PAPER
arXiv:2601.23048
DATASET
bwcao/ContextMATH

# From Abstract to Contextual: What LLMs Still Cannot Do in Mathematics

**Bowen Cao[1†] Dongdong Zhang[2*] Yixia Li[3] Junpeng Liu[1] Shijue Huang[4] Chufan Shi[5] Hongyuan Lu[6] Yaokang Wu[7] Guanhua Chen[3] Wai Lam[1] Furu Wei[2]**

[1]CUHK · [2]Microsoft · [3]SUSTech · [4]HKUST · [5]USC · [6]FaceMind · [7]CMU

[†]Internship at MSRA · [*]Corresponding author

bwcao@link.cuhk.edu.hk · dozhang@microsoft.com

# The Paradox of LLM Math Performance

- **Success on Benchmarks**
  - LLMs excel on abstract math problems (e.g., AIME, MATH).
- **The Real-World Gap**
  - This success doesn’t fully translate to real-world applications.
- **Our Core Question**
  - Where and why do LLMs fail when moving from abstract math problems to contextual problems?

# Defining the Underexplored Challenge

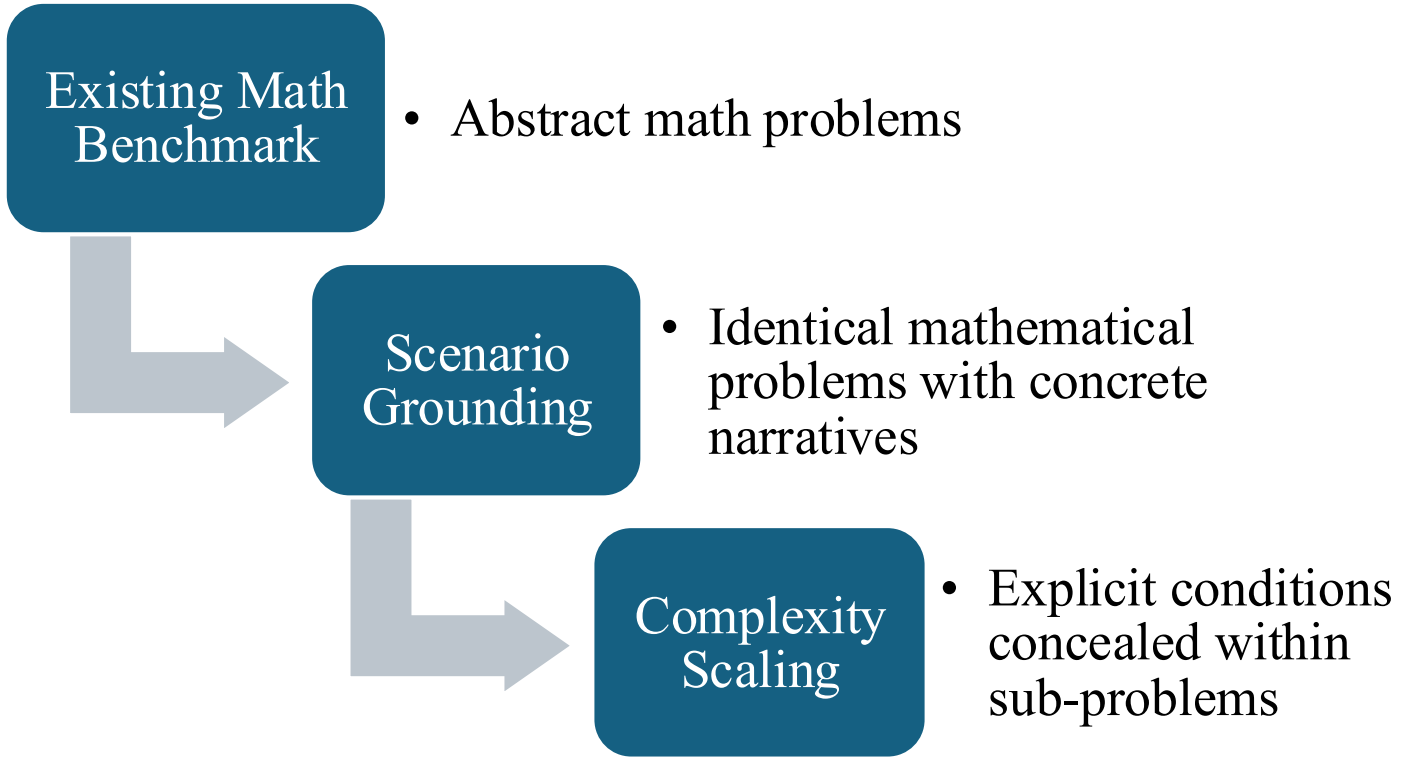
- **Contextual Mathematical Reasoning**
  - **Definition**: The ability to **formulate** and **solve** the mathematical core of a problem when it is embedded in a narrative scenario.
  - **Two Domains of Math**
    - **Abstract Math**: Clean, symbolic, and well-defined. (e.g., Solve for x).
    - **Contextual Math**: Narrative-driven, requiring interpretation and modeling. (e.g., Engineering design, financial analysis).

