AIME 2025 Benchmark: An Analysis of AI Math Reasoning

By Adrien Laurent, CEO at IntuitionLabs • 10/24/2025 • 30 min read

aime 2025 | ai benchmark | mathematical reasoning | llm performance | gpt-5 | chain-of-thought | open-source models | artificial intelligence | ai vs human

Executive Summary

The **AIME 2025 benchmark** – based on the 2025 American Invitational Mathematics Examination – has emerged as one of the most challenging AI tests of advanced mathematical reasoning. It consists of 30 "Olympiad-level" integer-answer math problems (15 per exam session) ([1] aiwiki.ai). In autumn 2025, frontier AI models routinely scored far above top humans on these problems: for example, OpenAI's newly released GPT-5 (behind the scenes of ChatGPT and Copilot) achieved **~94.6% accuracy** on AIME 2025 (pass@1, closed-book) ([2] www.snarful.com), whereas the median human (top high-school math competitors) solves only ~4–6 of 15 problems (~27–40%) ([3] aiwiki.ai). Table 1 (below) summarizes leading models' performance on AIME 2025; both proprietary and open-source systems now exceed 85% even without tools ([4] www.vals.ai) ([2] www.snarful.com). Access to external tools is being shown to *nearly* "solve" the exam: e.g. GPT-4 with Python reaches ~99.5% ([5] www.theregularizer.com), and GPT-5 Pro with code hit 100% ([6] venturebeat.com). Meanwhile, the latest research strategies (new prompting/inference methods and specialized fine-tuning) are pushing scores ever higher. However, even these leading systems still leave a few problems unsolved (~13% error for best closed models ([7] www.theregularizer.com)) and struggle on the *hardest* problems. ([8] epoch.ai) ([7] www.theregularizer.com). In short, AIME 2025 has become a **stress-test of multi-step reasoning**, highlighting both how dramatically AI math ability has advanced (frontier models now regularly beat top humans ([9] www.theregularizer.com)) *and* how significant challenges remain.

1. **Benchmark Definition:** AIME 2025 uses the official 2025 AIME contest problems (released Feb 6, 2025 ([10] aiwiki.ai)) with exact-match scoring (each answer is an integer 0–999, no partial credit) ([11] www.theregularizer.com) ([12] aiwiki.ai). It covers algebra, geometry, number theory, combinatorics, and probability ([1] aiwiki.ai). The benchmark report treats AIME as a closed-book test of purely *internal* reasoning (no examples or external tools), though researchers also evaluate tool-augmented modes.
2. **Model Performance:** By late 2025, the new top of the leaderboard is dominated by advanced LLMs. GPT-5 (full version) leads with ~94–95% ([2] www.snarful.com) ([4] www.vals.ai), followed closely by top "reasoning" models (e.g. xAI's Grok-4 at ~91–93% ([4] www.vals.ai)). Strong open models (e.g. Qwen-OSS 120B) reach ~92.6% ([4] www.vals.ai), and Claude 3.7's chain-of-thought variant attains ~52.7% ([13] www.theregularizer.com). The gap to perfection remains: even the best closed-book model left ~2–3 problems unanswered ([7] www.theregularizer.com). Table 1 (below) compares key models' AIME 2025 accuracies.
3. **Techniques & Trends:** Virtually all state-of-art models rely on *chain-of-thought* style reasoning or its equivalents. On top of that, new methods are driving rapid gains: e.g. **DeepConf** (confidence-based pruning of reasoning paths) reports ~99.9% accuracy ([14] www.researchgate.net); **Learning-from-Peers (LeaP)** (collaborative reasoning among parallel chains) adds ~5 points ([15] openreview.net); Reinforcement-Learning fine-tuning (e.g. GPPO in *Klear-Reasoner*) lifts smaller models to 83.2% ([16] www.researchgate.net); large curated reasoning datasets (OpenThoughts) yield ~53% on a 7B model ([17] huggingface.co); and hybrid strategies (e.g. DeepPrune's iterative pruning) drastically cut computation with minimal accuracy loss ([18] huggingface.co). The net effect is that models are learning more reliable multi-step math reasoning at breakneck pace.
4. **Implications:** AIME's difficulty has spurred proposals for even tougher benchmarks (e.g. Olympiad proofs, "AI decathlons" combining many skills ([19] www.theregularizer.com)). It also highlights alignment/safety needs: as reasoning gets stronger, verifying the chain of thought and preventing errors becomes crucial ([20] www.theregularizer.com). From an industry perspective, the AIME benchmark exemplifies the rapid commoditization of advanced reasoning: open models are closing the gap with proprietary ones, suggesting that soon commodity AI may solve high-school Olympiad math. In summary, AIME 2025 vividly illustrates

the explosion of AI reasoning power – with current models now routinely beating top students – while also underscoring that a "final frontier" of truly rigorous reasoning tasks still lies ahead.

Table 1. *Selected model performances on AIME 2025 (pass@1 accuracy, closed-book).* Top models (GPT-5, Grok-4, etc.) are in boldface; open-source models are italicized. Sources: leaderboards and reports ([4] www.vals.ai) ([2] www.snarful.com) ([21] www.theregularizer.com).

