RETHINKING LLM JUDGES: CHAIN-OF-THOUGHT AND MULTI-STEP PIPELINES FOR MATH GRADING

Eric Chen, Aryan Gulati, Brando Miranda, Zeyu Tang, Sanmi Koyejo
Department of Computer Science
Stanford University
Stanford, CA 94305, USA
{ericc27, aryangul, brando9, zeyu, sanmi} @cs.stanford.edu

ABSTRACT

LLM judges are promising for evaluating reasoned mathematical solutions, yet their scores can be prompt-sensitive, unstable, and opaque. Two assumptions are currently prevalent: that chain-of-thought (CoT) reasoning provides little or no improvement in agreement with human grades for LLM judges, and that multi-step pipelines—e.g., debate or ensembles—outperform single-step evaluation. We challenge both. For CoT, on Putnam-AXIOM-Grading and IMO-GradingBench, two human-graded competition-mathematics benchmarks, we find that *unconstrained* CoT often reduces *evaluation consistency* and doesn't significantly affect performance. In contrast, *deliberately structured* CoT recovers and often improves agreement with human grades relative to *single-pass CoT-absent scoring*. This pattern is strongest for reasoning-optimized models such as DEEPSEEK-R1. For multi-step pipelines, popular methods—G-Eval, Chain-of-Verification, and Debate—consistently underperform simpler strategies. Our most striking finding is that *comparative prompting*—a single-pass strategy that explicitly compares student solutions to reference answers—is the most consistently high-performing strategy we tested: it ranks in the top three for correlation with human grades on all five models, and achieves the best correlation for three of them. Our findings point to a simple prescription for difficult mathematical grading: compare against a reference, reason within constraints, and avoid unnecessary complexity.

1 INTRODUCTION

The LLM-as-a-judge paradigm offers an appealing alternative to expensive human evaluation for its scale and adaptability (Zheng et al., 2023; Gu et al., 2024). Prior benchmark-based evaluations demonstrated that strong LLM judges can correlate well with human preferences on open-ended dialogue, while also exhibiting characteristic failure modes such as position and verbosity biases (Liu et al., 2023; Zheng et al., 2023; Shi et al., 2025). Subsequent work expanded the design space of LLM judges beyond single scalar ratings toward rubric-based, multi-criterion assessments and structured elicitation protocols (e.g., form-filling with explicit criteria) to better align with human judgments (Liu et al., 2023; Gu et al., 2024).

Mathematical reasoning—increasingly viewed as a bellwether for general AI capabilities (Hendrycks et al., 2021; Glazer et al., 2025)—stresses LLM evaluation in a qualitatively different way. LLM judges exhibit well-documented weaknesses in mathematical grading, frequently failing to identify errors even in problems they can solve when asked separately (Zheng et al., 2023), and suffering from prompt sensitivity, infinite looping, and excessive overthinking when asked to verify reasoning traces (Khalifa et al., 2025). To build trust in model outputs, there is strong motivation to evaluate the validity of the reasoning trace itself (Xia et al., 2024; Lightman et al., 2023; Zheng et al., 2025). To completely evaluate a reasoned solution, a judge must award partial credit for promising approaches, identify subtle logical gaps, and assess intermediate steps. However, partial credit remains a key unsolved pain point for LLM-based grading: judges can often detect clear errors, yet their score allocations are still noisy and poorly calibrated on partially correct solutions (Mahdavi et al., 2025).

Two largely orthogonal assumptions now guide the field's response to these challenges.

1. A growing body of evidence suggests that Chain-of-Thought (CoT) reasoning—which is known to improve *solving* performance (Wei et al., 2022)—provides *little to no benefit*

for LLM judges. Yamauchi et al. (2025) find that CoT offers minimal gains in alignment with human judgments when clear evaluation criteria are present, recommending CoT-free scoring as the best cost-performance tradeoff. This is consistent with findings that longer rationales increase instability and prompt sensitivity without reliably improving calibration (Haldar & Hockenmaier, 2025b; Khalifa et al., 2025).

2. Second, a parallel line of work argues that multi-step pipelines—debate frameworks, ensemble panels, and verification cascades—should outperform single-pass evaluation by aggregating diverse perspectives and catching errors through iterative refinement (Chan et al., 2024; Verga et al., 2024; Dhuliawala et al., 2024). Multi-step debate has been shown to improve alignment with human judgment on open-ended NLG tasks such as dialogue and summarization (Chan et al., 2024), and panel-based evaluation reduces intra-model bias (Verga et al., 2024). However, neither assumption has been rigorously tested on domains requiring precise logical verification, such as competition-level mathematical grading.

We challenge both.

In this paper, we study a focused question: *Under what conditions does using CoT improve or degrade a judge's agreement with human judgment and scoring stability for challenging mathematical solutions?* We provide an extensive comparison of CoT prompting strategies and judge architectures on two benchmarks. First, we introduce **Putnam-AXIOM-Grading**[1], a new human-graded dataset released with this paper, constructed from problems drawn from the Putnam-AXIOM benchmark (Gulati et al., 2025). Second, we evaluate on **IMO-GradingBench**, a human-graded competition-mathematics grading benchmark introduced as part of IMO-Bench (Luong et al., 2025)[2]. We evaluate along two axes: (i) **agreement with human judgment**, measured by Pearson correlation with human grades averaged over repeated runs, and (ii) **stability**, measured by the variance of grades across repeated runs averaged over all questions.

Contributions. We provide the first controlled comparison (for math reasoning judges) spanning direct scoring, structured CoT templates, and multi-step pipelines on human-graded competition-level mathematics. Our evaluation reveals:

- **Comparative prompting dominates all multi-step pipelines.** Contrary to the prevailing assumption that multi-step pipelines outperform single-step evaluation (Chan et al., 2024; Verga et al., 2024), we find that comparative prompting—a simple single-step strategy that explicitly compares student solutions to reference answers—is the *most consistently high-performing* strategy we tested. When averaged across both benchmarks (Table 4), comparative prompting achieves the highest mean agreement with human grades for three of five judge models, second-highest for one, and third for one—while requiring only a single model call. Multi-step pipelines including G-Eval, CoVe, and Debate often increase variance while barely increasing or even decreasing agreement.
- **CoT can work for judges—when deliberately structured.** Contrary to the emerging consensus that CoT provides little to no benefit for LLM judges (Yamauchi et al., 2025), we show that the problem is not CoT per se but unconstrained CoT. Across Putnam-AXIOM-Grading and IMO-GradingBench, naive CoT often *increases score variance* and can *reduce agreement with human grades*. However, deliberately structured CoT strategies—especially comparative prompting—consistently outperform direct scoring. Reasoning-trained models like DeepSeek-R1 benefit most strongly, with comparative prompting improving agreement by up to 76 points over no-CoT on average. We expand on this nuance in Section 4.
- **Holistic reasoning, not any single component, drives comparative prompting's advantage.** Targeted ablations decomposing comparative prompting into its constituent ingredients—explicit reference comparison, template structure, task framing, and component ordering—show that no single factor explains its effectiveness. Instead, all high-performing variants share a common property: they allow the judge to reason holistically about both solutions without rigid intermediate output requirements. Strategies that fragment evaluation into mechanical subtasks (binary checklists, forced citation, structured verification) consistently degrade performance, suggesting that the active ingredient is keeping the judge focused on substantive assessment rather than procedural compliance (Section E).

[1]Dataset available at `https://anonymous.4open.science/r/cot-judge-6703/`.

[2]For details on these benchmarks, see Section 3.1

These results have immediate practical relevance for practitioners building reward models or automated grading systems for challenging mathematical evaluation.

2 Related Work

LLM judges and structured evaluation. Using LLMs as evaluators has become common for open-ended tasks (Dubois et al., 2025; Liu et al., 2023). Early work demonstrated that strong judges correlate with human preferences while exposing biases (Zheng et al., 2023), motivating transitions from unstructured ratings to rubric-based evaluation. G-Eval introduced explicit evaluation steps (Liu et al., 2023), and subsequent work has explored criterion-wise scoring (Hashemi et al., 2024), dynamic rubric generation (Du et al., 2024), and locked rubric mechanisms (Hong et al., 2026). Recent surveys emphasize measuring both accuracy and reliability (Gu et al., 2024; Li et al., 2024). While these efforts predominantly target open-ended natural language generation tasks such as summarization, dialogue, and chatbot evaluation, our work extends this line of inquiry to hard mathematical grading.

Chain-of-thought stability and verification. CoT prompting elicits multi-step reasoning and improves solving performance (Wei et al., 2022), with self-consistency further boosting accuracy (Wang et al., 2023). However, translating these benefits to *judging* is nontrivial: longer rationales can increase score variance and prompt sensitivity (Haldar & Hockenmaier, 2025a), and self-correction without external signal is unreliable (Huang et al., 2023). Recent empirical work has strengthened the case against CoT for judges: Yamauchi et al. (2025) find that CoT provides minimal gains when clear evaluation criteria are present, and Khalifa et al. (2025) document that thinking LLM judges are highly sensitive to instruction wording and prone to failure modes including infinite looping and excessive overthinking. This emerging consensus has pushed the field in two directions: abandoning CoT in favor of direct scoring, or embedding reasoning in increasingly complex multi-step pipelines. Verification-style approaches like CoVe (Dhuliawala et al., 2024) and process-level supervision (Lightman et al., 2024) separate drafting from checking, while multi-step debate (Chan et al., 2024) and panel-based evaluation (Verga et al., 2024) aggregate multiple judge perspectives. We test whether either direction is correct—and find that neither fully is: the problem is not CoT *per se*, but whether the reasoning stays anchored to the reference.

Mathematical reasoning benchmarks and evaluation beyond final answers. Math benchmarking has evolved from final-answer accuracy (Hendrycks et al., 2021) toward process-aware assessment. This shift is driven by the recognition that partial credit assignment remains a critical unsolved problem: Mahdavi et al. (2025) demonstrate that while LLMs can reliably flag incorrect solutions, they exhibit substantial calibration gaps in how partial credit is awarded, motivating agentic workflows with problem-specific rubrics. Putnam-AXIOM targets high-difficulty problems requiring partial credit and error detection (Gulati et al., 2025); IMO-GradingBench provides human-graded competition solutions with similar requirements (Luong et al., 2025). Unlike prior judge studies emphasizing general NLG (Liu et al., 2023; Zheng et al., 2023), we focus on this demanding setting where evaluation requires precise step-level reasoning—and where LLM judges are known to be particularly unreliable (Zheng et al., 2023; Khalifa et al., 2025).

3 Methods

3.1 Task and data

We study the efficacy of LLM judge pipelines for grading reasoned mathematical solutions using two benchmarks: Putnam-AXIOM-Grading and IMO-GradingBench (Luong et al., 2025). Each evaluation instance consists of: (i) a problem statement, (ii) a ground-truth solution, and (iii) a model-generated solution to be graded.

3.1.1 Putnam-AXIOM-Grading

We introduce Putnam-AXIOM-Grading, a collection of 500 problems randomly sampled from the William Lowell Putnam Mathematics Competition. Each problem in this benchmark consists of a LaTeX-formatted question and a fully reasoned ground-truth solution. Qwen2.5-72B-Math-Instruct gave a sample student model output to each of these questions. These responses are frozen and kept uniform across all candidate judge evaluations.

All 500 problem–solution pairs were graded on a 5-component rubric (Figure 2)[3] by an undergraduate mathematics major at a top-ten U.S. mathematics program with extensive competition mathematics experience. The cumulative score across these components serves as the final human grade for each question and is used for correlation calculations.

We employ a detailed analytic rubric with explicit criteria for each score level; rubric-based scoring substantially reduces rater variability compared to holistic grading (Jonsson & Svingby, 2007; Reddy & Andrade, 2010). Moreover, mathematical correctness—unlike subjective constructs such as writing quality—permits objective verification: a step is either valid or contains an identifiable error. Also, research on rater reliability in mathematics assessment indicates that trained raters using well-specified rubrics achieve high intra-rater consistency (Eckes, 2023).

3.1.2 PUTNAM-AXIOM-GRADING

We introduce Putnam-AXIOM-Grading, a collection of 500 problems randomly sampled from the William Lowell Putnam Mathematics Competition. Each problem in this benchmark consists of a LaTeX-formatted question and a fully reasoned ground-truth solution. Qwen2.5-72B-Math-Instruct produced a candidate response to each problem; these responses are frozen and kept uniform across all candidate judge evaluations.

All 500 problem–response pairs were graded by a primary annotator—an undergraduate mathematics major at a top-ten U.S. mathematics program with extensive competition mathematics experience—using a 5-component analytic rubric (Figure 2).[4] Each component is scored on a 0-to-1 scale with 0.5 increments, and the cumulative score (0 to 5) serves as the final human grade used in all correlation analyses. We use a detailed analytic rubric with explicit criteria for each score level; rubric-based scoring substantially reduces rater variability compared to holistic grading (Jonsson & Svingby, 2007; Reddy & Andrade, 2010; Eckes, 2023), and mathematical correctness—unlike subjective constructs such as writing quality—permits objective verification.

logiTo validate label stability, a second independent annotator re-graded a stratified subset of 60 items using the same rubric, without access to the primary annotator's scores. The subset was selected deterministically to ensure uniform coverage across six difficulty buckets and all 11 topic domains (Appendix A). The two annotators achieved 90.0% exact agreement on total score (54/60), with a quadratic weighted kappa of 0.983 and mean absolute difference of 0.067. Component-level exact agreement ranged from 95% to 100% across all five rubric dimensions. Full breakdowns by difficulty, topic, and item are in Appendix A.

