EQUILIBRIUM OF THOUGHTS: REASONING AS ITERATIVE STABILIZATION UNDER ADVERSARIAL PROBING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning frameworks from Chain-of-Thought to Tree-of-Thoughts are *constructive*: they build an answer step by step, so a single early error cascades and is rarely caught. These methods find an answer but cannot certify it. We propose **Equilibrium of Thoughts (EoT)**, which reframes reasoning as *stabilization under adversarial probing*: generate an answer, attack it for the single strongest objection, revise if the attack lands, and repeat until no valid objection remains. The surviving answer is the reasoning equilibrium, grounded in fixed-point, argumentation, and game theory. Our central finding concerns *who should attack*. A model cannot find the flaws it just made, so self-critique is unreliable. The bottleneck is the *critic*. Holding the reasoner fixed and varying only the adversary, we show a strong, decorrelated *cross-model* critic is the active ingredient: it rescues a weak reasoner ($+20$ on AIME for a 32B model) and lets critic capability substitute for reasoner capability, while a same-strength *same-family* critic does not, which isolates the contribution of decorrelation from that of raw strength. Recast as cross-model adversarial verification, this prompt-only recipe helps across math, science, code, and multi-hop QA, including benchmarks where multi-agent debate *degrades* accuracy. Finally, EoT *wraps* an initializer instead of replacing it: composed on Tree-of-Thoughts with the same critic, it beats debate on every hard benchmark (AIME-2024 93.3 vs. 83.3). Adversarial verification thus complements constructive reasoning as a primitive that amplifies it: the stronger the initializer, the higher EoT reaches.

1 INTRODUCTION

The dominant paradigm for language model reasoning is *constructive*: generate a chain of intermediate steps whose final step yields the answer (Wei et al., 2022; Nye et al., 2021). Subsequent work expands the construction from a chain to a tree (Yao et al., 2023), a graph (Besta et al., 2024), or a guided search (Sel et al., 2024), spending test-time compute to explore more paths. But constructive methods share a structural weakness: *errors cascade*. Because each step rests on the ones before it, a single mistake at step 3 silently corrupts steps 4 through 10, and searching over a flawed prefix rarely recovers. Exploring more paths raises the odds of finding a good one, but does little to certify that the path returned is sound. Construction can find the answer, yet on its own it cannot verify one.

How do humans guard against this? Not by deriving more carefully, but by *trying to break* what they have derived. A scientific claim earns confidence not because its derivation looked clean but because it has withstood concerted attempts at refutation (Popper, 1963), and a mathematical proof is hardened precisely by the counterexamples it survives (Lakatos, 1976). Correctness, on this view, is not a property of how an answer was *built*. It is a property of whether the answer can be *broken*.

We bring this principle to language-model reasoning. Instead of asking "how do I construct the right answer?", we ask "*what answer survives every attack?*" This reframes reasoning from construction to **stabilization**: an answer is a candidate to be stress-tested instead of a chain to be extended, and the output is whatever remains standing once no further objection lands. We call the surviving answer a *reasoning equilibrium* and the method that finds it **Equilibrium of Thoughts (EoT)**, a simple loop: generate an initial answer; *attack* it for the single strongest objection; if no valid objection exists the answer is at equilibrium and is returned; otherwise *revise* to address the objection and repeat. The method needs only two prompt templates (attack and revise) and no training, search tree, or problem-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

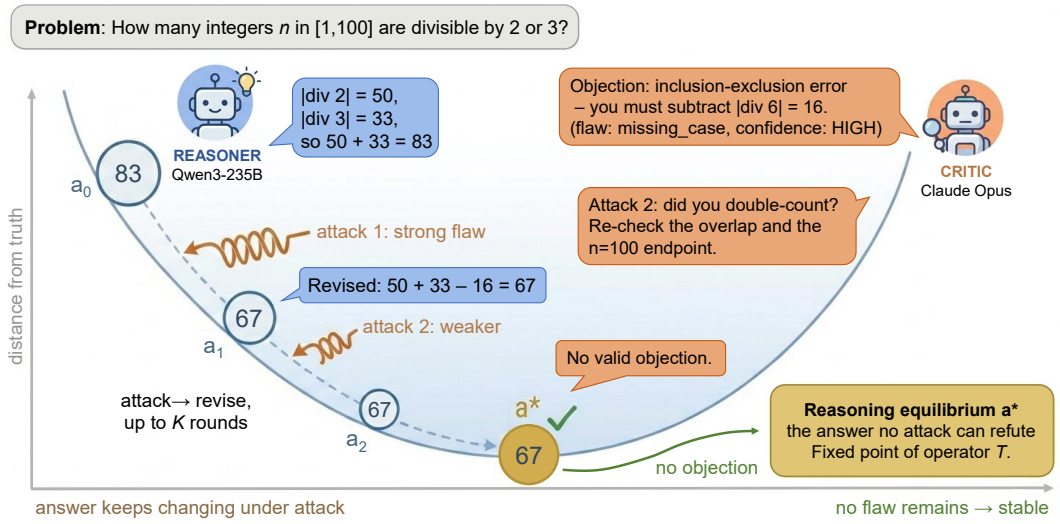

Figure 1: **Equilibrium of Thoughts (EoT)** reframes reasoning as stabilization under cross-model adversarial attack. A *reasoner* produces an initial answer, a *critic from a different model family* attacks it for the single strongest objection, the reasoner revises, and the loop repeats until the answer reaches the *reasoning equilibrium*, a fixed point of the attack–revise operator that no attack can refute. On the worked example, the reasoner's $50{+}33{=}83$ omits the inclusion–exclusion overlap, the critic flags it, and the revised $50{+}33{-}16{=}67$ survives further attack. Our central finding: the critic must be *decorrelated* from the reasoner (§5), because a model cannot reliably find the flaws it just made.

specific design, yet it induces reasoning dynamics fundamentally unlike construction: compute is spent on *certifying* an answer, where prior methods *extend* or *resample* it.

Why attack instead of self-check? A natural objection is that this is just self-correction by another name: "why not ask the model to check its own work?" The distinction is the crux of the method. Self-correction asks a *metacognitive* question ("is my answer good?"), and a growing body of evidence shows language models answer it unreliably: they cannot self-correct reasoning without external signals (Huang et al., 2024), their introspective reports are largely unfaithful (Singh et al., 2026; Wang, 2026), and they show no individuated metacognition (Moran & Whiting, 2026). EoT never poses that question. It asks an *adversarial* one ("find a specific flaw"), which exploits a basic asymmetry: *exhibiting one concrete counterexample is easier and more reliable than holistically judging quality*, just as verifying a proof is easier than producing it (Lightman et al., 2024).

But a model is a poor critic of itself. The asymmetry buys reliability only if the critic can actually find the flaw. Our experiments show that a model attacking *its own* answer largely cannot: self-play EoT adds nothing on problems whose errors are systematic instead of incidental slips, because the same blind spots that produce the error also hide it from the critic. The fix is not a better prompt but a better *critic*: holding the reasoner fixed and swapping in a strong critic from a *different model family* rescues a weak reasoner by up to $+20$ points, whereas an equally strong critic from the *same* family barely helps, because its errors are correlated with the reasoner's. The active ingredient is *decorrelation*: a critic whose failure modes differ from the reasoner's. This reframes EoT as **cross-model adversarial verification**, and it explains why multi-agent debate works at all, since debate manufactures a *weak* form of decorrelation through temperature-sampled copies of one model. With a single, genuinely decorrelated critic, EoT closes most of the gap to debate at lower cost, and because it only *stress-tests* an answer, it composes on a stronger initializer to overtake debate outright.

The equilibrium is principled. Stopping when "no objection lands" is not an arbitrary cutoff. The attack–revise loop is best-response dynamics for a two-player game between a reasoner and an adversary, and its fixed point, the answer the adversary can no longer object to, is simultaneously a *Nash equilibrium* of that game (Osborne & Rubinstein, 1994), a *preferred extension* in Dung's

abstract argumentation framework (Dung, 1995), and a fixed point of the attack–revise operator in the sense of classical fixed-point theory (Tarski, 1955). These three lenses agree on the same object (§3.2), and together they let us state when EoT converges and why. The same framing subsumes prior methods as special cases: Chain-of-Thought (CoT) is EoT with zero attack rounds, self-correction is EoT with one *metacognitive* round, and debate is EoT whose adversary must mount a full counter-position in place of a single objection. Our core contributions are:

- We introduce EoT, a reasoning primitive that reframes reasoning from construction to adversarial stabilization, with a unifying theory: the three equilibrium readings (game-theoretic, argumentation-theoretic, fixed-point) yield convergence conditions that predict its failure modes and subsume CoT, self-correction, and debate as special cases (§3.2).
- We identify the critic as the bottleneck and decorrelation as the lever: fixing the reasoner and varying only the adversary, we show a *strong, decorrelated cross-model critic* is what makes EoT work, and that a same-strength critic from a *different* family captures most of the gain, which isolates decorrelation as the active ingredient independent of raw strength.
- We demonstrate one recipe across four domains, and show that it composes: with a cross-model critic, a prompt-only EoT helps on math, science, code, and multi-hop QA with no per-benchmark tuning, and because it *wraps any initializer*, composing it on Tree-of-Thoughts (ToT) beats multi-agent debate on every hard benchmark, with the initializer alone setting the ceiling.
- We show that convergence speed is a free difficulty signal: because easy problems concede on the first attack and hard ones iterate, rounds-to-equilibrium predicts correctness *before* the gold answer is known, an unsupervised triage signal EoT yields at no extra cost.

2 Related Work

Constructive reasoning paradigms. CoT (Wei et al., 2022) and scratchpads (Nye et al., 2021) show that emitting intermediate steps improves reasoning. Self-Consistency (Wang et al., 2023) samples many chains and majority-votes, buying *coverage* through parallel diversity. ToT (Yao et al., 2023), Graph-of-Thoughts (GoT) (Besta et al., 2024), Algorithm of Thoughts (Sel et al., 2024), and Buffer of Thoughts (Yang et al., 2024) enrich the *topology* of construction with search, backtracking, and reusable templates, and later work selects this topology adaptively (Besta et al., 2025; Pandey et al., 2025; Fricke et al., 2026; Hao et al., 2025). All of these are *constructive*: they expand or reorganize the space of paths along which an answer is *built*, so all remain vulnerable to a sound-looking but wrong path. EoT is orthogonal and composable: it *stress-tests* a finished answer without building one, and wraps the output of any such method.

Iterating on an answer: self-correction and debate. A second family revisits an answer after producing it. Self-correction iterates "critique-then-improve" (Shinn et al., 2023), but its reliability is fragile: it *degrades* reasoning without external feedback (Huang et al., 2024) and helps mainly when an external signal is present (Kamoi et al., 2024). Multi-agent debate instead has agents argue toward consensus (Du et al., 2024; Liang et al., 2024). EoT differs from both in *what the critic does* and *who it is*. It replaces self-correction's metacognitive "did I err?" with the adversarial "produce a specific counterexample," isolated by our compute-matched self-correction baseline. Debate is *position-vs-position*, where each agent authors a full rival solution. EoT is *position-vs-objection*: it needs only one flaw and so yields a clean binary convergence signal debate lacks. We further argue debate's effectiveness owes less to argumentation than to the *decorrelation* between its temperature-sampled agents (§5), a stronger form of which EoT obtains directly from a single different-family critic. Both are recoverable as constrained instances of EoT.

Allocating test-time compute. A third family asks how to *spend* inference compute, which can rival scaling parameters (Snell et al., 2025; Muennighoff et al., 2025): by sampling more chains, training models to deliberate at length (DeepSeek-AI, 2025), or matching effort to difficulty via strategy selection (Sui et al., 2026), meta-CoT (Xiang et al., 2025; Gao et al., 2024), and adaptive depth (Shen et al., 2025). The payoff is uneven, concentrating on math and symbolic tasks (Sprague et al., 2025) and sometimes *hurting* (Zhou et al., 2026). EoT introduces a distinct axis, *adversarial verification of a single answer*, that is orthogonal to how the answer was produced, and so *composes* with the sampling and search methods above without replacing them, at lower cost than sampling

many chains (§5). It is also adaptive without a difficulty estimator, and the overthinking risk reappears concretely as an over-eager adversary inducing divergence. Both are measured directly (§5).

Verification and argumentation. Verifiers and process-reward models score steps or solutions to guide selection (Lightman et al., 2024), and Nash-style aggregation has been used for training-free step verification (Sinha et al., 2026). EoT performs *implicit, training-free* verification: the adversary's failure to find a flaw *is* the certificate, with no separately trained reward model. We make this rigorous through Dung's abstract argumentation (Dung, 1995), in which an equilibrium answer is a *preferred extension*, the maximal set of claims that withstands all attacks. To our knowledge, we are the first to operationalize this argumentation–reasoning correspondence for LLMs, grounded in the epistemology of knowledge-by-refutation (Popper, 1963; Lakatos, 1976).

3 Equilibrium of Thoughts

EoT is a short loop, but its stopping rule carries real content: terminating when "no attack lands" is well-defined only if we say what an attack, a revision, and surviving them are. We formalize the loop as iterating an *attack–revise operator* toward a fixed point (§3.1) and state when it converges and improves accuracy (§3.2). The payoff is concrete: the formalism predicts the three failure modes we later observe empirically.

3.1 Formalization

Let $\mathcal{P}$ be the space of problems and $\mathcal{A}$ the space of (answer, reasoning) pairs. EoT is built from two operators: one that looks for a flaw, and one that repairs it.