Model	Category	AIME 2025 Accuracy	Notes / Source
GPT-5 (full)	Closed (OpenAI)	**~94–95%**	Official report: 94.6% (no tools) ([2] www.snarful.com)
Grok-4 (Fast)	Closed (xAI)	**91–93%**	91.3% ([4] www.vals.ai) (leaderboard)
GPT-OSS-120B	Open-source	92.6%	Leaderboard ([4] www.vals.ai) </current_article_content>
GPT-5 Mini	Closed (OpenAI)	90.8%	Leaderboard ([4] www.vals.ai)
Grok-4 (standard)	Closed (xAI)	90.6%	Leaderboard ([4] www.vals.ai)
GLM 4.5 (DeepMind)	Closed (DeepMind)	86.7%	Leaderboard ([4] www.vals.ai)
o3 Mini (LLaMA-3 proj.)	Open-source	86.5%	Leaderboard ([4] www.vals.ai)
GPT-OSS-20B	Open-source	86.0%	Leaderboard ([4] www.vals.ai)
Claude 3.7 Sonnet	Closed (Anthropic)	52.7%	Leads Anthropic's chain-of-thought variant ([13] www.theregularizer.com)
DeepSeek R1 (Qwen 671B)	Open-source	74.0%	Leaderboard ([22] www.theregularizer.com)
OpenThinker3 (7B fine-tuned)	Open-source	53%	7B model trained on public data ([17] huggingface.co)
Baseline (random reasoning)	—	~20%	Non-reasoning baseline ([3] aiwiki.ai)
Human (median top competitor)	—	~26–40%	4–6 problems out of 15 ([3] aiwiki.ai)

> *Sources:* Leaderboard data compiled from Vals.AI and the Regularizer blog ([4] www.vals.ai) ([21] www.theregularizer.com); official OpenAI/VentureBeat report ([2] www.snarful.com); human baseline and context from AI wiki ([3] aiwiki.ai).

Table 1 shows that in the closed-book AIME 2025 test **GPT-5 and the best proprietary models now approach near-human saturation**, far exceeding typical human scores (4–6/15) ([3] aiwiki.ai). Open-source models (italicized) have also climbed into the high 80s or low 90s percent; for example, Qwen-OSS-120B scores ~92.6% ([4] www.vals.ai). Claude 3.7's "Sonnet" variant reaches ~52.7% ([13] www.theregularizer.com), trailing the leaders but demonstrating progress on Anthropic's side. The bottom row shows that a pure non-reasoning baseline would get only ~20% ([3] aiwiki.ai), underscoring that raw reasoning ability is needed. Note how the **headroom** above 86–90% remains significant: even top models in closed-book mode leave ~3 problems unsolved on average ([7] www.theregularizer.com).

Introduction and Background

Mathematical reasoning has become a key frontier for Large Language Models (LLMs). Starting around 2021-22, researchers introduced benchmarks like MATH, AMR (AI2 Reasoning Challenge), and bespoke exams to push

models beyond language patterns into genuine multi-step logic ([23] www.theregularizer.com) ([8] epoch.ai). The *American Invitational Mathematics Examination* (AIME) is a natural benchmark in this vein. The AIME is the second stage of the U.S. high-school math contest cycle (after AMC 10/12), used to select top students for the USA Math Olympiad ([24] medium.com). It consists of 15 integer-answer problems per session (AIME I and II) drawn from advanced college-prep math topics ([1] aiwiki.ai) ([24] medium.com). (The version used in AI benchmarking, "AIME 2025," combines both sessions for 30 problems total.) Each answer is an integer 0–999, and scoring is **exact-match** (no partial credit) ([11] www.theregularizer.com) ([25] aiwiki.ai). Because these problems are brand-new each year and highly nontrivial, AIME is effectively a *fresh* multi-step reasoning test: models trained before 2025 have seen none of these questions ([11] www.theregularizer.com). This makes AIME an excellent stress-test of "pure" reasoning, not mere pattern-matching.

From a contest standpoint, human performance on AIME is low – top students typically get 4–6/15 correct ([3] aiwiki.ai) (median score ~27–40% accuracy), and even perfect solvers miss the hardest problems. Benchmarker Vals.AI notes that baseline non-reasoning models (e.g. random guessers) get only ~20% ([3] aiwiki.ai). Thus, AIME's answer format and difficulty mean only systems with genuine reasoning can score well. As of 2025, AIME data is publicly available (e.g. via AoPS and official MAA/AIME releases), and multiple organizations have **immediately evaluated** leading LLMs on the 2025 problems. The latest data (through Sept 2025) show that LLM math abilities have skyrocketed: Table 1 and subsequent sections document these leaps.

In this report, we provide a **thorough analysis of AIME 2025 as an AI benchmark**. We first review the contest format and why it is so challenging ([11] www.theregularizer.com) ([1] aiwiki.ai). We then survey model performance – both closed-book and tool-augmented – contrasting top proprietary systems (GPT-5, Grok, Claude) with open models (Qwen, DeepSeek, etc). Key results are drawn from multiple sources (regularizer and leaderboard blogs ([26] www.theregularizer.com) ([4] www.vals.ai), OpenAI announcements ([2] www.snarful.com) ([6] venturebeat.com), and independent analyses ([27] medium.com) ([8] epoch.ai)). Next, we delve into **techniques and strategies** that have boosted AIME performance, from chain-of-thought prompting to reinforcement learning and confidence-based inference. We summarize research papers introducing new methods (e.g. *DeepConf* ([14] www.researchgate.net), *LeaP* ([15] openreview.net), *Klear* ([16] www.researchgate.net), *OpenThoughts* ([17] huggingface.co), etc.) and their impact on AIME scores. We include case studies of how specific models (e.g. GPT-5, DeepScaleR) fared and what that implies. Finally, we discuss the broader implications: what AIME's results say about the trajectory of AI reasoning, future benchmarks, and societal or safety considerations ([20] www.theregularizer.com) ([8] epoch.ai). Throughout, **all claims are supported by authoritative citations** from technical blogs, paper preprints, and authoritative data.

