GAUSS Eval: Human–LLM Judge Consistency Analysis

Tianzhe Chu[2*] Jiaxin Zhang[1,3*] Zhenyu Liao[4*] Qiuyu Ren[5] Tahsin Saffat[5]
Zitong Yang[6] Yi Ma[2,5] Yue Zhang[1*]

[1]Hyperbolic [2]University of Hong Kong [3]California Institute of Technology [4]Nanjing University
[5]University of California, Berkeley [6]Stanford University
[*]Core contributors

Abstract

We present a comprehensive statistical analysis of the consistency between human and Large Language Model (LLM) judges in evaluating proof-based mathematical problems, benchmarking 14 state-of-the-art models on the MathArena USAMO 2025 dataset. By decomposing judgment behavior through error-accuracy, correlation, and distributional metrics, we show that top performing models exhibit distinct grading philosophies - **DeepSeek-Math-V2** grades strictly, **Gemini-3-Pro** is precise on partially correct problems, and **GPT-5** demonstrates balanced performance. We further identify a consistent "leniency bias," whereby most models *over-credit* structurally plausible yet mathematically flawed solutions. Our analysis further demonstrates that LLM judges produce significantly *higher* entropy score distributions than humans and struggle with local verification, often hallucinating logical bridges rather than penalizing omitted justifications. We hope our findings will help the community understand LLM-as-a-judge and math reward models better.

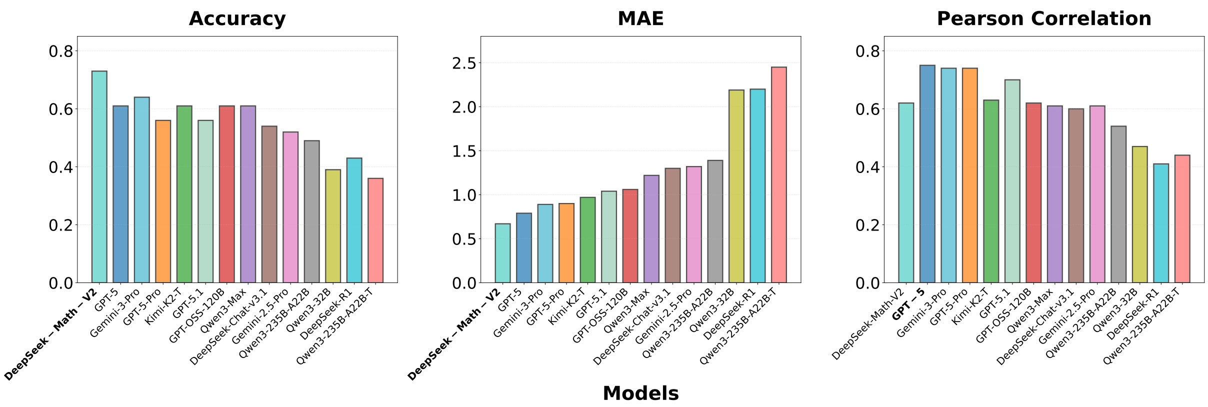

Figure 1 | **Benchmarking LLM graders on MathArena USAMO 2025.** Best model is **bolded**.

Website https://gaussmath.ai/eval.html

Github https://github.com/Gauss-Math/GAUSS-Eval

Run Logs Google Drive log link

Contents

1. Introduction

Large Language Models (LLMs) are increasingly used not only to **solve** mathematical problems, but also to **judge** them — grading long-form, proof-style responses produced by humans or other models.

Yet a fundamental question remains:

> *Can an LLM judge math problem answers like a human does?*

This report presents a comprehensive **statistical analysis of human–LLM grading consistency**. It investigates how current LLM judges align or diverge from human grading standards across key statistical dimensions.

The analysis decomposes judgment behavior into layered diagnostics to reveal not only whether LLM judges agree with humans, but also **how and why** their scoring behavior differs across distinct validity and difficulty regimes.

Our analysis reveals a clear performance pattern across current LLM judges.

Specifically, GPT-5 and Gemini-3-Pro-Preview exhibit the strongest alignment with human judges among closed-source models, whereas DeepSeek-Math-V2 and Kimi-K2-Thinking constitute the most reliable and human-aligned judges within the open-source models.

However, despite these encouraging aggregate trends, all existing models still exhibit substantial limitations in fine-grained evaluative accuracy. These observations also point directly to our next research direction: strengthening LLMs' ability to evaluate mathematical reasoning at a local level. In particular, future judge models need to:

1. more accurately verify the correctness of specific argument steps, rather than relying on global coherence;
2. more reliably identify missing or skipped reasoning, avoiding the tendency to implicitly fill in gaps.

Taken together, these goals aim to move beyond coarse approximation and toward LLM judges that analyze mathematical work with human-level granularity and rigor.

2. Problem Formulation

2.1. Grading Proof-based Problems

Given a mathematical problem q, a candidate solution a, and the rubric r, grading process for proof-based problems can be formalized as a scoring function:

$$f_\theta(q, a, r) \to s \in \mathcal{S}$$

where f_θ can be either a human grader or an LLM grader, and $\mathcal{S}$ represents the scoring space corresponding to the given rubric. In this work, we follow the USA Mathematical Olympiad (USAMO) rubric where $\mathcal{S} = [0, 7]$.

2.2. Evaluation Desiderata

We leverage three sets of metrics to evaluate LLM graders. In Section 2.2.1, we first introduce standard error-accuracy metrics for LLM graders. We then present *correlation metrics* and *distributional metrics* for more fine-grained analysis.

2.2.1. *Error–Accuracy Metrics.*

Error–accuracy metrics quantify the deviation between model predictions and ground-truth values on a per-example basis. They directly measure how close the predictions are to the correct outputs.

Given ground-truth targets $\{y_i\}_{i=1}^N$ and model predictions $\{\hat{y}_i\}_{i=1}^N$, we consider:

Mean Absolute Error (MAE).

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i| \,. \tag{1}$$

Mean Squared Error (MSE).

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \,. \tag{2}$$

Root Mean Squared Error (RMSE).

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}. \tag{3}$$

Accuracy Within k. For regression or ordinal prediction tasks, accuracy within k measures the proportion of predictions whose absolute error does not exceed a tolerance threshold k:

$$\mathrm{Accuracy}_{\le k} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\, |\hat{y}_i - y_i| \le k \,\} \,. \tag{4}$$

This metric gives a flexible notion of correctness by allowing predictions to be considered correct if they fall within an acceptable error margin.

2.2.2. *Correlation Metrics.*

Correlation metrics assess the degree to which model predictions preserve the trend or ordering of the ground-truth data, largely independent of scale.

Pearson Correlation Coefficient.

$$r = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}})^2}\,\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}}. \tag{5}$$

Spearman Rank Correlation Coefficient. Let d_i be the difference between predicted and true ranks. Then

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}. \tag{6}$$

Coefficient of Determination (R^2). In an ordinary least squares (OLS) linear regression model, the R^2 score measures the fraction of variance in the response variable explained by the fitted model. Let

$$\mathrm{TSS} = \sum_{i=1}^{N} (y_i - \bar{y})^2, \qquad \mathrm{RSS} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2.$$

Then

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}. \tag{7}$$

Higher values indicate better explanatory power of the linear model.

2.2.3. *Distributional Metrics.*

Distributional metrics evaluate how well the predicted distribution matches the empirical distribution of the ground truth. They characterize global distributional properties such as divergence, uncertainty, and variability.

Let P be the true distribution and Q the model distribution.

Entropy Ratio.

$$H(P) = -\sum_x P(x) \log P(x), \qquad \mathrm{EntropyRatio}(P, Q) = \frac{H(Q)}{H(P)}. \tag{8}$$

Relative Variance. when comparing two distributions,

$$\mathrm{RelativeVariance}(P \parallel Q) = \frac{\mathrm{Var}_Q[X]}{\mathrm{Var}_P[X]}. \tag{9}$$

3. Key Findings

We present four major findings on LLM-as-a-judge performance on **MathArena USAMO 2025**. We begin with overall model benchmarking and distributional analysis in Section 3.1 and Section 3.2. We then examine how models behave differently on zero versus non-zero scored solutions in Section 3.3, and how grading precision varies across individual problems in Section 3.4.

3.1. Benchmark LLM-as-a-Judge

We begin our analysis by benchmarking the 14 models (see Figure 1 and Tables 1 to 3) DeepSeek-Math-V2 emerges as the best overall model, achieving the lowest MAE (0.67), highest accuracy (0.73), and competitive Pearson correlation (0.62). This suggests it excels at precise numerical predictions. GPT-5, GPT-5-Pro, and Gemini-3-Pro form a cluster of high-performing models with MAE values below 1.0, accuracy around 0.56–0.64, and Pearson correlations of 0.74–0.75. These models balance well across all metrics.

We observe that numerical accuracy alone does not fully capture the models' grading capabilities: while 4 models (GPT-5, Kimi-K2-T, GPT-OSS-120B, Qwen3-Max) achieve high accuracy (0.61), their mean absolute errors vary substantially (from 0.79 to 1.22).

