MathPath AI™ . . . Math Made Possible.

ACCURACY VALIDATION REPORT

mathpathai.com

Independent verification of question-bank accuracy · Updated June 9, 2026

SUMMARY OF RESULTS

Questions sampled	**400**
Questions passed	**397**
Confirmed errors	**0**
Uncertain (genuine ambiguity)	**3**
Pass rate	**99.25%**
95% confidence interval (Agresti–Coull)	**97.7% – 99.9%**

WHAT WE TESTED

Every question in the MathPath AI question bank has a stored correct answer. To verify those answers are correct, we drew a random sample of 400 questions from the live bank — which at the time of this audit contained 5,374 questions — and had an independent AI model solve each one from scratch, without ever being shown the stored answer. We then compared what the model independently calculated to what the bank says the answer is.

HOW THE TEST WORKS

This is called a **blind-solve audit**. The auditing model receives only the question and the answer choices, exactly as a student would see them. It works through the problem on its own and records its answer. The process has two passes:

- **Pass 1** — the model solves the question independently. If it agrees with the stored answer, the question is marked PASS.
- **Pass 2** — if Pass 1 disagrees, a second independent solve is run. If both passes agree with each other — and both disagree with the stored answer — the question is marked FAIL (confirmed error).
- **Uncertain** — if Pass 1 and Pass 2 disagree with each other, the question is marked UNCERTAIN. This reflects genuine difficulty or ambiguity in the question, not a clear error in the stored answer.

Importantly, the auditing model is different from the model that generates questions. **No model ever validates its own output.**

THE RESULT

397 of 400 sampled questions passed. Zero were confirmed errors. Three were flagged as uncertain — meaning two independent AI solvers reached different conclusions on the same question, suggesting the question may be genuinely challenging or ambiguous rather than wrong. Uncertain questions are flagged for human review.

Pass rate: 99.25%

Using the Agresti–Coull method — a statistical technique specifically designed for accuracy rates near 100% — we can state with 95% confidence that the true accuracy of the full question bank falls between **97.7% and 99.9%**.

In plain terms: if you tested every single question in the bank, we are 95% confident the accuracy rate would be no lower than 97.7%.

WHY A SAMPLE OF 400 IS SUFFICIENT

The relationship between sample size and the precision of a confidence interval follows a square-root curve: to cut your margin of error in half, you need to quadruple the sample size. The jump from 100 to 400 questions is enormously productive; beyond 400, each additional question contributes diminishing returns.

At our observed pass rate of 99.25%, the variance of the estimate ($p \times (1-p)$) is approximately 0.0074 — roughly 34 times smaller than the worst-case variance that occurs at $p = 0.50$. This means our confidence interval at $n = 400$ is already very tight, and a larger sample would not materially change the conclusions.

CONSISTENCY ACROSS FIVE INDEPENDENT MEASUREMENTS

We have now measured accuracy five times — including two independent cross-validations using Google Gemini 2.5 Flash, a model from a completely separate AI company. All five results have converged at the same figure or better:

Measurement	Questions tested	Pass rate
Random sample	200	98.5%
Full-bank audit	3,634	98.7%
Random sample (5,374-question bank)	400	99.25%
Independent cross-validation (Google Gemini 2.5 Flash, May 30, 2026)	500	99.6%
Automated weekly cross-validation (Google Gemini 2.5 Flash — all new questions, ongoing from June 2026)	~300/week (all survivors)	Ongoing

When independent measurements taken at different times on different question sets keep producing the same answer, that consistency is itself strong evidence the rate is stable — not a lucky draw from a favorable slice of the bank. The slight improvement in the most recent measurement reflects questions identified and removed through prior audits, progressively raising the quality floor.

METHODOLOGY NOTE

- **Sampling strategy:** stratified random sample weighted 2:3:1 across Foundational, Test Ready, and Challenge difficulty tiers, reflecting student exposure patterns (Test Ready questions are the most frequently served).
- **Verification model:** Anthropic Claude Opus — a separate, higher-capability model from the one used to generate questions.
- **Confidence interval method:** Agresti–Coull (1998). This method adjusts the sample proportion by adding a mathematically determined number of pseudo-observations to the numerator and denominator before computing the interval, producing statistically reliable coverage even when the true proportion is near 0 or 1, where the standard Wald interval breaks down.

The audit toolchain is included in the platform codebase and can be re-run at any time to produce a fresh, independently verifiable result.

INDEPENDENT CROSS-VALIDATION — GOOGLE GEMINI 2.5 FLASH

Date: May 30, 2026

- **Sample:** 500 questions, stratified across 5 creation-date strata reflecting different generation/verification model combinations used over the platform's history.
- **Method:** Google Gemini 2.5 Flash independently blind-solved each question; responses compared to stored correct answers.
- **Result:** 485 of 500 questions returned valid responses (15 were malformed data rows excluded from analysis).
- **Match rate:** 99.6% (482 of 485 valid questions confirmed correct).

Findings:

- 1 genuine error found and deleted (statistics / Challenge tier — a median question with a wrong stored answer).
- 1 Gemini error identified (number-types / Test Ready tier — time arithmetic, where Gemini failed to carry seconds correctly; our stored answer was confirmed correct).

Questions created before the higher-rigor Opus verification layer was introduced achieved a 100% match rate in this audit — confirming that base question quality was already high before the additional verification was added.

Key conclusion: two independent AI families from different companies — Anthropic Claude Opus and Google Gemini — have now confirmed 99%+ accuracy across approximately 885 sampled questions combined. The Gemini audit specifically demonstrates that the accuracy result is not an artifact of using Claude to verify Claude-generated questions.

AUTOMATED WEEKLY VERIFICATION — A PERMANENT THIRD LAYER

Beginning June 2026, Gemini 2.5 Flash cross-validation runs automatically every Monday as part of the platform's weekly quality pipeline. This is not a periodic audit — it is a permanent, ongoing verification layer that runs without human intervention.

Each week, the pipeline works in two stages: first, a similarity audit removes questions that are too similar to others in the same slot, thinning the bank of near-duplicates. Second, Gemini independently blind-solves every question that survived the similarity filter — typically around 300 new questions per week generated by the auto-refresh system. Any mismatch between Gemini's answer and the stored answer triggers an email alert for manual review.

The practical effect: every question that enters the live bank has now been independently verified by two separate AI families — Anthropic Claude Opus (during generation, via IVQG™) and Google Gemini (after the similarity filter, via automated weekly cross-validation). Neither model ever sees the other's answer. This eliminates the shared-blind-spot concern that arises when a single AI family both generates and verifies its own output.

- **Cost:** approximately $0.05 per week at current Gemini 2.5 Flash pricing — essentially no operating cost, while providing an independently auditable quality record that accumulates week over week.
- **Implementation:** an automated weekly workflow runs the Gemini cross-validation against all new questions; results are logged and emailed to administrators.