.
In-house panel / mainstream panel	The panels contain 12 and 13 applicable task–grader cells, respectively. Current FN/FP intervals jointly cover all 25 cells. Historical multiplicity families are distinct: the in-house H1/H3a family used 24 tests (12 cells × {FN, FP}), the mainstream extension family used 26 tests (13 cells × {FN, FP}), and a separate historical H3a attenuation family used 12.

Table 3: Map of the gap quantities: what each answers and what it requires. The observed gap concerns grader verdicts. Reference-based estimators target their stated label convention; content bounds concern C. Neither becomes a semantic-accuracy guarantee without the corresponding label-validity or bridge assumptions.

Quantity	What it answers	What it requires
D_{obs}	The gap a study would report	Verdicts only; label-free, an exact census number on the fixed corpus
D	The gap a perfect grader would report under the reference convention	the fixed target S; never observed directly, only estimated or bounded by the rows below
D_{corr} (Equation (1))	D_{obs} with each mode's estimated FP/FN removed	Reference labels equal truth; unstable when the channel slope $J \approx 0$
D_{HT}	Reference-accuracy gap from the audit labels alone	Design-unbiased if the reference labels equal truth
D_{res} (Equation (8))	Reference-accuracy gap from full-corpus verdicts plus audit residuals	Same as D_{HT}
FP_a / FN_a	Per-mode grader error rates (the channels)	Reference labels and HT weights
FN gap / FN diff	Synonyms for the think−no-think difference of FN_a; used interchangeably across tables.	
Manski set (supplement Section S.5)	Range of reference-accuracy gaps compatible with the data under assumed caps	Assumed per-mode error-rate caps; a sensitivity device, not a confidence set
Bounded-overlap bound (Theorem 3)	Label-free content-level error bound	Code-audited hazard states, a safe-peer panel, and the content functional C; semantic reading costs the bridge δ
Decision rule (Theorem 2)	Direction call conditional on the set, abstention, or incompatibility flag	A joint confidence or sensitivity set for the channel quantities; coverage is only as valid as that set

B Proofs and Theoretical Details

This appendix collects the proofs of the results stated in the main text: the error decomposition, the feasibility theorem, the inversion rule, the bounded-overlap error bound, and the residual Horvitz–Thompson estimator. Extended readings, construction details, and verification evidence are in the supplement (Section S.5).

B.1 Error Decomposition, Feasibility, and Inversion: Proofs

All results rest on one identity: by the law of total probability, per mode a,

$$A_{\mathrm{obs},a} = A_a(1-\mathrm{FN}_a) + (1-A_a)\mathrm{FP}_a = \mathrm{FP}_a + J_a A_a, \qquad J_a = 1-\mathrm{FN}_a-\mathrm{FP}_a. \tag{4}$$

Proof of Equation (2). From Equation (4), $D_{\mathrm{obs}} = \Delta\mathrm{FP} + \Delta(JA)$. The exact product rule $\Delta(xy) = \bar{x}\,\Delta y + \bar{y}\,\Delta x$ gives $\Delta(JA) = \bar{J}\,\Delta A + \bar{A}\,\Delta J = \bar{J}\,D + \bar{A}(-\Delta\mathrm{FN} - \Delta\mathrm{FP})$. Substituting, $D_{\mathrm{obs}} = \bar{J}\,D - \bar{A}\,\Delta\mathrm{FN} + (1-\bar{A})\Delta\mathrm{FP}$. □

Proof of Theorem 1. (i) *Feasible region.* Parametrize the two modes' FN rates by $f_0 = \bar{f} - x/2$, $f_1 = \bar{f} + x/2$; then $f_0, f_1 \in [0,1]$ iff $\bar{f} \in [|x|/2,\ 1-|x|/2]$, which is nonempty iff $|x| \le 1$, and likewise $\bar{g} \in [|y|/2,\ 1-|y|/2]$ for the FP rates. The average-slope constraint $\bar{f} + \bar{g} = 1-j$ is attainable iff the two intervals admit values summing to $1-j$, i.e., iff $(|x|+|y|)/2 \le 1-j \le 2-(|x|+|y|)/2$; the right inequality is implied by the left one, so feasibility is equivalent to Equation (3). (ii) *Sign criterion.* By Equation (2), $D_{\mathrm{obs}} = jD - \ell_\pi(x,y)$. Multiplying by $s = \mathrm{sign}(D)$, $s\,D_{\mathrm{obs}} = j|D| - s\,\ell_\pi(x,y)$, which is positive, zero, or negative exactly under the stated conditions. (iii) *Sharp safety boundary.* The sign is preserved for every channel in K iff $\max_{(x,y)\in K} s\,\ell_\pi(x,y) < j|D|$. Since $s\,\ell_\pi(x,y) = \ell_\pi(sx, sy)$ and K is centrally symmetric, the maximum equals $\max_K \ell_\pi = h_K(w_\pi)$, giving the stated necessary and sufficient condition. For the axis rectangle, ℓ_π is increasing in x and decreasing in y, so with $p = x \ge 0$, $v = -y \ge 0$ the problem is the linear program $\max\ \pi p + (1-\pi)v$ over $p \le a$, $v \le b$, $p + v \le c$; the greedy solution that first saturates the larger coefficient is optimal by an exchange argument. For nonnegative input caps, define $a = \min(u_f, 1)$, $b = \min(u_g, 1)$, $c = 2(1-j)$ and $z_+ = \max(z, 0)$. Then

$$H_j(\pi; u_f, u_g) = \begin{cases} \pi\min(a,c) + (1-\pi)\min\{b, (c-a)_+\}, & \pi \ge 1/2, \\ (1-\pi)\min(b,c) + \pi\min\{a, (c-b)_+\}, & \pi < 1/2. \end{cases}$$

For example, $j = 0.2$, $\pi = 0.5$, $u_f = 1.5$, $u_g = 0$ gives $H_j = 0.5$, not the untruncated rectangle value 0.75. We have $H_j = B_\pi$ when the rectangle is feasible, so that $a = u_f$, $b = u_g$, and $a + b \le c$ makes the coupling constraint slack. Sharpness: if the maximizing corner (x^*, y^*) satisfies $|x^*| + |y^*| \le 2(1-j)$, the interval intersection $[|x^*|/2,\ 1-|x^*|/2] \cap [|y^*|/2 - j,\ 1-j-|y^*|/2]$ for $\bar{f}$ is nonempty (the two inequalities reduce to $(|x^*|+|y^*|)/2 \le 1-j$ and $(|x^*|+|y^*|)/2 \le 1+j$), so legal channel probabilities attaining $\ell_\pi = h_K(w_\pi)$ exist; with $s = \mathrm{sign}(D)$, the centrally symmetric point (sx^*, sy^*) satisfies $s\,\ell_\pi(sx^*, sy^*) = h_K(w_\pi)$, so at $h_K(w_\pi) = j|D|$ this point yields a tie, and beyond the boundary a reversal. The constructed mode channels are legal by this interval construction; at the tie point (sx^*, sy^*) their slopes are $J_0 = j + s(x^* + y^*)/2$ and $J_1 = j - s(x^* + y^*)/2$, both $\ge 2j - 1$, hence both positive when $j > 1/2$. If one additionally insists that both modes be positively oriented ($J_0, J_1 > 0$), the sharp value is the supremum of the same linear objective over $K \cap \{|x + y| < 2j\}$ (the rectangle keeps its own u_f, u_g caps) and need not be attained. We do not impose this restriction because negatively oriented modes, although pathological, are legal channels. □

Proof of Theorem 2. The working assumptions of Section 4.1 ($j > 0$ and both mode accuracies in $(0,1)$) place the true η^* in Θ_d with probability one: $|\phi_d(\eta^*)| = |D| \le 2\min(\pi, 1-\pi)$ is exactly the constraint $A_0, A_1 \in [0,1]$. On the event $\{\eta^* \in \mathcal{C}_\alpha\}$, which has probability at least $1-\alpha$, $D = \phi_d(\eta^*) \in \phi_d(\mathcal{C}_\alpha \cap \Theta_d) = \mathcal{I}_\alpha(d)$, so $\mathcal{I}_\alpha(d)$ covers D at level $\ge 1-\alpha$ (taking the smallest interval containing the image only widens it). A directional call is made only when the image excludes zero on one side. Its error probability is therefore at most $\mathbb{P}(D \notin \mathcal{I}_\alpha(d)) \le \alpha$. The rule abstains iff the nonempty image contains 0 or meets both open half-lines; an empty image is reported as *incompatible*, not as abstention. The direction rule is applied only after this nonempty check: physical feasibility is tested first, and the rectangular condition below then decides direction within

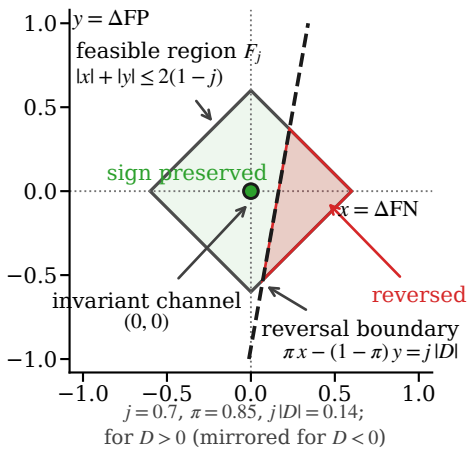

Figure 2: Schematic of Theorem 1. The feasible region F_j of channel differentials $(x, y) = (\Delta\mathrm{FN}, \Delta\mathrm{FP})$ is the diamond $\{|x| + |y| \le 2(1 - j)\}$ intersected with $[-1, 1]^2$; the comparison is preserved exactly when $\pi x - (1 - \pi)y < j|D|$ for $D > 0$. Here $j = 0.7$, $\pi = 0.85$, and $D = 0.20$, so the two accuracies are 0.75 and 0.95.

the relaxed problem (Figure 3 maps two such relaxation instances). Under the rectangular relaxation $q = \pi x - (1 - \pi)y \in [-B, B]$, the following condition is necessary and sufficient for the relaxed problem; for the exact physical set F_j the sharp boundary is the H_j of Theorem 1: $D = (d + q)/j$ with $j > 0$ has the sign of d whenever $|d| > B$, while $|d| \le B$ admits $q = -d$ (a zero true gap), so no strict sign is certifiable (both signs are compatible exactly when $|d| < B$). For $j \approx 0$ the divisor $1/j$ amplifies whatever numerator uncertainty the joint set leaves unresolved, while the Θ_d truncation caps the image; at $j = 0$, Equation (2) no longer involves D and the map is undefined, so such channels are excluded from Θ_d. □

Decision targets beyond a zero accuracy gap. The following is a consequence of confidence-set coverage, not a new calibration estimator. It also applies to a direct count-based interval, without estimating or inverting FN/FP rates.

Corollary 2 (Accuracy thresholds, equivalence, and fixed costs)**.** *Let $D = A_{\text{think}} - A_{\text{no-think}}$ be an accuracy difference for fixed candidate populations under a fixed binary reference convention. Suppose $I = [L, U]$ is nonempty and $\Pr\{D \in I\} \ge 1 - \alpha$ under the label-sampling design. For an accuracy threshold $\vartheta \ge 0$ and an equivalence tolerance $\epsilon \ge 0$, the implications*

$$L > \vartheta \implies D > \vartheta, \qquad U < -\vartheta \implies D < -\vartheta, \qquad [L, U] \subseteq [-\epsilon, \epsilon] \implies |D| \le \epsilon$$

are all correct on the same coverage event. If $c = C_{\text{think}} - C_{\text{no-think}}$ is an observed, fixed mean cost difference and $\lambda \ge 0$ converts cost units to accuracy units, the net-utility difference $V = D - \lambda c$ has interval $[L - \lambda c, U - \lambda c]$ on that event. The unconditional probability of any false assertion obtained through these implications is at most α.

Proof. On $\{D \in [L, U]\}$, $D \ge L$ and $D \le U$ imply each accuracy statement. Subtracting the fixed λc preserves both inequalities. Every false assertion is therefore contained in the complement of that one event, whose probability is at most α. These statements can be inspected together without an additional union bound when they use the same interval. They are not a conditional error guarantee given that a directional decision was made. □

These implications are complete only relative to the information that every $D \in [L, U]$ is possible: if $L \le \vartheta$, the endpoint $D = L$ prevents a uniform claim $D > \vartheta$; if $U \ge -\vartheta$, $D = U$ prevents $D < -\vartheta$. Equality at an endpoint does not certify a strict gain. A future random cost requires joint accuracy–cost inference; the corollary does not evaluate a routing policy or demonstrate savings.

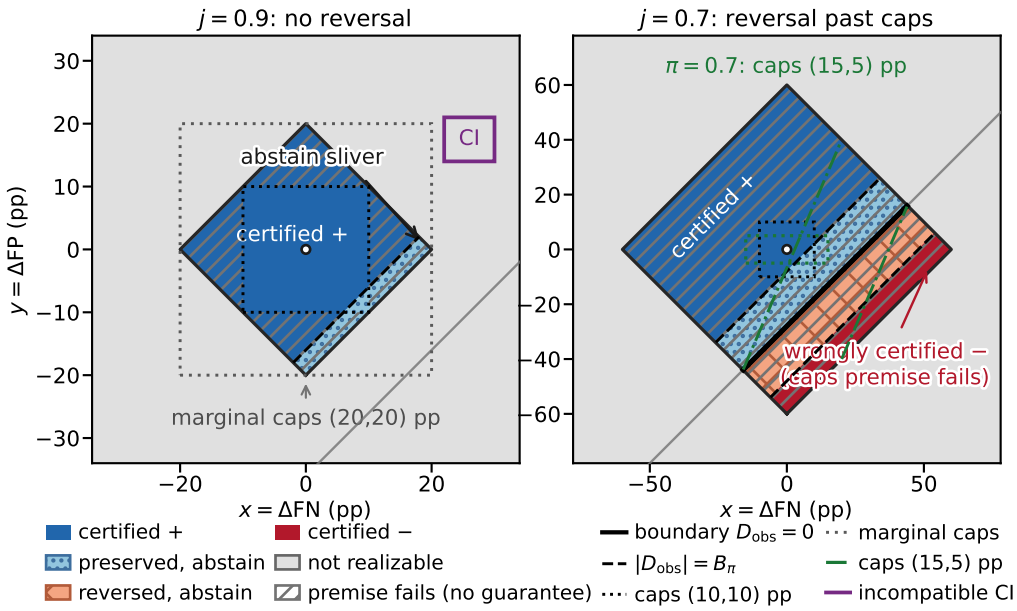

Figure 3: Synthetic decision maps for Theorems 1 and 2 at $\pi = 0.5$, $D = +20$pp, caps $u_f = u_g = 10$pp, with $j = 0.9$ (left) and $j = 0.7$ (right). The green overlay uses $\pi = 0.7$ and caps $(15, 5)$pp. Colors show the sign call or abstention from the rectangular relaxation; gray is outside the feasible region. Hatching marks violations of the base $(10, 10)$pp caps, not of the green overlay's $(15, 5)$pp caps. The dark-red corner is a wrong call only because the base caps premise is violated.

B.2 Validity targets

This subsection collects the definitions behind Section 4.1. Three validity targets organize the analysis. On a task where both truth classes have positive probability in each mode, a grader is *invariant* if $\mathrm{FN}_0 = \mathrm{FN}_1$ and $\mathrm{FP}_0 = \mathrm{FP}_1$. Equivalently, $g \perp a \mid S$, the equalized-odds condition (Hardt et al., 2016) with mode as the protected attribute. It is *gap-unbiased* if $\mathbb{E}[g - S \mid a{=}0] = \mathbb{E}[g - S \mid a{=}1]$. It is *rank-valid* for the fixed comparison if D_{obs} has the sign of D whenever $D \neq 0$. This last property is pointwise for this pair of modes, not uniform over prevalences.

B.3 Extended readings and realization

The supplement (Section S.5) collects extended readings of these results. It develops three readings of Theorem 1: invariance as the checkable sufficient condition for rank validity, the prevalence-dependent weights on the FN and FP channels, and the scaling of the safety margin with $j|D|$. It records the pitfalls that change the answer when the joint confidence set underlying Theorem 2 is constructed, and it states what the decomposition establishes beyond a pooled-accuracy bound.