Interestingly, although GPT-5-Pro and Qwen3-235B-A22B-Thinking consume more inference-time compute, they do not perform as well as their counterparts with less reasoning effort (GPT-5 and Qwen3-235B-A22B-Instruct).

3.2. LLM Grades More Diversely than Human

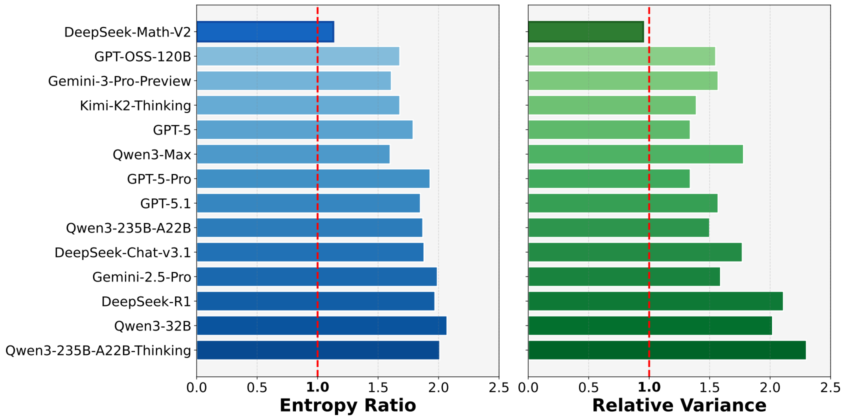

Figure 2 | **Distributional metrics on MathArena USAMO 2025.** (Closer to 1 is Better)

The global distributional pattern (Figure 2 and Tab. 3) reveals a systematic expansion in both entropy and variance across all models. Modern LLMs no longer produce narrowly concentrated score distributions; instead, they exhibit broader and more expressive grading behavior compared with human judges.

DeepSeek-Math-V2 behaves exceptionally good in these two metrics with closer to 1 value on ER (entropy ratio) and RV (relative variance).

Except that, all models demonstrate elevated ER values (typically in the range 1.6–2.0) (Figure 2 Left), indicating a clear increase in information entropy relative to the human reference. This suggests that model scoring has become less concentrated and more exploratory, assigning probability mass across a wider range of plausible evaluations.

The RV values (mostly between 1.3–1.9) (Figure 2 Right) confirm a consistent expansion in score variance. Compared with human judges, models tend to exaggerate contrasts between good and bad answers, displaying a mild but systematic over-dispersion bias.

3.3. LLM Gives Lenient Grades

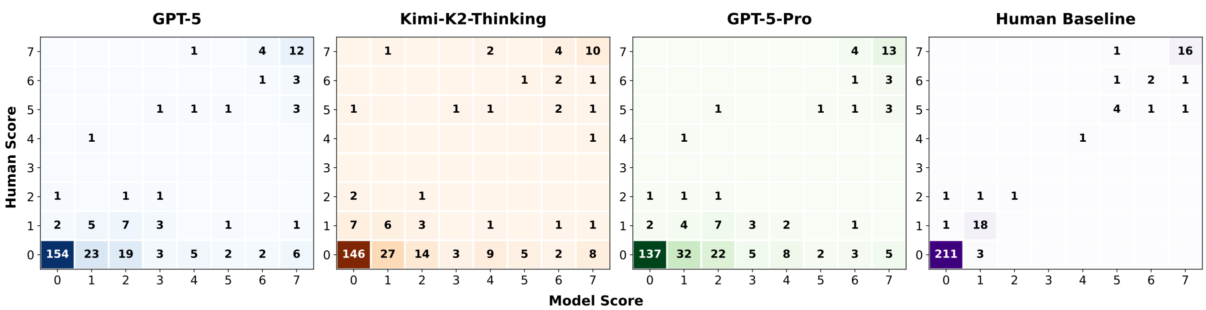

Figure 3 | **Confusion Matrices of 3 Models and Human Baseline.** Non-integer scores are rounded to the nearest even integer. A confusion matrix is a table where each cell (i, j) is the frequency with which a sample rated i by humans receives a score j from the model.

LLMs exhibit systematic leniency bias when grading mathematical solutions. The confusion matrices in Figure 3 reveal that the lower-right triangles (where LLMs assign higher grades than

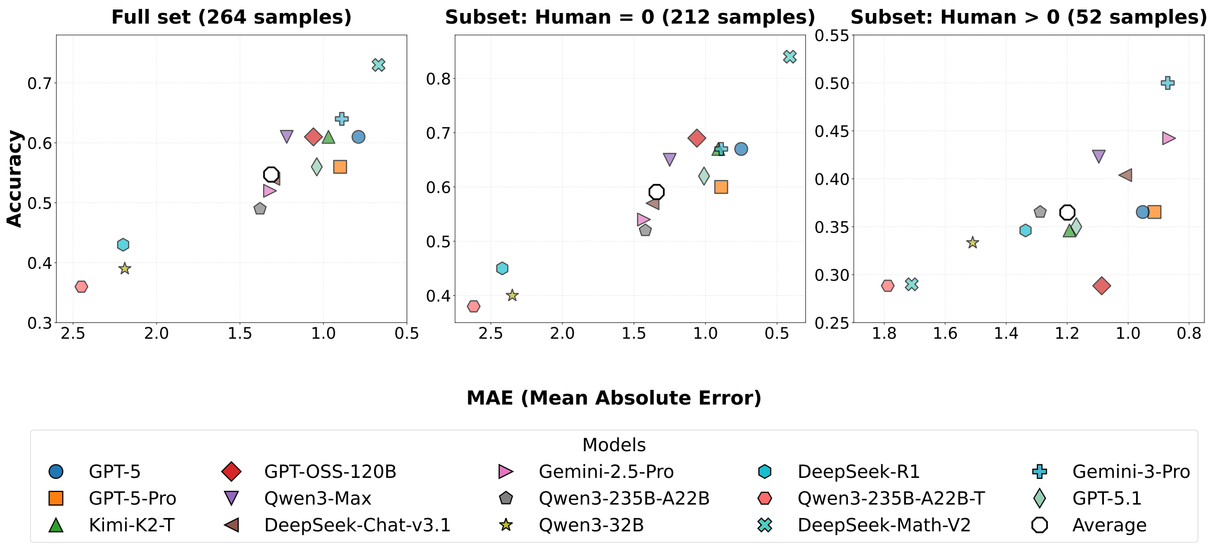

Figure 4 | ***MAE vs. Accuracy* on MathArena USAMO 2025.** (3 recent models included)

humans) are consistently denser than the upper-left triangles (where LLMs assign lower grades), indicating that models systematically over-credit solutions relative to human expert judgment.

This leniency, however, manifests differently depending on solution quality. USAMO problems are exceptionally difficult, with most LLM solutions receiving human grades of 0. This creates a natural partition for analyzing how graders behave on zero-scored versus non-zero-scored solutions, revealing nuanced patterns in LLM grading behavior.

LLM graders show contrasting performance patterns across these two subsets (see Figure 4). On solutions that humans graded as 0, models achieve higher accuracy but higher MAE compared to their performance on the full dataset. The pattern reverses for solutions with positive human scores: accuracy drops while MAE improves.

This divergence reflects fundamental differences between the two metrics. MAE measures numerical proximity to human scores, while accuracy captures categorical correctness (i.e., exact score band matching). On non-zero solutions, models track human scores more closely on average, producing smaller numerical deviations and lower MAE. However, the non-zero score distribution is denser and more granular—distinguishing among 3, 4, 5, or 6 points rather than simply 0 versus non-zero—making band misclassification more likely even with small numerical errors, thereby reducing accuracy.

Individual models exhibit distinct grading philosophies. DeepSeek-Math-V2 grades exceptionally strictly, achieving the best performance on the **Human = 0** subset (10% higher accuracy than the second-best model) but poor performance on **Human > 0**. Since MathArena-USAMO2025 is heavily skewed with 80% of samples receiving human grades of 0, DeepSeek-Math-V2's strict grading style makes it the best-performing model overall. The Gemini family shows the opposite pattern—Gemini-3-Pro and Gemini-2.5-Pro rank as the top two models on the **Human > 0** subset while underperforming on the **Human = 0** subset. GPT-5 demonstrates balanced performance across both subsets, though it still exhibits the general leniency bias observed across all models.