Definition 1 (Attack operator)**.** *$Atk : \mathcal{P} \times \mathcal{A} \to \mathcal{O} \cup \{\perp\}$ takes a problem and a candidate answer and produces either a specific objection $o \in \mathcal{O}$ or the null objection $\perp$ (indicating no valid flaw was found).*

Definition 2 (Revise operator)**.** *$Rev : \mathcal{A} \times \mathcal{O} \times \mathcal{O}^* \to \mathcal{A}$ takes the current answer, the objection, and the history of prior objections, and produces a revised answer that addresses the objection without reintroducing previously fixed flaws.*

Definition 3 (Attack-Revise operator)**.** *The composite operator $T : \mathcal{A} \to \mathcal{A}$ is defined as:*

$$T(a) = \begin{cases} a & \text{if } Atk(P, a) = \perp \\ Rev(a, Atk(P, a), H) & \text{otherwise} \end{cases}$$

where H is the accumulated history of prior objections.

Definition 4 (Reasoning Equilibrium)**.** *An answer $a^* \in \mathcal{A}$ is at* equilibrium *for problem P if $T(a^*) = a^*$, i.e., the attack operator produces no valid objection: $Atk(P, a^*) = \perp$.*

Three views of the same equilibrium. Definition 4 admits three reinforcing readings, which is what makes "no objection lands" a principled criterion instead of a heuristic. *(i) Game theory.* The reasoner (playing $a \in \mathcal{A}$) and adversary (playing $o \in \mathcal{O} \cup \{\perp\}$) form a zero-sum game in which the adversary gains from valid objections and the reasoner from surviving them. The pair $(a^*, \perp)$, where the reasoner's answer admits no objection and the adversary's best response is to concede, is a Nash equilibrium (Osborne & Rubinstein, 1994), and the EoT loop is best-response dynamics for it. *(ii) Argumentation theory.* Reading answers and objections as arguments with attack edges, a^* is a *preferred extension* of Dung's framework (Dung, 1995): a maximal set of claims that is conflict-free and defends itself against every attack. *(iii) Fixed-point theory.* a^* is a fixed point of T. Two distinct fixed-point principles apply to two distinct questions, and we keep them separate. *Existence* of an equilibrium follows from the argumentation reading: ordering answer-states by the attacks they survive makes T a monotone map, and Tarski's theorem guarantees a fixed point on the resulting complete lattice without any metric or contraction (Tarski, 1955). *Convergence* of the iteration to such a fixed point is a separate, stronger claim that needs a metric: if T is contractive in expectation (each revision moves the answer closer to the truth) then the iterates converge geometrically, a Banach-type argument (Banach, 1922). The two should not be conflated, since Tarski gives a fixed point but says nothing about reaching it, while contraction gives reachability. We make the contraction condition precise in §3.2 (where it is an *identified* condition that we do not derive from primitives) and measure it empirically in §5.

Algorithm. The loop is implemented entirely through prompting, with *generate*, *attack*, and *revise* each a single model call. The attack prompt elicits the *single strongest* objection or an explicit "NO VALID OBJECTION" (vague doubts are disallowed, giving a clean binary convergence signal), and the revision prompt is shown the objection plus the history of prior objections, which prevents *regression* (re-introducing a previously-fixed flaw, the cause of oscillation). We also detect oscillation directly: if an extracted answer returns to an earlier value after changing, the loop has cycled, and we resolve the cycle by majority vote over the trajectory. Pseudocode (Algorithm 1) and full prompt templates are in Appendix A.

3.2 Theoretical Properties

Claim 1 (Subsumption)**.** *Several established reasoning methods are recovered as special cases of EoT by restricting the iteration budget K and the form of the adversary, as summarized in Table 1. EoT generalizes them along two axes: it permits arbitrary iteration ($K \geq 1$) and it tasks the adversary with* finding a flaw*, as opposed to judging quality or authoring an alternative.*

Table 1: Prior methods as constrained instances of EoT. Each is obtained by limiting the iteration budget, the adversary's task, or both. EoT lifts these restrictions.

Method	Iteration budget	Adversary
CoT	$K = 0$	none (no probing)
Self-correction	$K = 1$	metacognitive ("did I err?")
Multi-agent debate	$K \geq 1$, distinct model	counter-position (a full rival solution)
Self-consistency	parallel (not iterative)	implicit (disagreement across samples)
Equilibrium of Thoughts	$K \geq 1$	explicit objection ("find the flaw")

Proposition 1 (Monotonicity under competent attack)**.** *Let accuracy be $Acc(a_i) = \mathbb{E}[\mathbf{1}[answer(a_i) = y^*]]$ over a problem distribution, and suppose the attack operator is* sound *(valid objections identify genuine errors, $P($flaw real $\mid o \neq \perp) > \frac{1}{2}$) and* complete *(genuine errors are eventually found, $P(o = \perp \mid a \neq a^*) < 1$). Then, provided the critic's false-objection rate on already-correct answers is below an accuracy-dependent threshold (made precise in Appendix D), expected accuracy is non-decreasing with iteration, $Acc(a_{i+1}) \geq Acc(a_i)$.*

Proposition 2 (Convergence)**.** *If the attack-revise operator T is contractive in expectation, i.e. $\mathbb{E}[d(T(a), a^*)] < \mathbb{E}[d(a, a^*)]$ for $a \neq a^*$, then the iteration converges almost surely to a fixed point within $O(\log(1/\epsilon))$ steps.*

The assumptions name the failure modes. The value of Propositions 1–2 is that violating each assumption predicts a *distinct, observable* pathology, and these are exactly the behaviors we find in §5: (i) violating *soundness* (a spurious objection) drives a revision away from a correct answer (*divergence*, the "overthinking" of Zhou et al., 2026); (ii) violating *completeness* (the adversary misses a real flaw) halts at a wrong answer (*premature convergence*); and (iii) violating *contractivity* (a revision that trades one flaw for another of equal weight) cycles between answers (*oscillation*), which is why we track objection history and detect cycles explicitly. Thus the theory is not decorative: each condition is a lever we can probe (attacker strength, cross-model adversaries, history ablation) and each failure is one we can measure. Appendix D proves both propositions under an explicit probabilistic model, expressing each assumption as a measurable critic parameter (false-objection rate ϵ, completeness defect c, contraction factor κ), so the conditions are checkable.

The critic's identity is the primary design lever. This theory has an immediate design consequence: since the conditions of both propositions are *critic* properties (ϵ, c, κ), the binding design choice is *who attacks* more than how. The attacker and reviser need not be the same model, and this is the central axis of the paper: §5 shows that a strong critic from a *different model family* is what makes EoT effective, because its errors are decorrelated from the reasoner's. We therefore treat the *critic's identity* as the primary knob: the secondary axes we ablate (Appendix C.12), attack *granularity* (full chain vs. per-step) and *strength* (lenient vs. strict), do not rival it.

4 EXPERIMENTAL SETUP

Benchmarks. We evaluate on eight benchmarks spanning four reasoning types, chosen to test both the difficulty axis our thesis predicts and whether the method generalizes beyond math. *Competition math*: grade-school arithmetic (**GSM8K**; Cobbe et al., 2021, near-saturated, limited headroom), competition mathematics (**MATH-500**; Hendrycks et al., 2021), and olympiad math (**AIME 2024** and **2025**; Mathematical Association of America, 2025), the close-but-flawed regime where long derivations make cascading errors dominant. *Graduate science*: **GPQA** (Rein et al., 2024), multi-hop reasoning under uncertainty. *Code*: **HumanEval** (Chen et al., 2021), graded by executing held-out unit tests, where verification is the canonical task and correctness is objective instead of string-matched. *Multi-hop factual QA*: **MuSiQue** (Trivedi et al., 2022) and the commonsense MCQ *control* **ARC-Challenge** (Clark et al., 2018), which stress *inductive* reasoning (inference over retrieved facts, as opposed to deductive derivation), where self-critique most catastrophically fails.

Models. We choose models so that each cuts a specific question beyond merely adding scale. As *reasoners*: open workhorses (Qwen3-235B, DeepSeek V3.2) and a *size ladder* (Qwen3-32B $\rightarrow$ Qwen3-Next-80B $\rightarrow$ Qwen3-235B) that traces the capability threshold, plus a reasoning model that already deliberates internally (Kimi K2 Thinking) to test whether external critique is redundant. As *critics*: same-model (self-play), a same-family stronger model (Qwen3-235B), and cross-family models of increasing strength (DeepSeek V3.2, Claude Sonnet 4.6, Claude Opus 4.8). A configured EoT run is specified by its *initializer* and *critic*, written **EoT (*initializer*, *critic*)**, e.g. EoT (CoT init., Opus critic), abbreviated EoT (CoT, Opus) in figures. This design makes three questions falsifiable:

1. Is the binding constraint the reasoner or the critic? (Fix the reasoner and vary only the critic.)
2. Is the gain driven by critic *strength* or by error *decorrelation*? (Compare a same-family strong critic against a different-family critic of equal strength.)
3. Does a cross-model critic help models that already reason at length (Kimi), or is its benefit redundant with internal deliberation?

Baselines and protocol. We compare against CoT, self-consistency, multi-round self-correction, multi-agent debate, best-of-N with a self-scoring verifier, and, as a strong constructive-search reference, ToT (Yao et al., 2023). Crucially, all methods are run at *matched token budgets*, so a comparison reflects how compute is *allocated*, holding the amount spent comparable (per-method budgets in Appendix B; *measured* tokens in §5). EoT caps iterations at $K{=}7$ with the strict attacker and revision history enabled, switching to the verification-style attacker (the critic re-solves and objects only on disagreement) on the *inductive* benchmarks (MuSiQue, ARC-Challenge) for optimal performance (§5, Appendix C.4). Accuracy is graded by `math_verify` (Kydlíček, 2025) for math, unit-test execution for code, and exact match for MCQ. Decoding settings, the full ablation menu, and per-experiment configurations are in Appendix B.

5 RESULTS

Main results. Table 2 reports the full method suite across all eight benchmarks, holding the reasoner fixed at Qwen3-235B. Our best configuration, EoT with a ToT initializer and a cross-model Opus critic, is the single strongest method on every *unsaturated* benchmark: 39.4 on GPQA, 93.3 on AIME-2024, and tied-best on AIME-2025 (80.0). Its margin over multi-agent debate is statistically significant where we have the power to test it: on GPQA ($+6.1$, $p{=}0.015$) and on the pooled two-year AIME set ($+11.7$, $p{=}0.007$). The single-year AIME-2024 gap ($+10$) is directionally identical but underpowered at $n{=}30$, so we read AIME pooled (Appendix C.10). On the near-saturated benchmarks (GSM8K, MATH-500, ARC) all strong methods cluster within noise, as expected where there is no headroom. Two comparisons inside the table preview the mechanism: EoT matches or improves over its *own* initializer on the hard benchmarks (GPQA 39.4 vs. 35.4, AIME-2024 93.3 vs. 86.7), and the same EoT on a weaker CoT initializer trails debate on AIME, so the initializer set that ceiling. The result holds on a second reasoner (DeepSeek V3.2; Appendix C.9).

The bottleneck is the critic. Fixing the reasoner at the weak Qwen3-32B and varying only the adversary (Figure 6, Appendix C.8), self-attack is worthless (AIME 30.0$\rightarrow$30.0): a model cannot

Table 2: **Main results** across all four domains (reasoner fixed at Qwen3-235B; per-cell accuracy, column-best in **bold**; HEval=HumanEval, ARC-C=ARC-Challenge). EoT (ToT init., Opus critic) leads every unsaturated deductive benchmark and ties the field on saturated GSM8K and MATH. The cross-model critic is essential, as the same loop with self-critique trails debate on AIME. †On the *inductive* benchmarks (MuSiQue, ARC-Challenge), the strict attacker already clears CoT (MuSiQue 72.7/71.3, ARC 96.7/95.7 for the CoT/ToT rows), and the verification-style attacker (the critic resolves and objects only on disagreement; Appendix C.4) further *optimizes* these, suppressing spurious provenance objections to reach the values shown (74.7/72.7, 98.0/98.3).