3.1.3 IMO-GRADINGBENCH

IMO-GradingBench is a benchmark consisting of 1000 student-written solutions to International Mathematical Olympiad (IMO)-level problems, annotated with human-assigned scores (Luong et al., 2025). Each example includes the original problem statement, a multi-step free-form solution authored by a model (the original authors do not specify which model), and one or more human-provided grades based on a rubric from 0-7 points that awards partial credit for reasoning quality, correctness of intermediate steps, and overall mathematical validity. Please see Luong et al. (2025) for more annotation details. The rubric used by our LLM judges to target this grading behavior is shown in Figure 3.[5] Like Putnam-AXIOM, the benchmark is designed to support evaluation of grading systems that must assess not only final answers but also the logical structure and correctness of multi-step reasoning.

3.2 RUBRIC AND JUDGE OUTPUT FORMAT

For both benchmarks, judges score each response under a fixed rubric, as described in Section 3.1. We prompt judges to emit their scores in a JSON format, enclosed by `<json>` tags to reduce parsing ambiguity, as was done in previous LLM judge works (Zheng et al., 2023; Liu et al., 2023).

3.3 JUDGE PROMPTING STRATEGIES

We implement a suite of single-step and multi-step judge prompting strategies, all sharing the same input context (problem, ground truth, model output) and the same rubric text. Strategies differ only in how they elicit and/or aggregate rationales before producing final scores.

[3]For details on the Putnam rubric, see Section F.

[4]For details on the Putnam rubric, see Section F.

[5]For details on the IMO rubric, see Section F.

Single-step strategies.

- **No-CoT**: direct JSON scoring with no rationale.
- **Brief-CoT**: 1–2 sentence assessment then JSON.
- **Full-CoT**: free-form step-by-step rationale then JSON, in the spirit of Chain-of-Thought prompting (Wei et al., 2022). In Putnam-AXIOM-Grading, this step-by-step rationale was given for every rubric component.
- **Structured-CoT**: a fixed three-stage template (summarize $\rightarrow$ check key steps $\rightarrow$ decide scores), intended to reduce unstructured drift.
- **Self-critique**: initial assessment followed by a critique-and-revise phase (motivated by the broader literature on self-critique/self-correction (Huang et al., 2023)).
- **Bullet-point**: criterion-wise bullet justification. Each bullet is very terse, typically a phrase or sentence fragment rather than a complete sentence.
- **Comparative**: explicit comparison between the model solution and the ground truth (related to findings that comparative judgments may be more reliable than absolute scoring in some settings (Zheng et al., 2023)).
- **Quote-forcing**: every evaluative claim must be supported by a direct quote from the model output. This evidence constraint is designed to mitigate hallucinated critiques, a known failure mode of unconstrained CoT (Turpin et al., 2023; Chen et al., 2025).

Multi-step strategies.

- **G-Eval** (Liu et al., 2023): (i) generate an instance-specific checklist of atomic evaluation steps; (ii) score using the checklist with explicit pass/fail reasoning.
- **Chain-of-Verification (CoVe)** (Dhuliawala et al., 2024): (i) draft evaluation; (ii) generate verification questions for draft; (iii) answer questions using only the original instance (to reduce anchoring); (iv) finalize scores conditioned on verification results.
- **Multi-judge ensemble**: three base judges with distinct personas (pedantic, holistic, teacherly) independently evaluate each solution with free-form CoT. The strict persona focuses on technical correctness and penalizes minor errors; the holistic persona weighs overall solution quality and mathematical insight; the teacherly persona considers partial credit generously and rewards demonstrated understanding. Our choice of personas is motivated by the general principle that ensemble diversity improves robustness (Verga et al., 2024), though we note that systematic optimization of persona selection remains an open problem.

  A chair judge then synthesizes these outputs by reviewing all three scores and rationales to produce a final score. This is inspired by the general observation that aggregating multiple evaluators can reduce idiosyncratic bias (Verga et al., 2024; Kalra & Tang, 2025).
- **Debate**: a pro-judge argues for maximal credit and a con-judge argues for minimal credit; an arbiter sees only the arguments from the pro-judge and con-judge to produces final scores. This resembles multi-step debate evaluators (Chan et al., 2024), but instantiated here as a deliberately polarized two-sided argument followed by arbitration.
- **Overseer-guided retry**: the judge produces a full-CoT evaluation, then an *overseer* model (the same model as the judge, but with a separate prompt focused on error detection in 5 categories) audits the evaluation for rubric-level errors; if the overseer flags any error, the judge is re-sampled up to five attempts and we select the first overseer-approved output (or, if none are approved, the attempt with the fewest flagged fields). This implements the pattern of using rationales as audit artifacts rather than directly trusting them (Lightman et al., 2024; Lanham et al., 2023; Chen et al., 2025).

For the explicit prompts we utilized, see Section H.1 for prompts used in the Putnam-AXIOM-Grading evaluations and Section H.2 for prompts used in the IMO-GradingBench evaluations.

3.4 Generation and execution

We evaluate four open-sourced models as judges: Qwen2.5-32B-Instruct, Deepseek-R1-Distilled-Qwen-32B, Qwen2.5-72B-Instruct, and Meta-Llama-3.1-70B-Instruct. All open-source models are

served with batched inference via `vLLM` (Kwon et al., 2023), using tensor parallelism across either two H200 GPUs or four A100 GPUs. For each open-source model and prompting strategy, we run five independent trials. We use temperature 0.7 to ensure sufficient stochasticity for meaningful variance estimation across repeated runs. We also use nucleus sampling at top_p=1 (VLLM default), maximum output length of 2048 tokens, and stop sequences at JSON closing tags to ensure well-formed outputs. Seeds 43-47 provide five independent trials per condition. For the overseer-guided retry strategy, we allow up to 5 retry iterations with overseer responses capped at 800 tokens.

Multi-step strategies (e.g., G-Eval, CoVe, debate, ensemble, overseer-guided retry) are implemented as explicit multi-call pipelines, whereas single-turn strategies issue one model call per instance.

To compare with proprietary models, we additionally evaluate GPT-5.1 (accessed via the OpenAI API as snapshot `gpt-5.1-2025-11-13`). Due to cost constraints, we run only three trials and restrict to a subset of strategies: No-CoT, Full-CoT, Comparative, Quote-Forcing, Multi-Judge, G-Eval, CoVe, and Overseer-guided Retry. This subset spans the range of performance observed in open-source experiments while including established baselines.

3.5 Evaluation metrics

We evaluate judge reliability along two axes, consistent with recent work on benchmarking and stress-testing LLM judges (Tan et al., 2024; Thakur et al., 2024; Haldar & Hockenmaier, 2025a):

Agreement with Human Judgment. For a judge to be reliable, its judgments should be similar to those of human experts. As such, we compute the Pearson correlation between each strategy's scores and the human-given grades on both benchmarks. This is a standard measurement of linear score calibration even when absolute score scales may differ across prompts or models (Zheng et al., 2023; Liu et al., 2023).

Within-condition Score Variance. For a judge to be reliable, it should assign (approximately) the same grade when presented with the same problem–solution pair. To quantify stability, we compute the variance of total scores across the five repeated runs for each instance under a fixed strategy. We then report the mean of these per-instance variances for each strategy. Strategies with lower mean variance produce more consistent scores. This directly targets the empirically observed randomness and prompt sensitivity of LLM judge scoring (Haldar & Hockenmaier, 2025a; Thakur et al., 2024).

To account for the wide variation in resource consumption across prompting strategies—and thereby enable a fair assessment of the cost-efficiency tradeoff for multi-step judges—we also track end-to-end token usage for every judgment strategy. Our token-sum formulae are described in Section C.

4 Results

Table 4 summarizes agreement (Pearson correlation with human grades, averaged across both benchmarks) and variance (mean within-condition variance, averaged across both benchmarks) for all strategies and models. Full per-benchmark results appear in Tables 4 and 5 in Appendix B. Plots of agreement vs. stability and the Pareto frontier for each model and benchmark are displayed in Figure 1.

Comparative prompting is consistently strong. Comparative prompting—which explicitly asks the judge to compare the student solution against the reference—is the single most reliable strategy across our entire evaluation. When averaged across both benchmarks (Table 4), it achieves the highest mean agreement for three of five models (Qwen2.5-72B, DeepSeek-R1-32B, and Llama-3.1-70B), second-highest for Qwen2.5-32B, and third for GPT-5.1. No other strategy achieves top-three performance this consistently across all models. On Putnam-AXIOM-Grading, it achieves the highest agreement for GPT-5.1 (0.87) and remains competitive for Qwen2.5-72B, DeepSeek-R1-32B, and Llama-3.1-70B (always within 2.0 points). On IMO-GradingBench, it achieves the highest agreement among all open-source models (0.45–0.50 Pearson; Table 5). Unlike multi-judge ensembles, which show benchmark-dependent performance, comparative prompting transfers robustly—and does so with only a single model call.

As shown in Section D, comparative prompting uses roughly $3 - 5\times$ fewer tokens than its closest competitors, which are multi-step strategies. This is a striking result: a single-step strategy consistently outperforms multi-step pipelines that require multiple model calls, directly challenging the assumption that pipeline complexity improves judge reliability (Chan et al., 2024; Verga et al., 2024).

Table 1: Average agreement and variance across Putnam-AXIOM-Grading and IMO-GradingBench. We report $\bar{\rho} = \frac{1}{2}(\rho_{\text{Putnam}} + \rho_{\text{IMO}})$ (higher is better) and $\overline{\sigma^2} = \frac{1}{2}(\sigma^2_{\text{Putnam}} + \sigma^2_{\text{IMO}})$ (lower is better). Marking is based on $\bar{\rho}$ *within each model*: best is **bold**, second-best is underlined, third-best is marked with †. For GPT-5.1, we show "–" where both source tables report "–".

		Qwen2.5-32B		Qwen2.5-72B		DeepSeek-R1-32B		Llama-3.1-70B		GPT-5.1	
Type	**Strategy**	$\bar{\rho}$	$\overline{\sigma^2}$	$\bar{\rho}$	$\overline{\sigma^2}$	$\bar{\rho}$	$\overline{\sigma^2}$	$\bar{\rho}$	$\overline{\sigma^2}$	$\bar{\rho}$	$\overline{\sigma^2}$
Baseline	No-CoT	.582†	**.162**	.558	.148	.510	.310	.491	.308	.768	**.045**
Single-Stage	Brief-CoT	.577	.390	.500	.213	.567	.319	.535	.440	–	–
	Full-CoT	.576	.314	.518	.168	.554	.300	.544	.361†	.777†	.139
	Structured-CoT	.561	.423	.468	.210	.581	.324	.545	.448	–	–
	Bullet-Point	.572	.285†	.552†	.224	.548	.291	.582	.532	–	–
	Comparative	.597	.286	**.613**	.198	**.586**	.312	**.593**	.564	.777†	.168
	Quote-Forcing	.549	.412	.497	.226	.561	.332	.540	.460	**.807**	.106
	Self-Critique	.549	.356	.491	.224	.556	.289†	.524	.503	–	–
Multi-Step	G-Eval	.470	.672	.477	.427	.520	.479	.474	.529	.737	.394
	CoVe	**.607**	.359	.532	.194	.549	.329	.568†	.424	.757	.203
	Multi-Judge	.582†	.220	.544	**.106**	.580†	**.239**	.556	**.286**	.780	.144
	Debate	.390	.760	.456	.585	.520	.736	.450	1.105	–	–
	Overseer	.578	.312	.514	.163†	.546	.287	.534	.379	.756	.134†

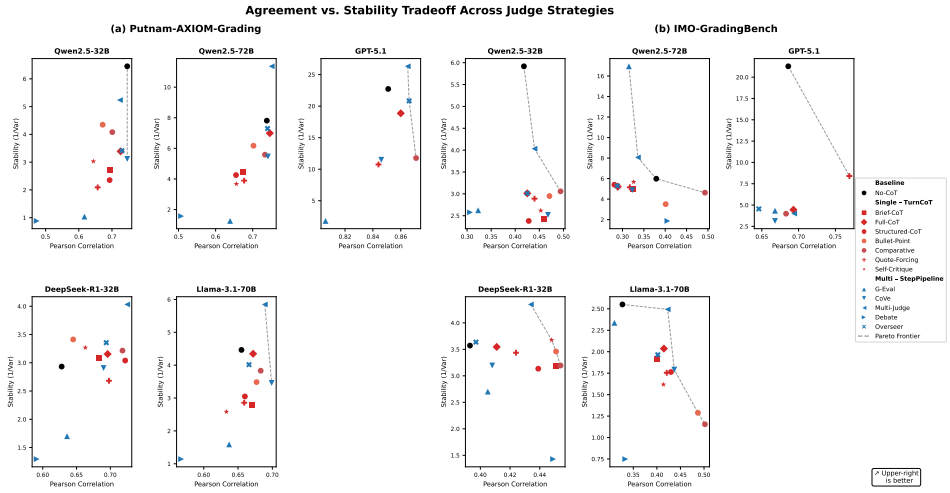

Figure 1: Agreement–stability tradeoff across LLM judge strategies on Putnam-AXIOM-Grading and IMO-GradingBench. Each panel corresponds to a different judge model. The x-axis reports Pearson correlation with human grades (agreement) averaged over repeated runs, and the y-axis reports stability measured as the inverse variance of total scores across repeated runs (*higher is more stable*) averaged across all problems. Markers denote prompting/pipeline strategies (single-turn CoT variants in red; multi-step pipelines in blue; No-CoT baseline in black). The dashed curve highlights the Pareto frontier within each model: strategies for which no other strategy achieves both higher agreement and higher stability.