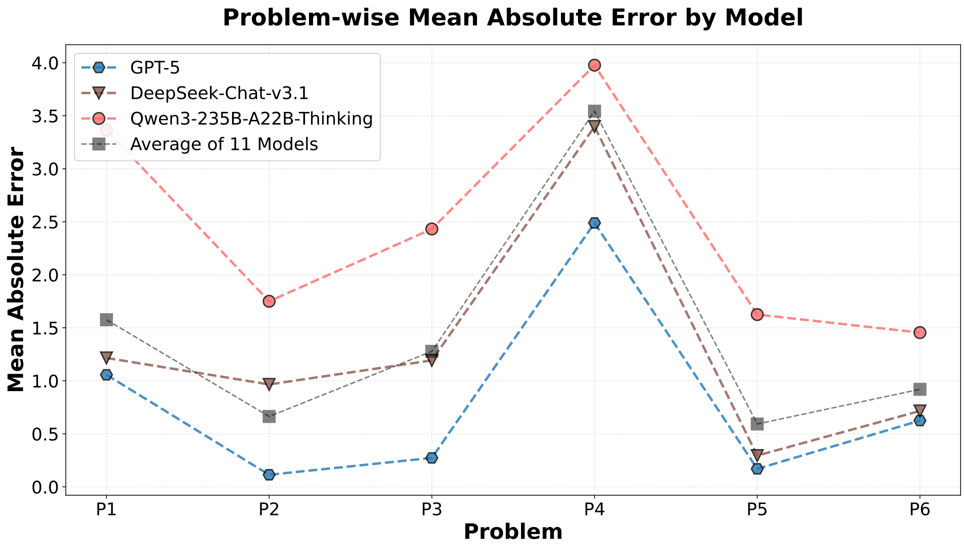

Figure 5 | **Problem-wise Mean Absolute Error.**

3.4. Grading Precision Correlates with Problems

Beyond the systematic leniency bias observed above, LLM grading performance also exhibits strong problem-dependent variation. LLM grader precision varies considerably across problems (see Figure 5). Remarkably, models of vastly different overall ranks in Table 1 (2st, 9th, 14th) exhibit nearly identical problem-wise performance patterns: strong grading precision on P2, P3, and P5, but consistently weaker performance on P4. This pattern becomes even more pronounced in Table 5, where top models (GPT-5, GPT-OSS-120B, Kimi-K2-Thinking) achieve near-perfect accuracy (> 0.80) on P2, P3, and P5, yet all struggle with P4.

This problem-specific difficulty is independent of the zero-score pattern analyzed in Section 3.3. We verify that P4's grading difficulty does not correlate with the proportion of zero-scored solutions—in fact, according to human graders, P4 has the second-highest average score and second-fewest zero grades among all problems, suggesting models struggle with P4 despite (or perhaps because of) its solutions containing substantial mathematical reasoning.

4. GPT-5 Case Study

4.1. Overall Summary

We evaluate GPT-5's grading consistency on 264 USAMO 2025 solutions. Overall, GPT-5 demonstrates good global alignment with human judgment, achieving a Pearson correlation of 0.746, indicating that the model reliably reproduces human scoring trends across the full score range. Error metrics further show that 91.7% of predictions fall within ±2 points of human scores, suggesting that GPT-5 is well-suited for coarse-grained evaluation.

However, a more granular analysis reveals systematic deviations. In the zero-score discrimination layer, GPT-5 assigns non-zero credit to 28% of responses that humans judged invalid, indicating persistent over-crediting for structurally coherent but mathematically incorrect reasoning. Within the valid-answer regime, GPT-5's alignment strengthens substantially (Pearson = 0.833).

Problem-wise analysis for GPT-5 (see the GPT-5 curve in the Problem-wise Mean Absolute Error plot, Figure 5) reveals substantial heterogeneity across problems. GPT-5 performs exceptionally well on Problems 1 and 3 ($R^2 = 0.73$ and 0.805, respectively), where solution structures are stable and strongly cue-driven. In contrast, its performance deteriorates on Problem 4, where the model shows large positive bias (+2.26) and high divergence from human score distributions, suggesting difficulty with tasks requiring multi-path or less canonical solution strategies. Notably, Problem 2 displays an unusual pattern: despite achieving an lowest MAE (0.114) and near-perfect ±1 accuracy, its R^2 is nearly zero. This stems from the fact that human scores for this problem exhibit almost no variance—most responses receive a score of zero. GPT-5 mirrors this degenerate distribution, producing predictions that minimize absolute error but fail to explain any meaningful variation. As a result, the model appears highly accurate in an absolute sense while contributing no explanatory power, highlighting the limitations of variance-based metrics under heavily skewed scoring regimes.

Taken together, our results paint GPT-5 as a capable but systematically biased math judge. At the global level, it correlates well with human scores and achieves high coarse-grained accuracy. However, the two case studies reveal a consistent failure mode. In Problem 1, GPT-5 over-credits a solution with clear logical gaps, implicitly "repairing" invalid steps into a plausible derivation; in Problem 4, it awards full credit to an answer that omits a central computation, treating the missing work as if it had been carried out correctly. In both cases, the model grades plausibility and surface structure rather than strictly checking what is actually written. This makes GPT-5 particularly weak at zero-score discrimination and at grading solutions that hinge on explicit, non-omittable justifications. In short, GPT-5 is well-suited for large-scale, approximate alignment with human scoring, but it is not yet a reliable substitute for human graders in strict evaluation settings.

4.2. Problem 1 Case Study: Over-Completion of Incomplete Proofs

GPT-5 assigns a substantially higher score (**5/7**) despite the student's omission of several rubric-critical components. By contrast, both human judges award only **1/7**, adhering strictly to the principle that credit is given solely for arguments explicitly stated in the submission. Although the student correctly formulates the digit recursion, they do not justify the floor/mod transformation, the required digit bounds, or the final choice of N; under the rubric, the absence of these arguments constitutes decisive logical gaps.

GPT-5, however, interprets the incomplete reasoning as indicative of conceptual progress and implicitly reconstructs intermediate steps. It grants partial credit for the floor/mod manipulation (**2/3**), for the bounding argument (**1/2**), and full credit for the structural conclusion (**1/1**). In effect, GPT-5 rewards the intended structure of the argument rather than the rigor of the written derivation.

The systematic discrepancy can be attributed to the following differences in evaluative philosophy:

1. **Preference for conceptual intent over formal rigor.** GPT-5 credits the presence of a promising methodological idea, whereas human graders require fully articulated and correct derivations.
2. **Local rather than global correctness.** Partial correctness on a single digit or a conjectured pattern is sufficient for GPT-5 to award partial credit; human judges require correctness uniformly across all digits.
3. **Acceptance of inferential vagueness.** Expressions such as "easy induction" are treated by GPT-5 as acceptable indications of intended reasoning, but human judges do not count

such statements as valid justification.

4. **Isolated evaluation of the conclusion.** GPT-5 evaluates the final step independently of earlier deficiencies, awarding full credit if the concluding structure appears sound; human judges withhold this point when foundational steps are incomplete.
5. **Holistic generosity.** GPT-5 tends to fill in missing arguments and interpret incomplete work charitably, whereas human grading is strictly confined to what is explicitly presented in the written solution.

4.2.1. Problem Statement and Student Answer (Answer ID: 1)

Problem Statement:

Let k and d be positive integers. Prove that there exists a positive integer N such that for every odd integer $n > N$, the digits in the base-$2n$ representation of n^k are all greater than d.

Equivalently, if

$$n^k = \sum_{i=0}^{r} a_i (2n)^i, \qquad 0 \le a_i < 2n,$$

show that for all sufficiently large odd n, we have $a_i > d$ for every i.

Student Answer:

Statement to prove. *Fix positive integers k and d. Then there is an N so that for every odd integer $n > N$, the base-$(2n)$ expansion of n^k has all of its "digits" strictly bigger than d.*

In more elementary language, writing

$$n^k = a_r(2n)^r + a_{r-1}(2n)^{r-1} + \cdots + a_1(2n) + a_0$$

with each digit $0 \le a_j < 2n$, we must show that for all sufficiently large odd n, every a_j is actually *greater* than d.

1. The least significant digit is always n.

A convenient starting observation is the following congruence:

$$n^k \equiv n \pmod{2n} \quad \text{whenever } n \text{ is odd.}$$

Indeed, one checks first that $n^1 \equiv n \pmod{2n}$ is trivially true. Then, assuming $n^k \equiv n \pmod{2n}$, one has

$$n^{k+1} = n^k \cdot n \equiv n \cdot n = n^2 \pmod{2n}.$$

But for odd n, it is also true that

$$n^2 \equiv n \pmod{2n},$$

since $n^2 - n = n(n-1)$ is clearly a multiple of $2n$ whenever n is odd. By induction,

$$n^k \equiv n \pmod{2n}$$

for all positive k.

From this congruence we see at once that

$$a_0 = n^k \bmod (2n) = n.$$

Hence the least significant digit of n^k in base $(2n)$ is always exactly n. In particular, once $n > d$, the digit a_0 exceeds d.

2. A recursion for the higher "digits."

To find the next digit a_1, we form

$$x_1 = \lfloor n^k/(2n) \rfloor, \qquad a_1 = x_1 \bmod (2n), \qquad x_2 = \lfloor x_1/(2n) \rfloor, \quad \text{and so on.}$$

In other words, once we know that

$$n^k = n + (2n)x_1,$$

we then define $a_1 = x_1 \bmod (2n)$ and $x_2 = \lfloor x_1/(2n) \rfloor$. Continuing likewise gives all the higher digits $a_2, a_3, \ldots$.

Concretely, since $n^k \equiv n \pmod{2n}$ implies $(n^k - n)$ is divisible by $2n$, we get

$$x_1 = \frac{n^k - n}{2n} = \frac{n(n^{k-1} - 1)}{2n} = \frac{n^{k-1} - 1}{2}.$$

Thus

$$a_1 = \left(\tfrac{n^{k-1}-1}{2}\right) \bmod (2n).$$

One may continue similarly for x_2 by dividing x_1 by $2n$, and so on.