The AIME 2025 Benchmark in Detail

Contest Format and Role as a Benchmark

The 2025 AIME exam (held Feb 6, 2025 ([10] aiwiki.ai)) consists of two fifteen-question sessions covering a wide range of advanced mathematics ([1] aiwiki.ai). The AI benchmark uses all 30 questions: each requires a single integer answer from 0 to 999 ([11] www.theregularizer.com). There is **no partial credit** – an answer is either exactly correct or not – so the metric is binary accuracy ("Pass@1" in ML terms). Typical AI evaluations report "accuracy" as the fraction of correct answers. For reference, human contestants historically average only ~4–6 correct (≈27–40%) ([3] aiwiki.ai).

The benchmark's rigorous format makes it extremely difficult: each problem requires multiple reasoning steps. As the Regularizer blog notes, "the human median is just 4–6 correct out of 15" ([28] www.theregularizer.com), and "each answer is a single integer; no wiggle room" ([11] www.theregularizer.com). Moreover, the 2025 problems were fresh and unseen by any model prior to Feb 2025 ([11] www.theregularizer.com), so any success indicates

true reasoning rather than memorization. In an AI context, AIME thus measures genuine problem-solving ability – analogous to an "Olympiad exam" for LLMs – far beyond simple arithmetic or formula retrieval.

Table 2 below summarizes this setting. The AIME categories (number theory, geometry, etc.) and strict answer format push models to chain together multiple inferences. The published difficulty ratings (Art of Problem Solving scale) confirm the challenge: problems #1–3 are relatively easy (AoPS rating ~2), but the hardest (#14–15) correspond to rating 6! ([29] epoch.ai). Epoch AI's analysis shows that public LLMs saturate the easiest AIME problems but fall off on the hardest ones ([8] epoch.ai) (*e.g.* GPT-5 solves >95% of problems rated ≤5 but only ~67% of rating-6 problems ([8] epoch.ai)). We return to these figures later in this report.

AIME 2025 Data and Metrics

Key technical details (from the AIwiki and related sources ([1] aiwiki.ai) ([30] aiwiki.ai)): AIME 2025 uses **30 examples** (AIME I & II). The primary metric is **exact-match accuracy**, or Pass@1. (Some analyses also report "Pass@K" for multi-sample settings, e.g. PACS reports 59.78% Pass@256 ([31] huggingface.co), but the standard closed-book score is Pass@1 as shown above.) The AIwiki notes that AIME 2025's *state-of-the-art (SOTA)* closed-book score is 94.6% by GPT-5 (as of Aug 2025) ([30] aiwiki.ai). Importantly, the wiki cites MAA/AoPS as sources, implying this benchmark is officially sanctioned by the math community. It also notes that AIME 2025 is *not* yet considered "saturated" (the bar for perfect AI performance) ([32] aiwiki.ai), leaving room for future gains.

Researchers generally treat this as a **closed-book** test of raw reasoning. However, many studies also explore "open-book" or tool-augmented performance. For example, allowing Python code execution during answer generation (the "calculator" or "tool" setting) dramatically boosts scores ([33] www.theregularizer.com) ([6] venturebeat.com). OpenAI itself reported that giving GPT-5 a coding toollets it reach 100% ([6] venturebeat.com); similarly, an o4-mini model with Python achieved ~99.5% ([33] www.theregularizer.com). These results indicate that with external help, the test is nearly trivial for modern systems – but closed-book performance (the focus of this report) remains the tougher challenge.

Data Analysis and Model Performance

Figure 1 below (conceptual) would illustrate the rapid climb of model scores. By early 2025, LLMs had already eclipsed the best human scores ([9] www.theregularizer.com). Table 1 above catalogues specific results: GPT-5 and Grok-4 are in the 90–95% range, while other top models cluster in the 85–90% range ([4] www.vals.ai) ([2] www.snarful.com). Open-source systems also show strong gains: e.g., an 8B Qwen model (DeepSeek R1) reached 74.0% ([22] www.theregularizer.com) and a 32B "AM-Thinking-v1" reached 74.4% ([34] huggingface.co), comparing to the 83–93% of the proprietary leaders.

A few broader patterns emerge from the data:

- **Rapid Progress:** The AI barrier has cracked wide open. Where GPT-4-era models might have managed ~50–60% on similar contest problems, the latest generation hits ~90%+ ([9] www.theregularizer.com) ([2] www.snarful.com). The Regularizer highlights that "frontier models could barely scratch AIME" in early 2023, but by 2025 they "routinely outperform top human contestants" ([9] www.theregularizer.com). In other words, AI is now far ahead of the median human in this domain.
- **No Single Solves All:** Even the best closed-book models leave a gap. The top score (e.g. o3-mini's 86.5% ([35] www.theregularizer.com) at the time of one leaderboard) means ~2–3 problems were still wrong on average. Different models tend to miss different questions. As the Regularizer notes, " [e]ven the best closed-book score (86.5%) leaves 2–3 problems unsolved. Different models miss different questions, suggesting no single system has a universal math strategy yet" ([7] www.theregularizer.com). This diversity of error patterns implies that we are not yet "saturated" – there is still diagnostic value in AIME as a benchmark.