Unconstrained CoT often degrades reliability. Across both benchmarks and most models, full chain-of-thought prompting yields worse agreement or higher variance than direct scoring. On Putnam-AXIOM-Grading, Qwen2.5-32B shows decreased Pearson correlation (2.0 points) with increased variance when using full-CoT versus no-CoT, while Qwen2.5-72B shows a modest gain in agreement (+0.8 points) but at increased variance. On IMO-GradingBench, Qwen2.5-72B drops from 0.38 (no-CoT) to 0.29 (full-CoT). Even for GPT-5.1, full-CoT provides negligible improvement in agreement (+0.7–0.9 points) at a stability cost (+0.9–17.8 points). This challenges the assumption that elicited rationales inherently improve evaluation quality.

G-Eval and Debate consistently underperform. Both methods exhibit high variance and low agreement across both benchmarks and all models tested. G-Eval variance ranges from 0.56–0.96 on Putnam and 0.06–0.43 on IMO, often exceeding other strategies by 2–5×. Debate shows even more extreme instability, with variance reaching 1.14 on Putnam (Qwen2.5-32B) and 1.34 on IMO (Llama-3.1-70B). These strategies appear unsuitable for difficult mathematical grading regardless of problem distribution, likely because dynamically generated checklists (G-Eval) and polarized argumentation (Debate) introduce compounding stochasticity across stages.

DeepSeek-R1 benefits from CoT across both benchmarks. DeepSeek-R1-Distill-Qwen-32B is a consistent exception to the pattern of CoT degrading reliability: in particular, singe-pass CoT-absent scoring always underperforms. On Putnam-AXIOM-Grading, structured-CoT improves agreement by 9.4 points over no-CoT (0.72 vs. 0.63). On IMO-GradingBench, comparative prompting achieves 0.45 versus 0.39 for no-CoT. Notably, the model exhibits a strong inductive bias toward deliberation: even when explicitly prompted to output only JSON, 99.4% of No-CoT outputs contain an internal reasoning scratchpad averaging over 2,000 characters—only 10% shorter than the Full-CoT scratchpads. The model *cannot help but reason*; the difference is what happens afterward.

Under Full-CoT, 86% of outputs exhibit a two-phase structure: a free-form scratchpad followed by a structured `<THINKING>` summary that distills observations into per-criterion assessments before scoring. Under No-CoT, the model proceeds directly from scratchpad to JSON with no consolidation step. This missing distillation phase produces a characteristic failure pattern: the No-CoT scratchpad frequently *identifies* discrepancies with the ground truth but then rationalizes them away, defaulting to leniency on surface-level criteria like methodological alignment and logical coherence, while applying error awareness nearly at random. The net effect is score compression toward the midrange (65% of No-CoT scores fall in 1.5–4.0 vs. 49% for Full-CoT and humans), reducing discrimination and lowering overall correlation. Without an explicit output-side reasoning structure, the model's exploratory deliberation passes through to the final score unchecked (see Appendix G and Figures 4–6 for detailed examples).

Frontier models compress strategy differences. GPT-5.1 substantially outperforms all open-source models (0.84–0.87 Pearson on Putnam vs. 0.70–0.75 for open-source; 0.65–0.77 on IMO vs. 0.28–0.50). However, the relative differences between strategies are much smaller: the gap between best and worst strategies (excluding G-Eval) is only 2–3 points for GPT-5.1, versus 10+ points for some open-source models. This suggests that strategy choice matters less as model capability increases, though comparative prompting remains among the top performers even at the frontier.

Multi-judge ensembles: strong on Putnam, weak on IMO. The most striking cross-benchmark divergence involves multi-judge ensembles. On Putnam-AXIOM-Grading, multi-judge achieves the highest agreement for Qwen2.5-72B (0.75) and DeepSeek-R1-32B (0.73), while maintaining the lowest or second-lowest variance across all models on average. On IMO-GradingBench, multi-judge *never* Pareto-dominates and often underperforms simpler baselines: for Qwen2.5-32B, it achieves lower agreement than comparative (0.44 vs. 0.49) with only modest variance improvement. This suggests that persona-based ensembling may be rubric- or distribution-specific rather than universally beneficial.

5 Discussion

Our results challenge several intuitions about LLM judge pipelines for mathematical evaluation.

Less is more. The most important finding is that *complex pipelines do not reliably outperform simple baselines*, directly challenging the assumption—supported by prior work on NLG tasks (Chan et al., 2024; Verga et al., 2024)—that multi-step methods should dominate. Multi-judge ensembles, which appeared robust on Putnam-AXIOM-Grading, failed to generalize to IMO-GradingBench. Verification-based methods like CoVe and G-Eval consistently increased variance without commensurate gains in agreement. In contrast, comparative prompting—a simple single-turn strategy—is the *most consistently high-performing strategy across both benchmarks* (Table 4), achieving top-three mean agreement for all five models while requiring only a single model call. This also challenges the emerging consensus that CoT is unhelpful for LLM judges (Yamauchi et al., 2025): comparative prompting *is* a CoT strategy—it elicits explicit reasoning about the relationship between solutions—but it succeeds because that reasoning is tethered to a concrete reference rather than left unconstrained.

Our ablation analysis (Appendix E) sharpens this conclusion: the common thread among high-performing variants is that they enable holistic side-by-side reasoning without imposing rigid intermediate output requirements.

Comparative Prompting Ablations. Why does comparative prompting succeed where elaborate verification fails? Our ablation results (Appendix E), in which we describe experiments A0-A6, suggest that the mechanism is more nuanced than simple reference-grounding. While explicit reference comparison provides a *grounding constraint* that reduces the degrees of freedom available to the judge, merely having the reference in context (A1, which does not explicitly instruct comparison) performs nearly as well as explicit comparison (A0). The critical factor appears to be enabling *holistic evaluation-focused reasoning*: comparative prompting lets the judge freely reason about the relationship between the two solutions without being forced into rigid intermediate structures. No single component—neither explicit reference-anchoring, template structure, nor task framing—explains comparative prompting's advantage on its own. Instead, strategies that fragment evaluation into mechanical subtasks (checklist completion, forced citation, binary verification, as in A3 and A5) consistently degrade performance, suggesting that the active ingredient is keeping the judge focused on substantive assessment rather than procedural compliance. Our finding parallels observations that pairwise comparison can better discern subtle quality differences (Zheng et al., 2023), but adds the insight that the comparison must remain holistic rather than decomposed.

The agreement–stability tradeoff. A second core finding is that agreement with human grades alone is an insufficient metric for judge reliability. Several strategies that achieve competitive correlation—including CoVe and some multi-step pipelines—exhibit substantially higher variance than simpler alternatives. For applications requiring reproducible scores (e.g., reward modeling, model selection), a high-variance judge can be worse than a lower-agreement but stable one: noisy reward signals destabilize RL training, and inconsistent grading undermines trust in automated evaluation. Our two-axis framework (Figure 1) makes these trade-offs explicit in a way that single-metric comparisons cannot. This framework also reveals that the agreement–stability Pareto frontier is neither model- nor benchmark-invariant. For Qwen2.5-72B on Putnam-AXIOM-Grading, multi-judge achieves the best of both axes simultaneously; on IMO-GradingBench, the same model's frontier fragments, with different strategies dominating each axis. Comparative prompting consistently appears on or near the Pareto frontier across both benchmarks—reinforcing its robustness.

CoT is not the problem; unconstrained CoT is. Our results offer a sharper view than the prevailing position that CoT is unhelpful for LLM judges (Yamauchi et al., 2025). While we confirm that *unconstrained* CoT (full-CoT, brief-CoT) often degrades reliability—consistent with findings that LLM-as-a-judge verification is sensitive to instruction wording and prone to failure modes like overthinking (Khalifa et al., 2025)—we show that *deliberately structured* CoT, particularly comparative prompting, is the strongest strategy we tested (Table 4). The distinction is not between "reasoning" and "no reasoning," but between reasoning that is tethered to a concrete evaluation task and reasoning that drifts into unconstrained deliberation. This matters for the broader problem of partial credit assignment in mathematical evaluation (Mahdavi et al., 2025): our results suggest that effective partial credit grading requires prompts that channel LLM reasoning toward substantive comparison with reference solutions, not prompts that ask for unconstrained analysis or mechanical verification procedures.

This two-part picture—structured CoT helps, unconstrained CoT hurts—holds cleanly for instruction-tuned models. DeepSeek-R1, trained via reinforcement learning on reasoning traces, is a consistent exception: it benefits from CoT regardless of structure, as we mention above (Section 4).

6 LIMITATIONS

Limited model coverage. We evaluated four open-source model families and one proprietary model (GPT-5.1). Results may not generalize to other architectures or training regimes. In particular, we did not evaluate recent reasoning-specialized models beyond DeepSeek-R1, which showed distinctive patterns that may extend to other similar systems.

Prompt sensitivity. Although we tested multiple prompting strategies, each strategy was instantiated with a single prompt template. LLM judges are known to be sensitive to prompt wording; different instantiations of the same conceptual strategy might yield different results. Our findings characterize the strategies as implemented, not the space of possible implementations. Furthermore,

our debate strategy uses explicitly polarized judges—one instructed to argue for maximal credit regardless of solution quality, the other for minimal credit. Though this tests whether arbitration can recover signal from deliberately extreme positions, a setting motivated by prior work on scalable oversight (Irving et al., 2018), we acknowledge that alternative debate topologies such as collaborative prompting may yield different results.

7 Conclusion

We present a systematic study of chain-of-thought (CoT) prompting strategies for LLM judges on difficult mathematical evaluation, a domain where partial credit remains largely unsolved. Our results overturn two prevailing assumptions: that CoT reasoning is unhelpful for LLM judges, and that multi-step pipelines outperform single-pass evaluation.

Comparative prompting—a single-pass strategy requiring only one model call—is the most consistently high-performing approach across both benchmarks and nearly all models tested. It achieves top-three agreement for all five judge models (Table 4), outperforming multi-step pipelines costing three to five calls. Ablations trace this advantage to holistic side-by-side reasoning rather than any single template component. Meanwhile, unconstrained CoT increases variance without improving agreement, and popular methods like G-Eval and Debate consistently underperform demonstrating that neither abandoning CoT nor adding pipeline complexity solves the problem.

For practitioners, we distill three guidelines: (1) default to comparative prompting—it is accurate, stable, cheap, and transfers across benchmarks; (2) measure stability alongside agreement, since high-variance judges may be unfit for deployment even when mean performance looks acceptable; and (3) match prompts to model class: reasoning-trained models benefit from explicit deliberation, while general-purpose models often perform best with simpler prompts.

References

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/25cc3adf8c85f7c70989cb8a97a691a7-Paper-Conference.pdf.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025. Also discussed in Anthropic research blog.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.212. URL https://aclanthology.org/2024.findings-acl.212/.

Hanyu Du, Parvez Rashid, and Qian Jia. Interactive rubric generator for instructor's assessment using prompt engineering and large language models. In *Proceedings of the IEEE Frontiers in Education Conference*. IEEE, 2024. URL https://ieeexplore.ieee.org/abstract/document/10893448.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators, 2025. URL https://arxiv.org/abs/2404.04475.

Thomas Eckes. *Introduction to many-facet Rasch measurement*. Peter Lang, 2023.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2025. URL https://arxiv.org/abs/2411.04872.

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. *arXiv preprint arXiv:2411.15594*, 2024. URL https://arxiv.org/abs/2411.15594.

Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Fronsdal, Bruno Dumont, Elyas Obbad, and Sanmi Koyejo. Putnam-axiom: A functional and static benchmark for measuring higher level mathematical reasoning in llms, 2025. URL https://arxiv.org/abs/2508.08292.

Ishita Haldar and Julia Hockenmaier. Rating roulette: Do language models give consistent ratings to different parts of a document? In *Proceedings of the 2025 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR)*, 2025a. doi: 10.1145/3701716.3715530. URL https://dl.acm.org/doi/10.1145/3701716.3715530.

Rajarshi Haldar and Julia Hockenmaier. Rating roulette: Self-inconsistency in llm-as-a-judge frameworks, 2025b. URL https://arxiv.org/abs/2510.27106.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13806–13834. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.745. URL http://dx.doi.org/10.18653/v1/2024.acl-long.745.

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

Yufei Hong, Hao Yao, Biao Shen, Wenxin Xu, and Hao Wei. Rulers: Locked rubrics and evidence-anchored scoring for robust llm evaluation. *arXiv preprint arXiv:2601.08654*, 2026. URL https://arxiv.org/abs/2601.08654.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023. URL https://arxiv.org/abs/2310.01798.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate, 2018. URL https://arxiv.org/abs/1805.00899.

Anders Jonsson and Gunilla Svingby. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, 2(2):130–144, 2007.

Nimit Kalra and Leonard Tang. VERDICT: A library for scaling judge-time compute. *arXiv preprint arXiv:2502.18018*, 2025.

Muhammad Khalifa et al. Process reward models that think. *arXiv preprint arXiv:2504.16828*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th ACM Symposium on Operating Systems Principles (SOSP)*, 2023. doi: 10.1145/3600006.3613165. URL https://arxiv.org/abs/2309.06180.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. LLMs-as-judges: A comprehensive survey on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024. doi: 10.48550/arXiv.2412.05579. URL https://arxiv.org/abs/2412.05579.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2305.20050.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2511–2522, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Thang Luong, Dawsen Hwang, Hoang H. Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, Alex Zhai, Clara Huiyi Hu, Henryk Michalewski, Jimin Kim, Jeonghyun Ahn, Junhwi Bae, Xingyou Song, Trieu H. Trinh, Quoc V. Le, and Junehyuk Jung. Towards robust mathematical reasoning, 2025. URL https://arxiv.org/abs/2511.01846.

Hamed Mahdavi et al. RefGrader: Automated grading of mathematical competition proofs using agentic workflows. *arXiv preprint arXiv:2510.09021*, 2025.

Y Malini Reddy and Heidi Andrade. A review of rubric use in higher education. *Assessment & evaluation in higher education*, 35(4):435–448, 2010.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge, 2025. URL https://arxiv.org/abs/2406.07791.