Method	GSM8K	MATH	GPQA	AIME-24	AIME-25	HEval	MuSiQue	ARC-C
CoT	95.7	91.8	28.8	33.3	30.0	93.9	70.7	97.0
Self-Cons. (8)	97.3	91.6	30.3	30.0	33.3	94.5	70.7	97.3
Self-Cons. (16)	97.3	92.6	28.8	33.3	33.3	94.5	72.3	97.3
Self-Corr.	96.0	95.2	31.3	46.7	53.3	95.7	70.7	97.0
Self-Corr. (multi)	94.0	97.0	33.3	73.3	60.0	95.1	67.7	97.0
Best-of-N (N=8)	97.0	91.4	28.3	33.3	26.7	93.9	66.0	97.3
Debate	95.3	**97.6**	33.3	83.3	66.7	95.1	68.0	96.3
ToT	**98.0**	97.0	35.4	86.7	**80.0**	93.9	66.7	97.3
EoT (CoT, self)	94.0	94.8	25.8	56.7	63.3	93.9	60.0	95.3
EoT (CoT, Opus)	97.0	94.0	37.9	73.3	63.3	**97.6**	**74.7**†	98.0†
EoT (ToT, Opus)	97.5	96.5	**39.4**	**93.3**	**80.0**	94.5	72.7†	**98.3**†

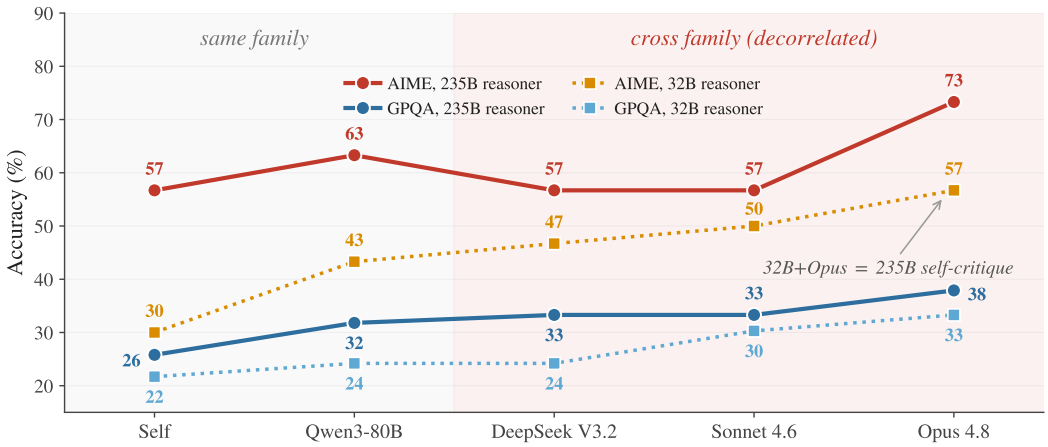

Figure 2: Holding the reasoner fixed (Qwen3-235B) and varying only the critic. Within-family critics (grey region) barely move accuracy, whereas cross-family critics (red region) lift it, with the strongest decorrelated critic (Opus) best on both benchmarks. The lift is not monotonic in raw critic strength (a critic must expose *this* reasoner's errors, beyond merely being strong), which is why decorrelation, more than strength alone, is the lever.

find the flaws it just made. A stronger *same-family* attacker (Qwen3-235B) helps only marginally (+0 on AIME, +3.1 on GPQA), consistent with *correlated* blind spots, whereas a strong *cross-family* attacker (Sonnet 4.6) rescues the same weak reasoner by +20 on AIME and +8.6 on GPQA. The reasoner can execute correct revisions once handed good objections, so generating them is the binding constraint: the generation–verification asymmetry made operational.

Decorrelation is the lever. The critic effect is *universal* and extends well beyond a weak-model crutch (Figure 2): even the capable Qwen3-235B reasoner improves by +16.7 on AIME and +12.1 on GPQA with an Opus critic, and stronger cross-family critics order the gains. Two consequences stand out. *(i) Substitution.* A weak reasoner with a strong critic matches a strong reasoner critiquing itself (AIME 56.7 either way), so critic capability trades against reasoner capability, and scaling either is a substitutable route past the threshold. *(ii) Strength is not the whole story.* Pairing Qwen3-235B with DeepSeek V3.2, a different family of merely comparable strength (GPQA 30.3 vs. 28.8),

lifts GPQA to 33.3 (+7.5 over self-attack), much of the way to the far stronger Opus critic (+12.1). Decorrelated errors drive a large part of the gain beyond what raw strength alone explains: the same mechanism that makes debate work, obtained from a single cross-family critic. *(iii) Not redundant with internal deliberation.* The benefit survives even for a reasoner that already "thinks" at length: Kimi K2-Thinking gains +9 on GPQA (33.3→42.3) from an Opus cross-critic, so external decorrelated critique adds something a model's own extended reasoning does not. (A full seven-critic sweep confirming Opus is best a priori on every benchmark appears in Appendix C.3.)

Composing on a stronger initializer beats debate. Since the residual gap to debate on AIME is set by the initializer, we swap the CoT initializer for ToT (Yao et al., 2023) under the *same* EoT loop and Opus critic (Table 2, last two rows). The lift is large and entirely attributable to the initializer: it raises AIME-2024 from 73.3 to 93.3, past debate's 83.3, and lifts GPQA and AIME-2025 above debate as well, a margin significant over debate on GPQA and pooled AIME (Appendix C.10). The gain requires the *cross-model* critic: EoT on a ToT initializer with *self*-critique *degrades* ToT (GPQA 35.4→27.8; Table 7), as the decorrelation thesis predicts, since a constructive search's errors are systematic. Nor is the gain single-reasoner: the recipe holds on DeepSeek V3.2 and on a third initializer, GoT (Appendix C.9).

One recipe across four domains. If cross-model verification were a math-specific trick, the gains would not transfer. Instead they hold across all four domains (Table 2, right columns). On execution-graded code (HumanEval), an Opus critic lifts Qwen3-235B from 93.9 to 97.6. The sharper test is the *inductive* benchmarks (MuSiQue, ARC-Challenge), where answers are inferred from retrieved facts and no deductive derivation exists, and where the critic's *attack mode* matters as much as its identity. The strict attacker, built for deductive proofs, objects that a factually-correct answer is not *provable* from the context, a soundness violation (§3.2): it sinks *self*-play EoT (MuSiQue 70.7→60.0) and even a cross-model critic only partly escapes it (Qwen MuSiQue 72.7; on DeepSeek it falls *below* CoT). Switching to a *verification* mode, where the critic independently re-solves and objects only on disagreement so that shared correct knowledge stands, resolves this directly, lifting MuSiQue to 74.7 (Qwen)/77.0 (DeepSeek) and ARC-Challenge to 98.3/98.0, matching or beating every baseline on both reasoners. This is a soundness correction to the attack prompt as opposed to a domain trick: a prompt that merely forbids objecting to retrieved facts only partially rescues the self-critic, whereas re-solving fixes the failure across reasoners and initializers (Appendix C.4).

Cross-model EoT is compute-efficient. EoT's advantage is not bought with extra compute, since the *initializer* sets the cost. With a CoT initializer, cross-model EoT is among the *cheapest* iterative/multi-sample methods on AIME (Qwen3-235B) at 43K tokens per problem, tying multi-round self-correction (43K) and undercutting debate (60K), self-consistency (58K), and best-of-N (59K). Composed on GoT it reaches 90.0 at debate's cost, and only the ToT initializer, whose tree search dominates the budget, is expensive, the price of its peak accuracy (Appendix C.5, Figure 4). Because EoT refines a *single* trajectory, without sampling or debating many, its cost scales with problem difficulty instead of a fixed sample budget. Notably the cross-model critic is cheaper than self-play (43K vs. 49K) despite higher accuracy: better objections converge in fewer rounds.

The three predicted failure modes, observed. The theory of §3.2 predicts that violating soundness, completeness, or contractivity yields a distinct, observable pathology. We find all three, and the cross-model critic suppresses the first two. *(i) Soundness violations (divergence).* A spurious objection drives a revision away from a correct answer, the "overthinking" regime. Under self-play this dominates on inductive QA: on MuSiQue the same-family critic broke 38 initially-correct answers while fixing only 6 (net −32), objecting to valid retrieved facts. A cross-model critic inverts the ledger (18 fixed, 12 broken, net +6), and switching to the verification-style attacker, which objects only when its own independent solve disagrees, nearly eliminates the spurious breaks (13 fixed, 1 broken; Appendix C.4), driving the soundness defect ϵ toward zero. *(ii) Completeness violations (premature convergence).* The attacker concedes on a wrong answer: the dominant error on competition math, what self-critique cannot escape and a stronger critic removes. *(iii) Contractivity violations (oscillation).* A revision trades one flaw for another and the answer cycles, which we detect and resolve by majority vote. Oscillation concentrates on GPQA, where objections are perpetually regenerable, and is the residual failure a single critic does not fully remove.

Convergence speed is a free difficulty signal. Because easy problems concede on the first attack and hard ones iterate, rounds-to-equilibrium is an inference-time difficulty estimate available *before*

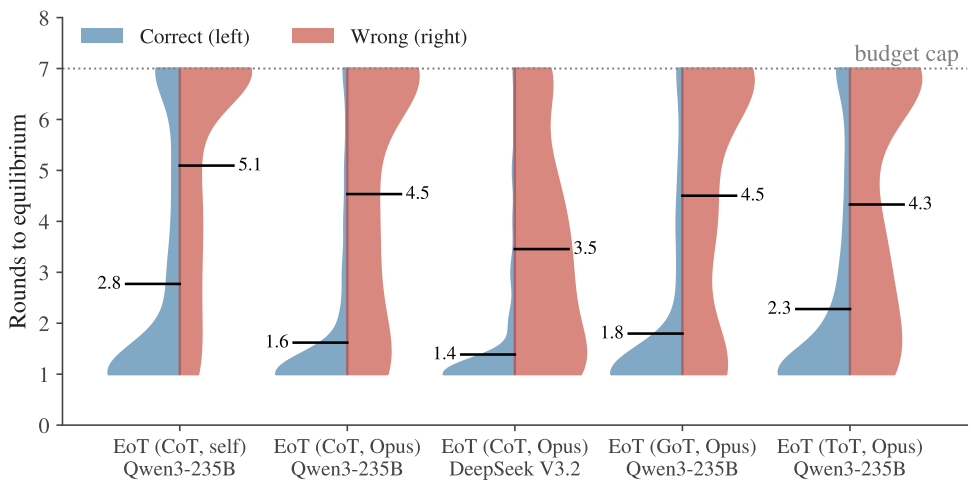

Figure 3: Rounds-to-equilibrium split by correctness across five EoT configurations (split violins: correct in blue on the left, wrong in red on the right; black bars are means). The separation is direction-consistent for every initializer, critic, and reasoner: correct answers concede fast (means 1.4–2.8, mode at 1) while wrong answers iterate far longer and pile up at the budget cap (means 3.5–5.1). Fixed-budget baselines cannot produce such a distribution (Appendix C.6).

the gold answer is known. Splitting the round distribution by final correctness (Figure 3) shows the signal holds across initializers (CoT, GoT, ToT), critics (self, Opus), and reasoners: correct answers concede fast (per-configuration means 1.4–2.8, 55–84% on the first attack) while wrong answers iterate roughly twice as long and pile up at the budget cap. Pooling all 4,800 traces, accuracy is 93.4% among problems that concede by round 1 versus 51.1% among those that exhaust the budget: a 42-point spread from round count alone (AUC 0.78; Appendix C.6). Only EoT produces such a distribution, since fixed-budget baselines spend the same compute on every problem, so it yields a free triage rule: route problems that fail to concede within k rounds to a stronger model or human.

6 Conclusion

We introduce Equilibrium of Thoughts, which reframes LLM reasoning from construction to stabilization: the output is the answer that *cannot be refuted*, the reasoning equilibrium. The method is simple (one loop, two prompts) and theoretically grounded (fixed-point, argumentation, and game theory), but its empirical lesson is sharper than "iterate and probe." A model is a poor critic of itself. The gains come from a *decorrelated* critic, and with a strong cross-family one a single refined trajectory matches or beats multi-agent debate at lower cost across math, science, code, and multi-hop QA, the only adjustment being that on *inductive* retrieved-fact tasks the critic re-solves and objects only on disagreement (in place of attacking the derivation), so shared correct knowledge is not argued away. Recast as cross-model adversarial verification, EoT subsumes CoT, self-correction, and debate as special cases and turns convergence speed into a free difficulty signal. The actionable insight: instead of asking a model "is my answer good?", ask a *different* model to break it.

Limitations. EoT's ceiling is set by the critic: where even a frontier cross-family critic finds no genuine flaw, premature convergence persists, and where objections are perpetually regenerable, oscillation does. It needs a second, family-decorrelated model, which a single-model deployment cannot provide, though the effect is not contingent on closed models: an *open* cross-family critic (DeepSeek V3.2) already captures much of the gain (+7.5 on GPQA over self-attack, §5), so it reproduces with open weights. Its inference cost is 2–13× a single CoT call, still below sampling and debate at matched accuracy. Finally, on problems whose entire initial *approach* is wrong, so that no local repair suffices, refining a single trajectory is weaker than exploring alternatives, a gap we leave to future work combining cross-model critique with exploration. A broader-impact note on EoT as training-free red-teaming is in Appendix E.

AI USE STATEMENT

In this work, we used generative AI tools for designing or providing feedback on research methodology or experiments, implementing methods, and interpreting results. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, or proposing or refining hypotheses. Assisting with translation, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for editing the paper to improve readability and for identifying relevant literature. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness, AI-suggested interpretations were checked against outputs, and AI-identified references were verified against their source records. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

EoT is implemented entirely through prompting: no fine-tuning, custom architectures, or additional training is required. The complete loop is specified as pseudocode (Algorithm 1) with the two verbatim prompt templates (attack and revise), and the oscillation-detection and majority-resolution rules are given formally in Appendix A. The prompt variants for every baseline and ablation (lenient/strict/factual/verification-style attackers, the domain prompt of Appendix C.4) are documented alongside, including the verbatim `verify` template (Appendix A) used for the inductive benchmarks. All experiments use publicly available benchmarks (GSM8K, MATH-500, AIME 2024/2025, GPQA, HumanEval, MuSiQue, and ARC-Challenge) with their splits and per-experiment problem counts in §4 and Appendix B, and grading is performed by `math_verify` (Kydlíček, 2025) for math, unit-test execution for code, and exact match for MCQ as described in §4. Unless otherwise noted, all runs cap iterations at $K{=}7$ with the strict attacker (the verification-style attacker on the inductive benchmarks MuSiQue and ARC-Challenge; §5) and revision history enabled, use greedy decoding (temperature 0) for deterministic components and temperature 0.7 for the sampling baselines (self-consistency, best-of-N, debate), and match token budgets across methods per Appendix B. The equilibrium-uniqueness study additionally samples at temperature 0.9 across fixed seeds (§4). We report the exact model versions used (Qwen3-32B, Qwen3-Next-80B, Qwen3-235B, DeepSeek V3.2, Kimi K2-Thinking, Claude Haiku 4.5, Claude Sonnet 4.6, Claude Opus 4.8), all accessed through a hosted inference API. As with any hosted model, exact outputs may shift across provider-side model updates. We therefore document all prompts, configurations, decoding settings, and statistical procedures (the bootstrap protocol of Appendix C.10) needed to reproduce our protocol independent of a specific endpoint, and note that the central decorrelation effect reproduces with an *open* cross-family critic (DeepSeek V3.2), so it does not depend on access to any proprietary model.