B.4 Joint-set inversion: elaborations

The supplement (Section S.5) elaborates on two points. The abstention of Theorem 2 is complete: it occurs exactly when a nonempty joint set cannot uniformly certify a strict ordering. A near-degenerate channel ($j \approx 0$) leaves the inversion weakly identified through the $1/j$ divisor, and an exactly degenerate one is excluded from the compatible set.

B.5 The Bounded-Overlap Error Bound

Setup. All quantities here are census quantities over the realized candidates of each mode; only the bridge audit (V2) carries sampling uncertainty. Fix a panel of K graders $g_k(x) \in \{0, 1\}$ and a *content functional* $C(x) \in \{0, 1\}$ that takes the last brace-balanced `\boxed{...}` content, or

otherwise the last numeric token of the comma-stripped text, as the committed answer and compares it to gold by decimal equality ($C := 0$ when no commitment exists; a non-numeric commitment is not code-decidable and the row is *uncovered* for every grader). The content functional is a reference ruler, not semantic truth: the bounds below certify agreement *with the ruler*, and the gap between the ruler and semantic truth Y is audited separately (the bridge δ_a of Corollary 3(ii)). From code facts alone, each grader k assigns one of four states: *safe* (code facts imply $g_k(x) = C(x)$), *probe-safe* (runtime probes check equality on decorated numerics), *hazard* (commitment-blind, deterministically wrong, or conservatively uncertified), or *uncovered* (mass τ_{ak}); a row–grader unit is *certified-safe* when safe or probe-safe, and "safe" below means certified-safe. Write $R(x)$ for the number of graders on which x is non-safe, and for $\rho = 1, \dots, K-1$ let $A_k(\rho) = (K-1)^{-1}\,\mathbb{E}\big[\sum_{\ell \neq k} \mathbf{1}\{g_k \neq g_\ell,\ \ell \text{ safe}\}\,\mathbf{1}\{R{=}\rho\} \,\big|\, a\big]$ be the disagreement mass of k against safe peers at overlap level ρ. Error targets are $e^c_{ak} = \mathbb{P}(g_k \neq C \mid a)$ (content level) and $e^Y_{ak} = \mathbb{P}(g_k \neq Y \mid a)$ against the reference-labeled semantic truth Y; "label-free" means that no candidate-level semantic Y-labels are used. On rows whose commitment is not code-decidable, C is not pinned by the rule; every bound stated below holds for all binary completions of C simultaneously, while any point value of e^c_{ak} is the exact identity value on the decided rows (the anchored disagreement mass plus the decidable $R{=}K$ error mass) and shifts by at most the uncovered mass under any completion (Table 17).

Theorem 3 (Bounded-overlap error bound)**.** *In the setup above, for each mode a and grader k, with $E_k = \mathbf{1}\{g_k \neq C\}$,*

$$e^c_{ak} \;=\; \sum_{\rho=1}^{K-1} \frac{K-1}{K-\rho}\, A_k(\rho) \;+\; \mathbb{P}(E_k{=}1,\, R{=}K \mid a) \;\;\leq\;\; \sum_{\rho=1}^{K-1} \frac{K-1}{K-\rho}\, A_k(\rho) \;+\; \mathbb{P}(R{=}K \mid a), \quad (5)$$

every term on the right of the final upper bound is label-free, and the constant $(K-1)/(K-\rho)$ is pointwise unimprovable.

Three refinements (a full-overlap split, a bridge from content to semantic truth at cost $\delta_a = \mathbb{P}(C \neq Y, R < K \mid a)$, and directional one-sided bounds at half the two-sided width) are stated as Corollary 3.

Proof of Theorem 3. Fix a mode a and grader k; all probabilities are conditional on a. On a row with $R(x) = \rho < K$, every safe grader ℓ satisfies $g_\ell(x) = C(x)$ (the safety property, established per grader by code-fact lemmas and re-verified row-by-row by check V1, both given in the supplement (Section S.5)). Hence if $g_k(x) \neq C(x)$ then k is non-safe on x (contraposition of safe $\Rightarrow$ correct), and exactly the $K-\rho$ safe graders disagree with $g_k(x)$; if $g_k(x) = C(x)$, a safe ℓ agrees with k while a non-safe ℓ may or may not. Therefore, pointwise on $\{R = \rho\}$, $\sum_{\ell \neq k} \mathbf{1}\{g_k \neq g_\ell,\ \ell \text{ safe}\} = (K-\rho)\,\mathbf{1}\{g_k \neq C\}$. Dividing by $(K-1)$, taking expectations, and summing over $\rho = 1, \dots, K-1$ gives the equality in Equation (5) on rows with at least one safe anchor; rows with $R = K$ carry no safe anchor and are accounted at their trivial mass $\mathbb{P}(E_k{=}1, R{=}K) \leq \mathbb{P}(R{=}K)$. *Sharpness.* Take a single candidate x_0 with $C(x_0) = 0$; let k and exactly $\rho - 1$ other graders be hazardous on x_0 with verdict 1, and the remaining $K - \rho$ graders safe with verdict 0. Then $e^c_{ak} = 1$ and the weighted disagreement term equals $\frac{K-1}{K-\rho} \cdot \frac{K-\rho}{K-1} = 1$: equality holds, so no smaller constant suffices at overlap level ρ. □

Corollary 3 (Split, bridge, and directional forms)**.** *(i) If* uncovered implies globally uncovered *(an uncovered row has $R{=}K$), a condition machine-checked with 0 violations in our instantiation, the trivial full-overlap bound splits exactly as $\mathbb{P}(R{=}K \mid a) = p^H_K(k,a) + \tau_{ak}$, where $p^H_K(k,a) = \mathbb{P}(k \text{ hazard},\ R{=}K \mid a)$ upper-bounds the label-dependent hazard-error mass. (ii) Let $\delta_a = \mathbb{P}(C \neq Y,\, R < K \mid a)$ be the bridge residual on rows with at least one safe anchor; then $e^Y_{ak} \leq \sum_\rho \frac{K-1}{K-\rho} A_k(\rho) + \mathbb{P}(R{=}K \mid a) + \delta_a$: disagreement upper-bounds a dangerous grader's deviation from C, the bridge bounds C's deviation from Y, and the full-overlap term covers rows with no safe anchor at trivial mass. (iii) Keeping the* direction *of disagreement separates the channels into one-sided bounds $\overline{\mathrm{FP}}_{ak}, \overline{\mathrm{FN}}_{ak}$ (definitions below), giving one-sided rankings for the* content-level *gap $D^C = \mathbb{P}(C{=}1 \mid a_1) - \mathbb{P}(C{=}1 \mid a_0)$:*

$$D^C \;\geq\; D^{(k)}_{\mathrm{obs}} - \overline{\mathrm{FP}}_{1k} - \overline{\mathrm{FN}}_{0k}, \qquad D^C \;\leq\; D^{(k)}_{\mathrm{obs}} + \overline{\mathrm{FN}}_{1k} + \overline{\mathrm{FP}}_{0k}, \quad (6)$$

whose width is exactly half of the two-sided bound. A directional semantic bound follows from part (ii)'s union bound: $D^Y \geq D^{(k)}_{\mathrm{obs}} - \overline{\mathrm{FP}}_{1k} - \overline{\mathrm{FN}}_{0k} - (\delta_0 + \delta_1)$; the audit of δ is in the supplement (Section S.5).

Directional bounds (definitions). With $A_k^{10}(\rho)$ the $\{g_k{=}1,\ g_\ell{=}0\}$ part of $A_k(\rho)$ (symmetrically A_k^{01}), the directional bounds of Corollary 3(iii) are $\overline{\mathrm{FP}}_{ak} = \sum_{\rho=1}^{K-1} \frac{K-1}{K-\rho} A_k^{10}(\rho) + \mathbb{P}(g_k{=}1,\ R{=}K \mid a)$ and $\overline{\mathrm{FN}}_{ak} = \sum_{\rho=1}^{K-1} \frac{K-1}{K-\rho} A_k^{01}(\rho) + \mathbb{P}(g_k{=}0,\ R{=}K \mid a)$; the same pointwise count gives $\overline{\mathrm{FP}}_{ak} \geq \mathbb{P}(g_k{=}1,\ C{=}0 \mid a)$ and $\overline{\mathrm{FN}}_{ak} \geq \mathbb{P}(g_k{=}0,\ C{=}1 \mid a)$. The width of Equation (6) is exactly half of the two-sided bound $|D_{\mathrm{obs}}^{(k)} - D^C| \leq B_k$, $B_k = \mathrm{bound}_{0k} + \mathrm{bound}_{1k}$ with $\mathrm{bound}_{ak} := \overline{\mathrm{FP}}_{ak} + \overline{\mathrm{FN}}_{ak}$. On $R{=}K$ rows the directional events are already charged at trivial mass inside $\overline{\mathrm{FP}}, \overline{\mathrm{FN}}$, which is why the directional semantic bound subtracts $(\delta_0 + \delta_1)$ only once.

Proof of Corollary 3. (i) Under globality, $\{R{=}K\}$ partitions into $\{k$ hazard, $R{=}K\}$ and $\{k$ uncovered$\}$, so the full-overlap mass splits exactly: $\mathbb{P}(R{=}K) = p_K^H(k,a) + \tau_{ak}$, where $p_K^H(k,a) = \mathbb{P}(k$ hazard, $R{=}K)$ is a hazard *mass*. The error residual splits as the exact event partition $\mathbb{P}(E_k{=}1, R{=}K) = \mathbb{P}(E_k{=}1,\ k$ hazard, $R{=}K) + \mathbb{P}(E_k{=}1,\ k$ uncovered), whose two terms are bounded above by $p_K^H(k,a)$ and τ_{ak} respectively. The hazard-error mass itself is label-dependent and not identified. Without the globality condition, rows uncovered only for k sit at $R < K$ yet remain charged in τ_{ak}, so $p_K^H + \tau$ overcounts $\mathbb{P}(R{=}K)$ and the exact split is lost; the condition is machine-checked on the full corpus (0 violations) rather than assumed. (ii) On rows with $R < K$, $g_k \neq Y$ implies $g_k \neq C$ or $C \neq Y$; union-bounding and adding the trivial $R{=}K$ accounting gives $e_{ak}^Y \leq \sum_\rho \frac{K-1}{K-\rho} A_k(\rho) + \mathbb{P}(R{=}K) + \delta_a$. (iii) Restrict the pointwise identity of Theorem 3 to the direction $\{g_k = 1, g_\ell = 0\}$: on $\{R = \rho\}$, $\sum_{\ell \neq k} \mathbf{1}\{g_k{=}1, g_\ell{=}0,\ \ell$ safe$\} = (K-\rho)\,\mathbf{1}\{g_k{=}1, C{=}0\}$, since safe peers agree with C and the direction forces $C = 0$. Expectations give $\overline{\mathrm{FP}}_{ak} = \sum_\rho \frac{K-1}{K-\rho} A_k^{10}(\rho) + \mathbb{P}(g_k{=}1, R{=}K) \geq \mathbb{P}(g_k{=}1, C{=}0)$, and symmetrically for $\overline{\mathrm{FN}}_{ak}$. Writing $q_a = \mathbb{P}(g_k{=}1 \mid a)$ and $s_a = \mathbb{P}(C{=}1 \mid a)$, we have $s_a \in [q_a - \overline{\mathrm{FP}}_{ak},\ q_a + \overline{\mathrm{FN}}_{ak}]$, and Equation (6) follows from $D^C = s_1 - s_0$ and $D_{\mathrm{obs}}^{(k)} = q_1 - q_0$; the interval is for the content-level gap. For the semantic gap, apply part (ii)'s union bound directionally: on $R < K$ rows $\{g_k{=}1, Y{=}0\}$ implies $\{g_k{=}1, C{=}0\}$ or $\{C \neq Y\}$, and on $R{=}K$ rows the event is contained in $\{g_k{=}1, R{=}K\}$, already charged at trivial mass inside $\overline{\mathrm{FP}}_{ak}$; hence $\mathbb{P}(g_k{=}1, Y{=}0 \mid a) \leq \overline{\mathrm{FP}}_{ak} + \delta_a$, and symmetrically $\mathbb{P}(g_k{=}0, Y{=}1 \mid a) \leq \overline{\mathrm{FN}}_{ak} + \delta_a$. It follows that $D^Y \geq D_{\mathrm{obs}}^{(k)} - \overline{\mathrm{FP}}_{1k} - \overline{\mathrm{FN}}_{0k} - (\delta_0 + \delta_1)$, and symmetrically for the upper side. (A direct lift $D^Y \geq D^C - (\delta_0 + \delta_1)$ is *not* valid under this definition of δ, which does not control C–Y deviation on $R{=}K$ rows; the grader-based form is the correct one.) The *content-level* interval width is $\overline{\mathrm{FP}}_{1k} + \overline{\mathrm{FN}}_{0k} + \overline{\mathrm{FN}}_{1k} + \overline{\mathrm{FP}}_{0k} = B_k$, half of the symmetric content width $2B_k$; the semantic-lifted interval is wider by the bridge masses, $B_k + 2(\delta_0 + \delta_1)$. □

Proposition 4 (Direct safe-anchor content bound)**.** *In the setup of Theorem 3, write* $c_a = \mathbb{P}(C{=}1,\ R{<}K \mid a)$ *for the content mass identified by safe anchors and* $u_a = \mathbb{P}(R{=}K \mid a)$ *for the anchor-free mass. Then*

$$\mathbb{P}(C{=}1 \mid a) \ \in\ [\,c_a,\ c_a + u_a\,], \qquad D^C \ \in\ [\,c_1 - c_0 - u_0,\ c_1 - c_0 + u_1\,]. \tag{7}$$

Both endpoints are attainable in the relaxation that treats the content labels of the $R{=}K$ rows as unconstrained binary values, so the interval is sharp relative to the row-wise safe states and verdicts alone. Imposing the $C{:=}0$ convention for no-commitment rows or other code-decidable labels on $R{=}K$ rows can only narrow it.

Proof. On $\{R < K\}$ every safe grader ℓ satisfies $g_\ell(x) = C(x)$ (the safety property established per grader by code-fact lemmas and re-verified row-by-row by check V1, both given in the supplement (Section S.5)), so C is identified row-wise and c_a is computable from the same audit records as Equation (5). On $\{R = K\}$ no safe anchor exists and C ranges freely in $\{0,1\}$, which gives $\mathbb{P}(C{=}1 \mid a) \in [c_a,\ c_a + u_a]$; the gap form follows from $D^C = \mathbb{P}(C{=}1 \mid 1) - \mathbb{P}(C{=}1 \mid 0)$. For sharpness, assign all $R{=}K$ rows of mode 1 the label $C{=}0$ and all $R{=}K$ rows of mode 0 the label $C{=}1$ to attain the lower endpoint, and exchange the assignments for the upper endpoint. *Dominance.* For every grader k and mode a, $q_a - \overline{\mathrm{FP}}_{ak} = \mathbb{P}(g_k{=}1, C{=}1, R{<}K \mid a) \leq c_a$ and, symmetrically, $q_a + \overline{\mathrm{FN}}_{ak} = c_a + u_a + \mathbb{P}(g_k{=}1, C{=}0, R{<}K \mid a) \geq c_a + u_a$, where $q_a = \mathbb{P}(g_k{=}1 \mid a)$ and the equalities expand Corollary 3(iii). Hence the interval $[c_a, c_a + u_a]$ lies inside the directional interval of Corollary 3(iii) for every grader, with slack exactly the grader's anchored directional-error masses. □

B.6 THE RESIDUAL HORVITZ–THOMPSON ESTIMATOR

The residual estimator D_{res} of Section 3.2 is the direct design-based form of the correction. Verdicts g_i are known on all N_a candidates of mode a, while truth S_i is known only on the audit sample $\mathcal{S}_a$. Let the audit inclusion probabilities be $\rho_i > 0$:

$$\hat{A}_a^{\mathrm{HT}} \;=\; A_{\mathrm{obs},a} \;+\; \frac{1}{N_a} \sum_{i \in \mathcal{S}_a} \frac{S_i - g_i}{\rho_i} \tag{8}$$

is design-unbiased for A_a with respect to the sampling design, provided the reference labels equal truth, the inclusion probabilities are known, and the design is executed as specified; the residual gap is $D_{\mathrm{res}} := \hat{A}^{\mathrm{HT}}_{\text{think}} - \hat{A}^{\mathrm{HT}}_{\text{no-think}}$ (the superscript HT marks the Horvitz–Thompson *residual*, not the label-only mean D_{HT}). D_{HT} is computed from labels alone with Horvitz–Thompson weights, while D_{res} anchors on the census verdicts and uses the labeled rows only for the bias-correction term. It requires no condition on the channel, no inversion, and no parametric assumption. Sampling positivity (every candidate auditable) and channel invertibility are distinct conditions and must not be conflated: Equation (1) needs $J_a \neq 0$ in each mode, while Theorem 2 divides by $j = \bar{J}$ and needs $\bar{J} \neq 0$. The audit is decision-oriented rather than significance-oriented; the sizing recipe is in the supplement (Section S.5).