# A New Tool to Probe the Gap

- **The ContextMath Benchmark**
  - **ContextMATH**: A **Context**ual Reasoning Evaluation for **Math** benchmark.
  - **Design Philosophy**: Systematically repurpose standard benchmarks (AIME, MATH-500) into controlled narrative variants.
  - **Two Controlled Settings**
    - **Scenario Grounding (SG):** Embeds abstract elements into a realistic narrative without changing the reasoning complexity.
    - **Complexity Scaling (CS):** Conceals explicit conditions as sub-problems that require an extra inference step.

# How We Make Abstract Problems Contextual

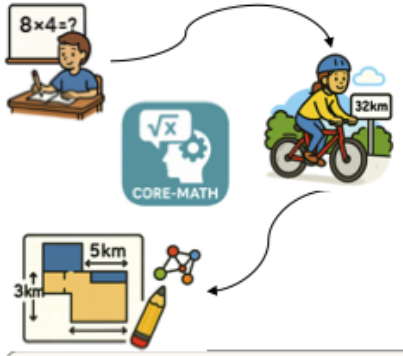
- **A Closer Look at ContextMath**
  - AIME 2024 (30 problems)
    - SG set
    - CS set
  - AIME 2025 (30 problems)
    - SG set
    - CS set
  - Math-500 (262 out of 500 problems with difficulty $\geq 3$)
    - SG set



# How We Make Abstract Problems Contextual

- **A Closer Look at ContextMath**
  - **Scenario Grounding** embeds abstract mathematical structures into concrete narratives with real-world entities and interactions. The reasoning core remains unchanged.

**(1) Original Problem:**
Let N denote the number of ordered triples of positive integers (a,b,c) such that $a, b, c \leq 3^6$ and $a^3 + b^3 + c^3$ is a multiple of $3^7$. Find the remainder when N is divided by 1000.




**(2) Scenario Grounding:**
…three distinct components within a sophisticated smart energy system…

…with positive values that do not exceed 729 units...

...raising their output metrics to the third power and summing these values, must yield a total that is perfectly divisible by 2,187…

…find the remainder when this number is divided by 1,000 for reporting purposes…

**(3) Complexity Scaling:**
In a simulation lab, three autonomous drones are sent out along separate axes — one moves north-south, one east-west, and one vertically. Each drone travels forward by a positive number of whole steps. These steps must remain below a certain threshold, which is known to be a power of 3. It is known that step counts as high as 2000 are not allowed, while 250 is within the allowed range. As part of a synchronization test, the system calculates a combined energy score by summing the cubes of the step counts of all three drones. Synchronization only succeeds if this total energy is also a multiple of a certain power of 3. From prior data, an energy total of 1458 fails to trigger the sync, but 8748 does trigger it. How many such ordered drone configurations satisfy both conditions? Report your answer as the remainder when this number is divided by 1000.

**Note**: Consistent color-coding highlights correspondence between mathematical components across the three versions.

# How We Make Abstract Problems Contextual

- **Scenario Grounding**
  - LLM-Driven Pipeline
    - Generation (**Prompt on the right**)
    - Verification (Omitted for brevity)
    - Revision (Omitted for brevity)

SG Generation Prompt

Your Goal: Convert an abstract math problem into a concrete, real-world story. The new story must be mathematically identical to the original.