- **Open vs Closed Models:** For years, proprietary models held a clear lead on math benchmarks. Now that gap is shrinking. For instance, DeepSeek R1 (an open-compute Qwen derivative) got 74.0% ([22] www.theregularizer.com) vs. GPT-5's 94.6% ([2] www.snarful.com). But new open methods (OpenThoughts dataset, DeepPrune inference, etc.) have produced open models reaching the mid-80s or higher. Table 1 shows several open models in the mid-80s, and even above 90% for the largest ones (GPT-OSS-120B at 92.6% ([4] www.vals.ai)). This trend suggests advanced math reasoning will become widely accessible.

- **Tool Use vs Pure Reasoning:** Access to external tools (e.g. code execution) essentially solves AIME. OpenAI and others report near-100% scores when coding is allowed ([6] venturebeat.com) ([33] www.theregularizer.com). This dichotomy is important: it shows *how close* we are to perfect accuracy if we relax the "closed-book" rule, but also indicates that future benchmarks may need to separate "calculator-augmented" performance from raw reasoning. (Indeed, the Regularizer calls it an "open-book vs closed-book" distinction ([33] www.theregularizer.com).)

In summary, the data show **unprecedented AI math performance** on AIME 2025. Top models are routinely above 90% accuracy, vastly surpassing historical human levels. Benchmarks like Vals.AI and the Regularizer blog have quantified this across dozens of models ([4] www.vals.ai) ([26] www.theregularizer.com). At the same time, the slight drop from 2024→2025 (15–20 points in some reports ([36] www.theregularizer.com)) demonstrates that question freshness still matters – models likely benefitted from some overlap in earlier years – and reinforces that fully novel problems still pose a challenge.

Technical Approaches and Results

Researchers have developed a diverse toolkit to tackle AIME. Key ideas include **chain-of-thought prompting**, **multiple-path ensembles**, **reinforcement learning fine-tuning**, **confidence filtering**, and **data augmentation**. Below we detail the most influential techniques and their reported effects on AIME performance. (Table 2 summarizes some representative approaches.)

- **Chain-of-Thought (CoT) and Prompting:** Simply prompting LLMs to "think step by step" is known to greatly boost problem-solving. Virtually all top model evaluations use CoT or few-shot chains-of-thought. For example, in OpenAI's GPT-4 technical report, a chain-of-thought prompt improved math accuracy dramatically (though exact AIME numbers were not given there). In practice, state-of-art runs often sample multiple CoT outputs (self-consistency ([11] www.theregularizer.com)) or use variants like Claude Sonnet (Anthropic's internal CoT mode, 52.7%) ([13] www.theregularizer.com). These methods alone yield sizable gains: one leaderboard notes that applying model-internal CoT (pass@256 sampling) can push accuracies from ~85% to ~90% on AIME ([36] www.theregularizer.com). CoT is essentially part of the baseline for all top accuracy runs.

- **Confidence-based Filtering (DeepConf):** A new test-time technique, exemplified by *Deep Think with Confidence* (DeepConf), uses the model's own token confidences to prune unlikely reasoning paths. In their paper, DeepConf achieves **99.9% accuracy on AIME 2025 ([14] www.researchgate.net)** by discarding low-confidence traces, all without any extra training. In effect, it runs many parallel chains and keeps only the ones the model trusts. DeepConf also reports an **84.7% reduction in total token usage** ([14] www.researchgate.net) (pruning wasteful paths) compared to naive self-consistency. This shows that intelligent filtering can essentially *perfect* the exam results on tested models. While DeepConf's 99.9% is higher than any model alone (since it ensembles many trials), it highlights how much latent capability exists in LLMs once inference is optimized.

- **Parallel Reasoning and Pruning (DeepPrune):** Relatedly, the *DeepPrune* framework dynamically prunes redundant chains during multi-trace inference. Researchers from Tsinghua report that over 80% of parallel reasoning paths often converge to the same answer, so DeepPrune uses a learned judge to stop duplicative chains. They show that on AIME (2024/2025) and GPQA, DeepPrune **cuts computation by ~80% (tokens)** while keeping accuracy within ~3 percentage points of full-consensus results ([18] huggingface.co). In other words, DeepPrune makes the same high-accuracy inference much more efficient. This work underscores that many AIME-model approaches are computationally redundant, and smart pruning can largely eliminate wasted work without losing score.

- **Learning From Peers (LeaP):** A recent ICLR paper introduces a training-time and inference strategy where multiple reasoning "paths" interact. Each path summarizes its intermediate steps and shares them with others, so that one path can recover from another's insight. The authors report that with this "peer reasoning" mechanism, a 32B model (Qwen-32B) saw its accuracy **jump by ~5 points** on average across AIME/HMMT/other benchmarks ([15] openreview.net), surpassing larger baselines. In concrete terms, a 32B Qwen model achieved 5% better AIME accuracy when using LeaP than its vanilla performance ([15] openreview.net). Even small models (7B LeaP-T) matched much bigger ones on AIME 2024. While this is cutting-edge research, it shows that **collaboration between parallel hypotheses can substantially boost math accuracy**. LeaP's results (reported on AIME 2025) suggest similar gains would apply on this benchmark.