Justification of Equation (8). By the Horvitz–Thompson identity (Särndal et al., 1992), $\mathbb{E}\big[\sum_{i \in \mathcal{S}_a} (S_i - g_i)/\rho_i\big] = \sum_{i=1}^{N_a} (S_i - g_i) = N_a(A_a - A_{\mathrm{obs},a})$ whenever every $\rho_i > 0$; adding $A_{\mathrm{obs},a}$ gives design-unbiasedness of $\hat{A}_a^{\mathrm{HT}}$. No property of the channel enters. □

B.7 INSTANTIATION ON GSM8K AND THE SEMANTIC BRIDGE

The supplement (Section S.5) reports the instantiation of the bound on the fixed GSM8K corpus with the in-house grader panel, including the safety verification pass and the resulting one-sided content-level bounds. It also covers the bridge audit that a lift from content correctness to semantic truth would require.

C PROTOCOL DETAILS

This appendix gives the protocol facts needed to read the main text: the grader panels, the item-selection rule, the reference-labeling design, and the mainstream panel's provenance and scope. Verbatim instruments, hash records, name mappings, and per-stratum count tables are in the supplement (Sections S.4 and S.6).

C.1 GRADER PANEL

Table 4: The in-house mechanism panel: seven answer-extraction graders from our own evaluation history, plus one appendix-only probe (`math_v4_normfix`, below the rule), for eight displayed rows. `mc_v2_buggy` reproduces a historical production bug (cross-line letter bleed) for calibration; `math_v4_normfix` is used only as a mechanism probe, never in multiplicity families.

Grader	Family	One-line definition
`mc_v1_fulltext`	MC	letter extracted from full candidate text
`mc_v2_buggy`	MC	v1 with the historical cross-line regex bug
`mc_v3_linesafe`	MC	line-safe horizontal-whitespace extraction
`math_v1`	math	boxed/final numeric-string match vs. gold
`math_v2_seg`	math	math_v1 on the answer-carrying final segment
`em_v1_firstans`	EM	first answer string, exact/alias match
`em_v2_seg`	EM	em_v1 on the final segment
`math_v4_normfix`	math (appendix)	v1 + boxed comma/currency normalization

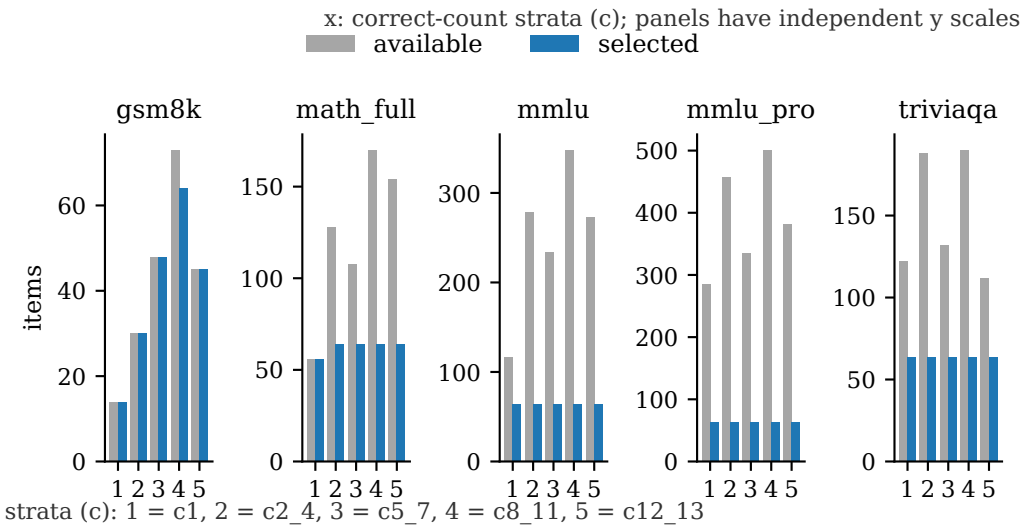

Figure 4: Item selection for the five benchmark-derived, difficulty-stratified subsets: available vs. selected items per screening stratum.

C.2 Item Sets and Selection

Items are grouped by screening-accuracy strata (pilot correct-count bands c1, c2–4, c5–7, c8–11, and c12–13) and up to 64 items per stratum are retained; strata with fewer available items are exhausted (Figure 4). The pilot screen used earlier generations of the same model; the mode and grading contract of that screen are not part of the frozen record, and we list this as a selection-effect limitation rather than reconstruct it from the production configuration. The seed used to pick 64 items within over-cap bands was not recorded, so within-band retention follows the frozen screening record and is not a probability sample. The cap limits concentration in the more populous screening strata (Section 3.1); it does not guarantee sufficient correct or incorrect reference labels within each stratum. The two modes share identical item sets; all estimands are conditional on these fixed sets. Per-band availability and retention counts, and the frozen item-list hashes, are in the supplement (Section S.4).

C.3 Study Flow and Candidate Corpus

The study flow is summarized in Figure 5.

Generation configuration. Both modes share the model checkpoint, the serving engine (vLLM 0.11.0), and the decoding parameters (temperature 0.6, top-p 0.95). The two modes differ in the chat-template thinking switch (`enable_thinking` on or off) and its associated prompt template, in the output-token cap (at most 8,192 tokens per no-think candidate and 32,768 per think candidate), and in the number of candidates per item (128 no-think, 8 think). Sampling seeds are per generation call, so the two modes draw disjoint seeded streams; per-candidate seeds were not recorded, individual candidates cannot be re-generated bit-exactly, and the frozen shards are the auditable record. The full seed scheme, the per-corpus generation protocol (including the five transfer models and the probe corpora), and the per-task output-token accounting are in the supplement (Section S.4).

C.4 Reference Labels

Sampling strata and weights. For MMLU and MMLU-Pro, we sampled 40 rows per mode from the all-agree stratum and 40 from the disagree stratum. For the math tasks and TriviaQA, we sampled 40–60 rows per mode and exhausted each small disagreement stratum. Rows carry Horvitz–Thompson weights (inverse of within-stratum inclusion probabilities (Horvitz & Thompson, 1952), which range from 0.10% to 100% across strata; the aggregate mode-level sampling fractions n/N are 0.1–0.2%

no-think and 1.6–3.1% think). Population sizes, sample sizes, and inclusion probabilities by stratum are in the supplement (Section S.6).

Prescreen judge and blinding. A commercial LLM (vendor withheld for review) labeled records in batches, seeing only the question, gold, and candidate text plus record header fields (sample id, task, item id, candidate index); explicit mode, accuracy, and grader-verdict fields were hidden. The judge is therefore metadata-blind but not necessarily mode-blind: candidate indices can reveal the mode through the 128-versus-8 counts, and thinking blocks remain visible. The verbatim prompt template, driver configuration, and hash records are in the supplement (Section S.4).

Second pass and human adjudication. A separate LLM agent (a different vendor, without access to the first pass's rationales) re-adjudicated the 68 semantic rows — all 32 label–grader mismatch rows and 36 stratified concordant rows — and a mechanical regular-expression re-extraction checked all 320 multiple-choice rows. Neither check found a discrepancy. The two semantic passes are not fully independent lineages (both are LLM-based), and the mechanical layer checks extraction consistency, not gold correctness; we report the outcome as no detected disagreement in the second pass (388 rows), not as human confirmation. A 132-row human adjudication sample — all flagged records plus a $\geq$ 20% seeded random draw per (task, mode, stratum) cell — confirmed every prescreen label (132/132, 0 changes). The adjudicator saw the prescreen verdicts and rationales, so this review is a confirmation review, not a blind re-labeling (anchoring caveat in Section 6). The 13 flagged rows comprise eight output-side issues (three truncations, all no-think; five conflicting or contradictory final answers) and five item-suspect rows; the underlying items are excluded wholesale in the item-exclusion sensitivity (supplement Section S.6). A labeling-instrument matrix covering every reference-label cohort of the paper is in the supplement (Section S.4).

Label-error safeguards. The semantic second-judge rows include all 32 label–grader mismatch rows, which carry the entire H1/H3a signal (case-level examples in Appendix D.12); the remaining second-pass checks mechanically re-extract multiple-choice labels.

C.5 Semantic-label extension

We sampled 400 additional candidates from three all-agree strata (150 GSM8K no-think, 90 GSM8K think, and 160 MATH think rows). A commercial code-oriented LLM (vendor withheld for review) supplied a prescreen; every row then received either a human confirmation field or a recorded adjudication by the study owner, yielding 393 binary semantic judgments and seven held records (ambiguous, inconsistent, or under-specified questions). Unlike the round-1 595-row rubric, these judgments may override a benchmark gold answer or recognize unit and option-format equivalence, so they target semantic correctness Y, not exactly the round-1 benchmark-gold reference variable S; they enter only the bridge and contract-mismatch sensitivity of Appendix D.16, never the primary Horvitz–Thompson estimate. For the bridge check, we recompute the recorded non-safe-grader count R on every observed $C \neq Y$ record (C, Y, R, K, and δ_a as defined in Appendix B.5 and Corollary 3). All five known violations satisfy $R < K{=}4$, so their binary-row count matches the event in $\delta_a = \mathbb{P}(C \neq Y, R < K \mid a)$. The seven held records are handled separately rather than assigned an event status.

C.6 Mainstream extension panel

Composition and provenance. The panel includes lm-evaluation-harness GSM8K strict-match and flexible-extract (Gao et al., 2024); the Qwen2.5-Math official math grader and multiple-choice extractor (Yang et al., 2024); `math_verify` in its Open R1 `accuracy_reward` configuration (Hugging Face, 2025c;d); the MMLU-Pro official extraction chain (Wang et al., 2024); the simple-evals MMLU extractor (OpenAI, 2023); and the official TriviaQA exact-match scorer (Joshi et al., 2017). Upstream grader source files are archived byte-verbatim, with SHA-256 re-verified at import and each file anchored to an upstream commit by blob-hash lookup. Graders run as whole files where possible and otherwise as function-level slices along definition boundaries. Four execution adaptations are recorded: two dependency shims (a `latex2sympy2` module shim and a `sympy` parser shim, both forced by ANTLR runtime conflicts), a TriviaQA adapter that extracts the text after the last `Answer:` marker before official normalization and alias matching, and a line-by-line

reimplementation of the lm-eval GSM8K strict and flexible pipelines from the vendored task YAML and filter source, regression-checked against fixed probes. Two further lm-eval MATH variants (Hendrycks and Minerva) were evaluated and excluded: their extraction contracts bind prompt templates absent from our chat-style candidates, so essentially every row falls back and no applicable channel exists to report. The three retained prompt-contract-degenerate cells stay in the fixed menu as registered-degenerate cells, reported as format-mismatch diagnostics with their count intervals, not as portability or correction evidence (Table 5). Display-name and artifact-key mappings, the full provenance ledger, the complete evaluated-contract table, and the adapter self-check are in the supplement (Section S.4).

Panel selection and inferential scope. The mainstream panel was an exploratory extension assembled after the in-house audit results were known. The retained set keeps graders whose extraction contracts apply to our chat-style candidates across at least one task, and reports their degenerate cells; letter-coverage restrictions also exclude applying A–D/A–E extractors to ten-option MMLU-Pro. These documented decisions do not establish that the broader grader search was independent of the labels. The current count intervals are therefore fixed-menu, reference-rule analyses. Their numerical coverage guarantee holds when the reported graders are treated as fixed independently of label sampling; it does not cover an undocumented label-dependent search or establish a prospective confirmatory result. The same limitation applies to post-hoc transfer panels. Census score differences do not require this sampling-inference assumption. Each extension grader re-scored all candidates in its applicable tasks (496,672 verdicts, zero timeouts). The reference strata were designed on in-house-panel disagreement, not mainstream-grader disagreement, so several extension intervals are wide, and our candidate text departs from some graders' prompt contracts, so the degenerate cells are a format-mismatch finding, not a defect claim about the grader.

D Extended Results

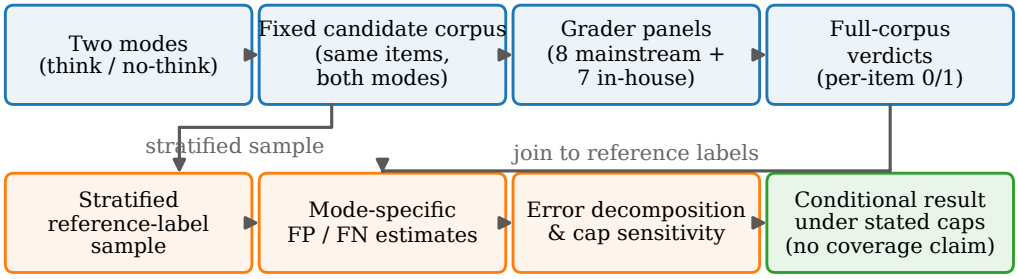

Figure 5: Study flow. Candidate generations from two modes form the fixed candidate corpus, which two grader panels (mainstream and in-house) score into full-corpus verdicts. A stratified reference-label sample yields mode-specific FP and FN estimates; these feed error decomposition and a conditional sensitivity analysis. The empirical cap-based decisions do not carry the coverage guarantee of the abstract confidence-set theorem.

Current fixed-population analyses, new generations, parser interventions, and transfer comparisons appear first. Per-stratum count tables are in the supplement (Section S.6) and meta-audit details in the supplement (Section S.7), pointed to where they arise. Fixed-menu interval guarantees do not adjust for earlier grader search (Appendix C.6).

D.1 Finite-population reference-label inference

Sampling target and validity. Let h index an original task–mode–agreement stratum and $g \in \{0, 1\}$ a fixed grader verdict. The complete corpus gives N_{hg}. Conditional on observing n_{hg} sampled records in that group, the k_{hg} reference-correct records follow $\mathrm{Hypergeom}(N_{hg}, K_{hg}, n_{hg})$, where K_{hg} is the unknown population count. The conditional probability of any sample with those group sizes is $\prod_g \binom{N_{hg}}{n_{hg}}^{-1}$, so group membership does not invalidate simple random sampling. We preserve the original agreement strata throughout.

For each group we retain the integer K satisfying both $\Pr_K(X \geq k) > \alpha_j/2$ and $\Pr_K(X \leq k) > \alpha_j/2$ ($[0, N]$ with no sampled records; $[k, k]$ in a census group). For a cell of J nonempty groups with cell budget α_c, set $\alpha_j = \alpha_c/J$ for every group $j = 1, \ldots, J$; the union bound gives at least $1 - \alpha_c$ coverage for all its counts jointly. If R and T denote reference-correct totals among rejected and accepted outputs, respectively, $\mathrm{FN} = R/(R + T)$; the analogous complementary counts define FP. Monotonicity yields the extreme ratios and mode differences over the count set. If zero denominators are compatible with it, the result is explicitly allowed to be undefined rather than assigned a zero error rate.

The per-cell (pointwise) column uses $\alpha_c = 0.05$. For simultaneous reporting, each of the 13 mainstream cells receives $\alpha_c = 0.05/26$ and each of the 12 in-house cells $\alpha_c = 0.05/24$ before group allocation; this retains the original factor of two for FN and FP differences, although both use the same cell count event. Each panel therefore has at least 97.5% coverage, and both panels together at least 95%. The per-grader guarantee holds conditionally on that grader's observed group counts, and the union bound makes it valid unconditionally over their draw; it is not a joint guarantee over all graders' margins, it excludes task-mean intervals and normal-approximation diagnostics, and it requires graders fixed independently of the audited labels.