REFERENCES

Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3:133–181, 1922.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Maciej Besta, Florim Memedi, Zhenyu Zhang, Robert Gerstenberger, Guangyuan Piao, Nils Blach, Piotr Nyczyk, Marcin Copik, Grzegorz Kwaśniewski, Jürgen Müller, Lukas Gianinazzi, Ales Kubicek, Hubert Niewiadomski, Aidan O'Mahony, Onur Mutlu, and Torsten Hoefler. Demystifying chains, trees, and graphs of thoughts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):10967–10989, 2025.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan,

Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, 2024.

Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.

Felix Fricke, Simon Malberg, and Georg Groh. Framework of thoughts: A foundation framework for dynamic and optimized reasoning based on chains, trees, and graphs. *arXiv preprint arXiv:2602.16512*, 2026.

Peizhong Gao, Ao Xie, Shaoguang Mao, Wenshan Wu, Yan Xia, Haipeng Mi, and Furu Wei. Meta reasoning for large language models. *arXiv preprint arXiv:2406.11698*, 2024.

Qianyue Hao, Sibo Li, Jian Yuan, and Yong Li. RL of thoughts: Navigating LLM reasoning with inference-time reinforcement learning. *arXiv preprint arXiv:2505.14140*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*, 2021.

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. In *International Conference on Learning Representations*, 2024.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

Hynek Kydlíček. Math-verify: Math verification library. `https://github.com/huggingface/math-verify`, 2025. Open-source LaTeX/expression equivalence checker used by lighteval.

Imre Lakatos. *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge University Press, 1976.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024.

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

Mathematical Association of America. American invitational mathematics examination (AIME) 2024–2025, 2025.

M. Moran and Mark Whiting. LLMs show no signs of individuated metacognition. *arXiv preprint arXiv:2605.24299*, 2026.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

Tushar Pandey, Ara Ghukasyan, Oktay Goktas, and Santosh Kumar Radha. Adaptive graph of thoughts: Test-time adaptive reasoning unifying chain, tree, and graph structures. *arXiv preprint arXiv:2502.05078*, 2025.

Karl Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, 1963.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-Proof q&a benchmark. In *Conference on Language Modeling*, 2024.

Bilgehan Sel, Ahmad Al-Tawaha, Vanshaj Khattar, Ruoxi Jia, and Ming Jin. Algorithm of thoughts: Enhancing exploration of ideas in large language models. In *International Conference on Machine Learning*, 2024.

Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, Zhaoxiang Liu, and Shiguo Lian. DAST: Difficulty-adaptive slow-thinking for large reasoning models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Shashwat Singh, Tal Linzen, and Shauli Ravfogel. Can LLMs introspect? a reality check. *arXiv preprint arXiv:2605.26242*, 2026.

Rohit Sinha, Kunal Tilaganji, Tanuja Ganu, Nagarajan Natarajan, Amit Sharma, and Vineeth N. Balasubramanian. A nash equilibrium framework for training-free multimodal step verification. *arXiv preprint arXiv:2605.20033*, 2026.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations*, 2025.

Yuan Sui, Yufei He, Tri Cao, Simeng Han, Yulin Chen, and Bryan Hooi. Meta-reasoner: Dynamic guidance for optimized inference-time reasoning in large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2026.

Alfred Tarski. A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics*, 5(2):285–309, 1955.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Jason Z Wang. MIRROR: A hierarchical benchmark for metacognitive calibration in large language models. *arXiv preprint arXiv:2604.19809*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, Louis Castricato, Jan-Philipp Franken, Nick Haber, and Chelsea Finn. Towards system 2 reasoning in LLMs: Learning how to think with meta chain-of-thought. *arXiv preprint arXiv:2501.04682*, 2025.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E. Gonzalez, and Bin Cui. Buffer of thoughts: Thought-augmented reasoning with large language models. In *Advances in Neural Information Processing Systems*, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *International Conference on Machine Learning*, 2024.

Shu Zhou, Rui Ling, Junan Chen, Xin Wang, Tao Fan, and Hao Wang. When more thinking hurts: Overthinking in llm test-time compute scaling. *arXiv preprint arXiv:2604.10739*, 2026.

A Algorithm and Prompt Templates

Algorithm 1 Equilibrium of Thoughts

Require: Problem P, max iterations K, model $\mathcal{M}$
Ensure: Answer a^*, convergence status, attack trace

$a_0 \leftarrow \mathcal{M}.\text{generate}(P)$ {Initial answer with reasoning}
$H \leftarrow \emptyset$ {Objection history}
for $i = 0, 1, \ldots, K-1$ **do**
  $o_i \leftarrow \mathcal{M}.\text{attack}(P, a_i)$ {Find strongest objection}
  if $o_i = \perp$ **then**
    return a_i, EQUILIBRIUM, H {No valid objection: equilibrium reached}
  end if
  if detect_oscillation$(a_0, \ldots, a_i, o_i)$ **then**
    return a_i, OSCILLATION, H {Cycle detected}
  end if
  $a_{i+1} \leftarrow \mathcal{M}.\text{revise}(P, a_i, o_i, H)$ {Revise to address objection}
  $H \leftarrow H \cup \{o_i\}$ {Track history to prevent regression}
end for
return a_K, BUDGET, H {Max iterations reached}

EoT uses two prompts, applied to any initializer's output. The attacker emits the single strongest objection or concedes, and the reviser addresses it while avoiding previously-fixed flaws.

Attack Template

```
You are a rigorous critic. Find the single strongest objection to the
following answer.
Problem: {P} Current answer and reasoning: {a_i}
Rules: - Identify a SPECIFIC flaw (not vague doubt) - The flaw must
be LOGICAL (not stylistic) - If no genuine flaw exists, respond: "NO
VALID OBJECTION" - State your confidence: [HIGH / MEDIUM / LOW]
Your strongest objection:
```

Revision Template

```
Your previous answer was challenged with a valid objection.
Problem: {P} Your answer: {a_i} Objection: {o_i} Previously fixed
issues (do NOT regress): {H}
Produce a revised answer that: 1. Directly addresses this specific
objection 2. Does not reintroduce any previously-fixed issues 3.
Maintains whatever was correct in the previous answer
Revised answer and reasoning:
```

Verification-style attacker (inductive benchmarks). On the inductive benchmarks (MuSiQue, ARC-Challenge) the default strict attacker misfires by objecting that a factually-correct answer is not *provable* from the retrieved context, a spurious objection that breaks correct answers (Appendix C.4). There we replace the attack prompt with the `verify` template below: the critic independently re-solves the problem and raises an objection only if its own answer *disagrees*. The revision template and the rest of the loop are unchanged. This keeps the attack adversarial (the objection is a concrete counter-answer) and cross-model, while letting the reasoner and critic's shared parametric knowledge stand when they agree.

Verify-Attack Template (inductive benchmarks)

```
You are a rigorous solver acting as an independent checker. Do NOT
trust the proposed answer's reasoning. Solve the problem yourself,
from scratch, then compare.
Problem: {P} Proposed answer under review: {a_i}
Steps: 1. Solve the problem yourself, independently, showing your
own work. 2. Compare your result to the proposed answer. 3. If
your independent answer DIFFERS, that discrepancy is the objection:
state your answer and the specific step where the proposed reasoning
goes wrong. 4. If your independent answer AGREES, respond with
exactly: "NO VALID OBJECTION"
OBJECTION: [your independent answer and where the proposed reasoning
errs] FLAW TYPE: [arithmetic_error | logical_leap | missing_case |
wrong_assumption | factual_error | invalid_step] CONFIDENCE: [HIGH /
MEDIUM / LOW]
```

Oscillation detection (formal). We flag oscillation when answer(a_i) = answer(a_j) for some $j < i - 1$ while answer$(a_i) \neq$ answer(a_{i-1}), i.e. the extracted answer has returned to an earlier state after changing. On detection we stop and return the majority answer over the visited trajectory.

B Experimental Setup Details

Compute-matched budgets. Every method is run at a comparable token budget so that comparisons isolate *allocation* from *spend* (Table 3). The critical test is EoT against self-consistency at the same nominal budget: same tokens, different strategy. These are target budgets, and the tokens *measured* at run time are reported in §5 (Table 10).

Ablations. We sweep (1) iteration count (1, 2, 3, 5, 7, 10 rounds, locating where improvement plateaus; Appendix C.7); (2) attack strength (lenient "find any issue", strict "find the most devastat-

Table 3: Target compute budgets per method. The critical Pareto test: EoT at ∼4K tokens vs. Self-Consistency at ∼4K tokens, i.e. the same budget under a different allocation strategy.

Method	Token Budget	What it Buys
CoT (greedy)	∼500	1 answer
Self-Consistency (k=8)	∼4,000	8 answers, majority vote
Self-Consistency (k=16)	∼8,000	16 answers, majority vote
Self-Correction (k rounds)	∼3,000	1 answer, k "check your work" passes
Multi-Agent Debate (3 rds)	∼6,000	2 agents, 3 exchanges, judge
Best-of-N + verifier	∼4,000	N answers, pick highest-scored
LATS (Zhou et al., 2024)	∼8,000	Tree search with reflection
EoT (ours, 3 rds avg)	∼3,000	1 answer, iteratively refined
EoT (ours, 7 rds max)	∼6,000	1 answer, heavily refined

ing objection", and an adaptive setting that escalates lenient→strict when objections stop landing; Appendix C.12); (3) granularity (answer-level vs. step-level; Appendix C.12); (4) revision history (prior objections shown to the reviser or not; Appendix C.12); (5) cross-model attacker (same vs. different model; §5, Appendices C.3 and C.8); (6) initialization sensitivity (whether the same problem run from k initial answers reaches the same equilibrium; Appendix C.15); and (7) EoT on a reasoning model that already deliberates internally (Appendix C.11).

Metrics. Beyond *accuracy* (exact match for math, accuracy for MCQ, unit-test execution for code), we report *convergence rounds* (mean iterations to equilibrium; Appendix C.6), *accuracy @ round k* (Appendix C.7), *tokens per problem* and the resulting accuracy-per-token Pareto frontier (Appendix C.5), *attack success rate* (fraction of attacks identifying a genuine flaw), *convergence rate* (fraction reaching equilibrium within budget), *oscillation rate* (fraction exhibiting cyclic behavior; the latter three diagnose the failure modes of §5), and *equilibrium consistency* (same answer from different initializations; Appendix C.15).

C Additional Results

This appendix collects results that support the claims of the main text but are not on its critical path: self-play headroom, the capability-threshold scaling ladder, the full critic sweep, the no-tricks ablation, the GoT composition point, and per-method token accounting.

C.1 Self-play EoT: gains track headroom

Before introducing a cross-model critic, we characterize *self-play* EoT (a model attacking its own answer) against its own greedy CoT (Table 4). The effect is governed by *headroom*: on saturated benchmarks (GSM8K at 95–98%, AIME for Kimi at 100%) EoT cannot help and a mild overthinking penalty appears, whereas on unsaturated AIME it adds $+13$ to $+23$ points. Adversarial refinement amplifies a model in proportion to how far its first answer sits from correct, and, as §5 shows, self-play leaves a large gap to multi-agent baselines on hard problems that only a cross-model critic closes.

Table 4: Self-play EoT accuracy gain over CoT (percentage points). Gains concentrate on the unsaturated hard benchmark (AIME), while saturated benchmarks show none. Kimi's AIME is already 100%, leaving no headroom.

Reasoner	GSM8K	MATH-500	GPQA	AIME
Qwen3-235B	-1.7	$+3.2$	-3.0	$\mathbf{+23.3}$
DeepSeek V3.2	-1.0	$+0.6$	-4.0	$\mathbf{+20.0}$
Kimi K2-Think	-2.9	$+0.1$	$+2.7$	$+0.0$
Claude Sonnet 4.6	$+0.7$	$+1.2$	$+2.0$	$\mathbf{+13.3}$

C.2 SCALING: A CAPABILITY THRESHOLD PAST WHICH SELF-CRITIQUE WORKS

Tracing self-play EoT up a size ladder (Qwen3-32B → Qwen3-Next-80B → Qwen3-235B) reveals a *capability threshold* (Table 5). On AIME the self-play gain over CoT is $+0.0$, $+16.7$, $+23.3$ respectively: the 32B model is below a self-critique floor and gains nothing, while larger models, able to find some of their own flaws, gain steeply. On saturated MATH the methods are indistinguishable. The threshold is a property of the *critic* more than the reasoner: the cross-model results of §5 show the same 32B reasoner, given a decorrelated critic, jumps from $+0$ to $+20$ on AIME (the same 32B critic ladder spans both AIME and GPQA in Figure 2). Scaling the reasoner and supplying a stronger critic are substitutable routes past the threshold.

Table 5: Full scaling ladder (self-play EoT vs. CoT and self-consistency). The self-play EoT gain on AIME emerges only past a capability threshold ($+0$ at 32B, $+16.7$ at 80B, $+23.3$ at 235B), and a decorrelated critic removes the threshold (§5, Figure 2).