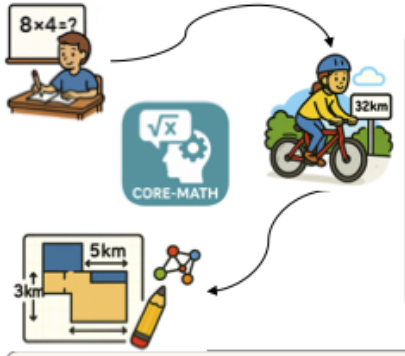
You can draft the story first, then use the following steps to ensure accuracy and format your final output correctly.

[Step 1] Map All Mathematical Components
…
[Step 2] Define the Real-World Rules of Interaction
…
[Step 3] Write the Final Problem
…

# How We Make Abstract Problems Contextual

- **A Closer Look at ContextMath**
  - **Complexity Scaling** conceals explicit conditions within sub-problems in the scenario.
  - These sub-problems yield exactly the information originally explicit, while reflecting how constraints are often expressed in real-world settings.

**(1) Original Problem:**
Let N denote the number of ordered triples of positive integers (a,b,c) such that $a, b, c \leq 3^6$ and $a^3 + b^3 + c^3$ is a multiple of $3^7$. Find the remainder when N is divided by 1000.




**(2) Scenario Grounding:**
…three distinct components within a sophisticated smart energy system…

…with positive values that do not exceed 729 units...

...raising their output metrics to the third power and summing these values, must yield a total that is perfectly divisible by 2,187…

…find the remainder when this number is divided by 1,000 for reporting purposes…

**(3) Complexity Scaling:**
In a simulation lab, three autonomous drones are sent out along separate axes — one moves north-south, one east-west, and one vertically. Each drone travels forward by a positive number of whole steps. These steps must remain below a certain threshold, which is known to be a power of 3. It is known that step counts as high as 2000 are not allowed, while 250 is within the allowed range. As part of a synchronization test, the system calculates a combined energy score by summing the cubes of the step counts of all three drones. Synchronization only succeeds if this total energy is also a multiple of a certain power of 3. From prior data, an energy total of 1458 fails to trigger the sync, but 8748 does trigger it. How many such ordered drone configurations satisfy both conditions? Report your answer as the remainder when this number is divided by 1000.

**Note**: Consistent color-coding highlights correspondence between mathematical components across the three versions.

# How We Make Abstract Problems Contextual

- **Complexity Scaling**
  - LLM-Driven Pipeline
    - Generation (**Prompt on the right**)
    - Verification (Omitted for brevity)
    - Revision (Omitted for brevity)

CS Generation Prompt

Task: Enhancing Math Problem Difficulty through Contextual Embedding

Your goal is to increase the difficulty of a given mathematical problem's real-world version, while preserving its core mathematical structure and the integrity of its real-world mapping. Employ the following strategies:

1. Maintain Core Alignment

…

2. Embed Conditions through Layered Obfuscation

…

3. Introduce Plausible, Irrelevant Information

...

4. Language Refinement and Simplicity:

...

# Key Finding 1: A Universal Bottleneck

- Accuracy of proprietary models on ContextMATH.
  - Even frontier models like GPT-5 see a 26% relative drop on AIME 2025-CS.