- **Reinforcement Learning (RL) & Policy Optimization:** Several teams have applied RL to further train models on reasoning rewards. For example, *Klear-Reasoner'* (8B) uses a "Gradient-Preserving PPO" to fine-tune on math and coding tasks. This yields **90.5% on AIME 2024 and 83.2% on AIME 2025** ([16] www.researchgate.net) for the 8B model – remarkable accuracy for that scale. Another approach, PACS (Proximal Actor-Critic with Softmax), outperforms standard PPO on math exams: it achieved 59.78% **pass@256** on AIME 2025 ([31] huggingface.co), a 13–14 point gain over vanilla PPO/GRPO. Other works (e.g. RLEP) show that replaying high-quality solution trajectories during RL can speed convergence and yield modest AIME improvements. In general, these studies imply that **post-training fine-tuning (especially on math problems or code) can add well over 10 percentage points** compared to off-the-shelf models ([37] huggingface.co) ([27] medium.com). They also reveal that careful RL design (handling clipped gradients, balancing math vs code data) can make even small models significantly competitive.

- **Data and Distillation Approaches:** Some efforts focus on generating better reasoning data. The *OpenThoughts* project created large public datasets of chain-of-thought reasoning examples and trained new models on them. Their latest 7B model (OpenThinker3) achieves **53% on AIME 2025** ([17] huggingface.co) – a new high for a model trained on 100% open data. This matches or exceeds much larger models (DeepSeek's distilled 32B) on similar benchmarks, indicating that careful data curation can partly fill the gap to proprietary systems. Other researchers note that distilled expert answers on math tasks (as used in GPT-4's training) are extremely powerful; replicating those results with public data remains an active challenge.

- **Tool-Use and Code Execution:** As mentioned, allowing LLMs to run code (e.g. through a Python interpreter) has a dramatic effect. For instance, giving GPT-4 ("o4-mini") a Python sandbox boosted its AIME accuracy to **~99.5%** ([5] www.theregularizer.com). At OpenAI's GPT-5 reveal, the "Pro" variant (with parallel compute) solved **100% of AIME 2025** when Python tools were enabled ([6] venturebeat.com). In other words, code-aided models essentially *solve* the test. This underscores the importance of distinguishing pure reasoning capabilities from engineered "open-book" solutions. It also foreshadows a future division of benchmarks into raw reasoning versus tool-augmented modes.

Table 2 (below) summarizes some of these approaches and their reported impacts on AIME. The takeaway is that **multiple orthogonal methods** – improved prompting, smarter inference, fine-tuning, data, and tools – are all being combined to push this benchmark. The result: closed-book accuracies in the low-to-mid 90s% are now common, and with tools we approach 100%. But each of these methods also reveals new trade-offs (e.g. efficiency, openness, verifiability) that are driving research directions.

Technique/Approach	Key Idea	Effect on AIME (2025)
Chain-of-Thought Prompting	Instruct model to "think step by step."	Fundamentally improves multi-step reasoning; underlies many baseline scores (e.g. GPT-5's ~94.6%). Enhances accuracy over naive prompting. ([11] www.theregularizer.com)
Majority-Vote / Self-Consistency	Sample many CoT traces and pick most frequent answer.	Improves reliability; typical practice for top models. Reduces variance of answers (used in GPT-4 tech report, etc.). Gains several points but at high cost.
Deep Confidence (DeepConf)	Prune low-confidence reasoning paths via model's own confidence.	*Reported:* 99.9% accuracy on AIME 2025 ([14] www.researchgate.net) (using GPT-OSS-120B), with ~84.7% fewer tokens. Essentially eliminates remaining errors via smart filtering.

Technique/Approach	Key Idea	Effect on AIME (2025)
Parallel Pruning (DeepPrune)	Dynamically stop duplicate chains during parallel inference.	*Reported:* ~80% reduction in wasted tokens vs. full parallel reasoning, with ≤3% accuracy loss on AIME ([18] huggingface.co). Makes high-accuracy inference far more efficient.
Peer Reasoning (LeaP)	Multiple chains periodically share intermediate summaries.	*Reported:* + ~5 percentage points accuracy on average across AIME/HMMT ([15] openreview.net) (LeaP boosted a 32B model's AIME accuracy by ~5%). Helps models recover from early mistakes.
RL Fine-Tuning (GPPO, PACS, etc.)	Optimize model on math-reward with advanced policy updates.	*Reported:* Large accuracy jumps, e.g. +14–17 points on AIME for 7B/14B models ([37] huggingface.co); 83.2% on AIME 2025 ([16] www.researchgate.net) for an 8B model; 59.8% pass@256 for 8B (PACS) ([31] huggingface.co). Substantial boost over supervised baselines.
Data Augmentation (OpenThoughts)	Build large open reasoning datasets and train models on them.	*Reported:* 53% accuracy on AIME 2025 with a 7B model ([17] huggingface.co), a state-of-the-art result for fully open-trained models. Demonstrates the value of specialized math/CoT data.
Expert Distillation (DeepSeek)	Distill large closed-book models into smaller ones for math tasks.	*Reported:* A 1.5B "DeepSeek-Qwen" model reached 43.1% on AIME 2024 ([27] medium.com) (improving 15 points over base) and outperformed GPT-4 on that set. Shows RL/distillation can empower small models.
Python Tool Use	Allow execution of Python/math code during reasoning.	*Reported:* ~99.5% accuracy for GPT-4 (with Python) ([5] www.theregularizer.com), and 100% for GPT-5 Pro (with Python) ([6] venturebeat.com). Essentially solves all problems by computation.
Distance-Limiting	Restrict model to solver-specific moves (e.g. linearization).	Used in some systems to guide stepwise solving; gains are task-specific. (Example: fine-tuning on step sequences.)
Others (e.g. Socratic CoT, reasoning decompositions)	Various prompt engineering and decomposition strategies.	These often yield incremental gains on individual problems. The field continues to explore creative prompting.