Reasoner	Method	AIME-24	GPQA	MATH-500
Qwen3-32B	CoT	30.0	22.7	87.0
	Self-Consistency	33.3	19.2	91.0
	EoT (self)	30.0	21.7	85.0
Qwen3-Next-80B	CoT	46.7	25.3	91.0
	Self-Consistency	46.7	26.3	93.0
	EoT (self)	63.3	24.2	91.5
Qwen3-235B	CoT	33.3	28.8	92.5
	Self-Consistency	36.7	28.8	92.5
	EoT (self)	56.7	25.8	96.0

C.3 FULL CRITIC SWEEP

The decorrelation account predicts *a priori* that the strongest, most family-decorrelated critic is best. Sweeping seven critics for a fixed Qwen3-235B reasoner (Table 6; reported in full without test-set selection), Claude Opus 4.8 is the best critic on *every* unsaturated benchmark (GPQA and *both* AIME years; on saturated MATH all critics tie within noise), and the across-year consistency is what rules out test-set fitting. Two caveats keep the story honest: critic quality is *not* monotonic in raw strength (Kimi K2-Thinking and Sonnet 4.6 underperform DeepSeek V3.2 and even the same-family Qwen3-Next-80B on AIME, so a critic must actually expose *this* reasoner's errors beyond merely being strong), and even the best critic only *narrows* the AIME gap to debate at the CoT initializer; the composition of §5 is what closes it.

Table 6: Critic sweep (fixed reasoner). Opus is best on every *unsaturated* benchmark (GPQA and both AIME years), predicted a priori by the decorrelation account (on saturated MATH all critics cluster at 95.5–98.0, within noise). Critic quality is not monotonic in strength (Kimi K2-Thinking, Sonnet 4.6 underperform), so the critic must genuinely expose the reasoner's errors beyond merely being strong.

Critic (reasoner = Qwen3-235B)	AIME-24	AIME-25	GPQA	MATH
Self (Qwen3-235B)	56.7	46.7	25.8	97.5
Qwen3-Next-80B (same fam.)	63.3	53.3	31.8	96.0
DeepSeek V3.2	56.7	53.3	33.3	97.0
Kimi K2-Thinking	46.7	36.7	32.3	95.5
Claude Haiku 4.5	50.0	46.7	30.3	96.0
Claude Sonnet 4.6	56.7	40.0	33.3	**98.0**
Claude Opus 4.8	**73.3**	**63.3**	**37.9**	96.5

The same a-priori ordering holds when the initializer is swapped from CoT to ToT (Table 7): Opus is again the best critic on both hard benchmarks, the same-family Qwen3-Next-80B does not beat self-critique on GPQA, and the cross-family ordering tracks decorrelation more closely than raw strength. The critic ranking is thus a property of the critic–reasoner pair, independent of the particular initializer it wraps.

Table 7: Critic sweep on a *ToT* initializer (Qwen3-235B). The critic ranking mirrors the CoT-init sweep (Table 6), with Opus best on both benchmarks and the same-family critic no better than self-critique, confirming that the best critic is determined a priori by the critic–reasoner pairing, independent of the initializer.

Critic (ToT init., reasoner = Qwen3-235B)	**AIME-24**	**GPQA**
Self (Qwen3-235B)	80.0	27.8
Qwen3-Next-80B (same fam.)	80.0	26.8
DeepSeek V3.2	70.0	28.8
Claude Sonnet 4.6	86.7	34.3
Claude Opus 4.8	**93.3**	**39.4**

C.4 Inductive benchmarks: domain prompt vs. the verification-style critic

The inductive benchmarks (MuSiQue, ARC-Challenge) are the regime where the strict attacker, designed for deductive proofs, misfires: it raises *provenance* objections ("the context does not state that X is Y's wife") against answers that are factually correct from the reasoner's parametric knowledge, and the reasoner complies by retracting to "cannot be determined", a *soundness* failure (the spurious-objection rate ϵ of Appendix D) that breaks correct answers. This subsection isolates the cause and gives the fix used for the inductive columns of Tables 2 and 18.

The domain prompt is neither necessary nor sufficient. We first confirm that the cross-model critic, and not a domain-specific prompt, is what rescues multi-hop QA, using a 2×2 ablation crossing the critic (self vs. Opus, the two rows of Table 8) with the attack prompt (the default *strong* attacker vs. a *factual* prompt that explicitly forbids objecting to retrieved facts, the two columns), on Qwen3-235B. The factual prompt partially helps the weak self-critic ($60.0\rightarrow65.0$, the self-critic row's *strong* $\rightarrow$ *factual* columns) but still trails CoT (70.7), whereas the strong cross-model critic already reaches 72.7 *without* the factual prompt (its *strong*-column entry), and adding it barely moves the number (74.2). The domain prompt is neither necessary nor sufficient, and the decorrelated critic alone already beats CoT.

Table 8: No-tricks ablation (2×2: critic $\times$ attack prompt; Qwen3-235B, MuSiQue). The *factual* column is the domain prompt that forbids objecting to retrieved facts, and the *strong* column is the default attacker. The factual prompt only partially rescues the self-critic ($60.0\rightarrow65.0$, still $<$ CoT 70.7), whereas the cross-model Opus critic already suffices at the strong prompt (72.7, no domain trick).

MuSiQue	**strong prompt**	**factual prompt**
Self-critic	60.0	65.0
Opus critic	72.7	74.2

The verification-style attacker fixes the soundness failure directly. The strong cross-model critic clears CoT on Qwen (72.7 vs. 70.7), but on DeepSeek V3.2, whose multi-hop answers lean more heavily on parametric knowledge, the same strict attacker *drops below* CoT (70.7 vs. 73.0; net -7 on the fixed/broke ledger), exactly the spurious-provenance failure above. Instead of patching this with a domain prompt (which only partly helps, and is a benchmark-specific trick), we address the soundness defect at its source with the *verification-style* attacker (Appendix A, `verify` template): the critic *independently re-solves* the question and objects *only if* its own answer disagrees, so when reasoner and critic agree on a fact, no objection is raised and the correct answer survives. This remains adversarial (the objection is a concrete counter-answer, as opposed to a metacognitive "is this good?" judgment) and remains cross-model; it simply stops the critic from demanding provenance the task never required. Table 9 shows it resolves the failure on *both* reasoners and *both* initializers: the broke-count collapses (MuSiQue: Qwen $38\rightarrow1$, DeepSeek $24\rightarrow4$), lifting every inductive cell to match or beat the strongest baseline, with MuSiQue at 74.7–77.0 (vs. self-consistency 77.3, a statistical tie at the top, and above CoT/debate/ToT) and ARC-Challenge at 97.7–98.3 (best or tied-best). We therefore use the verification-style attacker for the inductive columns of Tables 2 and 18 (marked †) and the strict attacker everywhere else. The two are the same loop differing only in the attack prompt.

Table 9: Strict vs. verification-style attacker on the inductive benchmarks (Opus cross-critic). The strict attacker, tuned for deductive proofs, raises spurious provenance objections that break correct retrieved-fact answers, dropping DeepSeek *below* its CoT baseline ($70.7 < 73.0$). The verify attacker (independently re-solve, object only on disagreement) suppresses these on both reasoners and both initializers, lifting every cell to match or beat the strongest baseline (self-consistency 77.3 on MuSiQue, debate 98.0 on ARC). These are the † cells in Tables 2 and 18.

		MuSiQue		ARC-Challenge	
Reasoner	**EoT init.**	**strict**	**verify**	**strict**	**verify**
Qwen3-235B	CoT	72.7	74.7	96.7	98.0
	ToT	71.3	72.7	95.7	**98.3**
DeepSeek V3.2	CoT	70.7	76.0	96.3	97.7
	ToT	71.7	77.0	95.3	98.0
best baseline (either reasoner)		self-cons. **77.3**		debate 98.0	

C.5 Compute accounting

Table 10 gives the full per-method token accounting behind §5 on AIME (Qwen3-235B), including the GoT- and ToT-initialized variants. With a CoT initializer, cross-model EoT is the cheapest iterative/multi-sample method, and the cross-model critic lifts accuracy from 56.7 to 73.3 at *lower* cost than self-play because better objections converge in fewer rounds. The GoT and ToT initializers buy higher accuracy (90.0 and 93.3) at their own search cost, with the initializer alone setting the budget.

Table 10: Compute vs. accuracy on AIME-2024 (Qwen3-235B). With a CoT initializer, cross-model EoT is the cheapest iterative/multi-sample method and the cross-model critic lifts it from 56.7 to 73.3 at *lower* cost than self-play, because better objections converge in fewer rounds. The initializer sets the cost: composed on GoT, EoT reaches 90.0 at debate's budget, while the ToT initializer's tree search dominates ($\sim$250K tokens), the price of its peak 93.3 (Figure 4).

Method (Qwen3-235B, AIME)	**Tokens/problem**	**Accuracy**
CoT	3.6K	33.3
Self-Consistency (16)	58K	36.7
Best-of-N (8)	59K	33.3
Multi-round Self-Corr.	43K	73.3
Multi-Agent Debate	60K	83.3
EoT (CoT init., self critic)	49K	56.7
EoT (CoT init., Opus critic)	**43K**	73.3
GoT	41K	73.3
EoT (GoT init., Opus critic)	61K	90.0
ToT	250K	86.7
EoT (ToT init., Opus critic)	251K	**93.3**

C.6 Convergence speed by correctness

Section 5 (Figure 3) reports the headline separation: accuracy 93.4% for round-≤ 1 concessions versus 51.1% at the budget cap, AUC 0.78 over all 4,800 traces. Here we record the supporting detail and explain why the effect is principled. The separation is direction-consistent in every one of the five configurations (three initializers CoT/GoT/ToT, two critics self/Opus, two reasoners Qwen3-235B/DeepSeek V3.2): per-configuration correct-answer means range 1.4–2.8 rounds (with 55–84% conceding on the first attack), wrong-answer means 3.5–5.1, the wrong distribution always shifted right and heavier at the cap.

Three points make this more than a curiosity. *(i) It is mechanistic.* The signal falls directly out of the equilibrium framing: an answer that is already correct presents no genuine flaw, so a sound attacker concedes immediately (round 1), whereas a wrong answer keeps generating real objections until

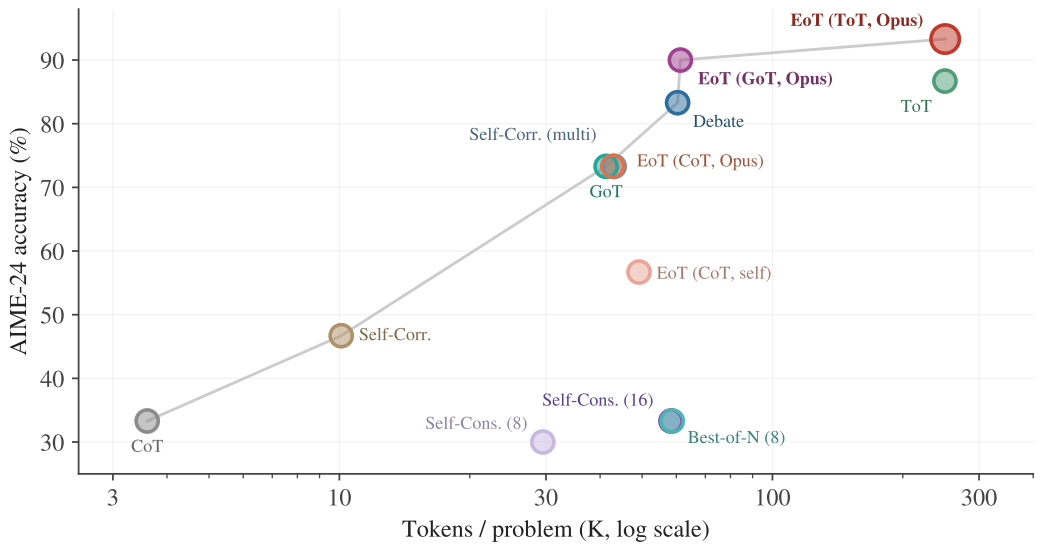

Figure 4: Compute vs. accuracy on AIME-2024 (Qwen3-235B; log-scale token axis, grey line traces the Pareto frontier). Cross-model EoT occupies the frontier at the high-accuracy end: composed on GoT it reaches 90.0 at debate's cost (61K vs. 60K), the best accuracy-per-token among all methods near that budget, and on a ToT initializer it attains the highest accuracy overall (93.3), though the tree search makes it the most expensive (∼250K tokens). With a CoT initializer, EoT ties GoT and multi-round self-correction (73.3 at ∼43K). The cheaper-than-debate regime is thus the CoT/GoT initializers, while the ToT initializer buys peak accuracy with compute.

either it is fixed or the budget is spent, exactly the contractivity dynamics of §3.2. The right tail at the budget cap is the empirical footprint of *premature convergence and oscillation* (the completeness and contractivity failures of Appendix D): hard wrong answers that the critic cannot conclusively break. *(ii) It is EoT-native.* Fixed-budget methods (CoT, debate, self-consistency) spend the same compute on every problem by construction and therefore cannot produce such a distribution, so the signal is unavailable to them without a separate difficulty estimator. *(iii) It is decorrelation-amplified.* The cross-model Opus critic sharpens the signal relative to self-play (correct-answer means drop from 2.77 (self) to 1.39–1.62 (Opus) and the first-round concede rate rises from 55% to 79–84%), because a decorrelated critic both fixes more wrong answers (shifting mass out of the right tail) and stops objecting once an answer is genuinely correct (concentrating the correct mass at round 1). The practical consequence is an abstention rule available at inference time for free: route problems that fail to concede within k rounds to a stronger model or to human review, capturing most errors at a controllable budget.

C.7 Accuracy by iteration round

Tracking accuracy as a function of the attack round k (Figure 5, Qwen3-235B self-critique) confirms Proposition 1's monotonicity prediction and locates the plateau. On MATH-500 accuracy climbs smoothly from 84.5 (round 0, the initial CoT answer) to 92.0 by round 4 and is flat thereafter, and on AIME-2024 the bulk of the gain (33.3 → 56.7) arrives by round 2. The plateau by round 4–5 is what justifies the $K{=}7$ iteration cap used throughout: the marginal round buys essentially nothing past that point, and on a self-critic the curve plateaus *below* what a cross-model critic reaches in fewer rounds (§5).