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. *arXiv preprint arXiv:2410.12784*, 2024.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. *arXiv preprint arXiv:2406.12624*, 2024. URL https://arxiv.org/abs/2406.12624.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023. URL https://arxiv.org/abs/2305.04388.

Patrick Verga, Sebastian Hofstätter, Nazneen Rajani, Mohit Iyyer, et al. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024. URL https://arxiv.org/abs/2404.18796.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2201.11903.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating mathematical reasoning beyond accuracy. *arXiv preprint arXiv:2404.05692*, 2024. URL https://arxiv.org/abs/2404.05692.

Yusuke Yamauchi, Taro Yano, and Masafumi Oyamada. An empirical study of LLM-as-a-judge: How design choices impact evaluation reliability. *arXiv preprint arXiv:2506.13639*, 2025.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Processbench: Identifying process errors in mathematical reasoning, 2025. URL https://arxiv.org/abs/2412.06559.

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. *arXiv preprint arXiv:2306.05685*, 2023. URL https://arxiv.org/abs/2306.05685.

A Inter-Annotator Agreement Study

A.1 Subset Construction

To evaluate the robustness of the benchmark labels, we constructed a fixed 60-item subset from the 500-item Putnam-AXIOM-Grading benchmark for independent re-annotation. We chose 60 rather than 50 so that the sample could be allocated evenly across six ordered difficulty buckets, with 10 items per bucket.

For modern Putnam identifiers (e.g., `2023_A6`), the trailing problem number was used as the difficulty index directly. Older historical identifiers do not follow the modern A/B convention and instead come from a single-sequence exam format with problems numbered from roughly 1 to 14. Their within-exam problem number was mapped into six ordered buckets using $\lceil \text{problem\_number} \times 6 \,/\, 14 \rceil$, preserving an approximate easy-to-hard ordering under a common six-level scale. Topic metadata was taken from the original Putnam-AXIOM dataset (Gulati et al., 2025), which tags each item with one or more of 11 benchmark domains.

A.2 Representation Analysis

The final subset contains exactly 10 items per difficulty bucket. In the full benchmark, bucket sizes range from 67 to 99, so within-bucket sampling fractions range from 10.1% to 14.9%. A purely proportional sample would have allocated fewer items to the hardest buckets, weakening the ability to test annotator stability on difficult partial-credit judgments.

Topic coverage is shown in Table 2. The subset covers all 11 domains. Low-frequency domains such as Differential Equations, Complex Numbers, and Probability are intentionally over-sampled relative to benchmark prevalence, ensuring sufficient representation for meaningful robustness claims.

Table 2: Topic frequencies in the benchmark (500 items) and the agreement subset (60 items). Items may carry multiple topic labels.

Topic	Benchmark	Subset
Algebra	178	17
Analysis	167	16
Calculus	136	15
Number Theory	127	14
Geometry	119	15
Combinatorics	77	14
Linear Algebra	40	11
Probability	26	11
Trigonometry	24	10
Complex Numbers	18	11
Differential Equations	18	11

Difficulty buckets 2–6 contain all 11 topics in the subset. Bucket 1 contains 10, which is the maximum possible because the benchmark itself has no Differential Equations items in that bucket.

A.3 Agreement Results

Overall. On the aggregate `score_total`, the two annotators achieved 90.0% exact agreement (54/60), 98.3% agreement within 0.5 points (59/60), and a mean absolute difference of 0.067. Pearson correlation was 0.983, Spearman correlation was 0.975, linear weighted kappa was 0.953, and quadratic weighted kappa was 0.983.

Rubric dimensions. Component-level exact agreement rates were: `score_logical_coherence` 100%, `score_methodological_alignment` 100%, `score_error_awareness` 98.3%, `score_faithfulness_to_task` 95.0%, and `score_intermediate_correctness` 95.0%.

By difficulty. Difficulty buckets 4–6 showed perfect exact agreement on total score. Buckets 1–3 had exact agreement of 70%, 90%, and 80%, respectively, with quadratic weighted kappa values of

0.966, 0.993, and 0.838. This pattern is consistent with the expectation that mild disagreement is most likely on easier or mid-range responses where partial-credit distinctions are more ambiguous.

By topic. Topic-level agreement was broadly strong. Some domains (e.g., Geometry, Trigonometry, Calculus) showed modestly lower exact agreement and weighted kappa than others (e.g., Number Theory, Combinatorics, Analysis). Because per-topic sample sizes are small and topic labels overlap, these differences should be interpreted cautiously. Even the weaker slices remained in a regime of high adjacent agreement and strong correlation.

A.4 Item-Level Disagreements

The six items on which the annotators assigned different total scores are listed below.

Table 3: Items with non-matching `score_total` in the 60-item agreement subset.

Problem ID	Grader 1	Grader 2
`1983_A2`	2.5	3.0
`1993_B3`	2.5	3.0
`1996_A1`	1.5	2.0
`1998_B3`	1.0	2.5
`2002_A1`	4.0	3.5
`2018_B1`	3.0	2.5

With one exception (`1998_B3`, a 1.5-point gap), all disagreements were limited to a single 0.5-point step on the 0-to-5 scale.

B PER-BENCHMARK TABLES

Table 4: Agreement and stability of judge strategies on **Putnam-AXIOM-Grading**. We report Pearson correlation with human grades (ρ, higher is better) and mean within-condition variance (σ^2, lower is better). Best results per model are **bolded**; second-best are underlined. Strategies are grouped by type: single-turn baselines, single-turn CoT variants, and multi-step pipelines. Some strategies in GPT-5.1 are omitted due to cost constraints.

		Qwen2.5-32B		**Qwen2.5-72B**		**DeepSeek-R1-32B**		**Llama-3.1-70B**		**GPT-5.1**	
Type	**Strategy**	ρ	σ^2	ρ	σ^2	ρ	σ^2	ρ	σ^2	ρ	σ^2
Baseline	No-CoT	**.746**	**.155**	.738	.128	.628	.341	.655	.224	.851	.044
	Brief-CoT	.695	.366	.675	.226	.683	.324	.670	.357	—	—
	Full-CoT	.726	.295	.746	.143	.696	.317	.672	.230	.860	.053
	Structured-CoT	.693	.425	.655	.235	.722	.329	.660	.328	—	—
Single-Turn	Bullet-Point	.672	.230	.702	.162	.645	.293	.677	.287	—	—
	Comparative	.701	.245	.733	.179	.718	.311	.683	.261	**.871**	.085
	Quote-Forcing	.657	.478	.677	.257	.698	.373	.659	.350	.844	.093
	Self-Critique	.644	.330	.656	.272	.663	.306	.633	.387	—	—
	G-Eval	.617	.962	.639	.795	.636	.588	.637	.629	.806	.556
	CoVe	.746	.321	.741	.183	.690	.344	**.699**	.289	.846	.087
Multi-Step	Multi-Judge	.724	.191	**.751**	**.088**	**.725**	**.248**	.689	**.171**	.865	**.038**
	Debate	.473	1.132	.507	.636	.591	.772	.567	.875	—	—
	Overseer	.730	.293	.740	.137	.694	.298	.666	.249	.866	.048

Table 5: Agreement and stability of judge strategies on **IMO-GradingBench**. We report Pearson correlation with human grades (ρ, higher is better) and mean within-condition variance (σ^2, lower is better). Best results per model are **bolded**; second-best are underlined. Strategies are grouped by type: single-turn baselines, single-turn CoT variants, and multi-step pipelines. Some strategies in GPT-5.1 are omitted due to cost constraints.

		Qwen2.5-32B		**Qwen2.5-72B**		**DeepSeek-R1-32B**		**Llama-3.1-70B**		**GPT-5.1**	
Type	**Strategy**	ρ	σ^2	ρ	σ^2	ρ	σ^2	ρ	σ^2	ρ	σ^2
Baseline	No-CoT	.418	**.169**	.379	.167	.393	.280	.327	**.392**	.686	**.047**
	Brief-CoT	.459	.414	.324	.199	.451	.313	.400	.522	–	–
	Full-CoT	.425	.332	.289	.192	.411	.282	.415	.491	.693	.225
	Structured-CoT	.428	.420	.281	.185	.439	.319	.430	.567	–	–
Single-Turn	Bullet-Point	.471	.339	.401	.285	.451	.289	.487	.776	–	–
	Comparative	**.494**	.327	**.493**	.216	**.454**	.313	**.502**	.866	.683	.251
	Quote-Forcing	.440	.346	.317	.194	.424	.291	.421	.570	**.769**	.119
	Self-Critique	.453	.382	.326	.176	.448	.272	.414	.618	–	–
	G-Eval	.323	.381	.315	**.059**	.405	.370	.310	.428	.668	.231
	CoVe	.468	.397	.323	.205	.408	.313	.437	.558	.668	.319
Multi-Step	Multi-Judge	.440	.248	.336	.124	.434	**.230**	.423	.401	.694	.250
	Debate	.306	.388	.405	.534	.449	.700	.333	1.335	–	–
	Overseer	.425	.332	.288	.189	.397	.275	.402	.509	.646	.220

C TOKEN CALCULATION FORMULAS

Token Cost Summary

Per call:

$$T_{\text{call}} = T_{\text{in}}(\text{call}) + T_{\text{out}}(\text{call}) \tag{1}$$

Per strategy (summed over problems i):

$$T_{\text{strategy}} = \sum_{i} T_{\text{strategy}}(i) \tag{2}$$

Per-problem token costs by strategy:

Single-turn (`no_cot`, `brief_cot`, `structured_cot`, `full_cot`, `self_critique`, `bullet_point`, `comparative`, `quote_forcing`):

$$T = T_{\text{in}}(\text{prompt}) + T_{\text{out}}(\text{output}) \tag{3}$$

G-Eval (two-stage checklist then scoring):

$$T = \big[T_{\text{in}}(\text{step}_1) + T_{\text{out}}(\text{checklist})\big] + \big[T_{\text{in}}(\text{step}_2(\text{checklist})) + T_{\text{out}}(\text{output})\big] \tag{4}$$

Debate (pro, con, then arbiter):

$$\begin{aligned} T = {} & \big[T_{\text{in}}(\text{pro}) + T_{\text{out}}(\text{pro\_arg})\big] \\ & + \big[T_{\text{in}}(\text{con}) + T_{\text{out}}(\text{con\_arg})\big] \\ & + \big[T_{\text{in}}(\text{arbiter}(\text{pro\_arg},\ \text{con\_arg})) + T_{\text{out}}(\text{output})\big] \end{aligned} \tag{5}$$

CoVe (draft, verify, answer, finalise):

$$\begin{aligned} T = {} & \big[T_{\text{in}}(\text{draft}) + T_{\text{out}}(\text{draft})\big] \\ & + \big[T_{\text{in}}(\text{verify}(\text{draft})) + T_{\text{out}}(\text{questions})\big] \\ & + \big[T_{\text{in}}(\text{answer}(\text{questions})) + T_{\text{out}}(\text{answers})\big] \\ & + \big[T_{\text{in}}(\text{final}(\text{draft},\ \text{Q\&A})) + T_{\text{out}}(\text{output})\big] \end{aligned} \tag{6}$$

Multi-Judge (three persona judges then chair):

$$T = \sum_{j=1}^{3} \big[T_{\text{in}}(\text{persona}_j) + T_{\text{out}}(\text{judgment}_j)\big] + T_{\text{in}}(\text{chair}(j_1, j_2, j_3)) + T_{\text{out}}(\text{output}) \tag{7}$$

Overseer-Retry (summed over all iterations k):

$$T = \sum_{k} \Big(\big[T_{\text{in}}(\text{judge}_k) + T_{\text{out}}(\text{judge\_out}_k)\big] + \big[T_{\text{in}}(\text{overseer}_k) + T_{\text{out}}(\text{overseer\_out}_k)\big] \Big) \tag{8}$$

where k iterates over all entries in `overseer_checks`.

D TOKEN USAGE ANALYSIS

Table 6: Total token usage of judge strategies across models on **Putnam-AXIOM-Grading**. We report total tokens consumed per strategy per model. Highest tokens per model are **bolded**; second-highest are underlined. Strategies are grouped by type: baseline, single-turn variants, and multi-step pipelines.

Type	**Strategy**	**DeepSeek-R1-32B** Tokens	**Qwen2.5-32B** Tokens	**Qwen2.5-72B** Tokens	**Llama-3.1-70B** Tokens
Baseline	No-CoT	866,526	850,780	851,797	844,746
Single-Turn	Brief-CoT	856,775	855,903	855,932	849,895
	Full-CoT	1,037,450	1,006,790	1,009,740	1,001,598
	Structured-CoT	988,835	933,031	932,164	928,397
	Bullet-Point	886,734	890,603	890,824	885,537
	Comparative	958,600	895,427	894,719	888,687
	Quote-Forcing	983,522	891,478	896,288	887,367
	Self-Critique	1,103,194	916,705	918,074	914,197
Multi-Step	G-Eval	2,363,471	1,960,149	1,905,160	2,085,316
	CoVe	5,125,495	3,774,706	3,799,769	3,860,194
	Multi-Judge	**6,070,101**	**4,804,855**	**4,937,384**	4,825,306
	Debate	4,744,275	3,532,677	4,190,834	3,695,543
	Overseer	4,545,332	2,884,088	3,499,382	**5,407,469**

Table 7: Total token usage of judge strategies across models on **IMO-GradingBench**. We report total tokens consumed per strategy per model. Highest tokens per model are **bolded**; second-highest are underlined. Strategies are grouped by type: baseline, single-turn variants, and multi-step pipelines.