Table 2. Recent techniques and their reported impact on AIME 2025. See cited sources for details: DeepConf ([14] www.researchgate.net), DeepPrune ([18] huggingface.co), LeaP ([15] openreview.net), Klear-Reasoner ([16] www.researchgate.net), OpenThoughts ([17] huggingface.co), DeepScaleR ([27] medium.com), etc. (Note: some entries like Python tool use appear in blog reports ([5] www.theregularizer.com) ([6] venturebeat.com).)

This survey makes clear that *there is no single "magic bullet"* – top performance arises from combining multiple ideas. In practice, the best systems use chain-of-thought prompting **plus** ensemble techniques (self-consistency, pruning, confidence filtering) and often specialized fine-tuning. OpenAI's GPT-5, for example, incorporates an enormous supervised training corpus (including likely AMC/AIME problems) and powerful inference tricks; the result is ~94.6% accuracy ([2] www.snarful.com) without an explicit math tool. Other groups have shown that similar benefits can be achieved along different axes (e.g. emphasizing RL or data). As a result, models are now capable of solving *most* AIME problems reliably.

Case Studies and Examples

Below are a few illustrative examples of how specific models and research projects fared on AIME 2025, highlighting different perspectives:

- **OpenAI's GPT-5 (Aug 2025 Event):** In August 2025, OpenAI publicly announced GPT-5's capabilities. They reported that GPT-5 achieved **94.6% accuracy on AIME 2025 (closed-book)** ([2] www.snarful.com). Importantly, this was emphasized as a major benchmark result in the press release. Even more striking, OpenAI noted that *with* code execution (a Python API), GPT-5 Pro "solved 100%" of AIME questions ([6] venturebeat.com). In short, GPT-5 essentially mastered the exam both in pure reasoning mode (94.6%) and in tool-augmented mode (100%). This case shows the cutting edge of closed-book math AI. OpenAI's dual reporting (closed vs. open) also underscores the need to distinguish "tool-augmented reasoning" from vanilla model skill.

- **Grok 3/4 (xAI Model):** Elon Musk's xAI team recently released "Grok" models for math reasoning. The Regularizer blog cites Grok-3 (175B) hitting **93.3% accuracy** on AIME 2025 ([21] www.theregularizer.com) (pending independent verification), and the Vals leaderboard lists Grok-4 at ~90–91% ([4] www.vals.ai). These results track similarly to GPT-5, confirming that multiple research labs are converging on ~90+% with large models. Grok's performance also exemplifies the "chain-of-thought" strategy: Musk's team has advertised that Grok uses an "improved reasoning style" akin to the Socratic method. Regardless of specifics, Grok's case illustrates that state-of-art closed-book models from different organizations now achieve very high AIME scores.

- **Open-Source Leader – Qwen/DeepSeek:** On the open side, Qwen-based models have climbed quickly. For instance, the *Klear-Reasoner* 8B model (trained openly by Zhenpeng Su et al.) scored **90.5% on AIME 2024** and **83.2% on AIME 2025** ([16] www.researchgate.net). Meanwhile, an 8B model fine-tuned by distributed RL (the "DeepScaleR" or DeepSeek-R1-distilled) achieved **43.1% on AIME 2024** compared to 28.8% base ([27] medium.com). Although 43.1% is far from GPT-level, it shows that a *1.5B model* can outperform GPT-4 on AIME with aggressive RL. More recently, the *OpenThoughts3-7B* model (trained on 1.2M reasoning examples) reached **53% on AIME 2025** ([17] huggingface.co). These cases demonstrate the "long tail" of community efforts: while closed models top the charts, open models are making rapid progress through clever training and fine-tuning.

- **Independent Analysis (Epoch AI):** A data-analytics firm, Epoch AI, released a public chart comparing LLM math performance. They note that publicly available models now solve virtually all AIME problems rated 1–5, but only **~67%** of those rated 6 (the hardest) ([8] epoch.ai). Their analysis highlights that GPT-5's remaining errors come on the most difficult problems, and that no model solves all high-difficulty items ([8] epoch.ai). This independent perspective aligns with our benchmark data and underscores that the AI "gap" remains only on the toughest challenges.

- **Human vs. AI Lift:** For context, recall that average human students get ~26–40% ([3] aiwiki.ai). All top AI models now surpass the top humans by a wide margin. Indeed, as one report puts it, students scoring in the top tier on AMC-level contests now face AI that routinely has near top math performance ([8] epoch.ai) ([9] www.theregularizer.com). This dramatic shift in performance is cause for both excitement (AI progress) and caution (the need to ensure AI truly understands its reasoning).