| Model | AIME 2024 (%) | | | | AIME 2025 (%) | | |
|---|---|---|---|---|---|---|---|
| | Ori | SG | SG Avg@3 | CS | Ori | SG | CS |
| Copilot | 30.0 | 30.0 ( -0% ) | 28.9 ( -4% ) | 23.3 ( -22% ) | 33.3 | 20.0 ( -40% ) | 16.7 ( -50% ) |
| gpt-4o-mini | 6.7 | 3.3 ( -50% ) | 6.7 ( -0% ) | 3.3 ( -50% ) | 10.0 | 6.7 ( -33% ) | 0.0 ( -100% ) |
| o1-mini | 60.0 | 53.3 ( -11% ) | 53.3 ( -11% ) | 40.0 ( -33% ) | 40.0 | 33.3 ( -17% ) | 23.3 ( -42% ) |
| gpt-4o | 16.7 | 13.3 ( -20% ) | 11.1 ( -33% ) | 3.3 ( -80% ) | 10.0 | 6.7 ( -33% ) | 0.0 ( -100% ) |
| gpt-4.1-mini | 46.7 | 26.7 ( -43% ) | 34.4 ( -26% ) | 30.0 ( -36% ) | 53.3 | 30.0 ( -44% ) | 16.7 ( -69% ) |
| gpt-4.1-nano | 26.7 | 23.3 ( -13% ) | 18.9 ( -29% ) | 6.7 ( -75% ) | 33.3 | 20.0 ( -40% ) | 13.3 ( -60% ) |
| R1 | **93.3** | 70.0 ( -25% ) | 70.0 ( -25% ) | 66.7 ( -29% ) | 86.7 | 73.3 ( -15% ) | 53.3 ( -38% ) |
| Doubao-1.5 | 90.0 | 70.0 ( -22% ) | 66.7 ( -26% ) | 56.7 ( -37% ) | 76.7 | 53.3 ( -30% ) | 43.3 ( -43% ) |
| Qwen-max | 23.3 | 16.7 ( -29% ) | 14.4 ( -38% ) | 6.7 ( -71% ) | 13.3 | 10.0 ( -25% ) | 0.0 ( -100% ) |
| QwQ-plus | 86.7 | 56.7 ( -35% ) | 60.0 ( -31% ) | 46.7 ( -46% ) | 73.3 | 53.3 ( -27% ) | 43.3 ( -41% ) |
| Grok3 | 43.3 | 23.3 ( -46% ) | 22.2 ( -49% ) | 23.3 ( -46% ) | 26.7 | 13.3 ( -50% ) | 20.0 ( -25% ) |
| Gemini 2.5 Flash | 70.0 | 63.3 ( -10% ) | 61.1 ( -13% ) | 53.3 ( -24% ) | 70.0 | 43.3 ( -38% ) | 30.0 ( -57% ) |
| Gemini 2.5 Pro | 83.3 | 73.3 ( -12% ) | 68.9 ( -17% ) | 76.7 ( -8% ) | 83.3 | 56.7 ( -32% ) | 50.0 ( -40% ) |
| o3 | 83.3 | 70.0 ( -16% ) | 73.3 ( -12% ) | 66.7 ( -20% ) | 76.7 | 70.0 ( -9% ) | 60.0 ( -22% ) |
| gpt-5 | 90.0 | **83.3** ( -7% ) | **82.2** ( -9% ) | **80.0** ( -11% ) | **90.0** | **80.0** ( -11% ) | **66.7** ( -26% ) |