Point estimates and linear comparisons. Published HT ratio points are kept for comparability. They do not condition on the grader-specific sample counts, whereas the count intervals do; a ratio point need not lie inside such a conditional interval (this occurs for the registered-degenerate MATH strict cell in Table 5). For task means, $\hat{\mu}_a = N_a^{-1} \sum_h N_{ah} \bar{S}_{ah}$ is design-unbiased. The residual estimate adds the exact population grader mean to the HT mean of $S - G$. Its estimated variance is

$$\widehat{V}_a(Z) = N_a^{-2} \sum_{h:\, n_{ah} > 0} N_{ah}^2 (1 - n_{ah}/N_{ah}) s_{Z,ah}^2 / n_{ah}, \qquad Z = S \text{ or } S - G,$$

over strata with $n_{ah} > 0$; empty strata do not enter the task-mean variance. The covariance of label and grader is retained through $S - G$; an additional independent grader resample would be inappropriate. A constant-residual sample can have estimated variance zero even when unobserved residuals vary, so a degenerate normal interval is not evidence of no error.

Verification and scope. All 595 sample keys are reproduced using the original allocation and seed, and each grader's complete candidate key set, binary verdicts, stratum sizes, and sample weights are checked. Exhaustive count inversion on all 2,469 designs with $N \leq 18$ gives a minimum coverage of 0.9502, and synthetic checks cover sparse and zero channels and the census limit (the supplement documents the calibration scenarios). These checks validate implementation boundaries, not reference labels. The original reference convention, the later blind-rule judgments, and the gold-revising semantic extension remain distinct targets.

Additional design calibration. Synthetic calibration tests the original MMLU agreement-stratum design and the follow-up's eight fixed verdict strata using their population sizes and baseline sample allocations, with synthetic labels fixed before sampling. Each of ten scenarios per design, including a perfect proxy, mode bias, thin denominators, nonrandom missing labels, an unsupported stratum, and a census, uses 5,000 draws. Joint count coverage ranges from 0.9696 to 1.0000 across these 20 cases;

undefined target rates remain undefined and missing labels are not deleted. These simulations test the specified implementations and designs, not human-label accuracy.

We separately execute the historical source functions on exact counterexamples: where the true fixed-population label-sampling variance is zero, the item and two-phase bootstrap FN-gap variances remain positive, and repeated resampling cannot repair that target mismatch; this is why the historical methods were stopped.

Table 5: Current fixed-population FN-gap inference in percentage points. The menu is treated as fixed; earlier grader search is not adjusted for. All 25 cells are retained, including prompt-contract-degenerate strict graders. Point estimates use the original HT ratios. Per-cell and simultaneous intervals use known grader group sizes and exact count inversion, so they need not be centered on those points. FN no-think and FN think are the per-mode HT-ratio point estimates; the FN gap is FN think minus FN no-think. An endpoint displayed as -0.0 is a strictly negative unrounded endpoint, not an exact zero; whether an interval contains zero is always decided on unrounded values. †: registered prompt-contract-degenerate cell (a format-mismatch diagnostic, not correction evidence; Appendix C.6). FP bounds and denominator flags, and the supporting group counts, are tabulated in the supplement (Section S.6).

Task	Grader	FN no-think	FN think	FN gap	Per-cell 95%	Simultaneous
Mainstream panel						
gsm8k	lm-eval flexible	58.6	75.0	$+16.4$	$[+3.3, +42.0]$	$[-4.6, +50.9]$
gsm8k	lm-eval strict†	100.0	100.0	$+0.0$	$[-0.1, +0.0]$	$[-0.1, +0.0]$
gsm8k	math_verify	0.0	0.0	$+0.0$	$[-12.3, +10.2]$	$[-18.4, +13.1]$
gsm8k	qwen_math	0.0	0.0	$+0.0$	$[-12.2, +10.2]$	$[-18.3, +13.1]$
math_full	lm-eval flexible†	95.7	100.0	$+4.3$	$[-0.8, +8.9]$	$[-0.9, +10.8]$
math_full	lm-eval strict†	100.0	97.4	-2.6	$[-0.7, -0.0]$	$[-0.7, -0.0]$
math_full	math_verify	56.5	23.7	-32.8	$[-48.6, +7.7]$	$[-62.2, +16.2]$
math_full	qwen_math	13.0	0.0	-13.0	$[-34.7, +6.7]$	$[-44.1, +9.1]$
mmlu	mmlupro_official	14.7	7.4	-7.3	$[-36.6, +23.4]$	$[-48.3, +31.5]$
mmlu	qwen_choice	5.3	50.9	$+45.6$	$[+11.8, +77.0]$	$[-5.8, +81.8]$
mmlu	simple_evals_mc	0.0	0.0	$+0.0$	$[-25.5, +21.4]$	$[-37.4, +29.3]$
mmlu_pro	mmlupro_official	13.1	8.7	-4.4	$[-35.3, +17.1]$	$[-45.9, +23.0]$
triviaqa	triviaqa_em_aliases	12.0	17.2	$+5.2$	$[-37.7, +31.3]$	$[-49.8, +39.1]$
In-house panel						
gsm8k	math_v1	37.9	0.0	-37.9	$[-49.9, -8.7]$	$[-56.6, -0.8]$
gsm8k	math_v2_seg	37.9	0.0	-37.9	$[-49.9, -8.7]$	$[-56.6, -0.8]$
math_full	math_v1	43.5	13.2	-30.3	$[-49.2, +3.7]$	$[-60.4, +10.5]$
math_full	math_v2_seg	43.5	13.2	-30.3	$[-49.2, +3.7]$	$[-60.4, +10.5]$
mmlu	mc_v1_fulltext	0.0	0.0	$+0.0$	$[-25.2, +21.4]$	$[-36.8, +29.1]$
mmlu	mc_v2_buggy	14.7	8.3	-6.4	$[-29.5, +18.6]$	$[-41.0, +27.0]$
mmlu	mc_v3_linesafe	0.0	0.0	$+0.0$	$[-25.2, +21.4]$	$[-36.8, +29.1]$
mmlu_pro	mc_v1_fulltext	0.0	0.0	$+0.0$	$[-23.7, +14.8]$	$[-33.8, +20.2]$
mmlu_pro	mc_v2_buggy	13.5	22.0	$+8.5$	$[-16.8, +25.5]$	$[-27.7, +31.6]$
mmlu_pro	mc_v3_linesafe	0.0	0.0	$+0.0$	$[-23.7, +14.8]$	$[-33.8, +20.2]$
triviaqa	em_v1_firstans	0.0	3.0	$+3.0$	$[-24.2, +18.6]$	$[-35.4, +25.1]$
triviaqa	em_v2_seg	0.0	2.9	$+2.9$	$[-24.4, +18.4]$	$[-35.6, +25.0]$

The per-stratum population counts, sampled counts, current FP bounds, and record disposition behind these intervals are in the supplement (Section S.6).

D.2 Separately sampled answer-extraction follow-up

Design and label targets. This follow-up is distinct from the earlier 400-record semantic extension and from the 4,160 new generations. Before receiving labels, seed 2026092104 selected 400 MMLU candidates by simple random sampling without replacement within eight fixed Qwen-choice verdict strata. Qwen3-8B contributes 60 candidates in each mode–verdict stratum (240 total); DeepSeek-V3.2 and MiMo-V2-Flash each contribute 40 per verdict among think candidates 8–15 (80 each). The

sample covers 236 distinct questions; repeated candidates share a question answer. There was no result-dependent top-up.

The return provides 400 final-answer extraction records and 236 question answers. We evaluate two separate targets: agreement of an extracted final letter with original benchmark gold, and agreement with the returned question-answer key; neither target assesses the full reasoning trace. Three candidates have no final answer and remain held under the frozen protocol, and one extraction transcription is corrected from C to B because the original answer segment is exactly "Answer: B".

The initial return had three conflicts between its summary and CSV, resolved in favor of the CSV answers; one A/B ambiguity and four questions marked truncated stay unresolved for that target. With the three held extractions, 386 labels are determined and 14 unresolved; the benchmark target has 397 determined and three unresolved (per-question resolutions in the supplement, Section S.6).

The submitter confirms that the question answers were produced by a human without seeing benchmark gold or candidate responses. This is self-reported annotation provenance, not an independently observed blinding procedure or evidence of agreement between independent raters. The results below are conditional on the supplied reference answers; their intervals do not account for errors in determined labels.

Unknown labels and simultaneous intervals. For stratum h, let k_h be the known-correct count and u_h the unresolved count among the original n_h sampled records. With $[L_h(k), U_h(k)]$ denoting the hypergeometric interval in Appendix D.1, use

$$[L_h(k_h), U_h(k_h + u_h)], \qquad \alpha_h = 0.05/8.$$

Monotonicity makes these the extremes over all binary completions, so the interval contains the one for the true completion and arbitrary label missingness cannot reduce its coverage; no missing-at-random assumption or complete-case reweighting is used. The eight-count event simultaneously bounds FN, FP, reference accuracy, and the Qwen mode differences, separately for each label target, conditional on fixed target labels and correct observed annotations; it does not cover annotation error or combine both targets into a single 95% family. Candidate-wise completions relax the shared-question-answer constraints, so these are conservative outer envelopes.

Table 6: Current follow-up analysis. Qwen-choice follow-up on MMLU. Intervals use all eight fixed sampling strata jointly for each reference target separately, including all unresolved-label completions. These are conditional reference results, not verified semantic-error rates. FN is in percent; mode differences are in percentage points. The first four rows give the unweighted sample marginals behind the Qwen estimates (n=240); the intervals below invert the stratum counts tabulated in the supplement (Section S.6).

Quantity	Benchmark gold	Returned question key
Qwen no-think FN (sample)	$1/61$ (1.6)	$12/67$ (17.9)
Qwen think FN (sample)	$38/85$ (44.7)	$36/81$ (44.4)
Qwen no-think ref. accuracy (sample)	$61/117$ (52.1)	$67/115$ (58.3)
Qwen think ref. accuracy (sample)	$85/120$ (70.8)	$81/120$ (67.5)
Qwen think–no-think FN gap	$[+31.7, +72.9]$	$[+11.6, +65.2]$
Qwen reference-accuracy gap	$[-8.2, +39.0]$	$[-23.7, +38.6]$
DeepSeek think 8–15 FN (%)	$[25.2, 43.2]$	$[25.5, 46.4]$
MiMo think 8–15 FN (%)	$[19.0, 38.3]$	$[25.0, 48.3]$

Results and scope. Qwen's benchmark-reference FN-gap HT ratio points span $[+56.2, +61.1]$ over held completions; under the returned key they span $[+38.7, +43.8]$. These point ranges exclude sampling uncertainty and are not confidence intervals. The simultaneous intervals in Table 6 remain positive for this FN gap, but both reference-accuracy intervals contain zero. DeepSeek and MiMo now have direct reference-key evidence for the previously uncovered later-think domain; this sample has no new no-think labels for those models and does not identify their mode gaps. As a sensitivity only, accepting the CSV answers even for the ambiguous and truncated questions gives a Qwen FN-gap interval of $[+11.6, +64.1]$; the conservative unresolved-label result is retained as the main returned-key analysis. Concrete checks caution against treating either key as semantic truth; marking

every candidate attached to four flagged questions unresolved gives a Qwen FN-gap interval of $[+10.3, +66.2]$ points (per-question checks in the supplement, Section S.6). This is a diagnostic on the same sample, not protection against undetected reference errors.

D.3 Paired reference evaluation of answer-segment extraction

We reuse the follow-up's 240 Qwen3-8B records, 60 per original mode–Qwen-choice-verdict stratum, to compare the fixed full-text rule with the same rule after removal of the thinking segment. Both rules continue to use benchmark gold; only the reference used to assess their errors changes. Both population verdict vectors were recorded before the follow-up draw and its returned labels, and all 43,520 verdict pairs match that earlier artifact, so the repair and its change sets do not use this follow-up's sampled identities or labels. This retrospective analysis reuses the follow-up; it is not another independent sample.

Every sampled text hash and both verdicts were recomputed against the cached population table. In the fixed population, 982 of 1,733 rejected think outputs become accepted, and 165 of 827 accepted outputs become rejected. The 40,960 no-think verdicts remain identical. Thus a score increase alone cannot decide whether the repair helps: the accepted-to-rejected changes also require reference checks.

Table 7: Current paired audit. Paired Qwen-choice audit of the answer-segment intervention on MMLU (an input-segmentation change; no generated answer is repaired). Counts are among the 51 changed sampled think verdicts, not unweighted population rates. The final column bounds the reduction in total grading-error rate on all 2,560 think outputs, in percentage points. Each target separately has a 95% joint count guarantee across four original sampling strata, conditional on its known reference labels. No-think reduction is exactly zero because its population verdicts do not change. Before/after marginal error rates on the labeled records are in the supplement (Section S.6).

Reference	Repaired	Introduced	Unknown	Error reduction
Benchmark gold	51	0	0	$[31.4, 44.8]$
Returned question key	47	4	0	$[20.1, 43.9]$
Flagged keys unresolved	47	2	2	$[20.1, 44.6]$

Writing R for repaired, I for introduced, and U for pairs erroneous under both rules (necessarily unchanged verdicts), the error counts on the labeled think sample satisfy $E_{\text{before}} = R + U$ and $E_{\text{after}} = I + U$, so $E_{\text{before}} - E_{\text{after}} = R - I$: for benchmark gold $R = 51$, $I = 0$, $U = 0$; for the returned key $R = 47$, $I = 4$, $U = 4$. These before/after totals are sample statements: the construction inverts only the changed-verdict subpopulation, so the population claim of this section is the error *reduction* interval, not a population before/after error rate. As a robustness budget for the returned key (`miaudit_repair_robustness_2026-09-25.json`), we recomputed the lower endpoint after adversarially reclassifying repairs as introduced errors, in question-consistent units: it stays positive through 8 flips and turns negative at 9. This is a sensitivity bound on undetected returned-key errors, not a claim that the key has them.

In the sample, 38 changes are rejected-to-accepted and 13 are accepted-to-rejected (per-stratum counts in the supplement, Section S.6). Against benchmark gold, all 51 fix grading errors: the latter 13 remove accidental acceptances. Against the returned question key, the respective repaired/introduced counts are $35/3$ and $12/1$, an FN-side versus FP-side transition split. Marking the four previously flagged questions unresolved changes two of the four introduced-error labels to unknown. On the labeled think records the segment channel has the fewest verdict errors under both keys ($0/120$ under benchmark gold and $8/120$ under the returned key, versus $51/120$ full text); the four-channel comparison is tabulated in the supplement (Section S.6).

Paired finite-population interval. Let g_i and g'_i be the original and repaired binary verdicts, both computed against benchmark gold, and let $Y_i^{(r)} \in \{0, 1\}$ indicate agreement with reference convention r. Inference assumes fixed potential labels over the full candidate population; it measures agreement with these reference labels, not what either parser would score against a different key. In original sampling stratum h, let $C_h = \{i : g_i \neq g'_i\}$ and $M_h = |C_h|$. Conditional on the observed

number n_h^c of sampled changed pairs, these records are a simple random sample without replacement from that subpopulation, and exactly one verdict of every changed pair agrees with $Y_i^{(r)}$. Writing $B_h = \sum_{i \in C_h} \mathbf{1}\{g_i \neq Y_i^{(r)}\}$, the population reduction in error count is exactly $2B_h - M_h$. For k_h known repairs and u_h unresolved changed pairs, invert the hypergeometric tails as in Appendix D.1 over population size M_h, sample size n_h^c, and $B_h \in \{0, \ldots, M_h\}$, using the lower endpoint at k_h and the upper at $k_h + u_h$, denoted $[L_h, U_h]$, with $\alpha_h = 0.05/4$. If mode a contains $N_a > 0$ outputs, its error-rate reduction lies in

$$\left[\frac{\sum_{h \in a}(2L_h - M_h)}{N_a}, \frac{\sum_{h \in a}(2U_h - M_h)}{N_a} \right].$$

The guarantee follows from conditional simple random sampling, the joint count event, and this linear transformation; unknown-label completions enlarge the event's image. Enumeration checks validate the interval arithmetic, not annotation accuracy.

D.4 What the original label budget resolves across tasks

The original 595 labels can estimate an accuracy difference directly, without dividing by an estimated grader sensitivity. We reconstruct the 15 original nonempty task–mode–agreement strata and invert a hypergeometric count interval in each at error level $0.05/15$. Summing lower and upper correct-count bounds within each mode and subtracting the resulting accuracy intervals gives simultaneous coverage of all five task differences. This is a separate target and family from the FN/FP panels and the later 400-record follow-up. It conditions on the original benchmark-reference labels, the selected items, and the fixed outputs, not on semantic truth or future tasks. The 13 historical review flags were resolved in the July 27 human confirmation record with no label changes (Appendix C.4); this was not blind adjudication.