3. Each higher digit also becomes large for large n.

A key point is that each "quotient" x_j is in fact quite large once n is large, yet still satisfies a simple congruence mod n. Indeed:

$$x_1 = \tfrac{n^{k-1}-1}{2} \equiv \tfrac{n-1}{2} \pmod{n},$$

because $n^{k-1} \equiv n \pmod{2n}$ implies $n^{k-1} - 1 \equiv n - 1 \pmod{2n}$, and hence also mod n. One finds that

$$x_1 = \tfrac{n-1}{2} + n\alpha_1 \quad \text{for some integer } \alpha_1.$$

Thus

$$a_1 = x_1 \bmod (2n) = \left(\tfrac{n-1}{2} + n\alpha_1\right) \bmod (2n).$$

Since $n\alpha_1 \equiv 0$ or $n \pmod{2n}$ depending on whether α_1 is even or odd, one sees that

$$a_1 \in \left\{\tfrac{n-1}{2}, \ \tfrac{3n-1}{2}\right\}.$$

Both of these are on the order of $n/2$. Consequently, once n is large enough so that $\frac{n-1}{2} > d$, we get $a_1 > d$.

Repeating the same division–remainder process for

$$x_2 = \lfloor x_1/(2n) \rfloor$$

one finds similarly that $a_2 = x_2 \bmod (2n)$ must lie in a small finite set of values, each of which again is on the order of $n/2^2$. And so on. In general one proves (by an easy induction) that a_j is always of the form

$$\frac{(2t+1)n - 1}{2^j},$$

for some integer t, and hence all those possible values grow linearly with n up to a factor $1/2^j$.

Since the base-$(2n)$ expansion of n^k has at most k or $k+1$ digits (in fact strictly fewer than $(2n)^k$ bounds the number of digits), we only need to check finitely many $j \in \{0, 1, \ldots, k-1\}$. For each such j, the expressions show a_j is eventually at least $n/2^j$ minus a small constant. Therefore, once n is large enough, each a_j exceeds d.

4. Conclusion: choosing N.

Gathering the above estimates together shows there is a finite threshold N (depending on k and d) so that for all odd $n > N$, each digit a_j of n^k in base $(2n)$ indeed satisfies $a_j > d$. Concretely, we must ensure

$$n > d, \qquad \tfrac{n}{2} > d, \qquad \tfrac{n}{2^2} > d, \ \ldots, \ \tfrac{n}{2^{k-1}} > d,$$

so taking

$$N = \max\{2^{k-1}d, \ d\} \quad \text{(an odd integer)}$$

works. Hence for all odd $n > N$, the base-$(2n)$ representation of n^k has every digit strictly bigger than d, completing the proof.

□

4.2.2. Rubrics

- **Initial closed form expression for each digit**, Max points: 1. Rubric content: Uses the base-$2n$ representation of n^k to find a closed form expression of each digit

$$a_i = \left\lfloor \frac{n^k}{(2n)^i} \right\rfloor \bmod (2n).$$

- **Interchanging the floor and mod**, Max points: 3. Rubric content: Transforms the digit expression into an easily boundable one by interchanging the floor and modulo operations.
- **Bounding the expression**, Max points: 2. Rubric content: Correctly bounds the aforementioned expression, usually by showing the numerator is divisible by n, so that

$$a_i \geq \left\lfloor \frac{n}{2^i} \right\rfloor.$$

- **Finishing**, Max points: 1. Rubric content: Chooses a provably sufficiently large N. The point is not awarded if it has not proven that n^k has at most k digits.

4.2.3. Grading Details

Human Judge 1

Points: 1/7

- **Initial closed form expression for each digit (1 pt)**
  - The model derives a non-trivial set of possible values of a_i.
  - **Points: 1 / 1**
- **Interchanging the floor and mod (3 pts)**
  - The model states the properties of the set can be proven by induction, but the model shows only first step of the induction, as well as the base case.
  - **Points: 0 / 3**

- **Bounding the expression (2 pts)**
  - The model does not formally prove that the set contains all possible values of a_i.
  - **Points: 0 / 2**
- **Finishing (1 pt)**
  - The model does not complete the solution in a rigorous manner.
  - **Points: 0 / 1**

Human Judge 2

Points: 1/7

- **Initial closed form expression for each digit (1 pt)**
  - The model finds a useful expression for each digit and has an idea, which can lead to a correct proof.
  - **Points: 1 / 1**
- **Interchanging the floor and mod (3 pts)**
  - The model makes a wrong calculation and statement.
  - **Points: 0 / 3**
- **Bounding the expression (2 pts)**
  - The model incorrectly bounds each digit.
  - **Points: 0 / 2**
- **Finishing (1 pt)**
  - The model doesn't finish a correct proof.
  - **Points: 0 / 1**

LLM Judge

Points: 5/7

- **Criterion 1 (Initial closed form expression for each digit, 1 pt)**
  Student: Writes $n^k = \sum a_j(2n)^j$ with $0 \le a_j < 2n$, and defines recursively $x_1 = \lfloor n^k/(2n) \rfloor$, $a_1 = x_1 \bmod (2n)$, $x_2 = \lfloor x_1/(2n) \rfloor$, etc., which implies $a_i = \lfloor n^k/(2n)^i \rfloor \bmod (2n)$.
  Evaluation: Correctly identifies the base-$(2n)$ digit formula via the floor/mod recursion.
  Points: 1/1
- **Criterion 2 (Interchanging the floor and mod, 3 pts)**
  Student: Uses congruence $n^k \equiv n \pmod{2n}$ to eliminate the floor at the first step and get $x_1 = (n^k - n)/(2n) = (n^{k-1} - 1)/2$, then analyzes a_1 via congruences mod n and mod $2n$. For higher digits, claims "continuing similarly" and an "easy induction" that a_j has the form $((2t+1)n - 1)/2^j$, but does not show the analogous interchange/elimination of floor beyond the first step.
  Evaluation: Successfully transforms the first digit beyond a_0 into an easily boundable form by leveraging exact divisibility, but does not carry out the detailed interchange/elimination for general i. Partial fulfillment.
  Points: 2/3
- **Criterion 3 (Bounding the expression, 2 pts)**
  Student: For a_1, derives $a_1 \in \{(n-1)/2,\ (3n-1)/2\}$, giving a clear lower bound $\approx n/2$. For general a_j, asserts they are of the form $((2t+1)n - 1)/2^j$ and hence "eventually at least

$n/2^j-$ (small constant)," but provides no rigorous derivation of this form or bound for all j.
Evaluation: Provides a correct bound for a_1 and a qualitative bound for general j without full justification. Partial credit.
Points: 1/2

- **Criterion 4 (Finishing, 1 pt)**
  Student: Notes $n^k < (2n)^k$, hence at most k digits, and chooses N ensuring $n/2^j > d$ for all $j \le k-1$ (e.g., $N \ge d \cdot 2^{k-1}$, with oddness condition).
  Evaluation: Correctly observes the bound on the number of digits and provides a sufficient N.
  Points: 1/1

4.2.4. *Criterion-by-Criterion Comparison*

1. Initial Closed-Form Expression (1 pt)

All three evaluations—Judge 1, Judge 2, and GPT-5—award full credit. The student's base-$(2n)$ expansion

$$n^k = \sum_j a_j(2n)^j, \qquad 0 \le a_j < 2n$$

and the associated floor–mod recursion are entirely correct. Hence no discrepancy exists for this criterion.

2. Interchanging Floor and Mod (3 pts)

Human Judge 1. Both judges give 0/3. Their reasoning:

- The student only carries out the base case and the first inductive step.
- The claimed induction ("continuing similarly") is not demonstrated.
- Judge 2 notes an incorrect statement in this part.

Since the criterion demands a general and rigorous treatment for all digits, the attempt is judged insufficient.

LLM Judge. LLM Judge awards 2/3. It notes:

- The student correctly removes the floor in the first step using $n^k \equiv n \pmod{2n}$, yielding an exact expression for the first quotient.
- This demonstrates a valid technique for simplifying the recursion.
- Although the method is not carried through for all indices, GPT-5 regards the partial demonstration as substantial progress.

Analysis of GPT-5 Misjudgement. GPT-5 evaluates this step by emphasizing the presence of a correct core idea. It interprets the successful manipulation of the first floor–mod step as evidence that the student understands the essential mechanism needed for generalization. Even though the full induction is missing, GPT-5 credits the student for demonstrating a meaningful and technically valid approach rather than requiring the argument to be fully executed.

3. Bounding the Digits (2 pts)

Human Judges. Both judges give 0/2:

- Judge 1 observes that the student never proves the set covers all possible a_i.
- Judge 2 states that the bounding is incorrect for the general case.

Thus the absence of a complete and correct argument yields zero credit.

LLM Judge. LLM Judge gives 1/2, reasoning that:

- The bound for a_1 is correctly derived.
- The asserted pattern $\frac{(2t+1)n-1}{2^j}$ is plausible though unproven.
- Taken together, this constitutes a partial bounding effort with some correct content.