These case studies illustrate the multipronged nature of advances: different models (GPT, Grok, Qwen, Claude) and different methods (RL, dataset creation, tools) are all contributing to today's high scores. They also hint at the future: GPT-5 sets a new bar, but open research means that others (e.g. DeepMind's "Deep Think" models) may soon match or exceed these results. Ongoing benchmarking (e.g. via MathArena or new leaderboards) will track these head-to-head, ensuring continued transparency.

Discussion: Implications and Future Directions

The rise of AIME 2025 as a benchmark has several broad implications for AI research, deployment, and ethics:

- **Advancing the Frontier:** The benchmark demonstrates that LLM reasoning is accelerating at an astonishing rate. As of 2025, systems that barely managed single-digit accuracy a couple of years ago now routinely score above 90%. The Regularizer blog observes that in early 2023 AIs "could barely scratch AIME," but by early 2025 they "routinely outperform top humans" ([9] www.theregularizer.com). This suggests that what was once considered a superhuman or "Olympiad-level" intelligence is now becoming commonplace. For researchers, this is both encouraging and sobering: it means that AI has mastered much of high-school math reasoning, but the remaining few problems (and the next level of proofs) are all that's left visible. Benchmarking will need to move on to stay ahead.

- **Future Benchmarks:** The obvious next step is **harder tasks**. Already, experts are proposing benchmarks like full Math Olympiad problems (proof-based) or multi-task challenges that combine reasoning with other domains. The Regularizer specifically mentions Olympiad proof problems and "AI decathlons" to keep pushing the frontier ([19] www.theregularizer.com). Indeed, if a well-engineered LLM+toolchain can solve AIME, the community will seek exams at the next level (USAMO, IMO, etc.). The Epoch AI analysis ([8] epoch.ai) shows that even GPT-5 Pro (with tools) solved all but one of the 2025 IMO problems. This trend implies that AI benchmarking will soon venture beyond high-school math into university-level or research-level problem solving.

- **Verifiability and Alignment:** As models get stronger, ensuring trust in their answers is critical. AIME-style questions have unique answers, so correctness is easy to verify *post hoc*. But in general, if AI outsources some reasoning to a Python tool, we need to be sure it doesn't silently code incorrectly. This has raised calls for verifiable chain-of-thought: e.g. providing step-by-step proofs that can be checked by humans or other tools. The Regularizer notes that stronger reasoning "magnifies the need for verifiable chains-of-thought and robust guardrails" ([20] www.theregularizer.com). In practice, many labs now require models to output their reasoning or to check their own work (as in DeepConf). Future benchmarks may include explicit verification steps (e.g. requiring the model to justify each answer). The broader alignment question looms: as AI solves harder math, how do we ensure it isn't just "gaming" the benchmark with shortcuts or hallucinations? The AIME benchmark itself is relatively objective, but applying similar AI to less structured domains (legal, medical, etc.) will demand similar rigor in reasoning transparency.

- **Democratization of Capability:** The rapid improvement of open models is lowering the bar for who can access this power. Just a year ago, only huge closed models could do advanced contest math. Now, public projects like OpenThoughts achieve 53% on AIME with a 7B model ([17] huggingface.co), and open 32B models (AM-Thinking) reach 74.4% ([34] huggingface.co). This suggests that advanced math reasoning will soon be a "commodity" feature of LLMs, not just a corporate secret. On the plus side, this democratization can accelerate innovation (everyone can build on these demos) and ensure accountability (results can be verified by outsiders). On the minus side, it means that applications requiring vetted reasoning (e.g. automated theorem proving, high-stakes decision aids) will be within reach of virtually any developer, raising issues of misuse or misunderstanding.

- **Practical Applications:** While AIME problems are academic curiosities, the same capabilities translate to real tasks: complex scheduling, engineering design, coding, or any domain requiring multistep logic. For example, better math reasoning aids in code generation (solving algorithmic problems), financial modeling, and scientific computing. Some companies are already integrating math-specialist LLMs into software tools (e.g. Copilot Advanced). In education, AIME-level AI tutors could emerge, helping students work through Olympiad problems. However, educators will face new challenges too: traditional math contests and exams will no longer differentiate humans from AI. The community may need to redesign assessments or focus on uniquely human skills (creativity, conceptual insights) rather than nuts-and-bolts problem solving.

- **Open Research Questions:** The AIME 2025 results spark many research questions. Why do certain problems still fool state-of-art LLMs? Analysis of human solvers suggests that AIME's hardest problems often involve novel insights or clever constructions; understanding how (or if) LLMs can rediscover those is an open area. Another question is *robustness*: how well do these models do on variations or perturbed questions? Given that AIME answers must be exact, even a tiny reasoning error causes a mark to fail – so AIME is a strict metric. Investigating why particular questions elicit errors could guide both model design and curriculum.

- **Future Directions:** The community is already moving on. We see proposals for combined benchmarks (math + code + language), more efficient reasoning algorithms, and initiatives for better evaluation. For instance, "AI Math Olympiad" style competitions (like OpenAI's recent V1 problem set) blend proof writing with multiple reasoning steps. As chain-of-thought techniques evolve (e.g. tree-of-thought, self-refinement), AIME will likely be one of the first tests those methods face. Given the pace, it is plausible that by 2026 a new set of metrics (safety, alignment, robustness) will be needed alongside simple accuracy.