Table 8: Direct reference-accuracy differences, think minus no-think, in percentage points. Point estimates use the original stratified weights; the two-mode columns give the Horvitz–Thompson reference accuracies behind each difference (%, same weights); differences are computed on unrounded rates, so the displayed margins can differ by 0.1; intervals share one 95% count event across 15 strata. All five comparisons remain unresolved under this conservative construction. This is a precision result for the existing audit.

Task	Labels	no-think acc.	think acc.	Point estimate	Simultaneous interval
GSM8K	80	72.5	90.0	+17.5	[−20.6, +50.3]
MATH	80	57.5	95.0	+37.5	[−2.8, +66.1]
MMLU	160	43.6	74.2	+30.6	[−15.6, +67.8]
MMLU-Pro	160	55.7	80.8	+25.0	[−18.7, +61.3]
TriviaQA	115	45.5	63.6	+18.2	[−22.3, +54.7]

The point estimates reproduce the prior direct estimates. MATH's previous per-task interval $[+6.5, +60.8]$ excludes zero; its interval under this five-task construction does not. None of the simultaneous intervals supports a gain above two points or equivalence within two points, used here as illustrative decision tolerances (Corollary 2). The construction is conservative and does not prove that every valid procedure must abstain; it shows neither absence of a thinking benefit nor a saturated benchmark. More reference information or a justified, more informative construction would be needed to resolve these comparisons at this joint level. The per-task MMLU interval quoted in Section 5 uses the same construction restricted to that task's four arm-by-agreement groups; its per-stratum counts are in the supplement (Section S.6).

D.5 New generations across temperatures and answer formats

We sampled 80 items uniformly without replacement from each complete MMLU and GSM8K test set, using seed 2026092107 before generation. These items were not filtered by earlier model outcomes; overlap with older corpora is possible. Qwen3-8B generated one candidate per mode at temperature 0 and four at each of 0.6 and 1.0, giving 2,880 outputs. A separate temperature-0.6 stage generated four candidates per item and mode with an added final-answer instruction, giving 1,280

outputs. The two stages use different seeds; candidate slots are not matched random trajectories. Both use top-p 0.95 and caps of 8,192 no-think and 32,768 think output tokens. All 4,160 expected keys and eight shard hashes passed integrity checks. There are six length stops, all in think mode; they remain in the reported scores.

Table 9: New-generation observed think-minus-no-think gaps (percentage points; checkpoint Qwen3-8B; per-condition seeds in the supplement, Section S.4). Each task uses the same 80 sampled items across conditions. The format stage uses independent generations with different seeds, so condition differences are not paired repairs to the same outputs. These are descriptive scores on the generated outputs, not confidence intervals for future generation or semantic correctness.

Task	Rule	$T = 0$	$T = 0.6$	$T = 1.0$	Format, $T = 0.6$
MMLU	Qwen full	-22.50	-17.19	-20.63	-12.50
	Qwen segment	$+1.25$	$+8.75$	$+6.25$	$+10.00$
	Qwen no-ICL	0.00	$+4.06$	$+3.75$	$+7.50$
	Simple-evals	$+1.25$	$+8.12$	$+6.25$	$+10.00$
GSM8K	Flexible full	-22.50	-22.81	-13.44	$+0.63$
	Final-answer rule	$+5.00$	$+1.56$	$+3.13$	$+1.56$
	Qwen math	$+1.25$	-0.94	0.00	$+0.31$

Temperature sensitivity. Full-text `qwen_choice` gives a negative MMLU gap at all three tested temperatures, whereas its answer-segment variant gives a positive gap (per-mode marginals and verdict-transition counts in the supplement, Section S.6). GSM8K flexible remains negative, but Qwen math is slightly positive at temperature 0, slightly negative at 0.6, and tied at 1.0. The positive-alternative part of the earlier GSM8K reversal therefore does not recur in every new condition. Dropping the six length-stopped outputs leaves the signs of the base full-text and MMLU segment gaps unchanged (a descriptive sensitivity, not an unbiased correction).

The final-answer instruction does not repair every extractor. The added MMLU instruction requests a final `Answer: <letter>` line; GSM8K requests a final boxed numerical answer with no subsequent text. On GSM8K, flexible scores fall from $113/320$ to $4/320$ in no-think and from $40/320$ to $6/320$ in think: the regex permits runs of dollar signs, and the final match contains only dollar signs in $315/320$ no-think and $308/320$ think format outputs, versus $179/320$ and $262/320$ in the base condition. The small positive format gap therefore reflects two nearly collapsed scores, not successful repair. The fixed final-answer rule and Qwen math move scores by at most three rows out of 320 in either mode, and MMLU segment scores decrease (per-rule counts in the supplement, Section S.6). These comparisons involve independently generated responses to the same items; they do not count repairs or damage to the same answer.

A small blinded single-rater check of the format experiment (37 unique questions, 40 paired keys, 80 responses) found no disagreement with the benchmark answer field; its simultaneous accuracy ranges are too wide to establish a semantic effect (the supplement documents the check and the per-layer inversions). This is a post-hoc, single-rater check, not an independent annotation study; one checkpoint, 80 items per task, and three temperatures do not establish a universal decoding result. The runtime manifest and the reporting-floor sensitivity are in the supplement (Section S.6).

D.6 Joint parser interventions and sampling sensitivities

We replayed 439,520 existing MMLU/GSM8K outputs, including all 368,664 complete-test-set outputs. Every baseline verdict matched the vendored adapter. The four MMLU variants were fixed before this replay: full text, post-thinking segment, no ICL truncation, and both changes. The GSM8K variants were full text, post-thinking segment, and last complete boxed answer in that segment (otherwise last explicit Answer line, otherwise the unchanged segment). These rules never consult gold when choosing text. This remains a post-hoc analysis of existing data. On the primary MMLU corpus the joint states are: truncated and sent to fallback, $2{,}034$ of $2{,}560$ think rows (79.45%) versus 261 of $40{,}960$ no-think rows (0.64%); truncated without fallback, 5 versus 2; fallback without truncation, 169 versus $35{,}675$; neither, 352 versus $5{,}022$.

On the complete MMLU set, the four observed gaps are -20.99, $+5.09$, $+3.42$, and $+5.02$ points (per-variant marginals in Table 10). Stripping changes 32,207 think verdicts from wrong to right and 2,901 from right to wrong (4.51% of previously accepted think verdicts).

Table 10: Per-variant observed accuracies (accepted/total, %) behind the four complete-MMLU gaps (`mechanism_factorial.json`). no-ICL disables the ICL-truncation rule; joint applies both changes. Transitions are wrong-to-right / right-to-wrong verdict changes against gold.

Variant	no-think acc.	think acc.	D_{obs} (pp)
full	78.20 (175,695/224,672)	57.21 (64,263/112,336)	-20.99
segment	78.20 (175,695/224,672)	83.29 (93,569/112,336)	$+5.09$
no-ICL	78.29 (175,885/224,672)	81.70 (91,783/112,336)	$+3.42$
joint	78.29 (175,885/224,672)	83.31 (93,585/112,336)	$+5.02$

On complete GSM8K, stripping changes no verdict; the boxed/Answer rule gives $+4.95$ points. It changes 11,448 no-think and 8,555 think verdicts from wrong to right, but also 27 no-think verdicts from right to wrong. On the primary GSM8K corpus the same rule moves $9{,}895$ no-think and $1{,}186$ think verdicts from wrong to right, and 80 no-think verdicts from right to wrong; no think verdict moves from right to wrong. These transition denominators differ from the earlier parser stress test, which selected proxy-judged false-negative rows and rewrote their answers.

Paired reference audit of the boxed/Answer rule. The paired construction of Appendix D.3 audits this rule on the primary GSM8K corpus against the same 595 reference labels ($\alpha_h = 0.05/2$ per arm). All 39 sampled verdict changes repair a full-rule error and none introduces one. Inverting the changed subpopulations ($9{,}975$ no-think, $1{,}186$ think) bounds the total grading-error reduction to $[+15.1, +38.8]$ points on the no-think outputs and $[+52.0, +73.8]$ on the think outputs (jointly at least 95%). Under the same 595-label design the reference FN gap moves from $+16.4$ $[+3.3, +42.0]$ under the full rule to -17.2 $[-30.2, +6.5]$ under the boxed/Answer rule, which straddles zero. These intervals measure verdict-change error against the fixed reference labels; they do not make the rule a validated generation-time repair.

The original raw corpus lacks recorded finish reasons. The complete corpus records 325 MMLU and 11 GSM8K length stops, and the main text reports first-eight comparisons both before and after removing these rows (counts in the supplement, Section S.6). Removal is a selection sensitivity, not an unbiased correction.

D.7 Transfer sampling design and current inference

The five transfer corpora each contain 212,112 candidates from 1,473 items (Appendices D.7 and D.8). Their reference studies eventually contained 1,296, 1,223, 1,400, 1,055, and 1,126 records, respectively. Extra records were allocated using observed reference-correct support, stopping when every formal FN support cell reached $n_{\text{eff}} \geq 70$ (rule and per-model stopping records in the supplement). Such a threshold does not establish inclusion probabilities or fixed-population coverage.

The old combined weight, stratum size divided by the final number of labels, is not generally valid under label-dependent stopping: a small exact counterexample rules out the proposed general justification. We withdraw the old adaptive-sample intervals and joint-q claims from current inference.

Each model's initial 550-record sample used a fixed allocation within task, mode, and grader-agreement strata. We replayed all five sample-key sets with their recorded seeds: 20260730 for GLM, Qwen3.6 and gpt-oss, and 2026081801/2026081802 for DeepSeek/MiMo. The first three sampled all 16 think candidates; the latter two sampled only candidates 0–7. Restricting all models to think 0–7 leaves 409, 416, 416, 550, and 550 initial reference records. We join final adjudicated labels by candidate key but use no newly sampled records or final adaptive weights.

For each grader, condition on the initial sample counts in the known domain, original stratum, and verdict groups; their samples remain uniform without replacement within each group. Apply the count inversion in Appendix D.1, including unsampled groups, then take ratio and mode-difference bounds. Points use the original sampling weights, so numerator and denominator target domain totals;

their ratio is not claimed to be unbiased. Per-cell intervals have at least 95% coverage. Allocating $0.05/125$ to each of the 125 cells gives simultaneous coverage for their FN and FP gaps under one count event. This coverage conditions on a fixed reference-label rule and the actual sampling design; it does not cover annotation error or future generation.

The supplement (Section S.6) reports GLM's in-house panel, and Table 11 gives the five TriviaQA first-answer comparisons. The fixed initial samples support less precise statements than the old supplements suggested. Qwen3.6's simultaneous FN difference remains positive; the other four TriviaQA simultaneous intervals include zero. This joint statement concerns five named corpora, not a population of model families.

On the common no-think 0–127 / think 0–7 domain, GLM MMLU `qwen_choice` pairs an observed gap of $+3.4$ points with an estimated FN difference of $+13.6$ points whose per-cell interval $[-65.6, +89.5]$ contains zero: the reference sample does not determine the sign of this error difference (table in the supplement, Section S.6). The older full-16 census gap was $+3.0$ points; it concerns a different candidate domain.

Table 11: Retrospective fixed-menu analysis, without correction for earlier grader search. TriviaQA first-answer FN gaps from fixed initial samples on the common candidate domain (pp). Per-cell intervals and intervals simultaneous over all 125 transfer cells use the same count construction; they need not be centered on the point estimates. FN no-think and FN think are the per-mode fixed-sample point estimates from the same artifact (`transfer_design.json`); the FN gap is FN think minus FN no-think. Point estimates use the initial full-frame design weights (initial stratum N/n), so they need not recompute from the common-domain group counts in the supplement (Section S.6), which drive the intervals. Earlier adaptive-sample bootstrap intervals and q-values do not support current inference.

Model	FN no-think	FN think	FN gap	Per-cell 95%	Simultaneous
GLM-4.5-Air	0.0	72.4	$+72.4$	$[+1.1, +99.7]$	$[-26.0, +99.9]$
Qwen3.6-35B-A3B	0.0	85.3	$+85.3$	$[+31.4, +100.0]$	$[+8.1, +100.0]$
gpt-oss-20b	0.2	92.5	$+92.3$	$[-42.5, +99.3]$	$[-60.7, +99.8]$
DeepSeek-V3.2	22.9	45.4	$+22.5$	$[-4.0, +78.7]$	$[-27.1, +92.4]$
MiMo-V2-Flash	0.0	19.1	$+19.1$	$[-27.1, +59.0]$	$[-46.8, +81.5]$

D.8 Four later model corpora

The supplement (Section S.6) tabulates the same fixed-initial-sample construction for all four later models. Their MMLU `qwen_choice` observed gaps on the common domain are -20.2, -14.1, -20.0, and -8.3 points. The FN estimates are positive, but every per-cell interval contains zero. The earlier statement that all four FN intervals excluded zero used the adaptive reference supplements and is withdrawn.

For DeepSeek, the first and last eight think candidates separately give observed MMLU gaps of -20.04 and -19.33 points under `qwen_choice`; MiMo gives -8.27 and -9.83. The observed reversal therefore occurs in both halves, but labels from the first half do not identify the second half's FN rate. New second-half reference labels are not included in these results.

The historical machine-assisted labeling included confirmation-style checks and later direct supplements. It is not independent blind human ground truth, and two later cohorts share an untraceable mixed same-vendor second-judge batch, so second-judge consistency cannot be audited per model at row level (Appendix C.4). Model-specific mode controls differ, and GLM lacks the recorded length-stop keys available for the four later models. The reference inference and observed-score comparisons should be read within these corpus and label-definition limits.

D.9 Complete-test-set anchor on MMLU and GSM8K

The label-dependent audit is conditional on the fixed 1,473 items by design (Section 3). To check whether the observed flip is a property of those subsets or of the benchmarks themselves, we ran a label-free anchor on the complete MMLU test split (14,042 items) and GSM8K test split (1,319 items): 15,361 items with 16 no-think and 8 think candidates per item (368,664 rows), generated with

the same model and decoding configuration as the primary corpus, with per-shard seeds $20260915+s$ (no-think) and $20270915+s$ (think). Graders are the same vendored mainstream implementations as in Appendix C.6; the observed gap is computed against gold only and needs no reference labels. The check is post-hoc and exploratory.

Table 12: Observed mode gap D_{obs} (pp) on the paper subsets and on the complete MMLU and GSM8K test sets. The middle column recomputes the paper subsets on the new generation batch, separating item-composition effects from generation drift. Bracketed intervals in the full-set column are 95% item-cluster bootstrap percentile intervals ($B = 2{,}000$, seed 20260915), resampling items independently within each arm; they are item-composition sensitivity intervals and do not estimate the sampling uncertainty of the paired same-item mode gap. Paired intervals that resample the same items jointly in both modes are narrower on 13 of 14 cells and change no sign (all 14 paired intervals, marginal accuracies, and matched-candidate counts in the supplement, Section S.6).

Task	Grader	Subset (paper)	Subset, new batch	Full set, new batch
GSM8K	lm-eval strict	$+0.1$	$+0.4$	$+1.4$ $[+1.1, +1.6]$
GSM8K	lm-eval flexible	-13.2	-13.1	-22.0 $[-24.4, -19.6]$
GSM8K	qwen_math	$+11.8$	$+12.3$	$+2.3$ $[+0.9, +3.8]$
GSM8K	math_verify	$+11.9$	$+12.4$	$+2.3$ $[+0.9, +3.8]$
MMLU	mmlupro_official	$+17.3$	$+15.4$	$+3.3$ $[+2.5, +4.1]$
MMLU	simple_evals_mc	$+17.0$	$+15.2$	$+4.9$ $[+4.1, +5.8]$
MMLU	qwen_choice	-15.7	-15.6	-21.0 $[-21.9, -20.1]$

Table 12 shows that the flip direction reproduces on the complete test sets with larger magnitude for the two flip cells: qwen_choice on MMLU moves from -15.7 to -21.0 points and lm-eval flexible on GSM8K from -13.2 to -22.0 points, while the positive gaps shrink but stay positive (GSM8K math graders $+11.8/+11.9 \to +2.3$; the other MMLU graders $+17.0/+17.3 \to +4.9/+3.3$). The sign structure is therefore present on these complete test sets for this model, configuration, and generation batch. Magnitudes move with item composition, so label-dependent estimates remain conditional on the 1,473 items. The extraction-state mechanism also reproduces on the full MMLU set: the trigger state share rises from 0.111 (no-think) to 0.272 (think; shares and accuracies as fractions), within-state accuracy is 0.724 (no-think) vs. 0.884 (think) on trigger rows and 0.789 vs. 0.455 on fallback rows, and a Kitagawa decomposition writes the full-set gap of -21.0 points as $+2.9$ points of composition change and -23.9 points of within-state change. The middle column recomputes the published subset numbers on the new batch (e.g., qwen_choice -15.6 vs. -15.7), so the subset-versus-full comparison within one batch isolates item composition; the cross-batch subset movement is smaller but nonzero, supporting a substantial composition contribution without excluding generation drift.