A natural question is whether the AIME curve's apparent late wiggle (50.0 → 60.0 over rounds 4–8) signals that the loop has not converged and a larger budget would help. It does not. The loop is self-terminating instead of budget-truncated: although the iteration cap here was set to $K{=}10$, *every* AIME problem halted on its own (by reaching equilibrium, by semantic convergence, or by oscillation detection) by round 8, and none was still running at the cap. The amber curve in Figure 5

(right axis) shows the fraction of problems still active collapsing from 100% to 17% by round 6 and 3% by round 8, so each accuracy datapoint past round 5 is moved by only one or two still-running problems (at n=30, a single problem is 3.3 points). The tail is therefore small-sample jitter in the frozen-answer average (a problem's answer is held fixed once its loop terminates) and does not reflect genuine ongoing improvement. The two problems responsible both terminate via oscillation detection and happen to resolve correctly. The smooth MATH-500 curve (n=200) confirms this reading, since with 6.7× the sample the plateau is clean. More attack rounds would change nothing, because by round 8 there is essentially nothing left to attack.

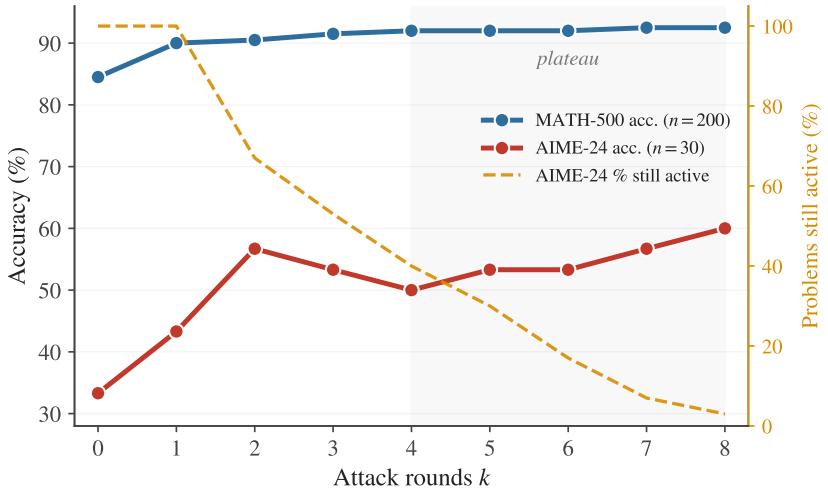

Figure 5: Accuracy vs. attack round k (Qwen3-235B, self-critique; round 0 is the initial answer, answers frozen after convergence). Accuracy is non-decreasing in expectation, as Proposition 1 predicts, and plateaus by round 4–5, justifying the K=7 cap. The bulk of the gain arrives in the first two rounds. The amber dashed line (right axis) is the fraction of AIME problems still active (loop not yet terminated): it falls to 3% by round 8, so the AIME tail past round 5 reflects 1–2 still-running problems at n=30, and all problems self-terminate by round 8, well under the cap.

C.8 Critic isolation on a fixed weak reasoner

Holding the reasoner fixed at the weak Qwen3-32B and varying only the adversary isolates the critic as the binding constraint (Table 11, Figure 6). Self-attack does nothing and a stronger *same-family* attacker barely helps, while a strong *cross-family* attacker rescues the same reasoner by +20 on AIME: decorrelation, more than raw strength, is the active ingredient.

Table 11: Holding the weak reasoner fixed and varying the critic. A strong, cross-family critic rescues a weak reasoner, whereas a same-family critic of equal-or-greater strength does not. The bottleneck lies in critique capability and *decorrelation*, whereas reasoning capability is secondary.

Attacker (reasoner = Qwen3-32B)	**AIME**	**MATH-500**	**GPQA**
None (CoT)	30.0	88.0	22.7
Self (Qwen3-32B)	30.0	86.5	21.7
Qwen3-235B (same family)	30.0	90.0	25.8
Claude Sonnet 4.6 (cross-family)	**50.0**	**94.0**	**30.3**

C.9 Composition: full table and second reasoner

Table 12 gives the full composition comparison behind the main-results table (Table 2), including standalone ToT and the CoT-initializer EoT variant, and a second reasoner. The same recipe holds on DeepSeek V3.2 (AIME-2024 90.0, AIME-2025 83.3, GPQA 36.9 with an Opus critic) and on

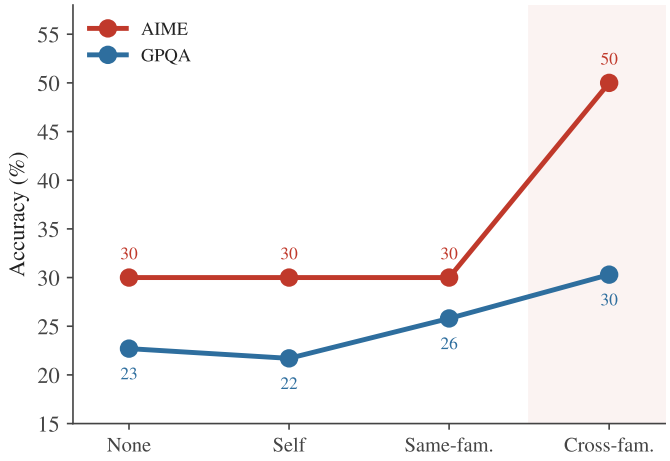

Figure 6: Critic isolation (reasoner fixed at Qwen3-32B). Accuracy is flat across no-attack, self-attack, and a stronger *same-family* attacker, and jumps only with a *cross-family* critic: decorrelation, more than raw strength, moves the needle.

a GoT initializer (Qwen3-235B: AIME-2024 90.0, GPQA 36.4), confirming the effect is neither single-reasoner nor specific to one initializer.

Table 12: EoT composes on a stronger initializer (Qwen3-235B, all EoT variants use the same Opus cross-critic). Replacing the CoT initializer with ToT lifts EoT past multi-agent debate on every hard benchmark, and EoT improves on its initializer instead of competing with it (EoT with a ToT initializer $>$ ToT alone). The two EoT columns share an identical loop and critic, and only the initializer differs.

Qwen3-235B	**Debate**	**ToT**	**EoT (CoT init.)**	**EoT (ToT init.)**
AIME-2024	83.3	86.7	73.3	**93.3**
AIME-2025	66.7	80.0	63.3	**80.0**
GPQA	33.3	35.4	37.9	**39.4**

C.10 Statistical significance

Because the AIME benchmarks have only 30 problems per year, single-year point gaps must be read with their uncertainty. We report 95% bootstrap confidence intervals (CIs) on per-cell accuracy (10K resamples, problems sampled with replacement) and, for the headline comparisons, paired bootstrap tests that resample the per-problem win/loss *difference* between two methods on the same problems, which is more powerful than comparing marginal CIs, since it controls for problem difficulty. All use the Qwen3-235B reasoner and the EoT (ToT init., Opus critic) configuration, and we report a one-sided p-value for the hypothesis that EoT exceeds the comparator.

Contamination. A concern for any benchmark evaluation is whether gains reflect training-data exposure as opposed to reasoning. Three features of our design argue against contamination driving the EoT effect. *(i) The effect is a within-reasoner delta.* Every headline number is EoT *minus* a baseline on the *same* reasoner and problems, so memorized problems inflate the baseline and the EoT run equally, and the *gap* we report is robust to leakage. *(ii) It holds on AIME-2025.* AIME-2025 was released after the training cutoff of most critics we use, yet the cross-model lift on AIME-2025 matches AIME-2024 (e.g. EoT (ToT, Opus) 80.0 on both; Table 2). If memorization drove the result, the post-cutoff year should behave differently. *(iii) The mechanism is benchmark-agnostic.* The decorrelation effect appears identically on GPQA and MuSiQue (whose answers are not short integers amenable to memorization) and on execution-graded HumanEval, where a memorized wrong answer simply fails the unit tests. We therefore attribute the gains to cross-model error decorrelation instead of contamination.

Table 13: Bootstrap 95% CIs on per-cell accuracy. On GPQA (n=198) the CIs are tight and EoT sits clearly above CoT. On AIME-24 (n=30) all CIs are wide, so single-year marginal CIs overlap and cannot by themselves separate the top methods, hence the paired tests in Table 14.

Benchmark	Method	Accuracy	95% CI
AIME-24 (n=30)	CoT	33.3	[16.7, 50.0]
	Debate	83.3	[70.0, 96.7]
	ToT	86.7	[73.3, 96.7]
	EoT (ToT, Opus)	93.3	[83.3, 100]
GPQA (n=198)	CoT	28.8	[22.7, 35.4]
	Debate	33.3	[27.3, 39.9]
	ToT	35.4	[28.3, 41.9]
	EoT (ToT, Opus)	39.4	[32.8, 46.0]

Table 14: Paired bootstrap tests for the headline comparisons. EoT's advantage over multi-agent debate is significant on GPQA (p=0.015) and on the *pooled* two-year AIME set (p=0.007). The single-year AIME-24 slice is directionally identical (+10) but underpowered at n=30 (p=0.108), which is why we report pooled AIME for the debate comparison in §5. EoT's gain over its own ToT *initializer* is positive but not significant (p=0.29): consistent with our claim that the initializer sets the ceiling and EoT certifies without re-deriving, it matches or modestly improves a strong initializer without beating it outright.

Comparison	Benchmark	Δ	95% CI	p (1-sided)
EoT (ToT, Opus) − Debate	AIME-24 (n=30)	+10.0	[−3.3, +23.3]	0.108
EoT (ToT, Opus) − Debate	AIME-24+25 (n=60)	+11.7	[+3.3, +21.7]	**0.007**
EoT (ToT, Opus) − Debate	GPQA (n=198)	+6.1	[+1.0, +11.6]	**0.015**
EoT (ToT, Opus) − CoT	GPQA (n=198)	+10.6	[+5.1, +16.2]	**< 0.001**
EoT (ToT, Opus) − ToT	AIME-24+25 (n=60)	+3.3	[−5.0, +13.3]	0.293

C.11 Cross-critique helps an internally-deliberating reasoner

The third falsifiable question of §4 asks whether a cross-model critic is redundant for a model that already reasons at length internally. It is not redundant: Kimi K2-Thinking, paired with an Opus cross-critic, improves from 33.3 to 42.3 on GPQA (+9) and from 85.7 to 100 on AIME-2025 over self-critique (this GPQA lift is the Kimi row of the cross-reasoner generalization sweep, Table 17, and is previewed in §5). (GPQA is scored on the 168/198 problems for which Kimi emits a parseable final answer. On the hardest items it exhausts its token budget and returns empty, a capability limit of the reasoner and not of EoT.) External, decorrelated critique thus supplies error-detection that a model's own extended deliberation does not, consistent with the correlated-blind-spots account.

C.12 Secondary ablations: granularity, attack strength, revision history

These ablations support the claim of §3.2 that the critic's *identity* is the primary lever and the remaining design axes are secondary. All use the Qwen3-235B reasoner.

Attack granularity (answer-level vs. step-level). Attacking the full chain at once (answer-level, our default) versus attacking individual steps before committing them (step-level) is roughly a wash, and answer-level is better on the benchmark that matters most (GPQA), so we keep the simpler answer-level loop (Table 15). The pattern holds under a cross-model Opus critic: step-level reaches 76.7/30.8 on AIME-2024/GPQA versus answer-level's 73.3/37.9, a small AIME gain bought at a clear GPQA loss.

Attack strength and revision history. *Attack strength* sets how aggressively the attacker prompt searches for a flaw, across three levels: *lenient* instructs the attacker to "find any issue" (raising an objection on the slightest doubt); *strict* instructs it to "find the most devastating objection" (demanding the single hardest-hitting flaw); and *adaptive* starts lenient and escalates to strict only on rounds where the previous objection was conceded or failed to land, spending aggression only when a light touch stalls. On the two math benchmarks with headroom (MATH-500, AIME-2024), the *strict*

Table 15: Attack granularity (Qwen3-235B, self-critique). Step-level attacking does not beat the simpler answer-level loop overall (it trades a small AIME gain for a MATH/GPQA loss), so EoT attacks the full chain by default.

Granularity (self-critique)	GSM8K	MATH-500	GPQA	AIME-24
Answer-level (default)	**95.0**	**95.0**	**28.8**	60.0
Step-level	94.0	89.7	24.7	**63.3**

attacker with revision history is best, decisively so on AIME (56.7 vs. 36.7 lenient, 50.0 adaptive, 50.0 no-history), which is why we adopt it as the default (Table 16). On GPQA the ordering *inverts*: the strict setting is worst (25.8 vs. 30.8/31.8), because under self-critique a more aggressive attacker manufactures more spurious objections on multiple-choice items it cannot conclusively break, the divergence failure of §3.2. This is not a contradiction but the same lesson from the other side: tuning the attacker prompt trades one benchmark against another within a narrow band, whereas swapping in a decorrelated critic lifts GPQA by +12 (§5), so the critic's identity, more than the prompt knobs, is the lever. We use the strict attacker with history enabled, the best setting on the math benchmarks and the configuration all body results use.

Table 16: Attack strength and revision history, Qwen3-235B self-critique. The three strength levels (lenient, adaptive, strict) are defined in the text above. Each sub-block varies one axis while holding the other fixed, and the default *strict attacker with revision history* (highlighted, 92.5/56.7/25.8) anchors both. It is best on the math benchmarks (MATH-500, AIME-2024), decisively so on AIME (56.7 vs. 36.7 lenient, 50.0 without history), but *worst* on GPQA (25.8), where a more aggressive self-critic raises more spurious objections on items it cannot break. Either way the spread is far smaller than swapping the critic (+12 to +20, §5): the critic's identity, more than these knobs, is the primary lever.