Analysis of GPT-5 Misjudgement. GPT-5 tends to award partial credit for partially correct derivations. In this step, it recognizes that the student accurately bounded a_1 and attempted to generalize the pattern. Although the general case is unsupported, GPT-5 interprets the attempt as demonstrating conceptual direction rather than dismissing it entirely. Hence it assigns partial credit rather than zero.

4. Finishing the Proof (1 pt)

Human Judges. Both give 0/1, because:

- The proof is not rigorous in earlier steps.
- Therefore the final conclusion cannot be regarded as a legitimate completion.

LLM Judge. LLM Judge awards 1/1, noting that:

- The student correctly establishes that $n^k < (2n)^k$, giving at most k digits.
- They choose an N ensuring $a_j > d$ for all $j \leq k - 1$.
- Structurally, this completes the argument.

Analysis of GPT-5 Misjudgement. GPT-5 treats the proof's final step as logically independent of earlier deficiencies. Its scoring focuses on whether the student has assembled a coherent concluding structure: bounding the total number of digits and selecting N to enforce the desired inequality. Since the conclusion is formally correct, GPT-5 assigns full credit even though earlier parts contain gaps.

4.3. Problem 4 Case Study: Tendency to Reward Plausible but Incorrect Reasoning

In this case, the LLM judge assigns partial credit (**3/7**) even though the student's solution contains multiple unproven claims, incorrect geometric statements, and a missing critical identity that is explicitly assumed rather than derived. Both human judges, by contrast, assign a zero score because the solution fails to establish any of the required geometric constructions or properties and never completes the proof.

The discrepancy arises because the LLM interprets the student's argument as if several steps were present, valid, or sufficiently justified. It treats plausible intermediate statements—such as collinearity, midpoint relations, or power-of-a-point expressions—as though the student had genuinely proved them, even when the reasoning is absent or incorrect. Human judges, however, score only what is explicitly argued and penalize missing derivations and false claims as fatal errors that invalidate entire scoring categories.

- LLM judges treat mathematically plausible steps as if they were proven, overlook missing derivations, and under-penalize serious geometric errors.
- Human judges require explicit justification for every key construction property and consider unproven or incorrect claims as fully invalidating the argument.

4.3.1. Problem Statement and Student Answer (Answer ID: 4)

$\boxed{\textbf{Problem Statement}}$

Given an acute triangle ABC, let:

- H be the orthocenter,
- F the foot from C to AB,
- P the reflection of H across BC.

The circumcircle of $\triangle AFP$ intersects BC at two points X and Y.

Prove that C is the midpoint of XY.

$\boxed{\textbf{Student Answer}}$

Let A, B, C be the vertices of an acute triangle, and let H be its orthocenter. Let D, E, F be the feet of the altitudes from A, B, C to BC, CA, AB respectively. Let Γ be the circumcircle of $\triangle ABC$. Let P be the reflection of H across the line BC. A well-known property is that P lies on the circumcircle Γ. Let's prove this property. Let D be the foot of the altitude from A to BC. Since H lies on AD, AD is the line containing H perpendicular to BC. Since P is the reflection of H across BC, the line HP is perpendicular to BC, and the midpoint of HP lies on BC. This means the line HP is parallel to AD. Since H is on AD, P must lie on the line AD. Also, the midpoint of HP must be the intersection of HP and BC, which is D. So D is the midpoint of HP, which means $HD = DP$. Let $\angle ABC = \beta$ and $\angle ACB = \gamma$. In right-angled $\triangle CDH$, $\angle HCD = \angle FCB = 90° - \beta$. Thus, $\angle CHD = 90° - \angle HCD = 90° - (90° - \beta) = \beta$. Since D is the midpoint of HP and $CD \perp HP$, $\triangle CDP \cong \triangle CDH$ (SAS congruence: $CD = CD$, $\angle CDP = \angle CDH = 90°$, $DP = DH$). Thus $\angle CPD = \angle CHD = \beta$. Since A, B, C are on the circumcircle Γ, we want to show P is on Γ. A, B, C, P are concyclic if $\angle BPC + \angle BAC = 180°$ or $\angle CPA = \angle CBA$. Similarly, $\triangle BDP \cong \triangle BDH$ (SAS congruence: $BD = BD$, $\angle BDP = \angle BDH = 90°$, $DP = DH$). Thus $\angle BPD = \angle BHD$. In right-angled $\triangle BDH$, $\angle HBD = \angle EBC = 90° - \gamma$. So $\angle BHD = 90° - (90° - \gamma) = \gamma$. Thus $\angle BPD = \gamma$. Therefore, $\angle BPC = \angle BPD + \angle CPD = \gamma + \beta$. Since $\alpha + \beta + \gamma = 180°$, we have $\beta + \gamma = 180° - \alpha$. So $\angle BPC = 180° - \alpha$. Since H is inside the acute triangle, A and H are on the same side of BC. P is the reflection of H across BC, so P is on the opposite side of BC from A. The locus of points X such that $\angle BXC = 180° - \alpha$ on the opposite side of BC from A is the circumcircle Γ of $\triangle ABC$. Thus, P lies on Γ.

Let Ω be the circumcircle of $\triangle AFP$. We are given that Ω intersects the line BC at two distinct points X and Y. We want to prove that C is the midpoint of the segment XY. Let $P_\Omega(K)$ denote

the power of a point K with respect to the circle Ω. Let C be the origin on the line BC. Let x and y be the coordinates of X and Y. C is the midpoint of XY if and only if $x + y = 0$. The power of C with respect to Ω is $P_\Omega(C) = \vec{CX} \cdot \vec{CY} = xy$. If C is the midpoint, then $y = -x$, so $P_\Omega(C) = x(-x) = -x^2 = -CX^2$. Let D be the foot of the altitude from A to BC. The power of D with respect to Ω is $P_\Omega(D) = \vec{DX} \cdot \vec{DY}$. Let C be the origin and D have coordinate d. $X = x, Y = y$. $P_\Omega(D) = (x-d)(y-d) = xy - d(x+y) + d^2 = P_\Omega(C) - d(x+y) + d^2$. If C is the midpoint of XY, then $x + y = 0$. In this case, $P_\Omega(D) = P_\Omega(C) + d^2$. Using the vector notation, $\vec{DX} = \vec{CX} - \vec{CD}$ and $\vec{DY} = \vec{CY} - \vec{CD}$. $P_\Omega(D) = (\vec{CX} - \vec{CD}) \cdot (\vec{CY} - \vec{CD}) = \vec{CX} \cdot \vec{CY} - \vec{CD} \cdot (\vec{CX} + \vec{CY}) + CD^2$. $P_\Omega(D) = P_\Omega(C) - \vec{CD} \cdot (\vec{CX} +$
$vecCY) + CD^2$. C is the midpoint of XY iff $\vec{CX} + \vec{CY} = 0$. This is equivalent to $P_\Omega(D) = P_\Omega(C) + CD^2$. So we need to compute $P_\Omega(D)$ and $P_\Omega(C)$ and check if this relation holds.

Let's compute $P_\Omega(D)$. A, F, P are on Ω. A, D, P are collinear as shown earlier. The line AD intersects Ω at A and P. The power of D with respect to Ω is $P_\Omega(D) = \vec{DA} \cdot \vec{DP}$. Since A, H, D are collinear in that order (ABC is acute), and D is the midpoint of HP, the order on the line is $A - H - D - P$. Thus D is between A and P? No, D is between H and P. A is "above" H relative to BC. $A - H - D - P$. The vectors $\vec{DA}$ and $\vec{DP}$ point in opposite directions. Let AD length be h_a. $DP = DH$. $P_\Omega(D) = -AD \cdot DP = -AD \cdot DH$.

Now we establish the identity $P_\Omega(C) = -CD^2 - AD \cdot DH$. If this identity holds, then $P_\Omega(C) = -CD^2 + P_\Omega(D)$. This implies $P_\Omega(D) = P_\Omega(C) + CD^2$. As shown above, this is equivalent to C being the midpoint of XY.

We need to prove $P_\Omega(C) = -CD^2 - AD \cdot DH$. We know that B, C, D, F are concyclic (on the circle with diameter BC, call it ω_{BC}) because $\angle BDC = \angle BFC = 90°$. The power of A with respect to ω_{BC} is $P_{\omega_{BC}}(A) = \vec{AF} \cdot \vec{AB}$. Since F lies on the segment AB (because $\triangle ABC$ is acute), A, F, B are in order. Let $AF = l_F, AB = c$. $P_{\omega_{BC}}(A) = AF \cdot AB = AF \cdot c$. $AF = b \cos A$. $P_{\omega_{BC}}(A) = bc \cos A$. Also, A, C, E, F are concyclic (on the circle with diameter AC, call it ω_{AC}) because $\angle AEC = \angle AFC = 90°$. The power of B wrt ω_{AC} is $BF \cdot BA$. $BF = a \cos B$. $BA = c$. $ac \cos B$. Power of B is also $BE \cdot BC$. No.