D.10 Motivating example: the MMLU sign reversal

Main-text summary: Section 5; the full six-grader census figure is Figure 1 in the introduction.

D.11 Mainstream panel: corroboration and a grader-dependent observed sign

Main-text summary: Section 5.

This subsection extends the audit (Section 3.3). Table 13 (Appendix C.6) re-runs the audit on the graders the field actually uses (Section 3.1); Figure 6 maps the observed gap of every cell.

Label-free comparison with mainstream GSM8K graders. The two standard math graders' observed gaps and error channels are summarized in Section 5. Because D_{obs} is computed without reference labels, this comparison is independent of the label pipeline, but not of the corpus or the panel selection. By Proposition 1 it cannot exclude a defect shared by the two graders. The bounded-overlap bound of Section 4.2 turns part of this comparison into a label-free content-level bound, bounding the content-level gap below for `math_v1` ($[+10.1, +37.8]$pp) and for both standard math graders ($[+10.1, +14.3]$pp) without candidate-level semantic labels, while lm-eval flexible correctly abstains (Table 17).

Table 13: Historical sensitivity record (retired analysis, descriptive only); its intervals and significance labels are not current design-based guarantees. Current reference-label inference is in Table 5; MMLU Qwen choice has per-cell $[+11.8, +77.0]$ and simultaneous $[-5.8, +81.8]$ FN-gap intervals in pp. Mainstream panel results (targeted 132-row confirmation review found no changes (Appendix C.4); this does not establish corpus-wide label validity). FN rates are Horvitz–Thompson-weighted against the same 595-row reference substudy; intervals are 95% item-cluster paired bootstrap ($B = 2{,}000$, seed 20260727). All differences are think−no-think, in pp. †: prompt-contract-degenerate cell (FN $\geq$ 95% in both modes; D_{corr} undefined)—a format mismatch with our chat-style candidate text, not a grader-quality claim. Historical flags (per-cell values not displayed) are Benjamini–Hochberg-adjusted Wald p-values from item-cluster paired-bootstrap standard errors, thresholded at $q \leq 0.05$ within the 26 extension mode-gap tests (13 cells × {FP, FN}); all other cells are unresolved. D_{obs} is reference-label-free. Full channels (FP, attenuation) in the supplement (Section S.4); per-mode marginal accuracies behind each gap and accepted/total census counts are in the supplement (Section S.6).

Task	Grader	$\text{FN}_{\text{no-think}}/\text{FN}_{\text{think}}$ (%)	FN diff (pp) [sens.]	D_{obs} (pp) [sens.]	D_{corr} (pp) [sens.]
gsm8k	`lmeval_strict`†	100.0/100.0	+0.0[+0.0, +0.0]	+0.1[+0.0, +0.2]	—
	`lmeval_flexible`	58.6/75.0	+16.4[−5.6, +40.4]	−13.2[−17.5, −8.6]	−10.4[−56.5, +49.0]
	`qwen_math`	0.0/0.0	+0.0[+0.0, +0.0]	+11.8[+8.2, +15.4]	+11.8[+8.2, +15.4]
	`math_verify`	0.0/0.0	+0.0[+0.0, +0.0]	+11.9[+8.3, +15.5]	+11.9[+8.3, +15.5]
math_full	`lmeval_strict`†	100.0/97.4	−2.6[−8.7, +0.0]	+0.5[+0.2, +0.9]	—
	`lmeval_flexible`†	95.7/100.0	+4.3[+0.0, +14.7]	−2.7[−4.1, −1.6]	—
	`qwen_math`	13.0/0.0	−13.0[−28.6, +0.0]	+40.3[+36.9, +43.7]	+32.6[+19.6, +40.8]
	`math_verify`	56.5/23.7	−32.8[−58.9, −5.7]	+31.3[+28.0, +34.7]	+1.7[−98.6, +40.8]
mmlu	`mmlupro_official`	14.7/7.4	−7.3[−20.6, +4.1]	+17.3[+14.0, +20.7]	+11.6[−5.2, +22.0]
	`simple_evals_mc`	0.0/0.0	+0.0[+0.0, +0.0]	+17.0[+13.4, +20.8]	+17.0[+13.4, +20.8]
	`qwen_choice`	5.3/50.9	+45.6[+25.2, +65.2]	−15.7[−20.4, −10.9]	−4.9[−227.2, +98.0]
mmlu_pro	`mmlupro_official`	13.1/8.7	−4.4[−15.6, +7.3]	+23.0[+19.8, +26.2]	+19.5[−2.8, +35.4]
triviaqa	`triviaqa_em_aliases`	12.0/17.2	+5.2[−11.3, +20.7]	+20.7[+16.6, +24.8]	+27.4[+16.4, +41.4]

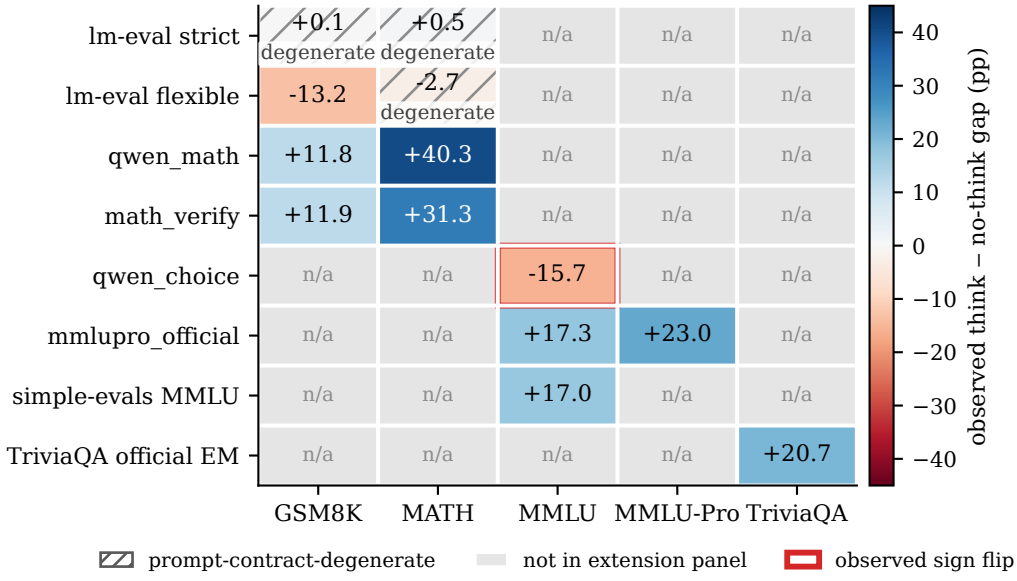

Figure 6: Observed think−no-think accuracy gap for every (grader, task) cell of the mainstream panel. Rows are mainstream graders, columns are tasks, and cell color encodes the observed think−no-think gap in percentage points. The red frame marks the `qwen_choice` × MMLU cell; hatched cells are prompt-contract-degenerate; gray cells are outside the panel, where the grader's extraction contract does not cover the task. Label-free observed-score comparison on raw verdicts.

The observed sign is grader-dependent. Section 5 reports the `qwen_choice` MMLU flip and its think-concentrated FN channel. The mechanism depends on the input contract: repeated trigger

phrases can cause first-paragraph truncation before the parser chooses a trigger or fallback branch (Appendix D.6). The reference FN point estimates are 50.9% for think and 5.3% for no-think. The lm-eval flexible extractor on GSM8K shows a second flip ($D_{\text{obs}} = -13.2$ $[-17.5, -8.6]$), though its FN mode gap is unresolved at the simultaneous level. Sign survival therefore held *within* the in-house panel (Appendix D.12) but not after adding this panel. The counterexample is label-free and visible on the raw verdicts. Section 5 discusses whether the *true* gap flips.

Math-task transfer and degenerate cells. The historical estimator comparison on MATH shows the same bias-removal-at-the-price-of-variance pattern as Appendix D.12. Three cells are prompt-contract-degenerate (lm-eval strict on both math tasks and lm-eval flexible on MATH: FN $\geq 95\%$ in both modes, D_{corr} undefined); we report them as format-mismatch diagnostics, not grader defects. The remaining mainstream cells are consistent with the corresponding full-text and line-safe in-house rules (within 2.3 points on each of MMLU, MMLU-Pro, and TriviaQA; the mechanism-probe in-house rule with a known defect differs more) with unresolved error channels (full panel in the supplement, Section S.4). The FP block of the full panel is mostly zero in this reference sample. No FP event was observed in these cells of the reference sample; the nonzero-width FP bounds are in Appendix D.1. The nonzero FP rows are all multiple-choice extractors, which output an option on every candidate and can land on the gold one by chance.

D.12 In-house mechanism panel: a shared defect invisible to standard safeguards

Main-text summary: Section 5.3. This subsection reports the in-house panel of Section 3.1; the primary analysis of Section 3.2 uses this panel. The headline GSM8K cell is discussed in the main text; its figure appears here (Figure 7), and the remaining cells are reported only here.

On GSM8K, the in-house panel's two math graders agree on 100% of the $25{,}728$ no-think-mode candidates (and all $69{,}768$ math candidates), yet both share an estimated 37.9% no-think-mode false-negative rate (Figure 7). Two graders can agree perfectly while sharing the same error.

Grading error is mode-specific on GSM8K, concentrated in the no-think mode (RQ1). The current GSM8K FN gap is -37.9 points, with fixed-menu, reference-rule per-cell interval $[-49.9, -8.7]$. The simultaneous interval is $[-56.6, -0.8]$ (Table 5). The point rates are 37.9% no-think and 0% think. For MATH, the corresponding point gap is -30.3 with per-cell interval $[-49.2, +3.7]$ and simultaneous interval $[-60.4, +10.5]$; its sign is unresolved. The GSM8K false negatives are normalization failures shared by both in-house math graders, a *common-mode* failure (Figure 7c): thousand separators and currency symbols (`\boxed{276,000}` versus `276000`; `\boxed{\$9.00}` versus `9`). The MATH ones are genuine equivalence failures (radical forms such as $2\sqrt{6} - 2\sqrt{3}$ versus $2\sqrt{3}(\sqrt{2} - 1)$, solution sets, matrices). A second LLM judge re-adjudicated all 21 no-think-mode FN rows behind these estimates (11 GSM8K, 10 MATH) and the full 32-row mismatch audit, with no disagreement (LLM concordance, not human confirmation; Section 3.3; instance cases in the supplement, Section S.7). False-positive point estimates in the in-house panel are near zero (at most 6.1%); multiple-choice and exact-match cells show smaller FN asymmetries (-6.4 to $+8.5$ points) whose current design intervals all cross zero (Table 5 and Appendix D.15).

The error propagates to conclusion-level effect sizes (RQ2). The current comparison in Figure 7 distinguishes the observed census gap, the direct reference estimate with a fixed-population interval, and the additive residual point estimate. Older inversion, repaired-grader, and item-bootstrap comparisons are historical sensitivity records; the same-labels estimator comparison is in the supplement (Section S.7). For GSM8K, the v1 observed gap is $+37.4$ points and the additive residual point estimate is $+9.9$. A positive point estimate alone does not resolve the semantic-truth sign.

Safeguards provide alarms, not guarantees (RQ3). Two widely used safeguards fail here. First, grader agreement: the two in-house GSM8K math graders agree on 100% of $25{,}728$ no-think-mode candidates while both injecting the same 37.9% false-negative rate. Agreement-only audits *can* miss common-mode error; here they miss all of it. Second, conclusion survival: all 12 applicable task–grader cells in Table 14 keep a positive observed gap with an interval excluding zero. Yet the quoted effect size may be overstated by 20–31 points (GSM8K) or understated by 11.7 points (MMLU-Pro

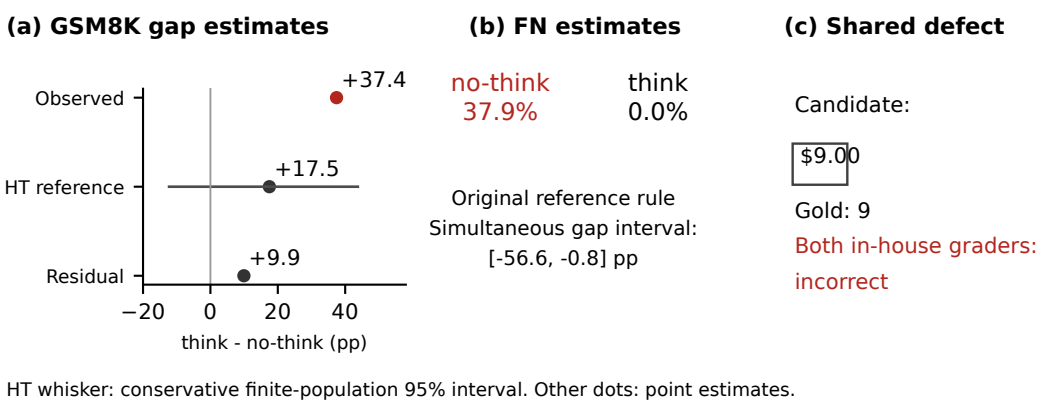

Figure 7: GSM8K observed gap, reference estimates, and a shared normalization defect. (a) The observed gap is a census quantity; the HT whisker is a conservative finite-population 95% interval. The residual dot is a point estimate, not a semantic certificate. (b) FN point estimates and a simultaneous interval conditional on the original reference rule. (c) Both shared-core in-house graders reject a currency-formatted answer that matches the numerical gold.

Table 14: Historical sensitivity record (retired analysis, descriptive only); its intervals and significance labels are not current design-based guarantees. Current reference-label inference is in Table 5. Conclusion survival across the in-house grader panel: observed think−no-think accuracy gap in pp [95% item-cluster bootstrap CI], with sign and CI-excludes-zero flags. $A_{\text{no-think}}/A_{\text{think}}$ are the per-mode census accuracies (%) behind each gap; mode populations (no-think/think) are $25{,}728/1{,}608$ (GSM8K), $39{,}936/2{,}496$ (MATH), and $40{,}960/2{,}560$ (MMLU, MMLU-Pro, TriviaQA). The sign is positive and every CI excludes zero in all 12 applicable task–grader cells (H2: no reversal; `math_v1`/`math_v2_seg` are identical on GSM8K and MATH, `mc_v1_fulltext`/`mc_v3_linesafe` are identical on MMLU and MMLU-Pro, and `em_v1_firstans`/`em_v2_seg` differ on four TriviaQA rows, so the 12 cells carry eight effectively distinct channels on the primary corpus); magnitudes vary up to $2\times$ across graders of the same task (e.g., mmlu_pro), so sign survival does not imply measurement invariance.