Variant (Qwen3-235B, self-critique)	MATH-500	AIME-24	GPQA
Attack strength (revision history enabled)			
Lenient ("any issue")	89.0	36.7	**30.8**
Adaptive	92.0	50.0	26.3
Strict (default)	**92.5**	**56.7**	25.8
Revision history (strict attacker)			
Without prior objections	91.5	50.0	31.8
With history (default)	**92.5**	**56.7**	25.8

C.13 Cross-model EoT generalizes across reasoners

The headline result (Table 2) fixes the reasoner at Qwen3-235B, raising the obvious question of whether the cross-model EoT win is specific to that one reasoner. Here, we demonstrate that it is not, by verifying generalization in two ways: a cross-reasoner sweep on a common benchmark (Table 17), and a complete same-reasoner method ladder on a second open reasoner (Table 18).

The cross-model lift is positive on every reasoner. Table 17 holds the configuration fixed (CoT initializer, Opus cross-critic) and the benchmark fixed (GPQA, the one we ran cross-model on all reasoners) while varying the reasoner across a 4× capability range and three model families (Qwen, DeepSeek, Kimi). The Opus cross-critic lifts *every* reasoner over its own CoT baseline: a weak open model (+10.6), two open workhorses (+9.1, +3.0), and an internally-deliberating reasoner (+9.0). The effect is therefore a property of the cross-model critique mechanism itself, beyond the Qwen3-235B reasoner we feature in the main text.

The full method ladder reproduces on a second reasoner. Table 18 runs the full method suite (the same baselines as the main table plus EoT on CoT and ToT initializers, all sharing the Opus cross-critic) on DeepSeek V3.2 across all eight benchmarks. The pattern matches Qwen3-235B exactly where the thesis predicts it should: a cross-model EoT variant is the strongest method on every *unsaturated* benchmark, with EoT (ToT init.) leading both AIME years and GPQA (beating debate by +10 on AIME-2024 and +20 on AIME-2025) and improving on its own ToT initializer, while

Table 17: Cross-model EoT generalizes across reasoners (GPQA, CoT initializer, Opus cross-critic). The decorrelated critic lifts every reasoner over its own CoT baseline, across a $4\times$ size range and three families (Qwen, DeepSeek, Kimi), so the effect is not specific to the Qwen3-235B reasoner of Table 2. †Kimi scored on the $168/198$ problems it parses (§C.11).

Reasoner (GPQA, CoT init.)	CoT	EoT (Opus critic)	Δ
Qwen3-32B	22.7	33.3	**+10.6**
DeepSeek V3.2	30.3	33.3	+3.0
Qwen3-235B	28.8	37.9	+9.1
Kimi K2-Thinking†	33.3	42.3	+9.0

EoT (CoT init.) tops execution-graded code (HumanEval 97.6). Where the thesis predicts no separation it finds none: on saturated MATH and GSM8K every strong method clusters within $\sim$2 points (self-consistency nominally tops MATH at 96.2), and on the inductive MuSiQue/ARC-Challenge a sampling or debate baseline edges it by 1–3 points, the same no-headroom regimes where the main-table cross-model critic also only ties. We report the full grid, and the win is concentrated exactly where reasoning depth, more than retrieval or saturation, decides the answer.

Table 18: Full method ladder on a second reasoner (DeepSeek V3.2, per-cell accuracy, column-best in **bold**; benchmark order and baselines as in Table 2; all EoT variants share the Opus cross-critic). Mirroring the picture for Qwen3-235B, EoT (ToT init., Opus critic) leads the *unsaturated deductive* benchmarks (AIME-2024 with +10 over debate, AIME-2025 with +20, and GPQA) and improves on its ToT initializer; on execution-graded code (HumanEval) the cross-model critic again tops the field (97.6 with the CoT initializer). On saturated MATH/GSM8K all strong methods cluster within $\sim$2 points (self-consistency nominally tops MATH). †On the *inductive* MuSiQue/ARC-Challenge the EoT rows use the verification-style attacker (Appendix C.4): EoT tops ARC-Challenge (98.0, tied with debate) and reaches 77.0 on MuSiQue, statistically tied with the best baseline (self-consistency 77.3) and above every other.

DeepSeek V3.2	GSM8K	MATH	GPQA	AIME-24	AIME-25	HEval	MuSiQue	ARC-C
CoT	96.0	93.6	30.3	53.3	46.7	90.2	73.0	97.3
Self-Cons. (8)	97.0	96.0	26.8	70.0	56.7	95.7	**77.3**	97.0
Self-Cons. (16)	96.7	**96.2**	33.3	73.3	56.7	96.3	76.0	97.3
Self-Corr.	95.0	94.2	30.8	73.3	63.3	93.9	72.3	97.7
Self-Corr. (multi)	94.7	95.2	31.3	80.0	63.3	90.9	68.3	97.7
Best-of-N (N=8)	96.3	94.6	30.3	73.3	66.7	95.1	72.0	96.3
Debate	96.3	95.0	32.3	80.0	63.3	64.0	72.7	**98.0**
ToT	96.3	94.5	35.9	86.7	66.7	95.7	72.3	96.7
EoT (CoT, Opus)	95.5	95.3	33.3	**90.0**	73.3	**97.6**	76.0†	97.7†
EoT (ToT, Opus)	**97.7**	95.5	**36.9**	**90.0**	**83.3**	95.7	77.0†	**98.0**†

C.14 EoT LOOP VARIANTS DO NOT RIVAL THE CRITIC

We explored many refinements to the *loop* itself (a revision-acceptance gate, confidence-gated oscillation handling, trajectory selection, a 3-attacker ensemble, early stopping, adaptive iteration, and regenerate-vs-revise), holding the (self) critic fixed on Qwen3-235B. None materially beats the base loop (Table 19): on AIME-24 the spread across nine variants is 46.7–60.0 with base at 56.7, and on GPQA 24.2–29.3 with base at 25.8. This is the control for our central claim: with the critic held fixed, loop engineering yields at most noise-level movement, whereas swapping in a decorrelated critic moves accuracy by +12 to +20 (§5). The lever is the critic, and the loop is secondary.

Two critics do not beat one. A natural extension is to use *multiple* decorrelated critics, e.g. alternating two cross-family critics round-robin so each round draws a fresh perspective. We test whether this compounds the decorrelation gain by pairing the strongest critic (Opus) with a second cross-family critic (Sonnet 4.6, or DeepSeek V3.2), against the single-Opus baseline (Qwen3-235B reasoner, CoT initializer; Table 20). It does not: both two-critic pools *match or trail* single-Opus on every benchmark, neutral on GPQA (37.4 vs. 37.9) and worse on AIME, where alternating in a

Table 19: EoT loop variants, self-critique on Qwen3-235B. No refinement to the loop materially beats the base loop (the spread is within noise), confirming that the critic's identity, more than loop engineering, is the active lever.

Loop variant (Qwen3-235B, self)	GSM8K	MATH-500	GPQA	AIME-24
Base EoT	94.0	94.7	25.8	56.7
Trajectory select	95.0	95.0	28.8	60.0
Ensemble (3 attackers)	94.0	92.7	28.8	56.7
EoT+ (gate+osc)	93.5	93.3	29.3	46.7
Revision gate	93.0	94.0	25.3	46.7
Confidence-osc. stop	94.0	94.3	24.2	56.7
Adaptive iteration	94.5	94.3	26.3	50.0
Early stop	91.5	94.0	26.8	56.7
Regenerate (vs. revise)	94.5	94.3	26.8	46.7

weaker second critic spends half the rounds on lower-quality objections. One strong, decorrelated critic already captures the gain, and a second adds cost without adding accuracy. This is consistent with the central account: what matters is that *some* critic decorrelated from the reasoner attacks each answer, regardless of how many do.

Table 20: Single vs. two-critic EoT (Qwen3-235B reasoner, CoT initializer; two-critic pools alternate the two critics round-robin). Adding a second decorrelated critic to the strongest one does not help: both pools match or trail single-Opus. One strong decorrelated critic suffices, and critic *identity*, more than critic *count*, is the lever.

Critic pool (Qwen3-235B, CoT init.)	AIME-24	AIME-25	GPQA
Opus (single)	**73.3**	**63.3**	**37.9**
Opus + Sonnet 4.6	70.0	46.7	37.4
Opus + DeepSeek V3.2	53.3	50.0	37.4

C.15 Equilibrium consistency: do different initializations reach the same answer?

The fixed-point view (§3.1) predicts that, when the attack–revise operator contracts, iteration converges to a *unique* equilibrium regardless of where it starts. We test this directly: for 100 MATH-500 problems we run EoT (Qwen3-235B *self*-critique) from five independently sampled initial answers (temperature 0.9, distinct seeds) and ask whether the five runs reach the same final answer. We use the self-critic deliberately: it is the *hardest* case for the contraction prediction, since sharing the reasoner's blind spots gives it the largest completeness defect c and the contraction factor κ closest to 1 (Appendix D), so convergence here is the conservative test. It also keeps the loop single-model, isolating the initial answer as the only varied factor, and we report the Opus cross-critic for comparison below. The initializations are genuinely diverse (all 100 problems produce five *different* initial chains), so agreement at the end reflects convergence and not shared starting points.

The equilibrium is largely but not perfectly initialization-invariant: on 77% of problems all five runs converge to the *same* answer (normalizing by math-equivalence), and the mean modal agreement (the largest cluster of identical answers among the five) is 91.8%. Correctness itself is more stable still: the five runs agree on whether the problem is solved correctly 89% of the time, at a pooled accuracy of 94.2%. The 23% of problems with residual disagreement are concentrated where the theory predicts the contraction fails: hard items on which different initializations land in different basins (premature convergence at distinct wrong fixed points) or cycle (oscillation). This is the expected signature of a contraction that holds in expectation but not pointwise: consistent enough to make the equilibrium a meaningful object, with the residual variance tracing exactly the failure modes of §3.2. Repeating the experiment with the Opus cross-critic gives a similar picture at the level that matters: *correctness*-agreement across the five seeds is 90.0% (vs. 89% self) at a pooled 94.4% accuracy (vs. 94.2%). Answer-identity convergence is marginally looser (67% all-agree, 87.8% modal, vs. 77%/91.8% self): the more active cross-critic raises more substantive objections,

so trajectories explore more before settling, yet land on the same *correctness* outcome. The decorrelated critic thus preserves equilibrium consistency where it counts (whether the problem is solved) without sharpening answer-identity agreement.

D CONVERGENCE PROOFS

This appendix makes Propositions 1–2 precise and proves them. Section 3.2 states them informally. The precise versions are *conditional*, and stating those conditions exactly is what gives the theory its predictive content (each violated condition corresponds to a named, observed failure mode). We first fix a probabilistic model of the loop, then prove a one-step accuracy bound (Proposition 1) and an almost-sure convergence-rate result (Proposition 2), and finally relate the constants to the decorrelation thesis.

D.1 PROBABILISTIC MODEL OF THE ATTACK–REVISE LOOP

Fix a problem P with gold answer y^*. Let $\textsc{ans} : \mathcal{A} \to \mathcal{Y}$ extract the final answer from a reasoning state, and call a state *correct* if $\textsc{ans}(a) = y^*$ and *wrong* otherwise. The loop produces a stochastic sequence $a_0, a_1, \ldots$ with $a_{i+1} = T(a_i)$, where the randomness is over the attacker's and reviser's sampled outputs. Write $\mathcal{F}_i = \sigma(a_0, \ldots, a_i)$ for the natural filtration and let

$$X_i := \mathbf{1}[\textsc{ans}(a_i) = y^*] \in \{0, 1\}, \qquad \text{so } \mathrm{Acc}(a_i) = \mathbb{E}[X_i].$$

The transition $a_i \mapsto a_{i+1}$ is governed by four conditional probabilities, which we take as the primitive parameters of the critic–reasoner pair (assumed time-homogeneous and, for the one-step result, dependent on a_i only through its correctness, a first-order Markov approximation we state explicitly as it is the one substantive idealization):

Assumption 1 (Loop parameters)**.** *There exist constants $s, c, \epsilon, r, \rho \in [0, 1]$ such that, conditional on $\mathcal{F}_i$,*

1. (Soundness) *given that an objection is raised, it identifies a genuine error with probability $s := P($flaw real $\mid o \neq \perp)$;*
2. (Completeness defect) *a wrong state draws* no *objection with probability $c := P(o = \perp \mid X_i = 0) < 1$;*
3. (Soundness defect / false-objection rate) *a correct state draws a (spurious) objection with probability $\epsilon := P(o \neq \perp \mid X_i = 1)$;*
4. (Repair rate) *a revision prompted by a* genuine *flaw on a wrong state reaches a correct answer with probability $r := P(X_{i+1} = 1 \mid X_i = 0,\ o \neq \perp,$ flaw real$)$;*
5. (Corruption rate) *a revision prompted by a* spurious *objection on a correct state lands on a wrong answer with probability $\rho := P(X_{i+1} = 0 \mid X_i = 1,\ o \neq \perp)$.*

Two remarks. First, soundness (s) and the soundness defect (ϵ) are *distinct* parameters (one conditions on an objection being raised, the other on the underlying answer being correct), so soundness alone does *not* bound ϵ, and the informal "sound $\Rightarrow$ monotone" of §3.2 is true only once ϵ is separately controlled, which we make explicit below. Second, a revision prompted by a real flaw on a wrong answer cannot make a correct answer wrong (there is no correct answer to lose), so the only downward transition is the $\epsilon\rho$ channel out of the correct state. Third, the parameters do not fully specify the wrong-state dynamics (a wrong answer may receive a genuine objection with probability $(1-c)s$, a spurious one with probability $(1-c)(1-s)$, or none with probability c), but the one-step accuracy result needs only the *upward* current, and every non-genuine branch from a wrong state keeps $X_{i+1} = 0$ (a wrong-to-wrong swap or a no-op), so it contributes zero to the accuracy drift regardless of its internal split. This is why Lemma 1 depends on the wrong-state parameters only through the product $(1-c)sr$, and we do not need to pin down the wrong-to-wrong transition law.