Type	**Strategy**	**DeepSeek-R1-32B** Tokens	**Qwen2.5-32B** Tokens	**Qwen2.5-72B** Tokens	**Llama-3.1-70B** Tokens
Baseline	No-CoT	5,132,342	4,440,810	4,440,810	4,424,912
Single-Turn	Brief-CoT	4,937,699	4,497,078	4,491,580	4,484,273
	Full-CoT	5,557,631	5,013,052	5,132,245	5,012,472
	Structured-CoT	5,418,253	4,886,272	4,833,874	4,805,231
	Bullet-Point	5,350,817	4,760,237	4,710,583	4,693,688
	Comparative	5,298,131	4,796,308	4,735,847	4,707,437
	Quote-Forcing	5,346,101	4,878,911	4,854,241	4,828,665
	Self-Critique	5,394,857	4,791,587	4,791,586	4,797,907
Multi-Step	G-Eval	11,221,644	9,251,876	9,158,026	9,152,668
	CoVe	20,402,280	14,964,140	15,242,634	14,813,446
	Multi-Judge	**24,068,464**	**20,357,440**	**21,004,026**	**20,412,120**
	Debate	17,228,546	14,088,592	14,980,337	14,175,135
	Overseer	16,709,961	9,853,854	10,179,871	10,510,512

E COMPARATIVE PROMPTING ABLATIONS

Given comparative prompting's consistently dominant performance (Table 4), we conduct targeted ablations to identify its active ingredients. We test five variants against the original (A0) on all four open-source models across both benchmarks (Table 8).

Template ablations We test three modifications to the comparative template: (A1) **Comparative no-compare** uses the same four-step template but replaces "compare to ground truth" with "evaluate the solution's correctness," while keeping the ground truth in context. This tests whether explicitly directing attention to the reference is the active ingredient. (A2) **Difference-only** strips the template to a minimal prompt: "List the differences between the model output and the ground truth, then score." This tests whether the bare act of difference-finding suffices without the template structure. (A4) **Comparative shuffled** uses the same four components as A0 but in a randomized order (error analysis $\rightarrow$ scoring rationale $\rightarrow$ approach comparison $\rightarrow$ differences). This tests whether the specific progressive-narrowing ordering matters.

Attention-direction ablations We test two variants that redirect the judge's focus: (A5) **Quote-forcing reference** requires the judge to cite direct quotes from the *ground truth* rather than the student solution, testing whether anchoring attention specifically to the reference drives performance. (A3) **Reference checklist** auto-generates 3–5 binary Yes/No verification questions from the ground truth, then has the judge answer them before scoring—similar to G-Eval but with checklist items derived from the reference rather than freely generated.

Results The ablation results (Table 8) fail to support any single-ingredient explanation for comparative prompting's effectiveness:

- **Attention anchoring is not the driver.** If explicitly directing attention to the reference were the key mechanism, A5 (quote-forcing reference) should outperform standard quote-forcing and approach A0. Instead, A5 consistently underperforms A0 across both benchmarks. Conversely, A1 performs competitively despite *not* explicitly referencing the ground truth in its instructions, suggesting that the reference need only be *present in context*, not explicitly called out.
- **Template structure is helpful but not sufficient.** A2 (difference-only), which removes all template structure, performs comparably to A0 on Putnam-AXIOM-Grading (matching or exceeding it for Qwen-72B) but degrades on IMO-GradingBench, indicating that structure contributes but is not the sole factor.
- **Task framing ("compare" vs. "evaluate") does not explain the gap.** A1 and A2 differ precisely on this dimension—A1 frames the task as evaluation while A2 frames it as difference-finding—yet perform similarly, ruling out cognitive framing as the primary mechanism.
- **Component ordering has limited effect.** A4 (shuffled order) performs as well as or better than A0 on Putnam-AXIOM-Grading, suggesting the progressive-narrowing structure of the original template is not critical. On IMO-GradingBench, A4 shows more variable results but remains competitive.
- **Structured reference decomposition hurts.** A3 (reference checklist) consistently underperforms across both benchmarks, mirroring G-Eval's poor showing in the main experiments—suggesting that decomposing the reference into binary checks destroys holistic reasoning.

Interpretation The pattern that emerges is that *all high-performing variants share a common property*: they allow the judge to holistically reason about both solutions with minimal output constraints. The strategies that degrade performance—A3 (reference checklist) and A5 (quote-forcing reference)—are precisely those that impose rigid intermediate output requirements, forcing the judge to fragment its reasoning into constrained artifacts (binary answers, quoted evidence) before scoring. We term this active ingredient *evaluation-focused holistic reasoning*: the prompt keeps the judge's attention on the substantive task of assessment rather than diverting it to mechanical subtasks. This also explains why G-Eval and Debate perform poorly in the main experiments (Table 4)—both impose intermediate reasoning structures that fragment the evaluation process.

Table 8: Ablation results for comparative prompting on Putnam-AXIOM-Grading and IMO-GradingBench. We report Pearson correlation with human grades (ρ, higher is better) and mean within-condition variance (σ^2, lower is better) across five independent trials. Best results per model–metric are **bolded**; second-best are underlined. Strategies are ordered by conceptual grouping: original comparative, template ablations, and attention-direction ablations.

	Pearson (ρ) ↑				Variance (σ^2) ↓			
Strategy	Qwen-32B	Qwen-72B	DS-R1-32B	Llama-70B	Qwen-32B	Qwen-72B	DS-R1-32B	Llama-70B
(a) Putnam-AXIOM-Grading								
BASELINE								
Comparative (A0)	.715	.733	.725	.677	.239	.179	.316	.293
TEMPLATE ABLATIONS								
Comp. no compare (A1)	.716	.714	.728	.681	**.219**	.184	.293	**.263**
Difference only (A2)	.718	.736	.704	.688	.254	**.158**	.322	.261
Comp. shuffled (A4)	**.732**	**.761**	**.740**	**.689**	.251	.165	**.279**	.297
ATTENTION-DIRECTION ABLATIONS								
Quote-force ref. (A5)	.690	.718	.684	.654	.361	.185	.334	.368
Ref. checklist (A3)	.678	.614	.660	.661	.419	.390	.427	.408
(b) IMO-GradingBench								
BASELINE								
Comparative (A0)	.493	**.505**	**.490**	.392	.332	.368	.317	.555
TEMPLATE ABLATIONS								
Comp. no compare (A1)	**.498**	.427	.465	.451	.367	**.252**	.339	**.429**
Difference only (A2)	.478	.460	.436	.463	**.265**	.265	**.257**	.730
Comp. shuffled (A4)	.486	.475	.460	**.465**	.352	.445	.338	.750
ATTENTION-DIRECTION ABLATIONS								
Quote-force ref. (A5)	.452	.450	.442	.429	.455	.387	.326	1.04
Ref. checklist (A3)	.341	.241	.411	.399	.398	.278	.435	.686

F GRADING RUBRICS USED

Reasoning-quality rubric (MODEL_OUTPUT vs. GROUND_TRUTH_SOLUTION).

Task: Evaluate the `MODEL_OUTPUT`'s *reasoning quality* against the `GROUND_TRUTH_SOLUTION`. Focus *only* on reasoning quality (not writing style).

Criteria (each scored 0–1; 0.5 allowed):

1. **Logical Coherence (0–1)**: Are steps locally valid and globally consistent?
2. **Faithfulness to Task (0–1)**: Does the reasoning address the question (no hallucinated subgoals or irrelevant paths)?
3. **Methodological Alignment (0–1)**: Is the approach compatible with (or reasonably equivalent to) the canonical method in the ground truth? If different, is it still valid?
4. **Intermediate Correctness (0–1)**: Are key intermediate results (derivations, lemmas, sub-calculations) correct?
5. **Error Awareness / Self-Correction (0–1)**: Does the reasoning notice and repair mistakes or check edge cases? (If model does not make any nontrivial mistakes, do not penalize here).

Scoring: Final score is the sum across the five criteria (0–5; half-points allowed).

Figure 2: Rubric used by judge LLM to evaluate student solution for Putnam problems.

IMO-style grading rubric (0–7).

Task: Grade the student's solution on the standard IMO 0–7 scale. Output a single integer score in $\{0, 1, \ldots, 7\}$.

Scale definitions:

- **7** Completely correct solution.
- **6** Correct solution with very minor slips (e.g., a sign error that doesn't affect the logic).
- **5** Main theorem/logic is established, but a significant case or detail is missed.
- **4** Significant progress toward the solution (e.g., reducing the problem to a known solved state), but main proof incomplete.
- **3** Non-trivial partial results that are steps toward the solution.
- **2** Some relevant formulas or observations, but no significant progress.
- **1** Fundamental misunderstanding or only trivial observations.
- **0** Completely incorrect or blank.

Output format: Return only the integer score (no extra text).

Figure 3: Rubric used by judge LLM to evaluate student solution for IMO problems.

G DEEPSEEK R1 ANALYSIS

Section 4 established that DeepSeek-R1-Distill-Qwen-32B is a consistent exception to the pattern of CoT degrading judge reliability: on Putnam-AXIOM-Grading, structured-CoT improves agreement by 9.4 points over no-CoT, and the model benefits from CoT across both benchmarks regardless of structure. We attributed this to the model's reinforcement-learning training on reasoning traces, which leaves it "undertrained for the terse responses required by direct scoring." Here we present a fine-grained analysis of *how* the Full-CoT and No-CoT conditions differ for this model, based on a comparison of all 500 Putnam-AXIOM-Grading outputs from a single representative trial.

DeepSeek-R1 cannot suppress reasoning. The most fundamental observation is that DeepSeek-R1-Distill-Qwen-32B generates extensive chain-of-thought reasoning in *both* conditions. Even under the No-CoT prompt—which instructs the model to output only a JSON score with no reasoning—497 of 500 outputs (99.4%) contain a `</think>` token demarcating an internal reasoning scratchpad. These scratchpads average 2,016 characters, only 10% shorter than the Full-CoT scratchpads (2,243 characters). After the `</think>` token, the No-CoT outputs contain only the JSON score block ($\sim$230 characters). This confirms the behavioral observation in Section 4.1: for reasoning-optimized models, "direct scoring" is not just less effective but behaviorally unnatural—the model *cannot help but deliberate*.

The critical difference lies not in whether the model reasons, but in what happens *after* the scratchpad. Under Full-CoT, 431 of 500 outputs (86.2%) exhibit a **two-phase structure**: a free-form scratchpad (enclosed within the model's internal think tokens), followed by a structured `<THINKING>` summary that distills the scratchpad's observations into per-criterion assessments, followed by the JSON score. Under No-CoT, the model proceeds directly from its scratchpad to the JSON score, with no intermediate consolidation step.

The distillation step corrects leniency on surface-level criteria. This architectural difference has measurable consequences. Comparing per-component score means across the two conditions reveals an asymmetric pattern: No-CoT assigns higher scores on *methodological alignment* ($+0.10$ shift, from 0.55 to 0.65) and *logical coherence* ($+0.05$ shift), while assigning *lower* scores on *error awareness* (-0.12 shift). The first two criteria are precisely those assessable from a surface reading of the student solution: "Does the method seem reasonable?" and "Do the steps flow logically?" The No-CoT scratchpad tends to answer these affirmatively based on the student's internal consistency, without weighing the ground truth.

The error awareness shift is a striking counterpoint: No-CoT is actually *harsher* on this criterion (mean 0.16 vs. 0.28 under Full-CoT), but this harshness appears nearly random—the Pearson correlation between No-CoT's error awareness scores and human total scores is only $r = 0.39$, compared to $r = 0.64$ under Full-CoT, the largest per-component gap we observe ($\Delta r = +0.25$). Substantial gaps also appear for logical coherence ($\Delta r = +0.08$) and methodological alignment ($\Delta r = +0.07$).

No-CoT compresses scores toward the midrange. The net effect of these per-component shifts is a compression of the total score distribution. Under No-CoT, 65.3% of scores fall in the 1.5–4.0 range, compared to 48.7% under Full-CoT and 48.4% for human grades. Correspondingly, No-CoT produces fewer extreme scores: only 15.9% of scores are ≤ 1 (vs. 21.2% for Full-CoT, 36.9% for humans) and only 17.8% are ≥ 4.5 (vs. 29.2% for Full-CoT). The reduced spread (standard deviation 1.37 vs. 1.56 for Full-CoT, 1.43 for humans) directly reduces Pearson correlation: on the 415 instances where both conditions parse successfully, Full-CoT achieves $r = 0.69$ versus $r = 0.64$ for No-CoT.

Qualitative mechanism: noticing errors without penalizing them. The quantitative patterns are explained by a consistent qualitative mechanism visible in individual outputs (Figures 4–6). In the No-CoT condition, the model's scratchpad frequently *identifies* discrepancies with the ground truth but then *rationalizes them away* in the same breath, defaulting to a lenient assessment of the student solution's internal logic. Three illustrative failure modes emerge:

1. **Internal-consistency bias** (Figure 4). The scratchpad notes that the student derives $I_k = I_{5-k}$ instead of the correct I_{4-k}, but then writes: "However, the steps are locally valid and globally consistent *within its own reasoning*, so I'll give a 1." The judge evaluates the solution as a standalone argument rather than against the reference.
2. **Premise-blindness** (Figure 5). The student's internal mathematics is correct (a quadratic is properly solved), but the entire argument rests on a false premise contradicted by the ground

truth. The No-CoT scratchpad never confronts this, writing: "The calculations are accurate. So, Intermediate Correctness is 1." The Full-CoT scratchpad immediately cross-references: "doesn't align with the ground truth, which shows all u's are zero."

3. **Failure to propagate errors across criteria** (Figure 6). The No-CoT scratchpad notices that the student's answer is wrong but confines the penalty to a single criterion (Intermediate Correctness), leaving Logical Coherence, Faithfulness, and Methodological Alignment at full marks. The Full-CoT <THINKING> summary propagates the error backward, recognizing that an incorrect conclusion undermines the coherence and faithfulness of the reasoning that led to it.