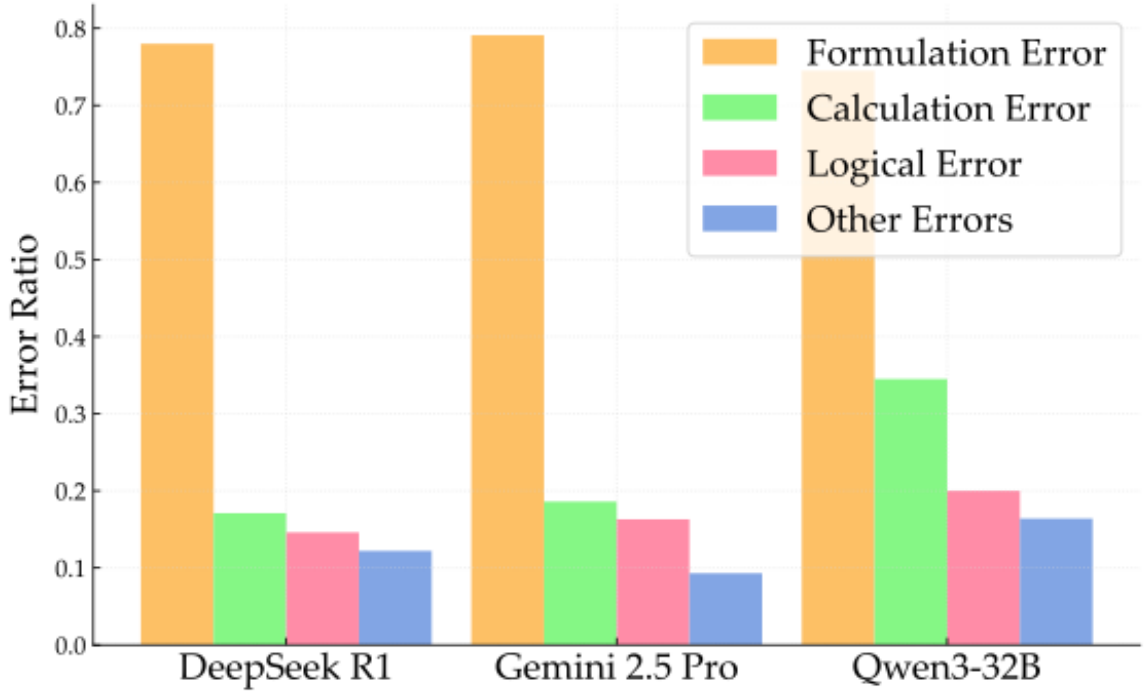
# Key Finding 1: A Universal Bottleneck

- Accuracy of open-source models on ContextMATH.
- **Consistent & Severe Drops**: Both open-source and proprietary models show a sharp decline in accuracy on ContextMATH.
- **By the Numbers**:
  - Open-source models drop by 13 (SG) & 34 (CS) points on average.
  - Proprietary models drop by 13 (SG) & 20 (CS) points on average.