D.2 One-step accuracy change (Proposition 1)

Lemma 1 (One-step accuracy drift)**.** *Under Assumption 1,*

$$Acc(a_{i+1}) - Acc(a_i) \;=\; \underbrace{(1 - X\text{-}mass)\,(1-c)\,s\,r}_{\text{repairs from wrong}} \;-\; \underbrace{(X\text{-}mass)\,\epsilon\,\rho}_{\text{corruptions from correct}} \;,$$

where X*-mass* $:= Acc(a_i) = P(X_i = 1)$*. Concretely, writing* $p := P(X_i = 1)$*,*

$$\Delta_i \;:=\; Acc(a_{i+1}) - Acc(a_i) \;=\; (1-p)\,(1-c)\,s\,r \;-\; p\,\epsilon\,\rho.$$

Proof. Condition on X_i and apply the law of total expectation to $X_{i+1} - X_i$. *Wrong state (*$X_i = 0$*, probability* $1 - p$*).* The state can only move up. An upward move requires (a) an objection, raised with probability $1 - c$; (b) the objection being genuine, probability s; and (c) a successful repair, probability r. A spurious objection $(1 - s)$ on a wrong state either leaves it wrong or swaps one wrong answer for another, contributing 0 to $\mathbb{E}[X_{i+1} - X_i \mid X_i = 0]$ in either case (the indicator stays 0). Hence $\mathbb{E}[X_{i+1} - X_i \mid X_i = 0] = (1-c)\,s\,r \geq 0$. *Correct state (*$X_i = 1$*, probability* p*).* The state can only move down, which requires a (necessarily spurious) objection, probability ϵ, followed by a corrupting revision, probability ρ; thus $\mathbb{E}[X_{i+1} - X_i \mid X_i = 1] = -\epsilon\rho \leq 0$. Taking the outer expectation, $\Delta_i = \mathbb{E}[X_{i+1} - X_i] = (1-p)(1-c)sr - p\,\epsilon\rho$. □

Proposition 3 (Monotonicity under a competent, calibrated critic)**.** *Under Assumption 1 with* $s > \frac{1}{2}$*,* $c < 1$*, and* $r > 0$*, expected accuracy is non-decreasing at step* i*, i.e.* $\Delta_i \geq 0$*, if and only if the false-objection rate satisfies*

$$\epsilon \;\leq\; \frac{1-p}{p} \cdot \frac{(1-c)\,s\,r}{\rho} \qquad (\textit{vacuous when } \rho = 0).$$

In particular, if the critic never corrupts a correct answer ($\rho = 0$*, or* $\epsilon = 0$*) then* $\Delta_i \geq 0$ *unconditionally, and* $\Delta_i > 0$ *whenever wrong-answer mass remains (*$p < 1$*); and the iteration is monotone at* every *step (hence* $Acc(a_{i+1}) \geq Acc(a_i)$ *for all* i*) provided the bound holds at the running accuracy* $p = Acc(a_i)$*, which is easiest to satisfy early (small* p*) and tightens as accuracy rises.*

Proof. Immediate from Lemma 1: $\Delta_i \geq 0 \iff (1-p)(1-c)sr \geq p\,\epsilon\rho$, which rearranges to the displayed bound when $\rho > 0$; when $\rho = 0$ the negative term vanishes and $\Delta_i = (1-p)(1-c)sr \geq 0$, strictly positive iff $p < 1$ (using $c < 1, s > 0, r > 0$). The monotonicity-at-every-step claim is the per-step inequality applied along the trajectory. □

This is the precise form of Proposition 1: soundness ($s > \frac{1}{2}$) and completeness ($c < 1$) guarantee a *positive* repair current out of the wrong state, but global monotonicity additionally requires the false-objection rate ϵ to sit below the threshold above; otherwise the $\epsilon\rho$ corruption current dominates once accuracy p is high. The failure of this inequality (large ϵ) is exactly *divergence* / "overthinking" (§5): a critic that manufactures objections to correct answers. Because the threshold scales as $(1 - p)/p$, divergence appears first on saturated benchmarks (high p, little wrong-answer mass to offset corruption), precisely the regime where Appendix C.1 reports a mild overthinking penalty under self-play.

D.3 Convergence and rate (Proposition 2)

For convergence we measure progress by a potential $V(a) := d(\textsc{ans}(a), y^*) \geq 0$, where d is any metric on the (finite, per-problem) answer set $\mathcal{Y}$ with $V(a) = 0 \iff \textsc{ans}(a) = y^*$; the binary choice $V = 1 - X$ is admissible and recovers the accuracy reading.

A clarification on what the next result does and does not establish. We do *not* prove that EoT's operator T is contractive from the primitive critic parameters of D.1; contraction is an *identified* condition (the precise hypothesis $(\star)$ under which the loop provably converges) and is not asserted for every critic. The content of Proposition 4 is therefore conditional ("*if* $(\star)$, *then* geometric convergence"), and its value is that it pins convergence to a single measurable critic property and lets us read off the rate. Whether $(\star)$ *holds* is exactly what §5 probes empirically: the round-distribution evidence (Appendix C.6) is a direct estimate of the contraction factor, and the failure of $(\star)$ is what

produces the pathologies below. The link to D.1 is one-directional but informative: the binary potential $V = 1 - X$ has $\mathbb{E}[V(a_{i+1}) \mid \mathcal{F}_i] = (1 - X_i)(1 - (1 - c)sr) + X_i\,\epsilon\rho$, so on the wrong state $(\star)$ reduces to the repair current $(1 - c)sr > 0$ (completeness and soundness), while a non-zero corruption current $\epsilon\rho$ on the correct state is what can push the conditional ratio above $\kappa \geq 1$; i.e. $(\star)$ is the same balance of repair-over-corruption that Lemma 1 already exhibited, now required to hold *contractively* instead of merely in net.

Proposition 4 (Geometric convergence in expectation and almost surely)**.** *Suppose the attack–revise operator is* contractive in expectation*: there is* $\kappa \in [0, 1)$ *with*

$$\mathbb{E}[V(a_{i+1}) \mid \mathcal{F}_i] \;\leq\; \kappa\, V(a_i) \qquad \textit{for all } i. \tag{$\star$}$$

Then: (a) $\mathbb{E}[V(a_i)] \leq \kappa^i\, \mathbb{E}[V(a_0)]$*, so* $\mathbb{E}[V(a_i)] < \delta$ *after* $i \geq \log(\mathbb{E}[V(a_0)]/\delta)/\log(1/\kappa) = O(\log(1/\delta))$ *steps;* (b) $V(a_i) \to 0$ *almost surely, i.e. the iterates reach and remain at a correct equilibrium with probability one; and* (c) *if in addition* V *is integer-valued with* $V \geq 1$ *off the target (e.g.* $V = 1 - X$*, or any integer answer-distance), the expected hitting time* $\tau := \inf\{i : V(a_i) = 0\}$ *satisfies* $\mathbb{E}[\tau] \leq \mathbb{E}[V(a_0)]/(1 - \kappa)$.

Proof. (a) Taking expectations in $(\star)$ and using the tower property, $\mathbb{E}[V(a_{i+1})] \leq \kappa\, \mathbb{E}[V(a_i)]$; iterating gives $\mathbb{E}[V(a_i)] \leq \kappa^i \mathbb{E}[V(a_0)]$. Setting the right side $< \delta$ and solving for i yields the stated step count.

(b) The process $V(a_i) \geq 0$ together with $(\star)$ makes $V(a_i)$ a non-negative supermartingale: $\mathbb{E}[V(a_{i+1}) \mid \mathcal{F}_i] \leq \kappa V(a_i) \leq V(a_i)$. By the supermartingale convergence theorem it converges almost surely to an integrable limit $V_\infty \geq 0$. To identify $V_\infty = 0$: from (a), $\mathbb{E}[V(a_i)] \to 0$, and since $V(a_i) \to V_\infty$ a.s. with $V \geq 0$, Fatou's lemma gives $\mathbb{E}[V_\infty] \leq \liminf_i \mathbb{E}[V(a_i)] = 0$, so $V_\infty = 0$ a.s. Because $\mathcal{Y}$ is finite, V takes values in a finite set bounded below by a gap $v_{\min} > 0$ off the target; $V_\infty = 0$ then forces $V(a_i) = 0$ for all large i almost surely, i.e. absorption at a correct equilibrium.

(c) Assume V is integer-valued with $V(a) \geq 1$ whenever $\textsc{ans}(a) \neq y^*$, so that $V(a_i) \geq \mathbf{1}[V(a_i) \geq 1]$ pointwise. Condition $(\star)$ then gives the drift

$$\mathbb{E}[V(a_{i+1}) - V(a_i) \mid \mathcal{F}_i] \leq -(1 - \kappa)V(a_i) \leq -(1 - \kappa)\,\mathbf{1}[V(a_i) \geq 1],$$

so $M_i := V(a_i) + (1 - \kappa)\sum_{j<i} \mathbf{1}[V(a_j) \geq 1]$ is a non-negative supermartingale. We first bound $\mathbb{E}[\tau]$, then apply optional stopping (it does not apply to τ directly without an integrability check). For the bounded stopping time $\tau \wedge n$, optional stopping for the bounded case gives $\mathbb{E}[M_{\tau\wedge n}] \leq \mathbb{E}[M_0] = \mathbb{E}[V(a_0)]$. Since $M_{\tau\wedge n} \geq (1 - \kappa)\sum_{j<\tau\wedge n} \mathbf{1}[V(a_j) \geq 1] = (1 - \kappa)\,(\tau \wedge n)$ (every step before absorption has $V \geq 1$), we get $(1 - \kappa)\,\mathbb{E}[\tau \wedge n] \leq \mathbb{E}[V(a_0)]$ for all n. Letting $n \to \infty$, monotone convergence yields $(1 - \kappa)\,\mathbb{E}[\tau] \leq \mathbb{E}[V(a_0)]$, i.e. $\mathbb{E}[\tau] \leq \mathbb{E}[V(a_0)]/(1 - \kappa)$; in particular $\tau < \infty$ a.s., which is the integrability fact that retroactively licenses the limit. This also sharpens the step count of part (a) from an ϵ-accuracy horizon to a finite expected hitting time. □

Why the two failure modes are exactly the two ways $(\star)$ fails. Condition $(\star)$ bundles the requirements on the critic, and each can break in one way, matching the empirics of §5. *Premature convergence* is the case where the loop halts at a *wrong* fixed point $\hat{a}$: $\textsc{ans}(\hat{a}) \neq y^*$ yet $\textsc{Atk}(P, \hat{a}) = \perp$, so the operator stops with $V(\hat{a}) > 0$. This is a *completeness* failure (c is effectively 1 at $\hat{a}$), and $(\star)$ fails because no contraction acts: the chain is absorbed above the target. *Oscillation* is the case where V fails to contract on a recurrent set ($\kappa \geq 1$ along a cycle $\textsc{ans}(a_i) = \textsc{ans}(a_j)$, $j < i - 1$): a revision trades one flaw for another of equal potential. EoT's countermeasures attack exactly these: passing the objection history H to the reviser forbids re-introducing a fixed flaw (raising r and lowering the chance of a non-contractive swap), and explicit cycle detection (§3) truncates any residual oscillation and returns the trajectory majority, restoring termination when $(\star)$ would not hold globally.

D.4 Interpretation: reading the constants through the decorrelation lens

The two results are conditional by design, and the conditions are critic properties: monotonicity (Prop. 3) needs a small false-objection rate ϵ, and convergence (Prop. 4) needs a contraction factor $\kappa < 1$. This subsection is *interpretive* and proves nothing new: D.1–D.3 contain no decorrelation parameter, so the claims below are a reading of how the constants *should* move with critic–reasoner

correlation, motivating the empirics of §5 without following from the proofs. With that caveat, the heuristic is the following. A *self*-critic is plausibly doubly disadvantaged. Sharing the reasoner's blind spots should inflate the completeness defect c on exactly the problems the reasoner gets wrong (the flaw the model made is the flaw it cannot see), pushing κ toward 1 and producing the premature-convergence tail, and on correct answers a self-critic's calibration errors should raise ϵ, eroding the monotonicity margin. A *decorrelated cross-model critic* should improve both constants: with errors closer to independent of the reasoner's, it both finds residual flaws the reasoner missed (smaller c, hence $\kappa < 1$ and faster geometric decay) and is less prone to inventing objections to answers the reasoner got right (smaller ϵ, hence a wider monotonicity threshold). We do not derive these dependencies from a model of correlation here (doing so would require formalizing the joint law of reasoner and critic errors, which we leave to future work), but they are exactly the qualitative predictions our experiments test: the round-distribution evidence of Appendix C.6 (correct answers conceding at round 1, the cross-model critic shrinking the budget-cap tail) is the empirical signature we would expect of a smaller κ and ϵ, and it is what we observe.

E Broader Impact: EoT as Training-Free Red-Teaming

EoT's adversarial probing is structurally red-teaming, and the decorrelation finding has a safety reading: a model is a poor auditor of its own outputs precisely because it shares its own blind spots, so robust verification should come from a decorrelated external critic. Outputting only answers that survive cross-model attack is a lightweight, training-free check that needs no separately trained reward model or verifier.