Task	Grader	$A_{\text{no-think}}/A_{\text{think}}$ (%)	D_{obs} (pp) [CI] (sign, excl. 0)
gsm8k	math_v1	50.4/87.8	+37.4 [+33.1, +41.5] (+Y)
	math_v2_seg	50.4/87.8	+37.4 [+33.1, +41.5] (+Y)
math_full	math_v1	44.2/81.7	+37.5 [+33.1, +41.8] (+Y)
	math_v2_seg	44.2/81.7	+37.5 [+33.1, +41.8] (+Y)
mmlu	mc_v1_fulltext	47.8/64.4	+16.6 [+13.2, +20.1] (+Y)
	mc_v2_buggy	41.9/58.5	+16.6 [+12.9, +20.5] (+Y)
	mc_v3_linesafe	47.8/64.4	+16.6 [+13.2, +20.1] (+Y)
mmlu_pro	mc_v1_fulltext	56.3/77.0	+20.7 [+17.5, +24.1] (+Y)
	mc_v2_buggy	49.2/59.8	+10.5 [+6.8, +14.4] (+Y)
	mc_v3_linesafe	56.3/77.0	+20.7 [+17.5, +24.1] (+Y)
triviaqa	em_v1_firstans	39.0/61.5	+22.5 [+18.2, +26.5] (+Y)
	em_v2_seg	39.0/61.6	+22.5 [+18.2, +26.5] (+Y)

`mc_v2_buggy`). Sign survival does not identify either the semantic direction or the magnitude. Reference labels inform that comparison subject to their definition and sampling uncertainty. Both safeguards use only grader verdicts: against a shared defect, their failure is principled blindness, not lack of power (Proposition 1).

The historical nested two-phase bootstrap demoted the GSM8K FN gap from BH-passing to non-passing; that variance model is superseded by the fixed-population construction of Appendix D.1, and the claim-scope ledger in the supplement (Section S.1) marks its outputs historical.

D.13 THE DECOMPOSITION AT WORK: POOLED ACCURACY AND MODE-SPECIFIC ERROR

Main-text summary: Section 5.

Table 15 (Appendices D.13 and D.17) evaluates the decomposition of Equation (2) on five illustrative cells; Section 5 states the precision check and the flip anatomy. (i) *Anatomy of the flips*: High task prevalence places most weight on the FN term. The FP channel of `qwen_choice` partially *offsets* its FN channel; the flip would have been larger otherwise. (ii) *Pooled accuracy and the comparison margin*: high pooled accuracy alone need not protect a small gap. Its equal-mode-weight 2ϵ bound is given in Appendix B.3; the mode-specific decomposition retains finer information in Section 4.1. `mmlupro_official` rejects 13–16% of proxy-accepted candidates in both modes. Its two channels nearly cancel, and the conclusion survives ($+0.3$ points of bias). By contrast, `qwen_choice` rejects 0.8% of proxy-accepted no-think outputs and flips the sign, with a 59.7% proxy-relative rejection rate on think outputs (census rates relative to the proxy, Table 15; the reference-label estimates of Appendix D.11 differ in reference source and denominator, not direction). A proxy comparison on no-think text alone would miss this disparity; it would not establish semantic validity. (iii) *Degeneracy is real*: the lm-eval strict extractor rejects essentially everything, with FN $\approx 100\%$ in both modes and $J \approx 0$. Theorem 2 says it is correctable only with extreme variance inflation, and it was excluded. (iv) *The design-based correction needs no inversion*: Equation (8) uses full-corpus verdicts plus audit residuals and is design-unbiased for the same estimand as the plain $D_{\mathrm{HT}} = +17.5$ (estimator comparison in the supplement, Section S.7); its GSM8K point estimate is reported in Appendix D.12. The instance behind Proposition 1 is also here: `math_v1` and `math_v2_seg` produce identical verdicts on all 69,768 math candidates while sharing the GSM8K defect, and `mc_v1_fulltext` $\equiv$ `mc_v3_linesafe` on all 87,040 multiple-choice candidates (the fold of Section 3.2). The historical channel-cap certificates and the original PPI threshold claim are superseded; see the claim-scope ledger in the supplement (Section S.1).

D.14 MECHANISMS: EXTRACTION STATES AND PARSER STRESS TESTS

Main-text summary: Section 5.

Why are the channels mode-specific at all? Table 16 (Appendix D.17) decomposes the three flip/inflation cells by the extraction state exposed by each grader's code. A Kitagawa split (Kitagawa, 1955) separates each FN mode gap into a composition shift over states and a within-state channel change; the split is exact on unrounded values in all three cells (displayed terms may differ by 0.1 after rounding) under the convention that a state with no support in one mode enters with within-state FN rate 0 (shown as — in Table 16): `math_v1` $-33.6 = -15.1 - 18.5$, lm-eval flexible $+19.6 = -1.3 + 20.8$, and `qwen_choice` $+58.9 = -1.5 + 60.3$ (gap = composition + within-state, pp; `miaudit_contract_decomp_2026-07-27.json`). Three distinct extraction contracts harm opposite modes. (i) `math_v1` keys on the `\boxed` contract: the no-think mode places 31.3% of its proxy-accepted mass into boxed answers carrying `\$` or thousand separators. The normalizer scores this state 100% wrong, while the think mode never enters it ($< 0.01\%$). (ii) `lmeval_flexible` takes the *last* number-like token. Among proxy-accepted outputs, the plain-integer state dominates. Last-match disagreements occur in 84.4% of think versus 63.0% of no-think in that state. A number-like regex match can be a recap number or punctuation-only text; these rates do not isolate recap causality. (iii) `qwen_choice` counts both "choice is" and "answer is" before truncating on repeated triggers. Only then does it choose the trigger or fallback branch. The former state aggregation mixed rows that originally lacked triggers with rows whose triggers were removed by truncation. Appendix D.6 reports these paths separately and tests the segmentation and truncation rules jointly.

Parser stress test (surface rewrites). Correlation between state and error is not yet evidence that the surface form *causes* the error. The probe design and its recovery and breakage rates are summarized in Table 16. An LLM analysis agent (vendor withheld for review; not a human rater) inspected each of 84 seeded (before, after) pairs: 83 preserve the model's committed answer, and the single exception is a degenerate self-contradictory text with no committed option whose verdict is unchanged. Surface changes can change verdicts on pairs that retain a committed answer, so the mode-specific channels are properties of the grader–text interaction, not of the answers' correctness. Rewriting answers into a grader's extraction contract therefore removes most of the measured mode-specific FN mass (the recovery rates of Table 16); the repaired `math_v4_normfix` grader's GSM8K gap ($+13.8$

$[+9.9, +17.7]$) then lies within the range of the label-based point estimates, while the historical paired-difference intervals include zero, so equivalence is not established. What remains open after such a repair is the *true* gap, which is unresolved with or without the repair — the decision rule's abstention on the corrected sign is unchanged.

D.15 Robustness of the audit itself

Main-text summary: Section 5.

The audit includes checks for label quality, systematic label bias, variance choice, item exclusions, and matched arm widths.

(i) *Label quality*: 388 targeted second-pass checks (semantic re-adjudications and mechanical re-extractions, Appendix C.4) detected no discrepancy. This is targeted checking, not a population error rate.

(ii) *Historical sensitivities*: a random-relabeling grid and a two-phase versus item-bootstrap variance comparison were recorded as descriptive diagnostics under historical procedures. Neither is a worst-case bound or a coverage validation; both are archived, and current inference uses Appendix D.1.

(iii) *Item-exclusion sensitivity*: five items were flagged as suspect during adjudication. We excluded every candidate of both modes and all reference rows for these items and re-ran the full estimation chain on the reduced item sets. No key estimate moves by more than 1.1 points, and no interval changes its excludes-zero status, in both panels (the supplement, Section S.6).

(iv) *Matched-sample-count widths*: the primary design uses 128 no-think and 8 think candidates per item. Restricting both arms to the first 8 candidates per item (a matched 8v8 design in candidate count, same verdicts) shifts every observed gap by at most 1.8 points and preserves the sign of all 25 cells: the `qwen_choice` MMLU flip survives (-15.2 $[-20.0, -10.3]$) and the GSM8K patterns are intact. The 128-vs-8 arm widths therefore do not drive the reported patterns; all 25 cells are listed in the supplement (Section S.6).

(v) *Machine-augmented reference labels (PPI)*: withdrawn as evidence of restored significance. Its mixed target weights and changed label definitions are audited in Appendix D.19. Current fixed-population inference uses Appendix D.1.

D.16 Semantic bridge extension

Main-text summary: Section 4.2.

The all-agree extension produced 393 binary semantic judgments and seven held records. Dropping the seven held rows gives optimistic complete-case hypergeometric sensitivities of 5.18% (no-think) and 6.97% (think), whose sum already exceeds the 10.1pp content-level margin; counting every held record as a bridge violation instead gives one-sided finite-population 95% design bounds of 7.87% and 14.68%, so the sum exceeds the margin under either reading (per-stratum descriptive rates, frame-exclusion variants and the arithmetic are in the supplement, Section S.7). The extended audit identifies why C and Y can differ; its effect on the semantic certificate is stated in Corollary 3 and Section 4.2. We do not merge these semantic judgments with the round-1 595 labels because the two rounds use different target-label definitions.

D.17 Main exhibit tables

Table 15 instantiates the decomposition of Equation (2); Table 16 reports extraction-state mechanisms; and Tables 17 to 19 give the content-level and decision exhibits. The reference-label confusion tables, the item-exclusion sensitivity, and the observed-versus-corrected gap figure are in the supplement (Section S.6).

Notes on Table 17. The identified per-mode content rates behind the shared 13.907pp contrast are $c_{\text{no-think}} = 73.9\%$ and $c_{\text{think}} = 87.8\%$, with anchor-free masses $u_{\text{no-think}} = 3.774$pp and $u_{\text{think}} = 0.373$pp, so the per-mode content-rate ranges are $[73.9, 77.7]\%$ and $[87.8, 88.2]\%$. The disagreement mass $A(3){=}0$ (the corpus histogram has no $R{=}3$ rows), and the bound is formed by the $(K-1)/(K-$

Table 15: The error decomposition, instantiated. Each row evaluates $D_{\text{obs}} - D = \text{FN channel} + \text{FP channel}$ in pp, where FN channel $= A_0\text{FN}_0 - A_1\text{FN}_1$ and FP channel $= (1 - A_1)\text{FP}_1 - (1 - A_0)\text{FP}_0$; this joint-mass bookkeeping is algebraically identical to Equation (2) (the $(\bar{J} - 1)D$ attenuation is absorbed into the two channels) and holds to machine precision in every row (generation script asserts $< 10^{-9}$). All values are computed from unrounded inputs and rounded for display. D reference: row 1 uses the misclassification correction of Equation (1) (a point estimate; its interval $[-26.1, +24.4]$pp, 95% item-cluster bootstrap, historical sensitivity, is too wide to determine the corrected sign, supplement Section S.6); rows 2–4 use the task's error-free-on-audit mainstream grader named in the Proxy column (FN=FP=0 on all audited rows of the task) as a *surrogate proxy* and compute full-corpus channel rates against its labels — zero detected errors on the audit sample bound but do not prove the proxy's error rate (zero-count upper bounds, supplement Section S.5), so rows 2–4 decompose relative to the proxy, not to certified semantic truth, and "sign flip" is versus the proxy's sign; row 5 is prompt-contract degenerate (FN $\approx 100\%$ in both modes), so no channel correction is identifiable (Theorem 2). Per-mode census accuracies and proxy accepted/total counts are in Table 14 and the supplement (Section S.6); row-1 per-mode reference counts are in the supplement (Section S.6).

Task	Grader	Proxy	D (ref./proxy)	D_{obs}	bias	FN ch.	FP ch.	outcome
gsm8k	`math_v1` (panel)	Equation (1)	+6.7	+37.4	+30.8	+30.8	+0.0	inflated (Thm. 1)
gsm8k	lm-eval flexible	`math_verify`	+11.9	-13.2	-25.1	-24.9	-0.2	sign flip
mmlu	`qwen_choice`	simple-evals	+17.0	-15.7	-32.7	-38.1	+5.4	sign flip (FP offsets)
mmlu	`mmlupro_official`	simple-evals	+17.0	+17.3	+0.3	-3.9	+4.2	preserved (channels cancel)
gsm8k	lm-eval strict	— (degenerate)	—	+0.1	—	FN≈100% both modes		degenerate, excluded

Table 16: Extraction-state mechanisms (full corpus; channel rates against a task-specific proxy grader with no observed errors on the audit sample). For each mechanism: the extraction state that carries the error, the share of each mode's proxy-accepted candidates landing in that state (no-think/think, %), the within-state FN rate (no-think/think, %), and the parser stress test: the fraction of proxy-FN error rows flipped to correct by a deterministic surface transformation intended to preserve the answer (paired; gold-free, correctness judged by the proxy), and the transformation's break rate on previously proxy-accepted rows (rows accepted by both the proxy and the original grader). The lethal-state share denominators are the proxy-accepted counts (no-think/think, `miaudit_contract_states_2026-07-27.csv`): $19{,}523/1{,}412$ (`math_verify` on GSM8K) and $19{,}445/1{,}650$ (simple-evals on MMLU). Answer preservation was agent-checked on 84 pairs (83 determinate), not verified for every transformation. Transformations: for `math_v1`, delete dollar signs and thousands commas inside `\boxed` (e.g., `\boxed{\$9.00}` → `\boxed{9.00}`; `\boxed{276,000}` → `\boxed{276000}`); append the strict-marker phrase with the model's own final number for `lmeval_flexible`; prepend the trigger phrase with the model's own final letter for `qwen_choice`.

Mechanism cell	harmed mode	correct mass in lethal state (%)	within-state FN (%)	stress test flip rate	break on correct rows
gsm8k × `math_v1` (`\boxed` w/ `\$,`)	no-think	31.3/0.0	100.0/—	92.5% (6,071/6,566)	0.00% (0/14,369)
gsm8k × `lmeval_flexible` (last-number)	think	94.8/99.4	63.0/84.4	86.1% (11,944/13,866)	0.42% (30/7,069)
mmlu × `qwen_choice` (no-trigger fallback)	think	87.7/83.2	0.7/70.9	94.3% (1,078/1,143)	0.29% (57/19,952)

ρ)-weighted sum plus the τ and p_K^H accounting. Safety is constructive: check V1 reports 0 violations and 0 demotions over the 73,097 certified-safe row–grader pairs (of $27{,}336 \times 4$ units in total), and the uncovered⇒globally-uncovered premise of the split is machine-checked with 0 violations. The e^c column is the true code-level error against C as a fraction of the full mode population, evaluated through the exact identity of Equation (5): the anchored-row disagreement mass plus the $R{=}K$-row error mass (for `math_verify` no-think, $e^c = 0.0002$ is the $R{=}K$ term—4 rows—not the $A(\rho)$ columns). It is not the conditional error rate over decidable rows (for `math_v1` no-think that conditional rate is $6{,}073/24{,}886 = 24.4\%$). Rows whose commitment is not code-decidable (842 no-think and 6 think, inside τ) pin no binary C, each binary completion of C shifts a grader's e^c by at most the uncovered mass, and the completion-independent bounds already charge that mass. `math_v1`'s two-sided content-level bound $[+9.7, +65.1]$pp contains the repaired grader's observed gap (13.8pp) and the label-based HT estimate (17.5pp); the 6.7pp point estimate sits below it, but its item-cluster interval $[-26.1, +24.4]$pp overlaps it (no formal contradiction). A content-certified

Table 17: Label-free bounded-overlap error bounds (Theorem 3) and one-sided ranking bounds (Corollary 3(iii)) on the fixed GSM8K corpus (27,336 candidates, census quantities; $K=4$ after panel curation). *Top*: bound components (uncovered mass τ, full-overlap hazard mass p_K^H, disagreement masses $A(\rho)$), the bound of Equation (5), and the true code-level error e^c against the content functional. *Bottom*: one-sided content-level bounds (Equation (6)); no Y-labels enter. Reading notes follow the table.

Mode	Grader	τ	p_K^H	$A(1)$	$A(2)$	$A(3)$	bound	e^c (full pop.)
					(fraction)			
no-think	math_v1	0.0333	0.0044	0.0003	0.1568	0.0000	0.2732	0.2354
	lmeval_flexible	0.0333	0.0044	0.2797	0.1335	0.0000	0.5178	0.4800
	qwen_math	0.0333	0.0044	0.0000	0.0000	0.0000	0.0377	0.0000
	math_verify	0.0333	0.0044	0.0000	0.0000	0.0000	0.0377	0.0002
think	math_v1	0.0037	0.0000	0.0000	0.0000	0.0000	0.0037	0.0000
	lmeval_flexible	0.0037	0.0000	0.7419	0.0000	0.0000	0.7456	0.7419
	qwen_math	0.0037	0.0000	0.0000	0.0000	0.0000	0.0037	0.0000
	math_verify	0.0037	0.0000	0.0000	0.0000	0.0000	0.0037	0.0000

Grader	D_{obs} (pp)	D^C one-sided (pp)	B_k (pp)	decision
math_v1	+37.4	[+10.1, +37.8]	27.7	**content-certified** +
lmeval_flexible	-13.2	[-64.2, +62.1]	126.3	abstain
qwen_math	+11.8	[+10.1, +14.3]	4.1	**content-certified** +
math_verify	+11.9	[+10.1, +14.3]	4.1	**content-certified** +

decision certifies agreement with the content rule C only; it does not certify Y. Lifting to semantic truth requires a separate C–Y bridge (Corollary 3(ii)). The original 80-row audit bounds only C–S benchmark-reference disagreement (7.21% / 7.15% separately at one-sided 95%), so it does not bound the semantic bridge. The 400-row semantic extension observes five C–Y violations across four items and leaves the semantic-level guarantee a sensitivity region (Appendix D.16); content-level bounds do not involve δ. Per-mode observed accuracies are in Table 14 and the supplement (Section S.6).