| Model | AIME 2024 (%) | | | AIME 2025 (%) | | | Math-500 (%) | |
|---|---|---|---|---|---|---|---|---|
| | Ori | SG | CS | Ori | SG | CS | Ori | SG |
| *≤4B* | | | | | | | | |
| Qwen3-0.6B | 7.9 | 6.2 (-21%) | 0.4 (-95%) | 15.4 | 5.4 (-65%) | 2.5 (-84%) | 62.4 | 42.6 (-32%) |
| Qwen2.5-Math-1.5B | 11.5 | 2.5 (-79%) | 0.4 (-96%) | 4.6 | 0.8 (-82%) | 0.0 (-100%) | 32.5 | 25.8 (-21%) |
| ↪R1-Distil-Qwen-1.5B | 28.3 | 20.0 (-29%) | 7.1 (-75%) | 19.6 | 9.6 (-51%) | 5.8 (-70%) | 74.0 | 57.6 (-22%) |
| ↪DeepScaleR-1.5B-Preview | 41.5 | 23.3 (-44%) | 7.5 (-82%) | 30.4 | 15.4 (-49%) | 7.5 (-75%) | 80.1 | 64.2 (-20%) |
| ↪OpenMath-Nemotron-1.5B | 62.5 | 34.2 (-45%) | 14.6 (-77%) | 50.4 | 21.2 (-58%) | 11.2 (-78%) | 87.3 | 70.1 (-20%) |
| ↪DeepMath-Omn-1.5B | 62.9 | 33.8 (-46%) | 13.8 (-78%) | 55.8 | 20.8 (-63%) | 12.5 (-78%) | 87.1 | 73.0 (-16%) |
| Qwen3-1.5B | 46.2 | 29.6 (-36%) | 12.5 (-73%) | 34.6 | 24.6 (-29%) | 11.2 (-67%) | 84.1 | 67.4 (-20%) |
| Qwen3-4B | **70.4** | **52.5** (-25%) | **34.6** (-51%) | **64.2** | **39.6** (-38%) | **33.8** (-47%) | **90.9** | **78.8** (-13%) |
| *7B/8B* | | | | | | | | |
| Qwen2.5-Math-7B | 10.8 | 6.7 (-38%) | 1.5 (-85%) | 5.0 | 3.3 (-33%) | 0.8 (-83%) | 44.8 | 36.7 (-18%) |
| ↪OpenMath-Nemotron-7B | 72.9 | 52.1 (-29%) | 30.0 (-59%) | 60.0 | 40.4 (-33%) | 29.2 (-51%) | 90.5 | 78.5 (-13%) |
| ↪R1-Distil-Qwen-7B | 48.8 | 40.0 (-18%) | 23.3 (-52%) | 41.5 | 22.5 (-46%) | 15.0 (-64%) | 87.4 | 73.9 (-15%) |
| ↪AceMath-RL-Nemotron-7B | 69.2 | 48.8 (-30%) | 32.9 (-52%) | 54.2 | 26.7 (-51%) | 22.5 (-58%) | 89.9 | 79.0 (-12%) |
| Qwen3-8B | 73.8 | **61.5** (-16%) | **42.9** (-42%) | 64.6 | **48.3** (-25%) | 35.8 (-45%) | 91.0 | 81.0 (-11%) |
| ↪R1-0528-Qwen3-8B | **75.0** | 55.0 (-27%) | 39.6 (-47%) | 65.8 | **48.3** (-27%) | 32.9 (-50%) | 90.7 | 77.2 (-15%) |
| ↪AReaL-boba-2-8B | 74.2 | 58.3 (-21%) | 41.5 (-44%) | **67.9** | 47.9 (-29%) | **37.1** (-45%) | **91.5** | **82.0** (-11%) |
| *14B* | | | | | | | | |
| Qwen2.5-14B | 6.2 | 3.8 (-40%) | 1.5 (-73%) | 3.3 | 1.5 (-50%) | 0.0 (-100%) | 48.5 | 34.9 (-28%) |
| ↪R1-Distil-Qwen-14B | 67.5 | 47.1 (-30%) | 35.4 (-48%) | 50.8 | 26.2 (-48%) | 25.8 (-49%) | 89.5 | 76.9 (-14%) |
| ↪OpenMath-Nemotron-14B | 73.8 | 51.5 (-30%) | 42.1 (-43%) | 63.8 | 42.5 (-33%) | 29.2 (-54%) | 90.8 | 80.8 (-11%) |
| Qwen3-14B | 80.0 | 64.6 (-19%) | 50.8 (-36%) | 72.9 | 49.2 (-33%) | **42.1** (-42%) | **92.6** | 81.9 (-12%) |
| ↪AReaL-boba-2-14B | **82.9** | **65.8** (-21%) | **53.8** (-35%) | **73.3** | 52.1 (-29%) | 39.2 (-47%) | 91.9 | 82.9 (-10%) |
| Phi-4-reasoning-plus | 80.4 | 60.4 (-25%) | 52.9 (-34%) | 71.5 | **55.4** (-22%) | 39.6 (-45%) | **92.6** | **83.1** (-10%) |
| *≥32B* | | | | | | | | |
| Qwen2.5-32B | 11.2 | 6.7 (-41%) | 3.3 (-70%) | 3.8 | 2.9 (-22%) | 0.0 (-100%) | 45.2 | 37.8 (-16%) |
| ↪OpenMath-Nemotron-32B | 57.1 | 42.5 (-26%) | 27.9 (-51%) | 52.1 | 34.6 (-34%) | 27.5 (-47%) | 75.8 | 61.8 (-18%) |
| ↪R1-Distil-Qwen-32B | 69.6 | 52.5 (-25%) | 39.2 (-44%) | 56.2 | 39.6 (-30%) | 30.0 (-47%) | 89.4 | 78.9 (-12%) |
| Qwen3-32B | 81.2 | **67.9** (-16%) | 57.1 (-30%) | 70.0 | 54.4 (-22%) | **45.0** (-36%) | 92.1 | 82.7 (-10%) |
| ↪AReaL-boba-2-32B | **81.5** | 65.4 (-20%) | **58.3** (-29%) | **77.1** | **55.0** (-29%) | 43.8 (-43%) | 92.3 | **82.9** (-10%) |
| QwQ-32B | 80.4 | 58.3 (-27%) | 53.3 (-34%) | 66.2 | 53.3 (-20%) | 39.2 (-41%) | **92.5** | **82.9** (-10%) |
| R1-Distill-Llama-70B | 65.4 | 48.8 (-25%) | 38.8 (-41%) | 50.0 | 38.8 (-22%) | 29.2 (-42%) | 89.8 | 77.3 (-14%) |

# Key Finding 2: The Formulation Bottleneck

- **Error type analysis**
  - Incorrect Problem Formulation accounts for ~80% of failures.
  - What This Means: Models fail to correctly map narrative descriptions to mathematical equations and constraints.