In each case, the Full-CoT condition avoids the failure because its two-phase structure—scratchpad followed by <THINKING> summary—forces the model to *consolidate* its observations into rubric-calibrated judgments. The scratchpad serves as exploratory analysis; the summary serves as deliberate assessment. Without this distillation step, the model's initial impressions pass through to the final score unchecked.

Implications. These findings refine the claim in Section 4.1 that reasoning-trained models benefit from CoT because they are "undertrained for terse responses." The mechanism is more specific: DeepSeek-R1 deliberates regardless of the prompt, but the *quality of the transition from deliberation to scoring* depends on whether the prompt elicits an explicit output-side reasoning structure. The Full-CoT prompt does not make the model think more—the scratchpads are similar in length and content—but it provides a scaffold for converting messy, exploratory reasoning into calibrated judgments. For practitioners using reasoning-specialized models as judges, this suggests that the key design choice is not whether to allow CoT (the model will reason anyway) but whether to require a structured output format that forces distillation of that reasoning before scoring.

Problem (Putnam #227)

Let $p(x) = 2 + 4x + 3x^2 + 5x^3 + 3x^4 + 4x^5 + 2x^6$. For k with $0 < k < 5$, define $I_k = \int_0^\infty \frac{x^k}{p(x)}\,dx$. For which k is I_k smallest?

Ground Truth Solution (abridged)

Substitute $x = 1/t$ to show $I_k = I_{4-k}$. Then $I_k = (I_k + I_{4-k})/2 = \int_0^\infty \frac{(x^k + x^{4-k})/2}{p(x)}\,dx > \int_0^\infty \frac{x^2}{p(x)}\,dx = I_2$ by AM–GM, since $(x^k + x^{4-k})/2 \geq x^2$. Answer: $k = \boxed{2}$.

Student Solution (abridged)

Notes $p(x) = x^6 p(1/x)$ (correct). Claims $I_k = I_{5-k}$ (**should be** I_{4-k}). Approximates $p(x) \approx 23$ near $x = 1$, argues minimum at $k = 5/2$. **Assumes k must be integer** (not stated in problem). Checks $k = 2$ vs. $k = 3$, concludes $k = 2$. Gets the right answer through flawed reasoning.

No-CoT Judge (score: 4.5 / 5)

"The model's reasoning seems to flow well [...] I notice that the model assumes symmetry around 5, not 4 as in the ground truth. **However, the steps are locally valid and globally consistent within its own reasoning, so I'll give a 1.**"

"The model correctly identifies $p(x)$ as symmetric, but the substitution in the ground truth relates I_k to I_{4-k}, not I_{5-k}. The model's symmetry is slightly off. **However, the main conclusion k=2 is correct, so intermediate steps are mostly correct. Score: 1.**"

"Both methods are valid but different. The model's approach is reasonable, so **alignment is present. Score: 1.**"

LC: 1 — FT: 1 — MA: 1 — IC: 1 — EA: 0.5 — **Total: 4.5**

Full-CoT Judge (score: 1.0 / 5)

Scratchpad:
"**Wait, that's a discrepancy.** The ground truth uses $4-k$, but the model says $5-k$. Hmm, that might be an error."

"**Wait, but why does the model assume k must be an integer?** The problem states $0 < k < 5$, but k doesn't have to be an integer [...] That could be a mistake."

⟨THINKING⟩ summary:
"Methodological Alignment (**0**): The approach differs from the ground truth [...] leading to invalid steps."
"Intermediate Correctness (**0**): Key steps, like the substitution and integer assumption, are incorrect, undermining the reasoning."

LC: 0.5 — FT: 0.5 — MA: 0 — IC: 0 — EA: 0 — **Total: 1.0**

Figure 4: **"Locally valid within its own reasoning" (Human: 1.0).** Both judges notice that the student derives $I_k = I_{5-k}$ instead of the correct I_{4-k}. The No-CoT judge excuses this: because the steps are internally consistent, it awards full marks on four criteria. The Full-CoT judge's scratchpad flags the same discrepancy ("*Wait, that's a discrepancy*") and the unjustified integer assumption, then the structured `<THINKING>` summary consolidates these into zero-score penalties on methodology and correctness—matching the human grade exactly. The key failure mode: **without a distillation step, No-CoT evaluates the student solution as a standalone argument rather than against the reference standard.**

Problem (Putnam #92)

Suppose $u_0, u_1, u_2, \ldots$ is a sequence of real numbers with $u_n = \sum_{k=1}^{\infty} u_{n+k}^2$ for all $n \geq 0$. If $\sum u_n$ converges, find the maximum value of u_k for all k.

Ground Truth (abridged)

Shows $\{u_n\}$ is non-increasing and non-negative. Chooses p with $\sum_{n>p} u_n < 1$, then derives $u_p \leq u_p \cdot \sum_{n>p} u_n < u_p$, a contradiction unless $u_p = 0$. Hence all u's are zero. Answer: $\boxed{0}$.

Student Solution (abridged)

Sets $n = 0$: $u_0 = \sum_{k=1}^{\infty} u_k^2$. Sets $n = 1$: $u_1 = \sum_{k=2}^{\infty} u_k^2$. Subtracts to get $u_0 = u_1^2 + u_1$, a quadratic. Solves: $u_1 = (-1 + \sqrt{1 + 4u_0})/2$. **Claims $u_0 = 1/4$ is possible**, giving $u_1 = 1/2$. Answer: $\boxed{1/2}$ (incorrect—**the correct answer is 0**).

No-CoT Judge (score: 4.5 / 5)

"The model correctly derives the quadratic equation and solves it, finding $u_1 = 1/2$ when $u_0 = 1/4$. **The calculations are accurate. So, Intermediate Correctness is 1.**"

"The model correctly identifies that the maximum occurs at u_1 when u_0 is $1/4$, leading to $u_1 = 1/2$. **This directly addresses the task without irrelevant paths. So, Faithfulness is 1.**"

"The ground truth uses induction and an inequality approach, while the model uses a quadratic equation [...] **Both methods aim to find the maximum value, and the model's approach is valid. So, Methodological Alignment is 1.**"

LC: 1 — FT: 1 — MA: 1 — IC: 1 — EA: 0.5 — **Total: 4.5**

Full-CoT Judge (score: 1.0 / 5)

Scratchpad:
"However, the reasoning **doesn't align with the ground truth, which shows all u's are zero**. The model's approach is inconsistent because it arrives at a non-zero maximum, **contradicting the correct solution**. So, Logical Coherence is 0."

"The model's intermediate steps, like solving the quadratic, are mathematically correct. However, **the premise that u_0 can be $1/4$ is flawed because the ground truth shows all terms must be zero**."

⟨THINKING⟩ summary:
"Logical Coherence (**0**): The argument is inconsistent. It derives a maximum of $1/2$ but **fails to consider that all terms must be zero**."

LC: 0 — FT: 0.5 — MA: 0 — IC: 0.5 — EA: 0 — **Total: 1.0**

Figure 5: **The internal-consistency trap (Human: 1.5).** The student solution is internally valid mathematics—the quadratic really does yield $u_1 = 1/2$—but the entire argument rests on the false premise that $u_0 = 1/4$ is achievable. No-CoT evaluates the solution as a standalone argument ("the calculations are accurate"; "the model's approach is valid") without confronting the ground truth's proof that all terms are zero. Full-CoT's scratchpad immediately cross-references: "*doesn't align with the ground truth, which shows all u's are zero*," then the `<THINKING>` phase assigns consequences. **This example illustrates the core failure: No-CoT checks whether the student's math is self-consistent, not whether its premises are correct.**

Problem (Putnam #62)

Let $f(x)$ be a cubic polynomial with roots x_1, x_2, x_3. Assume that $f(2x)$ is divisible by $f'(x)$. Compute the ratios $x_1 : x_2 : x_3$.

Ground Truth (abridged)

Writes $f(2x) = f'(x)(px + q)$, compares coefficients to get $b = 4a^2/3$, $c = -16a^3/27$. Factors $f(x)$ to find roots $1, -2 + 2\sqrt{3}\,i, -2 - 2\sqrt{3}\,i$ (up to scaling by a). Ratios involve **complex numbers**.

Student Solution (abridged)

Notes that roots of $f'(x)$, when doubled, must be roots of $f(x)$. Uses Vieta's formulas. **Assumes all roots are real**, concludes $x_1 : x_2 : x_3 = 2 : 1 : 1$ without justification. Does not consider complex roots.

No-CoT Judge (score: 3.5 / 5)

"The MODEL_OUTPUT's steps are locally valid and globally consistent. It outlines a logical progression [...] **Logical Coherence: 1.**"

"The approach is compatible with the canonical method. Both the model and ground truth involve setting up integrals [sic] and solving for c [sic]. **Methodological Alignment: 1.**"

"The model concludes the ratio is 2:1:1. However, the ground truth shows the roots are $1, -2+2\sqrt{3}\,i, -2-2\sqrt{3}\,i$, which don't align [...] The model's intermediate step about the ratio might be incorrect, so I'll give **0.5**."

LC: 1 — FT: 1 — MA: 1 — IC: 0.5 — EA: 0 — **Total: 3.5**

Full-CoT Judge (score: 1.5 / 5)

Scratchpad:
"However, the model uses Vieta's formulas [...] but **it doesn't apply them as thoroughly as the ground truth**. The ground truth sets up equations by comparing coefficients, which is a more precise approach [...] **0.5**."

⟨THINKING⟩ summary:
"Logical Coherence (**0.5**): [...] **the leap to the ratio 2:1:1 is abrupt and lacks detailed justification.**"
"Faithfulness (**0.5**): [...] **oversimplifies by assuming real roots** and a specific ratio without considering complex roots as in the ground truth."
"Methodological Alignment (**0.5**): Uses Vieta's formulas [...] but **applies it less rigorously** than the ground truth."

LC: 0.5 — FT: 0.5 — MA: 0.5 — IC: 0 — EA: 0 — **Total: 1.5**

Figure 6: **Failure to propagate error backward through rubric criteria (Human: 1.5).** The No-CoT judge evaluates each criterion independently: "Does the reasoning flow logically?" (yes, on the surface)—and awards 1. It even *notices* the wrong answer ("the ground truth shows the roots are $1, -2+2\sqrt{3}\,i, \ldots$ which don't align") but confines the penalty to Intermediate Correctness (0.5) without revising the upstream criteria. Full-CoT recognizes that the wrong answer retroactively undermines Logical Coherence ("the leap to 2:1:1 is abrupt"), Faithfulness ("oversimplifies by assuming real roots"), and Methodology ("applies it less rigorously")—propagating the error backward to produce a correctly calibrated total. Note also that No-CoT *hallucinates* problem content: it refers to "setting up integrals and solving for c," neither of which appear in this problem.

H PROMPTS

H.1 PUTNAM-AXIOM-GRADING PROMPTS

Base Context (Common to All Strategies)

Note: All strategies typically begin with this context block.
<CONTEXT> [Rubric Text Omitted] </CONTEXT>
<PROBLEM> {problem_statement} </PROBLEM>
<GROUND_TRUTH_SOLUTION> {ground_truth_solution}
</GROUND_TRUTH_SOLUTION>
<MODEL_OUTPUT> {model_output} </MODEL_OUTPUT>

SINGLE-TURN STRATEGIES

Strategy: No CoT

<INSTRUCTIONS> Provide your scores DIRECTLY in the specified JSON format inside <json> tags. Do NOT include any reasoning.
Your response MUST include a <json>...</json> block and NOTHING after it.
Here is an example response: <json> { "score_logical_coherence": 1, "score_faithfulness_to_task": 1, "score_methodological_alignment": 1, "score_intermediate_correctness": 1, "score_error_awareness": 1, "score_total": 5 } </json> </INSTRUCTIONS>

Strategy: Brief CoT

<INSTRUCTIONS> First, write ONLY 1-2 SHORT sentences inside <THINKING> tags summarizing your overall assessment. Do NOT explain your entire process. Just state your assessment directly. Then provide your scores in the specified JSON format inside <json> tags.
Example format: <THINKING> The solution correctly applies integration by parts but makes an arithmetic error in step 3. Otherwise reasoning is sound. </THINKING> <json> { "score_logical_coherence": 1, "score_faithfulness_to_task": 1, "score_methodological_alignment": 1, "score_intermediate_correctness": 0, "score_error_awareness": 0, "score_total": 3 } </json> </INSTRUCTIONS>

Strategy: Structured CoT

<INSTRUCTIONS> You MUST follow this exact three-step structure inside <THINKING> tags:
STEP 1 - SUMMARIZE: In 2-3 sentences, summarize what the model's solution attempts to do.
STEP 2 - CHECK KEY STEPS: List the 2-4 most critical steps/calculations. For each, note if it is correct (✓) or incorrect (×).
STEP 3 - DECIDE SCORES: Based on your analysis, determine each rubric score with one-line justification.
Then provide your final scores in <json> tags.
Example: <THINKING> STEP 1 - SUMMARIZE: The model attempts to solve the integral using substitution, converting the expression and evaluating bounds.
STEP 2 - CHECK KEY STEPS: - Substitution $u = x^2 + 1$: ✓Valid transformation - Computing $du = 2xdx$: ✓Correct derivative - Final evaluation at bounds: ×Arithmetic error (should be 15, got 12)
STEP 3 - DECIDE SCORES: - Logical coherence: 1 (steps follow logically) - Faithfulness: 1 (addresses the problem) - Methodological alignment: 1 (standard approach) - Intermediate correctness: 0 (arithmetic error) - Error awareness: 0 (error not noticed) </THINKING> <json> { "score_logical_coherence": 1, "score_faithfulness_to_task":