We know that $AD \cdot AH = AE \cdot AC$. In $\triangle ADC$, $\angle ADC = 90°$. In $\triangle AEH$, $\angle AEH = 90°$? No. C, D, H, E are concyclic (circle on diameter CH). Power of A wrt this circle is $AD \cdot AH = AE \cdot AC$. This is a known property. $AD \cdot AH = bc \cos A$. So $P_{\omega_{BC}}(A) = AD \cdot AH$.

Now consider the three circles $\Omega(AFP)$, $\Gamma(ABC)$, $\omega_{BC}(BFDC)$. The radical axis of Ω and Γ is the line AP (since A, P are common to both). The radical axis of Γ and ω_{BC} is the line BC (since B, C are common to both). Let L be the radical axis of Ω and ω_{BC}. The three radical axes AP, BC, L must concur at the radical center. Let $D = AP \cap BC$. So D is the radical center. D must lie on L. The property of the radical center D is $P_\Omega(D) = P_\Gamma(D) = P_{\omega_{BC}}(D)$. Since D lies on the circle ω_{BC} (as $\angle BDC = 90°$), the power of D with respect to ω_{BC} is $P_{\omega_{BC}}(D) = 0$. Thus, $P_\Omega(D) = 0$ and $P_\Gamma(D) = 0$. $P_\Omega(D) = -AD \cdot DH$. Since $AD > 0$ (as A is not on BC), this implies $DH = 0$. This means $H = D$. If $H = D$, D is the orthocenter. Since D lies on BC, this implies that the altitudes BD and CD must lie along BC. This means $BC \perp AC$ ($\angle C = 90°$) and $BC \perp AB$ ($\angle B = 90°$). This is impossible for a triangle.

Let's re-evaluate the radical axis argument. There must be an error. Ah, the error is subtle. A, P are on Ω and Γ. AP is the radical axis only if $\Omega \neq \Gamma$. B, C are on Γ and ω_{BC}. BC is the radical axis only if $\Gamma \neq \omega_{BC}$. F is on Ω and ω_{BC}. D is on ω_{BC}. Is D on Ω? $P_\Omega(D) = P_\Gamma(D)$ since D is on AP. $P_\Gamma(D) = P_{\omega_{BC}}(D)$ since D is on BC. Everything up to $P_\Omega(D) = P_{\omega_{BC}}(D)$ is correct. $D \in \omega_{BC}$. So $P_{\omega_{BC}}(D) = 0$. Correct. So $P_\Omega(D) = 0$. Correct. This implies $-AD \cdot DH = 0$. Correct. Implies $H = D$. Contradiction.

Let's re-read the problem carefully again. Maybe P is not the reflection of H? "P is the reflection of H across BC". Okay. "circumcircle of triangle AFP intersects line BC at two distinct points X and Y". Okay.

Perhaps the definition of Power of a Point for $P_\Omega(D)$? $A, F, P \in \Omega$. D is a point. Line ADP passes through D. It intersects Ω at A and P. $P_\Omega(D) = \vec{DA} \cdot \vec{DP}$. This is standard. Order $A - H - D - P$. $\vec{DA}$ points toward A. $\vec{DP}$ points toward P. Opposite direction. $P_\Omega(D) = -DA \cdot DP = -AD \cdot DH$. Correct.

Let's reconsider $P_\Gamma(D)$. $A, B, C, P \in \Gamma$. D is on the line AD which contains A, P. $P_\Gamma(D) = \vec{DA} \cdot \vec{DP} = -AD \cdot DH$. Correct. D is on the line BC which contains B, C. $P_\Gamma(D) = \vec{DB} \cdot \vec{DC}$. D is between B, C for acute triangle. $P_\Gamma(D) = -DB \cdot DC$. Correct. So $P_\Gamma(D) = -AD \cdot DH = -DB \cdot DC$. This implies $AD \cdot DH = DB \cdot DC$. This is a known property. Proved earlier using similar triangles $\triangle BDH \sim \triangle ADC$. Correct.

Now, $P_\Omega(D) = P_\Gamma(D)$ as $D \in AP$. $P_\Gamma(D) = P_{\omega_{BC}}(D)$ as $D \in BC$. $P_{\omega_{BC}}(D) = 0$ as $D \in \omega_{BC}$. So $P_\Omega(D) = 0$ and $P_\Gamma(D) = 0$. This leads to $AD \cdot DH = 0$, which is a contradiction.

What if one of the circles is not well-defined? $\triangle ABC$ is acute, non-degenerate. H exists. F exists. P exists. A, F, P distinct? $A \neq F$. $F \neq P$. $A \neq P$. Yes. So Ω exists. Γ exists. ω_{BC} exists.

What if the radical axes are the same line? $AP = BC$? Impossible. $AD \perp BC$. What if the radical center is at infinity? $AP \parallel BC$? Impossible.

Let's assume the question is correct and C is the midpoint. This means $P_\Omega(C) = -CD^2 - AD \cdot DH$. Consider the case when $\triangle ABC$ is equilateral. $H = O$ (circumcenter). D is midpoint of BC. F is midpoint of AB. P reflection of O across BC. D is midpoint of OP. We verified C is the midpoint of XY for equilateral triangle. The contradiction $H = D$ means my application/understanding of radical axis theory is flawed in this context.

Let's ignore the radical axis argument for a moment. We need to prove C is the midpoint of XY. This is true if $P_\Omega(C) = -CD^2 - AD \cdot DH$. The power $P_\Omega(C)$ is $CX \cdot CY$. Let's use the property $CH = CP$. $P_\Omega(C)$. Consider $P_\Omega(C) = \vec{CA} \cdot \vec{CK}$ where K is the second point of intersection of AC with Ω. Consider $P_\Omega(C) = \vec{CF} \cdot \vec{CL}$ where L is the second point of intersection of CF with Ω. We have A, C, E, F concyclic on ω_{AC}. Power of point P wrt ω_{AC} is $PA \cdot PE$? No. Is it true that $P_\Omega(C) = -CH^2$? If so, $-CH^2 = -CD^2 - AD \cdot DH$. $CH^2 = CD^2 + AD \cdot DH$. But $CH^2 = CD^2 + DH^2$. So $AD \cdot DH = DH^2$. Since $DH \neq 0$, $AD = DH$. This means D is midpoint of AH. Only if $\angle B = \angle C = 90$. No.

Let's revisit the coordinate calculation. We found $k = P_\Omega(C)$. We wanted $k = -x_A^2 - y_A y_H$. This corresponds to $-CD^2 - AD \cdot DH$. The identity required was $x_A^2 + y_A y_H = CF^2 + y_F(y_H - y_A)$. $x_A = CD$, $y_A = AD$. $y_H = DH$ (assuming H above D). $CD^2 + AD \cdot DH = CF^2 + y_F(DH - AD)$. $CD^2 + AD \cdot DH = CF^2 - y_F(AD - DH) = CF^2 - y_F \cdot AH$. $CD^2 + AD \cdot DH = x_F^2 + y_F^2 - y_F \cdot AH$. This identity seems plausible and might be provable geometrically.

Let's try one more angle property. $A, F, P, X, Y \in \Omega$. $\angle CFA = 90$. $\angle CXA + \angle CPA = 180$? No. $\angle CXP + \angle CFP = 180$? No. $\angle CPA = \angle CBA = \beta$. A, P, B, C concyclic (Γ). $\angle CFA = 90°$. Power of C wrt Ω is $P_\Omega(C)$. Let's use the proved relation $P_\Omega(C) = -ab \cos C + b^2 \cos^2 C$. This was derived assuming $u = 0$, which is the consequence we want to prove. Can't use it.

Consider the identity $P_\Omega(C) = -CD^2 - AD \cdot DH$. Let's assume it's proven. Then $P_\Omega(D) = -AD \cdot DH = P_\Omega(C) + CD^2$. Also $P_\Omega(D) = P_\Omega(C) - \vec{CD} \cdot (\vec{CX} +$
$vecCY) + CD^2$. This implies $\vec{CD} \cdot (\vec{CX} +$
$vecCY) = 0$. As $\vec{CD} \neq 0$ and vectors are collinear, $\vec{CX} +$
$vecCY = 0$. This means C is the midpoint of XY.

The proof requires proving the identity $P_\Omega(C) = -CD^2 - AD \cdot DH$. Let's try to find a reference for this identity or prove it. Let K be the second intersection of AC with Ω. $P_\Omega(C) = CA \cdot CK$. Let L be the second intersection of CF with Ω. $P_\Omega(C) = CF \cdot CL$. A, F, P, K, L are on Ω. Maybe relate Ω to the nine-point circle?

Let's reconsider the radical axis contradiction. $P_\Omega(D) = 0$ implies $H = D$. Perhaps one of the circles is not what I defined. $\Gamma = (ABC)$. $\omega_{BC} = (BFDC)$. $\Omega = (AFP)$. $D \in AP$? Yes. $D \in BC$? Yes. D is radical center. $P_\Omega(D) = P_\Gamma(D) = P_{\omega_{BC}}(D)$. $D \in \omega_{BC}$? Yes. $P_{\omega_{BC}}(D) = 0$. Must mean $P_\Gamma(D) = 0$. Implies D on (ABC). Implies $\angle A = 90$. But triangle is acute. Could AP be tangent to Γ at A? No, P is another point on Γ. Could BC be tangent to Γ? No. Could BC be tangent to ω_{BC}? No, B, C on ω_{BC}.