Table 18: Joint design-effective-direction certificates from exact count inversion, in percentage points (pp). All 25 cells are retained, including prompt-contract-degenerate strict graders. In every row the pointwise (pw) line sits above the familywise (fw) line. For each cell the four supergroup counts $K^{(a,v)}$ (reference-correct records of arm a in verdict group v; arm 0 = no-think, arm 1 = think) vary independently over the exact joint count events behind Table 5: the per-cell pointwise 95% event and the panel-simultaneous familywise 95% event (family sizes 26 mainstream, 24 in-house, as recorded in the source analysis). $\eta = (\pi, x, y, j)$ ranges over its coordinate-wise box, the extrema of η over the 2^4 count corners (a box relaxation of the joint image: the coordinate extrema need not be simultaneously attainable). $\psi = d + \pi(x + y) - y$ is the numerator of ϕ_d in Theorem 2, so when $j > 0$ the sign of ψ is the sign of the implied true gap. Its exact extrema over the η-box intersected with $\{|x| + |y| \leq 2(1 - j_{\mathrm{lo}})\}$ are computed by vertex enumeration and cross-checked by linear programming; the diamond uses the box lower bound j_{lo} and is a necessary condition for $(x, y) \in F_j$. The physical-accuracy constraint $|\phi_d(\eta)| \leq 2\min(\pi, 1 - \pi)$ of Theorem 2 is *not* imposed, so the reported ranges are conservative. Decision: certify + when $\min \psi > 0$, certify − when $\max \psi < 0$, otherwise abstain; incompatible when $j_{\mathrm{hi}} \leq 0$ or the diamond is disjoint from the box (the count event then contradicts the channel model); undefined when a rate denominator can vanish at a count corner (per-cell reasons in the machine-readable output). D_{obs} is the census observed gap (think − no-think). Computed from unrounded inputs; decisions use unrounded values. Per-mode observed accuracies behind the D_{obs} values are in Table 14 and the supplement (Section S.6).

Task	Grader	D_{obs} (pp)	ψ range (pp)	Decision
Mainstream panel				
gsm8k	lm-eval flexible	-13.2	pw $[-44.6, +28.8]$	abstain
			fw $[-65.0, +42.9]$	abstain
gsm8k	lm-eval strict	$+0.1$	pw $[-1.0, +0.1]$	abstain
			fw $[-2.8, +0.1]$	abstain
gsm8k	`math_verify`	$+11.9$	pw $[-19.5, +32.4]$	abstain
			fw $[-33.3, +43.2]$	abstain
gsm8k	`qwen_math`	$+11.8$	pw $[-19.5, +32.3]$	abstain
			fw $[-33.3, +43.1]$	abstain
math_full	lm-eval flexible	-2.7	pw $[-32.0, +7.7]$	abstain
			fw $[-48.0, +10.7]$	abstain
math_full	lm-eval strict	$+0.5$	pw $[-22.5, +0.5]$	abstain
			fw $[-39.1, +0.5]$	abstain
math_full	`math_verify`	$+31.3$	pw $[-39.4, +55.8]$	abstain
			fw $[-51.7, +71.8]$	abstain
math_full	`qwen_math`	$+40.3$	pw $[-17.8, +56.1]$	abstain
			fw $[-29.0, +65.8]$	abstain
mmlu	`mmlupro_official`	$+17.3$	pw $[-24.9, +41.4]$	abstain
			fw $[-40.4, +50.7]$	abstain
mmlu	`qwen_choice`	-15.7	pw $[-50.0, +49.7]$	abstain
			fw $[-70.5, +57.9]$	abstain
mmlu	`simple_evals_mc`	$+17.0$	pw $[-13.6, +40.2]$	abstain
			fw $[-29.0, +49.8]$	abstain
mmlu_pro	`mmlupro_official`	$+23.0$	pw $[-32.0, +48.2]$	abstain
			fw $[-46.1, +58.3]$	abstain
triviaqa	`triviaqa_em_aliases`	$+20.7$	pw $[-14.0, +46.1]$	abstain
			fw $[-27.2, +54.6]$	abstain
In-house panel				
gsm8k	`math_v1`	$+37.4$	pw $[-22.2, +50.9]$	abstain
			fw $[-34.8, +68.3]$	abstain
gsm8k	`math_v2_seg`	$+37.4$	pw $[-22.2, +50.9]$	abstain
			fw $[-34.8, +68.3]$	abstain
math_full	`math_v1`	$+37.5$	pw $[-32.6, +57.3]$	abstain
			fw $[-43.3, +71.9]$	abstain
math_full	`math_v2_seg`	$+37.5$	pw $[-32.6, +57.3]$	abstain
			fw $[-43.3, +71.9]$	abstain

continued on next page

Table 18 (continued)

Task	Grader	D_{obs} (pp)	ψ range (pp)	Decision
mmlu	mc_v1_fulltext	+16.6	pw $[-13.9, +40.2]$ fw $[-28.9, +49.5]$	abstain abstain
mmlu	mc_v2_buggy	+16.6	pw $[-14.9, +38.5]$ fw $[-29.5, +47.9]$	abstain abstain
mmlu	mc_v3_linesafe	+16.6	pw $[-13.9, +40.2]$ fw $[-28.9, +49.5]$	abstain abstain
mmlu_pro	mc_v1_fulltext	+20.7	pw $[-16.8, +44.9]$ fw $[-32.3, +54.6]$	abstain abstain
mmlu_pro	mc_v2_buggy	+10.5	pw $[-20.9, +39.7]$ fw $[-37.1, +48.7]$	abstain abstain
mmlu_pro	mc_v3_linesafe	+20.7	pw $[-16.8, +44.9]$ fw $[-32.3, +54.6]$	abstain abstain
triviaqa	em_v1_firstans	+22.5	pw $[-2.5, +38.9]$ fw $[-14.5, +45.8]$	abstain abstain
triviaqa	em_v2_seg	+22.5	pw $[-2.4, +38.9]$ fw $[-14.4, +45.8]$	abstain abstain

Table 19: Current design-based analysis. Direct count-inversion of the reference-accuracy gap $D = A_{\text{think}} - A_{\text{no-think}}$ on the same joint count events as Table 5 (miaudit_direct_count_decision_2026-09-25.json). Each cell's four (or more) group counts take two-sided hypergeometric intervals at level $\alpha_j = \alpha_c/J$; D's interval is the extreme composition $(R_1+T_1)/N_1 - (R_0+T_0)/N_0$ over that event, with no η parameter box. At the per-cell level ($\alpha_c = 0.05$), 5 of the 25 cells certify a positive reference-accuracy gap, shown above; the other 20 cells abstain (their intervals do not exclude zero). At the familywise level ($\alpha_c = 0.05/26$ mainstream, $0.05/24$ in-house) the direct inversion abstains on all 25 cells, as does the box relaxation of Table 18: the envelope's per-cell abstentions are a property of the relaxation, not of the counts. The five certified cells correspond to two distinct task-level reference gaps (math_full and triviaqa). Convention check: the same construction gives the MMLU qwen_choice FN-gap interval $[+11.7, +77.0]$pp, matching Table 5's $[+11.8, +77.0]$ at 0.1pp display precision. D_{obs} is the census observed gap (think − no-think).

Task	Grader	D_{obs} (pp)	Direct D interval, per-cell (pp)	Decision
math_full	qwen_math	+40.3	$[+7.6, +58.3]$	certify +
math_full	lm-eval flexible	−2.7	$[+2.4, +65.3]$	certify +
math_full	lm-eval strict	+0.5	$[+3.6, +61.9]$	certify +
triviaqa	em_v1_firstans	+22.5	$[+3.3, +42.0]$	certify +
triviaqa	em_v2_seg	+22.5	$[+3.3, +42.0]$	certify +

D.18 Prescreen-blind relabeling of the reference labels

The 132-row human adjudication of Section 3.3 saw the prescreen verdicts, so anchoring was not ruled out. We therefore relabeled 194 round-1 records blind to the prescreen labels (gold answers and candidate texts remained visible): the false-negative-carrier rows of the two headline cells plus a stratified random draw (design, stratum manifest, and labeler–gold comparison in the supplement, Section S.7). The blind labels disagree with the prescreen labels on 0 of 194 records. Because the design mixes certainty carriers with random draws, a zero-discovery inversion under the actual mixed design gives a one-sided 95% upper bound of 10 of 595 records (1.68%) on label disagreement within the original audit frame; this is not a bound on semantic error, corpus-wide label error, or agreement between independent human annotators. Swapping the blind labels into the reference set leaves both headline false-negative estimates unchanged to four decimals.

D.19 Audit of the machine-augmented sensitivity analysis

The original machine-augmented (PPI) sensitivity analysis claimed a large corrected MMLU gap by combining the original 595 records with two later machine-labeled batches (G1 relabels and a G2

Table 20: The input contract is part of the measurement channel (Appendix D.20 for per-path detail). Same fixed outputs, same parser; only the input segmentation or a single rule changes. *full*: status-quo contract. *seg*: only the post-`</think>` segment is scored. *no-ICL*: the vendored parser with its ICL-truncation rule disabled (single-rule counterfactual). *joint*: both changes at once (the `seg_noicl` rule of Appendix D.6; its reference FN diff is not estimated — the repair-audit family evaluates *seg*, the candidate remedy, not the joint variant). ICL share: rows whose extraction takes the ICL-truncation path. Per-mode observed accuracies behind the status-quo contract are in the supplement (Section S.6) ($48.1/32.3\%$ no-think/think for this cell); per-mode FN rates are in Table 13. The `qwen_choice` think-side accuracies behind the four channels are $32.3/64.2/62.3/64.2\%$ (full/seg/no-ICL/joint, from the segment-contract artifact): *seg* leaves no-think verdicts unchanged (no think blocks) and *no-ICL* raises the no-think side by 0.21pp. FN diff is the HT estimate on the 595 reference rows (Appendix D.1). The official end-to-end flows (own prompt templates) are not reproducible from fixed outputs; the official strict contract is degenerate on our generation template ($\leq 0.1\%$ both modes). The 79.6% share in the first row counts all rows taking the truncation path; the 79.5% reported in the main text is the narrower joint truncation-and-fallback share.

Grader × channel	D_{obs} (pp)	think extraction path	FN diff (pp)
`qwen_choice` full	-15.7	79.6% ICL-truncated	$+45.6$
`qwen_choice` seg	$+16.2$	96.1% no-trigger	-5.1
`qwen_choice` no-ICL	$+14.1$	93.4% trigger	$+4.4$
`qwen_choice` joint	$+16.0$	segment scored, ICL off	—
lm-eval flexible full	-13.2	last-number rule	$+16.4$
lm-eval flexible seg	-13.2	verdicts identical to full	$+16.4$

sample that permits gold revision). It is withdrawn from current inference: the batches mix weight targets and label conventions, so their combined weights do not estimate a single common-target quantity, and unbiased correction of a mean does not extend to FN/FP ratios under the fixed population (Angelopoulos et al., 2023). No interval or significance claim from that construction is current. Its estimator and outputs remain in the research archive for reproducibility. A same-labels, same-design comparison of the observed, label-only, multiplicative-inversion, and additive-residual gap estimators is in the supplement (Section S.7).

D.20 Input-segmentation channel comparison

The singleton mechanism — the MMLU think row whose gold and final answer are both C, scored A on the full text because seven "answer is" mentions fire the vendored parser's ICL-truncation heuristic — generalizes to a channel comparison on the fixed outputs (`miaudit_segment_contract_2026-09-18.*`). Channels were frozen before running: *full* (status-quo contract), *seg* (only the post-`</think>` segment scored), and *no-ICL* (the vendored parser with the truncation rule disabled, a single-rule counterfactual). Table 20 gives the headline numbers. The earlier path summary classified ICL truncation and fallback as exclusive states. That summary concealed their sequential overlap; Appendix D.6 reports the actual joint states. MMLU no-think rows ($n = 40{,}960$) contain no think blocks, so segmentation leaves their verdicts unchanged. Disabling the ICL rule is a different intervention: it raises no-think accuracy by 0.21pp. The upstream end-to-end flows ship their own generation prompts and cannot be reproduced from fixed outputs; the official strict contract is degenerate on our template, so byte-verbatim extraction functions on our generations do not stand in for the official end-to-end pipelines. For lm-eval flexible, segmentation changes no verdict. Its failures can occur in the answer segment, including recaps and nonnumeric regex matches (Appendix D.5).

Compliant feeding does not close the problem. The "user error" reading predicts that feeding inputs the way officials recommend — reasoning stripped to the post-`</think>` answer segment, as lm-eval's `think_end_token` and vendor guides do — makes every mainstream grader behave. The supplement (Section S.6) tabulates this test on all seven applicable cells (`miaudit_compliant_channel_2026-09-20.*`): the `qwen_choice` flip disappears ($-15.7 \to +16.2$; FN gap $+45.6 \to -5.1$) and lm-eval flexible's flip persists bit-for-bit (-13.2 observed, FN gap $+16.4$ unchanged); stripping does not remove answer-side parsing hazards. Com-

pliant feeding also removes the only nonzero think-arm FP rates (`qwen_choice` $18.2\% \to 0\%$, `mmlupro_official` $9.1\% \to 0\%$, observed point rates); no-think rows are identical by construction, and no FP events are observed in the four GSM8K cells in either channel (zero-count intervals in the supplement, Section S.6, are not point zeros). This input-segmentation change is insufficient for general validity; its own right-to-wrong transitions are in Appendix D.6.

D.21 Public incidents of the full-transcript failure family

The full-transcript feeding regime is documented in public pipelines, not only in our controlled corpus. Public records include leaderboard score discrepancies, reported extraction failures on reasoning-model outputs, and released evaluation scripts that feed full transcripts to the extractor; the vendor-official end-to-end flows we checked strip reasoning before extraction. The incident table and the environment details are in the supplement (Section S.7).

D.22 Simple worst-case content envelope

The bounded-overlap rule of Theorem 3 discards known directional errors when forming its worst-case bound. A simpler envelope keeps them: Proposition 4 states the resulting direct content interval in general form, using the per-row safety states and verdicts already retained by the audit. Per grader, correcting the observed gap by the audited directional false-negative masses and false-positive means identifies the contrast $c_1 - c_0$ exactly, and the residual width is the shared anchor-free mass. The direct interval is never wider than any per-grader directionally corrected one-sided bound (Proposition 4, dominance). The bounded-overlap identity of Theorem 3 remains the per-grader attribution form with its pointwise-optimal constant once row-level information is reduced to the aggregate overlap counts $A_k(\rho)$.

On the envelope analysis frozen 2026-07-30 and repackaged 2026-09-18 ($K{=}4$ GSM8K panel), the identified contrast is $c_1 - c_0 = 13.907$pp; with the anchor-free masses (supplement Section S.6) the direct interval gives $[+10.1, +14.3]$pp for every grader — including lm-eval flexible, for which the one-sided rule of Theorem 3 abstains. The direct baseline of Proposition 4's premise is computable here in full: its all-completions interval is $[+10.6, +14.2]$pp, nearly the safe-anchor interval $[+10.1, +14.3]$pp (row-level counts in the supplement, Section S.6). In this instance the multi-grader safety analysis therefore adds little identification power over computing C directly; its contribution is grader attribution — locating which grader and state carries the error mass — rather than a tighter comparison. These values are corpus-specific and post-hoc: the construction is general, but it is not an independent validation, and it does not lift to semantic truth Y (Corollary 3).