1, "score_methodological_alignment": 1, "score_intermediate_correctness": 0, "score_error_awareness": 0, "score_total": 3 } </json> </INSTRUCTIONS>

Strategy: Full CoT

<INSTRUCTIONS> You MUST output your response in this EXACT format: 1. FIRST, write your step-by-step reasoning inside <THINKING>...</THINKING> tags 2. THEN, output the scores inside <json>...</json> tags
For instance, a valid grading response to a correct answer could look like this: "<THINKING> Logical Coherence (1): The argument proceeds in a clear sequence, with each step following from the last; no leaps or contradictions detected. Faithfulness to Task (1): The response stays focused on the stated goal and uses only information relevant to answering the problem. The response never strays away from the given task. Methodological Alignment (1): The chosen method is standard for this type of problem and is applied appropriately. Intermediate Correctness (1): Intermediate claims match known identities/definitions and are used consistently. Error Awareness / Self-Correction (1): No mistakes observed; scope conditions are implicitly respected and edge cases would not change the conclusion. </THINKING> <json> { "score_logical_coherence": 1, "score_faithfulness_to_task": 1, "score_methodological_alignment": 1, "score_intermediate_correctness": 1, "score_error_awareness": 1, "score_total": 5 } </json>"
A valid grading response to an incorrect answer could look like this: "<THINKING> Logical Coherence (0): The reasoning is inconsistent: it mixes unrelated steps, reverses implications, and reaches conclusions that don't follow from prior claims. Faithfulness to Task (1): The response attempts to answer the stated question and references the required quantities, but stays narrowly focused while still arriving at the wrong result. Methodological Alignment (0): The chosen approach conflicts with standard methods for this problem and misapplies key definitions/identities. Intermediate Correctness (0): Several intermediate statements are false (e.g., misuse of a determinant identity and an invalid simplification), so later conclusions are unsupported. Error Awareness / Self-Correction (0): No checks, sanity tests, or alternative cases are considered; contradictions and edge conditions go unaddressed. <json> { "score_logical_coherence": 0, "score_faithfulness_to_task": 1, "score_methodological_alignment": 0, "score_intermediate_correctness": 0, "score_error_awareness": 1, "score_total": 1 } </json>" </INSTRUCTIONS>

Strategy: Self-Critique

<INSTRUCTIONS> Follow this two-phase evaluation inside <THINKING> tags:
PHASE 1 - INITIAL ASSESSMENT: Quickly evaluate the solution and propose initial scores.
PHASE 2 - SELF-CRITIQUE: Challenge your initial assessment. Ask: "Am I being too harsh or too lenient?" "Did I miss anything?" Adjust scores if needed.
Then provide your final scores in <json> tags.
Example: <THINKING> PHASE 1 - INITIAL ASSESSMENT: The solution uses the correct method but has a sign error. Initial scores: coherence=1, faithfulness=1, method=1, intermediate=0, awareness=0.
PHASE 2 - SELF-CRITIQUE: Wait - I should check if the sign error propagates. Looking again... yes, it affects the final answer but the method is still valid. The model did attempt to verify the answer at the end, showing some error awareness even if unsuccessful. Adjusting error_awareness to 0.5. </THINKING> <json> { "score_logical_coherence": 1, "score_faithfulness_to_task": 1, "score_methodological_alignment": 1, "score_intermediate_correctness": 0, "score_error_awareness": 0.5, "score_total": 3.5 } </json> </INSTRUCTIONS>

Strategy: Bullet Point

<INSTRUCTIONS> Use bullet points inside <THINKING> tags to systematically evaluate each criterion:
• Logical Coherence: [observation] → [score] • Faithfulness to Task: [observation] → [score] • Methodological Alignment: [observation] → [score] • Intermediate Correctness: [observation] → [score] • Error Awareness: [observation] → [score] • Total: [sum]
Then provide scores in <json> tags.
Example: <THINKING> • Logical Coherence: Steps follow clear sequence with proper transitions → 1 • Faithfulness to Task: Directly addresses the integral as asked → 1 • Methodological Alignment: Uses substitution, same as ground truth → 1 • Intermediate Correctness: Calculation error in step 4 (8x3=24, not 26) → 0 • Error Awareness: No verification attempted → 0 • Total: 3 </THINKING> <json> { "score_logical_coherence": 1, "score_faithfulness_to_task": 1, "score_methodological_alignment": 1, "score_intermediate_correctness": 0, "score_error_awareness": 0, "score_total": 3 } </json> </INSTRUCTIONS>

Strategy: Comparative

<INSTRUCTIONS> Inside <THINKING> tags, perform a comparative analysis:
1. KEY APPROACH COMPARISON: - Ground Truth approach: [brief description] - Model approach: [brief description] - Match? [Yes/No/Partial]
2. CRITICAL DIFFERENCES: List any significant differences in reasoning or results.
3. ERROR ANALYSIS: Note any errors in the model output that differ from ground truth.
4. SCORING RATIONALE: Brief justification for each score based on the comparison.
Then provide scores in <json> tags.
Example: <THINKING> 1. KEY APPROACH COMPARISON: - Ground Truth: Uses integration by parts twice - Model: Uses integration by parts twice - Match? Yes
2. CRITICAL DIFFERENCES: - Model skips showing the second application explicitly - Final constant differs (model has +C, GT shows specific value)
3. ERROR ANALYSIS: - Sign error in step 2: -cos(x) written as cos(x) - This propagates to incorrect final answer
4. SCORING RATIONALE: - Coherence: 1 (logical flow maintained) - Faithfulness: 1 (addresses the integral) - Method: 1 (same technique) - Intermediate: 0 (sign error) - Awareness: 0 (not caught) </THINKING> <json> { "score_logical_coherence": 1, "score_faithfulness_to_task": 1, "score_methodological_alignment": 1, "score_intermediate_correctness": 0, "score_error_awareness": 0, "score_total": 3 } </json> </INSTRUCTIONS>

Strategy: Quote Forcing

<INSTRUCTIONS> You MUST output your response in this EXACT format: 1. FIRST, write your reasoning inside <THINKING>...</THINKING> tags, citing evidence from <MODEL_OUTPUT> for EVERY evaluative claim with short direct quotes like: "..." 2. THEN, output the scores inside <json>...</json> tags
Your response MUST start with <THINKING> and include quotes before the JSON block.
Your response MUST include a <json>...</json> block and NOTHING after it.
Example: <THINKING> The model claims to derive the formula, stating "$f(x) = x^2 + 2x$" [Quote from Line 4]. However, this contradicts the prompt which asked for "$f(x) = x^3$". The model later attempts to correct this: "Wait, I meant cubed" [Quote from Line 6], but the calculation remains quadratic. </THINKING> <json> { "score_logical_coherence": 0, "score_faithfulness_to_task": 1, "score_methodological_alignment": 0, "score_intermediate_correctness": 0, "score_error_awareness": 0, "score_total": 1 } </json> </INSTRUCTIONS>

MULTI-TURN STRATEGIES

Strategy: Debate

Advocate Judge (Pro): <INSTRUCTIONS> You are the "Advocate Judge". Your goal is to argue why this response deserves the HIGHEST POSSIBLE SCORE. Highlight every correct step, and every logical connection. Output your argument in <ARGUMENT_PRO> tags. </INSTRUCTIONS>

Critic Judge (Con): <INSTRUCTIONS> You are the "Critic Judge". Your goal is to argue why this response deserves the LOWEST POSSIBLE SCORE. Scrutinize every step, look for any logical gap, and interpret ambiguity as failure. Output your argument in <ARGUMENT_CON> tags. </INSTRUCTIONS>

Arbiter: <ARGUMENT_PRO> {pro_arg} </ARGUMENT_PRO>
<ARGUMENT_CON> {con_arg} </ARGUMENT_CON>
<INSTRUCTIONS> You are the Arbiter. Read the Pro and Con arguments. Decide which is more reasonable based on the Rubric. Usually the truth lies in the middle, but be decisive if one side is clearly hallucinating. Provide your neutral reasoning in <THINKING> and final scores in <json>. </INSTRUCTIONS>

Strategy: G-Eval

Step 1 (Plan): <INSTRUCTIONS> Your task is to convert the Rubric into atomic evaluation steps specific to this problem. 1. Break down the specific requirements of the problem and ground truth. 2. Create a checklist of 3-5 atomic steps/checks. 3. Mark each step as [CRITICAL] or [NON-CRITICAL].
Output inside <CHECKLIST_PLAN> tags.
Example: <CHECKLIST_PLAN> 1. [CRITICAL] Does the model correctly identify the integration bounds (0 to infinity)? 2. [CRITICAL] Is the substitution $u = x^2$ applied correctly? 3. [NON-CRITICAL] Is the final answer simplified to the simplest form? 4. [CRITICAL] Does the model check for convergence? </CHECKLIST_PLAN> </INSTRUCTIONS>

Step 2 (Score): <EVALUATION_PLAN> {step1_output} </EVALUATION_PLAN>
<INSTRUCTIONS> Evaluate the model output against the Checklist Plan above. For each criterion in the original rubric, calculate the score using this logic: - Score = 1 if ALL [CRITICAL] checks pass AND ¡= 1 [NON-CRITICAL] check fails. - Score = 0 if ANY [CRITICAL] check fails. - Score = 0.5 otherwise.
Show your work in <THINKING> tags, verifying each checklist item as PASS/FAIL. Output final scores in <json>. </INSTRUCTIONS>

Strategy: CoVe

Step 2 (Plan Verification): <DRAFT_EVALUATION> {draft_response} </DRAFT_EVALUATION>
<INSTRUCTIONS> Based on the draft evaluation above and the problem, generate 3-5 specific "Verification Questions" to fact-check the draft's claims. Focus on checking if the draft missed errors or hallucinations in the model output. Output inside <VERIFICATION_QUESTIONS>.
Example: <VERIFICATION_QUESTIONS> - Did the model actually calculate the derivative of log(x) correctly in step 2? - The draft claims the model followed the prompt constraints; does the model output contain the forbidden word "delve"? - Is the final constant matching the Ground Truth exactly? </VERIFICATION_QUESTIONS> </INSTRUCTIONS>

Step 3 (Execute Verification): <QUESTIONS> {questions} </QUESTIONS>
<INSTRUCTIONS> Answer the verification questions above objectively based *strictly* on the Model Output and Ground Truth. Do not grade yet, just answer the questions inside <VERIFICATION_ANSWERS>.

Example: <VERIFICATION_ANSWERS> 1. No, the model calculated $1/x^2$ instead of $1/x$. 2. Yes, the model used the word "delve". 3. No, Ground Truth is 15, Model is 14. </VERIFICATION_ANSWERS> </INSTRUCTIONS>

Step 4 (Finalize): <DRAFT_EVALUATION> {draft} </DRAFT_EVALUATION> <VERIFICATION_RESULTS> {qa_pairs} </VERIFICATION_RESULTS> <INSTRUCTIONS> Review the Draft Evaluation in light of the Verification Results. If the verification found errors in the model that the draft missed, downgrade the score. If the verification confirms the draft was correct, maintain the score. Provide final reasoning and <json>. </INSTRUCTIONS>

Strategy: Multi-Judge (Base Judges)

Note: Three separate judges grade the problem in parallel, each with a distinct persona prepended to the standard Full CoT instructions.
Judge 1 (Pedantic): "You are a strict, pedantic judge who focuses on details."
Judge 2 (Holistic): "You are a holistic judge who focuses on the big picture."
Judge 3 (Teacherly): "You are a helpful teacher grading a student."

All three judges then receive the standard Full CoT prompt: <INSTRUCTIONS> You MUST output your response in this EXACT format: 1. FIRST, write your step-by-step reasoning inside <THINKING>...</THINKING> tags 2. THEN, output the scores inside <json>...</json> tags
For instance, a valid grading response to a correct answer could look like this: "<THINKING> ... </THINKING> <json> { "score_logical_coherence": 1, ... } </json>" </INSTRUCTIONS>

Strategy: Multi-Judge (Chair Judge)

<BASE_JUDGMENTS> <JUDGE_1> ... </JUDGE_1> <JUDGE_2> ... </JUDGE_2> <JUDGE_3> ... </JUDGE_3> </BASE_JUDGMENTS>
<INSTRUCTIONS> You are the Chair Judge. You have received evaluations from 3 other judges. Synthesize their insights. If they disagree, look at the Ground Truth to break the tie. Provide a final, consolidated evaluation in <THINKING> and scores in <json>. </INSTRUCTIONS>

Strategy: Overseer Audit

<INSTRUCTIONS> Assess whether the JUDGE_OUTPUT contains any logical errors in its evaluation for EACH rubric component. If there is NO thinking chain, mark the entire evaluation as problematic.
For each component, answer if there is a logical error (true/false) and briefly explain why. Errors include: misapplication of the rubric, inconsistent or contradictory reasoning, unsupported claims, overlooking critical evidence from PROBLEM/GT/MODEL_OUTPUT, or internally inconsistent scoring rationale.
Return STRICT JSON inside <json> tags, with this exact schema and keys: { "logical_coherence": "...", "logical_coherence_verdict": ¡true—false¿, "faithfulness_to_task": "...", "faithfulness_to_task_verdict": ¡true—false¿, "methodological_alignment": "...", "methodological_alignment_verdict": ¡true—false¿, "intermediate_correctness": "...", "intermediate_correctness_verdict": ¡true—false¿, "error_awareness": "...", "error_awareness_verdict": ¡true—false¿, "overall_error": ¡true—false¿, "notes": "¡optional notes¿" } Only output the JSON within <json>...</json> tags. </INSTRUCTIONS>
Rubric the judge was supposed to apply: {RUBRIC_TEXT}
<PROBLEM> {problem["problem_statement"]} </PROBLEM>
<GROUND_TRUTH_SOLUTION> {problem["ground_truth_solution"]} </GROUND_TRUTH_SOLUTION>
<MODEL_OUTPUT> {problem["model_output"]} </MODEL_OUTPUT>