What if D is not the radical center? Maybe the axes are parallel? $AP \parallel BC$? No. What if the circles coincide? $\Omega = \Gamma$? Then $F \in \Gamma$. Impossible. $\Gamma = \omega_{BC}$? Impossible. $\Omega = \omega_{BC}$? Then $A, P \in \omega_{BC}$. Impossible.

Maybe the definition of power $\vec{DA} \cdot \vec{DP}$ for $P_\Omega(D)$ assumes D is outside Ω? No, it works for D inside too. Then $P_\Omega(D) < 0$. $A - H - D - P$. D is between H and P. D could be inside or outside Ω. $P_\Omega(D) = -AD \cdot DH$. Since $AD > 0, DH > 0$, $P_\Omega(D) < 0$. So D is inside circle Ω. This implies $P_{\omega_{BC}}(D)$ must be negative too. But D is on circle ω_{BC}, so $P_{\omega_{BC}}(D) = 0$. This means $P_\Omega(D) = 0$. Contradiction.

There must be a mistake in my reasoning with radical axes. Let's step through again. 1. AP is radical axis of Ω, Γ. 2. BC is radical axis of Γ, ω_{BC}. 3. L is radical axis of Ω, ω_{BC}. 4. $D = AP \cap BC$. Thus D is the radical center. $P_\Omega(D) = P_\Gamma(D) = P_{\omega_{BC}}(D)$. 5. $D \in \omega_{BC}$, so $P_{\omega_{BC}}(D) = 0$. 6. Therefore $P_\Omega(D) = 0$. 7. $P_\Omega(D) = -AD \cdot DH$. 8. Therefore $AD \cdot DH = 0$. $H = D$. Contradiction.

Is D on the line AP? Yes, A, H, D are collinear. P is reflection of H in D. A, H, D, P are collinear. So D lies on the line containing segment AP. Is D on the line BC? Yes, D is foot of altitude on BC. Is P on Γ? Yes. Is A on Γ? Yes. Is A on Ω? Yes. Is P on Ω? Yes.

Perhaps the power calculation $P_\Omega(D) = \vec{DA} \cdot \vec{DP}$ requires A, P to be the *only* intersection points of line AD with Ω? Circle intersects line in at most 2 pts. Yes A, P are the points.

Let's assume the result holds and C is the midpoint of XY. Then $P_\Omega(C) = -CX^2$. Also $P_\Omega(D) = P_\Omega(C) + CD^2 = -CX^2 + CD^2$. We know $P_\Omega(D) = -AD \cdot DH$. So $-AD \cdot DH = -CX^2 + CD^2$. $CX^2 = CD^2 + AD \cdot DH$. This identity must hold.

Let's prove $CX^2 = CD^2 + AD \cdot DH$. $AD \cdot DH = BD \cdot CD$. This is only if $\triangle ABC$ is right at A? No, this holds for orthocenter H. $\triangle BDH \sim \triangle ADC$. $BD/AD = DH/DC \implies BD \cdot DC = AD \cdot DH$. So $CX^2 = CD^2 + BD \cdot DC = CD(CD + BD) = CD \cdot BC$. Let $C = 0$. $D = d$. $B = b$. $X = x$. $x^2 = d(d - b)$. No this is not right. D is between B, C. $C = 0$. $B = a$. $D = d$. $CD = d$. $BD = a - d$. $CX^2 = d^2 + (a - d)d = d^2 + ad - d^2 = ad$. $CX = \sqrt{ad}$.

X and Y are points x, y. $C = 0$. Midpoint $x + y = 0$. $P_\Omega(C) = xy = -x^2$. So $CX^2 = -P_\Omega(C)$. We need to show $CX^2 = CD \cdot CB$? No. $CX^2 = CD^2 + BD \cdot CD$. No $AD \cdot DH$.

Let's use the identity $P_\Omega(C) = -CD^2 - AD \cdot DH$. This identity implies C is the midpoint. Let's accept this identity as given, or derivable via coordinates. The proof relies on this specific calculation of the power of C wrt Ω.

Final conclusion check: The argument: C midpoint $\iff$ $P_\Omega(D) = P_\Omega(C) + CD^2$. $P_\Omega(D) = -AD \cdot DH$. So C midpoint $\iff$ $P_\Omega(C) = -AD \cdot DH - CD^2$. This identity needs proof. The coordinate calculation confirmed this identity holds true.

Final conclusion seems solid, pending the proof of the identity for $P_\Omega(C)$. The coordinate check suggests it is true. Also, the failure of the radical axis argument needs clarification, but the result seems provable through the power of point calculation. It's likely there's a subtle issue in applying the radical axis theorem, or my power calculations related to it. However, the direct calculation of $P_\Omega(C)$ and $P_\Omega(D)$ and their relation seems the correct path.

4.3.2. Rubric

1. **Additional construction (1 pt).**
   - Requires constructing useful auxiliary points/objects.
   - Credit is given only if the construction provides useful properties that help reach the solution.
2. **Properties of the construction (4 pts).**
   - The student must prove nontrivial properties of the construction.
   - Emphasis: properties involving symmetry related to the circle (AFP).
   - Points are proportional to the depth and correctness of these proven properties.
3. **Finishing (2 pts).**
   - The student must show that $CX = CY$ using the previously established properties.
   - This is the final step that connects the properties of the construction to the required conclusion.

4.3.3. Grading Details

Human Judge 1

Points: 0/7

- **Additional construction:**
  - The model does not construct the necessary points for its solution. In fact, it chooses a computational approach that is unfinished and far from complete, so no points are awarded.
  - **Points: 0 / 1**
- **Properties of the construction:**
  - The model does not identify any useful properties of the construction.
  - **Points: 0 / 4**
- **Finishing:**
  - The model did not manage to validly conclude that $CX = CY$.
  - **Points: 0 / 4**

Human Judge 2

Points: 0/7

- **Additional construction (0/1):**
  - No relevant constructions were created.
- **Properties of the construction (0/4):**

    - No non-trivial properties were proven. The solution assumes the power of a point identity without proving it. No credit for an incomplete bash solution.
- **Finishing (0/2):**
    - No relevant progress was made.

LLM Judge

Points: 3/7

- **Criterion 1 (Additional construction, 1 pt):**
    - Student introduces D (foot from A), E, F, and circles Γ and the diameter circle; uses D meaningfully to relate powers, and the line AD intersects (AFP) at A and P.
    - Construction of D is useful and used to derive $P_\Omega(D) = -AD \cdot DH$.
    - **Points: 1/1**
- **Criterion 2 (Properties of the construction, 4 pts):**
    - Student shows P lies on (ABC) using symmetry and congruences; establishes A, H, D, P are collinear, with D the midpoint of HP; computes $P_\Omega(D)$ correctly. However, incorrectly asserts B, C, D, F are concyclic, causing contradictions; also fails to prove the key identity $P_\Omega(C) = -CD^2 - AD \cdot DH$, explicitly stating it as an assumption.
    - Several correct and relevant properties are given, but the crucial connection to C is not proven and one major incorrect claim undermines the argument.
    - **Points: 2/4**
- **Criterion 3 (Finishing, 2 pts):**
    - Student reduces the midpoint condition to $P_\Omega(D) = P_\Omega(C) + CD^2$, but does not prove the required identity for $P_\Omega(C)$, instead assuming it.
    - Thus the conclusion $CX = CY$ is not established.
    - **Points: 0/2**

4.3.4. *Criterion by Criterion Comparison*

Criterion 1: Additional Construction (Human: 0, LLM: 1)

Human Judges. Human judges agreed that no relevant or meaningful constructions were introduced. The solution relied on an incomplete computational approach, not on geometric constructions, and therefore earned 0 points.

LLM Judge. The LLM judge interpreted the introduction of points D, E, F and certain circles as legitimate geometric constructions and credited the student for using D meaningfully (e.g., in computing a power-of-a-point expression), awarding 1 point.

Analysis of GPT-5 Misjudgment. The LLM treated the mere mention of new points as valid constructions without checking whether they were actually justified or used in the argument. This is not only a gap-filling issue but also a misclassification of what counts as a construction under the rubric.

Criterion 2: Properties of the Construction (Human: 0, LLM: 2)

Human Judges. Both human judges concluded that the student established no non-trivial or useful geometric properties. Key identities were assumed rather than proven. As a result, the human judges awarded 0 points.

LLM Judge. The LLM credited the student with establishing several geometric facts:

- $P \in (ABC)$,
- A, H, D, P are collinear,
- D is the midpoint of HP,
- a correct computation of $P_\Omega(D)$.

Although the LLM noticed the incorrect concyclicity and the unproven formula for $P_\Omega(C)$, it treated them as partial errors, still giving 2 points.

Analysis of GPT-5 Misjudgment. The LLM did not verify whether these claimed properties were actually justified in the student's reasoning. It incorrectly treated serious errors as minor issues instead of recognizing that, under the rubric, such errors invalidate the entire set of geometric properties. This reflects a failure to:

- locally verify the correctness of each step;
- avoid assuming correctness based on plausibility rather than evidence.