# Case Study (DeepSeek R1)

- **Original question**
  - Rectangles ABCD and EFGH are drawn such that D,E,C,F are collinear. Also, A,D,H,G all lie on a circle. If BC=16,AB=107,FG=17, and EF=184, what is the length of CE?
- **Generated Scenario**
  - At a large construction site, two rectangular platforms are being assembled for a robotics demonstration. The first platform, called the Alpha Pad, is 107 meters long and 16 meters wide. The second platform, the Echo Deck, is 184 meters long and 17 meters wide. The platforms are being arranged in a very specific layout: The back-right corner of the Alpha Pad, the front-left corner of the Echo Deck, the front-right corner of the Alpha Pad, and the front-right corner of the Echo Deck all lie along this straight track. Meanwhile, the back-left and the back-right corners of both platforms are all positioned along the edge of a circular fountain that marks the center of the plaza. Given this setup, how far apart are the front-right corner of the Alpha Pad and the front-left corner of the Echo Deck?

# Case Study (cont'd)

- **R1 response**
  - The platforms are arranged such that the back-right corner of the Alpha Pad (BR_A), the front-left corner of the Echo Deck (FL_E), the front-right corner of the Alpha Pad (FR_A), and the front-right corner of the Echo Deck (FR_E) lie on a straight track. Additionally, the back-left and back-right corners of both platforms lie on a circular fountain centered at the plaza's center.
  - The Alpha Pad is 107 meters long and 16 meters wide, and the Echo Deck is 184 meters long and 17 meters wide. The straight track is treated as the x-axis in a coordinate system, with the fountain center at (h, k) and radius r. The back corners on the fountain and the points on the track are used to determine the positions.
  - For the Alpha Pad:
    - BR_A is at (0, 0).
    - FR_A is at (107, 0), as the distance along the length is 107 meters.
    - BL_A is at (0, 16), assuming the platform is above the track ($y > 0$).
  - This gives k = 8 from the circle conditions, and $r^2 = h^2 + 64$.
  - For the Echo Deck:
    - FL_E is at (b, 0).
    - FR_E is at (b + 17, 0), as the width is 17 meters. Incorrect mathematical interpretation
    - BL_E is at (b, 184), assuming the platform is above the track ($y > 0$).

The subsequent reasoning process is correct, but based on wrong conditions.

# Quantitative Evidence

- **The Formulation Bottleneck**
  - Formulation accuracy decreases significantly as problem difficulty increases. (i.e., MATH→AIME24→AIME25).
  - Scaling improves performance (e.g., Qwen3-0.6B averages 42.8% vs. 75.0% for Qwen3-32B), but even GPT-5 remains below 75% on AIME25.

**Formulation Accuracy**

An LLM-based judger determines whether the model's output is mathematically equivalent to the original problem.

| Model | Formulation Accuracy (%) | | | |
|---|---|---|---|---|
| | MATH | AIME24 | AIME25 | Avg. |
| R1-Distill-Qwen-1.5B | 62.3 | 48.8 | 40.4 | 48.1 |
| R1-Distill-Qwen-7B | 75.7 | 57.9 | 47.1 | 57.1 |
| R1-Distill-Qwen-14B | 85.7 | 68.8 | 51.7 | 65.3 |
| R1-Distill-Qwen-32B | 87.3 | 71.7 | 59.6 | 70.0 |
| Qwen3-0.6B | 57.4 | 37.9 | 40.4 | 42.8 |
| Qwen3-1.7B | 82.5 | 57.1 | 56.7 | 62.0 |
| Qwen3-4B | 84.8 | 64.6 | 47.1 | 61.6 |
| Qwen3-8B | 86.3 | 77.9 | 63.3 | 73.8 |
| Qwen3-14B | 87.1 | 72.9 | 64.6 | 72.4 |
| Qwen3-32B | 89.3 | 77.1 | 65.8 | 75.0 |
| gpt5 | 90.5 | 85.0 | 73.3 | 81.4 |

# Key Finding 3: The Dual Bottleneck

- **Formulation is a Strong Prerequisite for Answer Correctness**
  - On average, necessity is consistently above accuracy across all models. For example, Qwen3-4B achieves 61.6% accuracy vs. 79.2% necessity, and GPT-5 81.4% vs. 85.6%.