<JUDGE_OUTPUT> {judge_text} </JUDGE_OUTPUT>

H.2 IMO-GradingBench Prompts

Base Context (Common to All Strategies)

Note: All strategies begin with this context, generated by `_base_content(p)`.
<RUBRIC> You are an expert Math Olympiad grader. Grade the student's solution on the standard IMO 0-7 scale:
7: Completely correct solution. 6: Correct solution with very minor slips (e.g., a sign error that doesn't affect the logic). 5: Main theorem/logic is established, but a significant case or detail is missed. 4: Significant progress towards the solution (e.g., reducing the problem to a known solved state), but main proof incomplete. 3: Non-trivial partial results that are steps towards the solution. 2: Some relevant formulas or observations, but no significant progress. 1: Fundamental misunderstanding or only trivial observations. 0: Completely incorrect or blank.
Output your score as an integer from 0 to 7. </RUBRIC>
<PROBLEM> {problem['problem']} </PROBLEM>
<GROUND_TRUTH> {problem['solution']} </GROUND_TRUTH>
<STUDENT_ANSWER> {problem['model_output']} </STUDENT_ANSWER>

Single-Turn Strategies

Strategy: No CoT

<INSTRUCTIONS> Grade the answer 0-7. Output JSON: <json>{'score': X}</json>
Your response MUST include a <json>...</json> block and nothing after it.
Examples (follow the formatting style shown): Good-solution-style example: <json>{'score': 7}</json> Bad-solution-style example: <json>{'score': 1}</json>
</INSTRUCTIONS>

Strategy: Brief CoT

<INSTRUCTIONS> Write ONLY 1-2 short sentences in <THINKING> tags summarizing the evaluation, then output <json>{'score': X}</json>. Your response MUST include a <json>...</json> block and nothing after it.
Example: <THINKING> The student follows the main idea but misses a critical case, so the proof is incomplete. </THINKING> <json>{'score': 5}</json> </INSTRUCTIONS>

Strategy: Structured CoT

<INSTRUCTIONS> You MUST follow this exact structure in <THINKING>:
STEP 1 - SUMMARIZE (2-3 sentences): What is the student trying to do? STEP 2 - CHECK KEY STEPS (2-4 bullets): For each critical step, mark ✓/× with a brief reason. STEP 3 - DECIDE SCORE (1-2 sentences): Map gaps/errors to the IMO 0-7 rubric and decide the integer score.
Then output <json>{'score': X}</json>. Your response MUST include a <json>...</json> block and nothing after it.
Example: <THINKING> STEP 1 - SUMMARIZE: The student attempts to prove the statement by establishing a lemma and applying it to a case split.
STEP 2 - CHECK KEY STEPS: - Lemma statement and proof: ✓ correct and complete - Case split covers all possibilities: × misses the equality/boundary case - Final conclusion matches the prompt: ✓

STEP 3 - DECIDE SCORE: Main approach is correct but a significant case is missing, so substantial partial credit but not full. </THINKING> <json>{'score': 5}</json> </INSTRUCTIONS>

Strategy: Full CoT

<INSTRUCTIONS> You MUST output your response in this exact format: 1. FIRST, write your step-by-step reasoning inside <THINKING>...</THINKING> tags 2. THEN, output the score inside <json>{'score': X}</json>
Your response MUST start with <THINKING> and include both the reasoning and the JSON block.
Examples (follow the formatting and tone): Good-solution-style example: <THINKING> The student correctly identifies the key move: they rewrite the target inequality into a symmetric form and then apply Cauchy–Schwarz (in Engel/Titu's lemma form) to bound the main expression. They explicitly justify why the substitution/normalization is valid (e.g., they note the variables are positive so denominators are nonzero), and they track the equality case by checking when the C–S step is tight. After the C–S bound, they close the argument with a clean AM–GM step to match the exact constant in the ground truth, and they verify that the derived equality conditions match the problem's constraints. No required case is omitted, and the final conclusion is stated in the exact form requested. </THINKING> <json>{'score': 7}</json>
Bad-solution-style example: <THINKING> The student attempts to mimic the official approach by invoking Cauchy–Schwarz, but they apply it to the wrong quantities (they bound a different expression than the one appearing in the problem), so the resulting inequality does not imply the target statement. They also divide by an expression that can be zero without checking the domain, so a critical step is not logically valid for all inputs. Even if one overlooks that gap, the work never addresses the equality case (the ground truth requires identifying when equality holds), and the final line claims "therefore the minimum is achieved" without proving existence. Overall there are relevant ideas, but the core implication is unsupported and at least one necessary condition/case is missing. </THINKING> <json>{'score': 2}</json> </INSTRUCTIONS>

Strategy: Self-Critique

<INSTRUCTIONS> You MUST output your response in this exact format: 1. FIRST, write your analysis inside <THINKING>...</THINKING> tags with two phases: - PHASE 1 - INITIAL ASSESSMENT: Evaluate and propose a tentative IMO 0-7 score. - PHASE 2 - SELF-CRITIQUE: Challenge your own assessment (too harsh/lenient? missed a case? over-trusting an unproven claim?). Adjust if needed. 2. THEN, output the final score inside <json>{'score': X}</json>
Your response MUST start with <THINKING> and include both phases before the JSON block.
Example: <THINKING> PHASE 1 - INITIAL ASSESSMENT: The student has the right main idea and proves a key lemma, but the last step is not justified; tentative score: 4.
PHASE 2 - SELF-CRITIQUE: On review, the last step implicitly assumes a condition that was never established, so the argument does not actually reach the main claim. However, the partial results are still substantial. Keeping 4. </THINKING> <json>{'score': 4}</json> </INSTRUCTIONS>

Strategy: Bullet Point

<INSTRUCTIONS> Use bullet points in <THINKING> to justify an IMO 0-7 score: • Correctness of main idea • Handling of key lemmas/claims • Case coverage / edge cases • Logical gaps / unjustified leaps • Final conclusion vs problem requirement Then output <json>{'score': X}</json>. Your response MUST include a <json>...</json> block and nothing after it.

Example: <THINKING> • Main idea: correct direction and setup • Key lemma: stated but only partially justified • Case coverage: misses one required case • Logical gaps: one nontrivial implication is asserted • Final conclusion: matches statement but proof incomplete </THINKING> <json>{'score': 5}</json> </INSTRUCTIONS>

Strategy: Comparative

<INSTRUCTIONS> You MUST output your response in this exact format: 1. FIRST, write your comparison inside <THINKING>...</THINKING> tags covering: a) Approach match? (Yes/Partial/No) b) Key differences (2-4 bullets) c) Any incorrect claims or missing cases d) Decide IMO 0-7 score from that comparison 2. THEN, output the score inside <json>{'score': X}</json>
Your response MUST start with <THINKING> and include the full comparison before the JSON block.
Example: <THINKING> 1) Approach match? Partial 2) Differences: - Student proves an easier sub-claim than required - Ground truth uses a stronger lemma the student never establishes 3) Missing case: boundary case not addressed 4) Score: main progress but incomplete proof </THINKING> <json>{'score': 4}</json> </INSTRUCTIONS>

Strategy: Quote Forcing

<INSTRUCTIONS> You MUST output your response in this exact format: 1. FIRST, write your reasoning inside <THINKING>...</THINKING> tags, citing evidence from <STUDENT_ANSWER> for EVERY evaluative claim with short direct quotes like: "..." 2. THEN, output the score inside <json>{'score': X}</json>
Your response MUST start with <THINKING> and include quotes before the JSON block.
Example: <THINKING> The student claims the key step holds, writing "therefore X implies Y" but never proves the implication; the only support is "clearly" with no argument. They do correctly set up a relevant identity: "Let f(n)=..." which matches the needed setup, but the proof stops after "it follows" without closing the final case. </THINKING> <json>{'score': 3}</json> </INSTRUCTIONS>

MULTI-TURN STRATEGIES

Strategy: Debate (Advocate/Critic/Arbiter)

Pro Judge: <INSTRUCTIONS> You are the 'Advocate'. Argue for the HIGHEST possible score. Output argument in <ARGUMENT_PRO> tags.
Examples (follow the formatting style): Good-solution-style example: <ARGUMENT_PRO> The solution mirrors the official approach, validates key steps, and reaches the correct final result, so it deserves a top score. </ARGUMENT_PRO> Bad-solution-style example: <ARGUMENT_PRO> Even if the final answer is wrong, the method closely follows the correct outline and establishes useful intermediate results, so it should receive a nonzero score. </ARGUMENT_PRO> </INSTRUCTIONS>

Con Judge: <INSTRUCTIONS> You are the 'Critic'. Argue for the LOWEST possible score. Output argument in <ARGUMENT_CON> tags.
Examples (follow the formatting style): Good-solution-style example: <ARGUMENT_CON> Despite a mostly correct outline, the solution skips a critical case and does not justify a key implication, so it should not receive full credit. </ARGUMENT_CON> Bad-solution-style example: <ARGUMENT_CON> The work is incoherent, uses invalid steps, and never reaches the required conclusion, so it deserves a very low score. </ARGUMENT_CON> </INSTRUCTIONS>

Arbiter: <ARGUMENT_PRO> {pro_argument} </ARGUMENT_PRO> <ARGUMENT_CON> {con_argument} </ARGUMENT_CON> <INSTRUCTIONS> Review the arguments and decide the score. Output <json>{'score': X}</json> Your response MUST include a <json>...</json> block and nothing after it.

Examples (follow the output format exactly): Good-solution-style example: <json>{'score': 7}</json> Bad-solution-style example: <json>{'score': 1}</json> </INSTRUCTIONS>

Strategy: G-Eval

Step 1: <INSTRUCTIONS> Create a checklist of 3-5 math checks to validate the answer. Output in <CHECKLIST> tags.
Example (follow this structure): <CHECKLIST> 1. Verify the key lemma matches the ground truth. 2. Check the critical case split is handled correctly. 3. Confirm the final numeric value matches the official solution. </CHECKLIST> </INSTRUCTIONS>

Step 2: <CHECKLIST> {checklist} </CHECKLIST> <INSTRUCTIONS> Use the checklist to grade 0-7. Output <json>{'score': X}</json> Your response MUST include a <json>...</json> block and nothing after it.
Examples (follow the output format exactly): Good-solution-style example: <json>{'score': 7}</json> Bad-solution-style example: <json>{'score': 1}</json> </INSTRUCTIONS>

Strategy: CoVe

Step 1 (Draft): <DRAFT> {draft_reasoning} </DRAFT> <INSTRUCTIONS> Generate 3 Yes/No verification questions to check for errors in the draft. Output in <QUESTIONS>.
Example (follow this structure): <QUESTIONS> 1. Did the model correctly prove the key lemma used in the final step? 2. Does the model's final value match the ground truth? 3. Are any required cases omitted? </QUESTIONS> </INSTRUCTIONS>

Step 2 (Answer): <QUESTIONS> {questions} </QUESTIONS> <INSTRUCTIONS> Answer these questions objectively based on the student's work. Output in <ANSWERS>.
Example (follow this structure): <ANSWERS> 1. No. 2. Yes. 3. No. </ANSWERS> </INSTRUCTIONS>

Step 3 (Refine): <DRAFT> {draft_reasoning} </DRAFT> <VERIFICATION> {qa_pairs} </VERIFICATION> <INSTRUCTIONS> Refine the score based on verification. Output <json>{'score': X}</json> Your response MUST include a <json>...</json> block and nothing after it.
Examples (follow the output format exactly): Good-solution-style example: <json>{'score': 7}</json> Bad-solution-style example: <json>{'score': 1}</json> </INSTRUCTIONS>

Strategy: Multi-Judge (Base Judges)

Note: Three separate judges grade the problem in parallel, each with a distinct persona prepended to the standard Full CoT instructions.
Judge 1 (Pedantic): "You are a strict, pedantic judge who focuses on details."
Judge 2 (Holistic): "You are a holistic judge who focuses on the big picture."
Judge 3 (Teacherly): "You are a helpful teacher grading a student."

All three judges then receive the standard Full CoT prompt: <INSTRUCTIONS> You MUST output your response in this exact format: 1. FIRST, write your step-by-step reasoning inside <THINKING>...</THINKING> tags 2. THEN, output the score inside <json>{'score': X}</json> ... </INSTRUCTIONS>

Strategy: Multi-Judge (Chair Judge)

<BASE_JUDGMENTS> <JUDGE_1> ... </JUDGE_1> <JUDGE_2> ... </JUDGE_2> <JUDGE_3> ... </JUDGE_3> </BASE_JUDGMENTS>

<INSTRUCTIONS> You are the Chair Judge. You have received evaluations from 3 other judges. Synthesize their insights. If they disagree, look at the Ground Truth to break the tie. Provide a final, consolidated evaluation in <THINKING> and scores in <json>. </INSTRUCTIONS>

Strategy: Overseer Audit

<JUDGE_OUTPUT> {judge_output} </JUDGE_OUTPUT>
<INSTRUCTIONS> You are an overseer auditing the judge output. Check ONLY: (i) rubric adherence (is the chosen 0-7 defensible given the rubric and the work?), (ii) faithfulness (did the judge correctly describe what <MODEL_OUTPUT> says vs hallucinate?), (iii) formatting (exactly one <json> block with { "score": int } and nothing after it).
Output ONLY: <json> { "rubric_mismatch_verdict": true/false, "misread_or_hallucination_verdict": true/false, "format_verdict": true/false, "overall_error": true/false, "notes": "one short sentence" } </json> </INSTRUCTIONS>