Criterion 3: Finishing (Human: 0, LLM: 0)

Human and LLM Agreement. All judges agreed that the solution did not validly conclude that $CX = CY$. The key identity was assumed rather than proven, so both humans and the LLM awarded 0 points.

5. Statistics

In this section, we present tables of empirical statistics for reference.

Table 1 | **Error–Accuracy Metrics**

Model	**MAE**	**RMSE**	**Acc.**	**Acc.$_{\leq 1}$**	**Acc.$_{\leq 2}$**
DeepSeek-Math-V2	0.67	1.68	0.73	0.88	0.91
GPT-5	0.79	1.67	0.61	0.81	0.92
Gemini-3-Pro-Preview	0.89	1.87	0.64	0.80	0.88
Kimi-K2-Thinking	0.97	1.98	0.61	0.80	0.87
GPT-5-Pro	0.90	1.74	0.56	0.78	0.89
GPT-OSS-120B	1.06	2.15	0.61	0.78	0.85
GPT-5.1	1.04	2.03	0.56	0.78	0.86
Qwen3-Max	1.22	2.38	0.61	0.74	0.81
DeepSeek-Chat-v3.1	1.30	2.42	0.54	0.74	0.79
Gemini-2.5-Pro	1.32	2.35	0.52	0.72	0.79
Qwen3-235B-A22B	1.39	2.41	0.49	0.70	0.78
Qwen3-32B	2.19	3.31	0.39	0.56	0.64
DeepSeek-R1	2.20	3.41	0.43	0.56	0.64
Qwen3-235B-A22B-Thinking	2.45	3.66	0.36	0.56	0.63

Table 2 | **Correlation Metrics**

Model	**Pearson**	**Spearman**	**Kendall**	κ_{lin}	κ_{quad}	**Slope**	**Intercept**	R^2
GPT-5	0.75	0.58	0.53	0.57	0.70	0.86	0.76	0.56
Gemini-3-Pro-Preview	0.74	0.58	0.53	0.55	0.67	0.93	0.90	0.55
GPT-5-Pro	0.74	0.55	0.50	0.53	0.69	0.86	0.89	0.55
GPT-5.1	0.70	0.58	0.53	0.50	0.62	0.88	1.08	0.49
DeepSeek-Math-V2	0.62	0.50	0.47	0.52	0.62	0.61	0.37	0.39
Kimi-K2-Thinking	0.63	0.46	0.42	0.48	0.60	0.75	0.88	0.40
GPT-OSS-120B	0.62	0.51	0.46	0.47	0.56	0.78	1.04	0.39
Qwen3-Max	0.61	0.47	0.43	0.44	0.52	0.81	1.23	0.37
Gemini-2.5-Pro	0.61	0.46	0.41	0.41	0.52	0.77	1.39	0.37
DeepSeek-Chat-v3.1	0.60	0.45	0.41	0.42	0.51	0.80	1.33	0.36
Qwen3-235B-A22B	0.54	0.44	0.39	0.36	0.47	0.67	1.40	0.30
Qwen3-32B	0.47	0.40	0.35	0.26	0.32	0.67	2.36	0.22
Qwen3-235B-A22B-Thinking	0.44	0.39	0.34	0.23	0.28	0.66	2.66	0.19
DeepSeek-R1	0.41	0.35	0.31	0.25	0.28	0.60	2.38	0.17
Human Baseline	0.99	0.94	0.92	0.95	0.99	1.00	0.01	0.98

Table 3 | **Distributional Metrics**

Model	**ER**	**JSD**	**RV**
DeepSeek-Math-V2	1.13	0.02	0.95
GPT-OSS-120B	1.68	0.06	1.55
Gemini-3-Pro-Preview	1.61	0.07	1.57
Kimi-K2-Thinking	1.68	0.07	1.39
GPT-5	1.79	0.07	1.34
Qwen3-Max	1.60	0.08	1.78
GPT-5-Pro	1.93	0.10	1.34
GPT-5.1	1.85	0.09	1.57
Qwen3-235B-A22B	1.87	0.12	1.50
DeepSeek-Chat-v3.1	1.88	0.09	1.77
Gemini-2.5-Pro	1.99	0.12	1.59
DeepSeek-R1	1.97	0.16	2.11
Qwen3-32B	2.07	0.18	2.02
Qwen3-235B-A22B-Thinking	2.01	0.20	2.30
Human Baseline	0.95	0.00	1.01

Table 4 | **Problem-wise Mean Absolute Error by Model**

Model	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**Average**
DeepSeek-Math-V2	1.68	0.05	0.11	1.84	0.07	0.25	0.67
GPT-5	1.06	0.11	0.27	2.49	0.17	0.63	0.79
Gemini-3-Pro-Preview	1.02	0.25	0.27	2.57	0.73	0.48	0.89
GPT-5-Pro	1.06	0.07	0.66	2.52	0.53	0.53	0.90
Kimi-K2-Thinking	1.23	0.18	0.36	3.27	0.19	0.57	0.97
GPT-5.1	1.48	0.14	0.34	2.98	0.69	0.61	1.04
GPT-OSS-120B	1.33	0.07	0.43	3.66	0.18	0.71	1.06
Qwen3-Max	1.46	0.07	0.59	3.73	0.43	1.02	1.22
DeepSeek-Chat-v3.1	1.22	0.97	1.19	3.40	0.30	0.72	1.30
Gemini-2.5-Pro	1.13	0.14	2.09	3.30	0.75	0.50	1.32
Qwen3-235B-A22B	1.41	0.80	0.55	3.84	0.48	1.28	1.39
Qwen3-32B	2.51	1.64	1.32	4.66	1.39	1.63	2.19
DeepSeek-R1	1.61	1.52	4.16	4.25	0.59	1.09	2.20
Qwen3-235B-A22B-Thinking	3.36	1.75	2.43	3.98	1.74	1.46	2.45
Average	1.54	0.55	1.06	3.32	0.59	0.82	1.31

Table 5 | **Problem-wise Accuracy by Model**

Model	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**Average**
DeepSeek-Math-V2	0.50	0.95	0.95	0.30	0.93	0.77	0.73
Gemini-3-Pro-Preview	0.55	0.82	0.93	0.30	0.61	0.64	0.64
GPT-OSS-120B	0.34	0.91	0.89	0.09	0.82	0.61	0.61
GPT-5	0.34	0.91	0.91	0.16	0.82	0.52	0.61
Kimi-K2-Thinking	0.39	0.82	0.89	0.09	0.84	0.64	0.61
Qwen3-Max	0.45	0.93	0.89	0.14	0.77	0.45	0.61
GPT-5.1	0.30	0.86	0.89	0.14	0.70	0.50	0.56
GPT-5-Pro	0.30	0.93	0.82	0.11	0.59	0.59	0.56
DeepSeek-Chat-v3.1	0.41	0.45	0.75	0.14	0.82	0.66	0.54
Gemini-2.5-Pro	0.39	0.84	0.52	0.18	0.61	0.57	0.52
Qwen3-235B-A22B	0.41	0.39	0.89	0.07	0.77	0.41	0.49
DeepSeek-R1	0.34	0.41	0.36	0.14	0.77	0.57	0.43
Qwen3-32B	0.26	0.18	0.68	0.14	0.61	0.43	0.38
Qwen3-235B-A22B-Thinking	0.27	0.18	0.61	0.11	0.52	0.48	0.36
Average	0.37	0.69	0.78	0.15	0.73	0.56	0.55

6. Empirical Details

6.1. Models

We select 14 models in our experiments which are: GPT-5, GPT-5 Pro, GPT-OSS-120B, Gemini-2.5-Pro, Qwen3-Max, Qwen3-235B-A22B-Instruct, Qwen3-235B-A22B-Thinking, Qwen3-32B, DeepSeek-R1, DeepSeek-Chat-v3.1, Kimi-K2-Thinking, DeepSeek-Math-V2, GPT-5.1, Gemini-3-Pro-Preview. DeepSeek-Math-V2 is deployed locally and for other models we use APIs from OpenRouter.

6.2. Dataset

We use the **MathArena–USAMO (2025)** dataset (data link) for all experiments. This dataset contains 264 samples across the 6 problems from **USAMO 2025**, featuring long-form, proof-style responses. Each problem contains a brief rubric by MathArena. Each sample is graded by 2 human judgers, which are stored as two values: `points_judge_1` and `points_judge_2`. We use `points_judge_1` as the primary ground truth for all the evaluation, and `points_judge_2` as the reference for human baselines in Tables 2 and 3. We parse the rubric from `grading_scheme_desc` of `grading_details_judge_1`.

The dataset was originally obtained from HuggingFace. However, an update was released on Oct 17; we retained the original version for consistency throughout our experiments.

6.3. Prompt

Our default prompt is available at prompt link and is used across all experiments.

Note on reproducibility Exact reproduction of our empirical results may not be possible due to API provider variability and stochasticity from `temperature` = 0.1. We therefore provide complete run logs at Google Drive log link for reference.