Conclusion

The AIME 2025 benchmark has proven to be a pivotal milestone in AI research. In just a few years, LLMs have transformed from barely scratching the surface of these problems to casually outperforming human contenders ([9] www.theregularizer.com) ([3] aiwiki.ai). This rapid escalation underscores both the power and the maturity of

chain-of-thought reasoning approaches, advanced inference methods, and model scaling. Key findings of this report include: reliable demonstration of **≤95% closed-book accuracy** by top models (e.g. GPT-5 at ~94.6%) ([2] www.snarful.com); nearly perfect accuracy with tools (GPT-5 Pro with Python at 100%) ([6] venturebeat.com); and significant contributions from novel techniques such as confidence filtering and peer reasoning ([14] www.researchgate.net) ([15] openreview.net). Table 1 and Table 2 collected evidence from a wide range of sources to support these points.

At the same time, AIME 2025 remains an *active* test: it is not yet solved universally, and different approaches still yield different strengths. The fact that accuracy improvements continue (e.g. open models climbing further) means that AIME will keep driving innovation. Moreover, the benchmark's existence highlights important themes: the necessity of verifiable reasoning (as the Regularizer and others emphasize ([20] www.theregularizer.com)), the value of banning external aids to isolate raw ability, and the realization that complex reasoning is now within reach of open, widely accessible systems.

Looking forward, we expect new benchmarks to appear that push beyond AIME: Olympiad proof exams, multi-modal reasoning tasks, or "AI Olympiads" explicitly crafted to stump today's models. For now, AIME 2025 stands as a **snapshot of AI reasoning** at a critical moment – revealing a landscape where machine intelligence has nearly mastered high-school math contests. As one summary noted: "by early 2025… frontier models are routinely outperforming top humans" ([9] www.theregularizer.com). The path ahead will be set by the remaining hard problems and how AI researchers choose to challenge them.

In closing, all claims in this report are backed by contemporary data and analysis (cited above). We have aimed to provide a comprehensive, deeply-researched perspective on AIME 2025 as an AI benchmark. The cited sources include official model announcements ([6] venturebeat.com) ([2] www.snarful.com) community benchmarks ([21] www.theregularizer.com) ([3] aiwiki.ai), and cutting-edge research papers ([15] openreview.net) ([16] www.researchgate.net). Together, they paint a detailed picture of where AI stands on this test, and where it seems headed in the near future.

External Sources

[1] https://aiwiki.ai/wiki/AIME_2025#:~:AIME%...

[2] https://www.snarful.com/news/openai-highlights-gpt-5-scores-on-math-coding-and-health-benchmarks-94-6-on-aime-2025-without-tools-74-9-on-swe-bench-verified-46-2-on-healthbench-hard-carl-franzen-venturebeat/#:~:OpenA...

[3] https://aiwiki.ai/wiki/AIME_2025#:~:Human...

[4] https://www.vals.ai/benchmarks/aime-2025-09-26#:~:1GPT%...

[5] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:%3E%2...

[6] https://venturebeat.com/ai/openai-launches-gpt-5-not-agi-but-capable-of-generating-software-on-demand/#:~:This%...

[7] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:4,a%2...

[8] https://epoch.ai/data-insights/math-competitions#:~:Publi...

[9] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:In%20...

[10] https://aiwiki.ai/wiki/AIME_2025#:~:Relea...

[11] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:,%28V...

[12] https://aiwiki.ai/wiki/AIME_2025#:~:Numbe...

[13] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:5%20%...

[14] https://www.researchgate.net/publication/394830749_Deep_Think_with_Confidence#:~:trace...

[15] https://openreview.net/forum?id=qImzresc4C#:~:LeaP....

[16] https://www.researchgate.net/publication/394439826_Klear-Reasoner_Advancing_Reasoning_Capability_via_Gradient-Preserving_Clipping_Policy_Optimization#:~:sampl...

[17] https://huggingface.co/papers?q=AIME-2025#:~:data%...

[18] https://huggingface.co/papers?q=AIME+2025#:~:a%20s...

[19] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:Looki...

[20] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:probl...

[21] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:Rank%...

[22] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:2%20%...

[23] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:Why%2...

[24] https://medium.com/%40annie_7775/deepseeks-abysmal-performance-with-the-aime-2025-math-benchmark-688bb8598d12#:~:The%2...

[25] https://aiwiki.ai/wiki/AIME_2025#:~:Numbe...

[26] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:Rank%...

[27] https://medium.com/%40annie_7775/deepseeks-abysmal-performance-with-the-aime-2025-math-benchmark-688bb8598d12#:~:%3E%2...

[28] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:no%20...

[29] https://epoch.ai/data-insights/math-competitions#:~:2%20%...

[30] https://aiwiki.ai/wiki/AIME_2025#:~:Human...

[31] https://huggingface.co/papers?q=AIME-2025#:~:stron...

[32] https://aiwiki.ai/wiki/AIME_2025#:~:Basel...

[33] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:%3E%2...

[34] https://huggingface.co/papers?q=AIME-2025#:~:innov...

[35] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:time,...

[36] https://www.theregularizer.com/blog/aime-2025-benchmark-results#:~:What%...

[37] https://huggingface.co/papers?q=AIME-2025#:~:train...

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies including Scilex Holding Company (SCLX) and leading CROs across North America.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at https://intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.