Formulation Necessity $= P(F = \text{True} \mid R = \text{True})$.

| Model | Formulation Necessity (%) | | | |
|---|---|---|---|---|
| | MATH | AIME24 | AIME25 | Avg. |
| R1-Distill-Qwen-1.5B | 70.6 | 52.6 | 58.3 | 58.5 |
| R1-Distill-Qwen-7B | 82.9 | 67.0 | 60.8 | 67.7 |
| R1-Distill-Qwen-14B | 90.1 | 76.2 | 79.8 | 80.4 |
| R1-Distill-Qwen-32B | 91.8 | 78.5 | 73.9 | 79.3 |
| Qwen3-0.6B | 69.2 | 76.4 | 29.2 | 56.1 |
| Qwen3-1.7B | 90.1 | 79.1 | 75.8 | 80.0 |
| Qwen3-4B | 90.3 | 85.5 | 67.4 | 79.2 |
| Qwen3-8B | 91.3 | 86.7 | 77.2 | 83.8 |
| Qwen3-14B | 90.4 | 78.5 | 88.4 | 84.8 |
| Qwen3-32B | 92.0 | 77.2 | 81.4 | 81.9 |
| gpt5 | 94.5 | 87.6 | 79.2 | 85.6 |

# Key Finding 3: The Dual Bottleneck

- **Reasoning is the Second Bottleneck**
  - Sufficiency improves with scale but lags behind both accuracy and necessity.
  - A correct formulation does **not** guarantee a correct answer. The subsequent reasoning steps can still fail.

Formulation Sufficiency $= P(R = \text{True} \mid F = \text{True})$.

| Model | Formulation Sufficiency (%) | | | |
|---|---|---|---|---|
| | MATH | AIME24 | AIME25 | Avg. |
| R1-Distill-Qwen-1.5B | 65.0 | 13.8 | 12.5 | 23.5 |
| R1-Distill-Qwen-7B | 81.1 | 38.7 | 24.5 | 41.5 |
| R1-Distill-Qwen-14B | 80.8 | 47.9 | 37.0 | 50.1 |
| R1-Distill-Qwen-32B | 83.5 | 48.8 | 42.7 | 53.3 |
| Qwen3-0.6B | 52.2 | 6.0 | 1.7 | 13.5 |
| Qwen3-1.7B | 73.8 | 28.6 | 20.8 | 34.5 |
| Qwen3-4B | 83.6 | 57.1 | 54.2 | 61.3 |
| Qwen3-8B | 86.3 | 57.7 | 51.0 | 60.7 |
| Qwen3-14B | 85.2 | 61.0 | 62.2 | 66.3 |
| Qwen3-32B | 85.0 | 63.0 | 56.6 | 64.9 |
| gpt5 | 86.9 | 84.2 | 79.2 | 82.7 |

# Improving Contextual Mathematical Reasoning

- **Path 1: End-to-End Fine-Tuning**
  - **Setup**:
    - **Models**: Qwen3-Base series (4B, 8B, 14B).
    - **Data**: A new dataset of 50k contextual math problems, created by using Qwen3-32B to generate scenarios for DeepMath-103K and filtering them for mathematical equivalence.
    - **Three fine-tuning strategies**
      - $SFT_{ori}$: Using the original abstract problems only.
      - $SFT_{syn}$ : Using the synthetic scenario problems only.
      - $SFT_{mix}$ : Using a balanced mixture of both.

# Improving Contextual Mathematical Reasoning

- **Path 1: End-to-End Fine-Tuning**
  - The $SFT_{mix}$ approach consistently yields the best performance.
  - Explicitly training on narrative scenarios is effective, but it only partially closes the performance gap.

| Model | AIME 2024 (%) | | | AIME 2025 (%) | | | Math (%) | | Math-P (%) | | AMC23 | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ori | SG | CS | Ori | SG | CS | Ori | SG | Simple | Hard | | |
| Qwen3-4B-Base | 9.8 | 6.2 | 3.3 | 8.1 | 4.0 | 0.8 | 51.7 | 40.3 | 53.5 | 34.6 | 43.4 | 23.3 (+ 0.0%) |
| + $SFT_{Ori.}$ | 32.5 | 27.3 | 14.2 | 27.9 | 18.5 | 11.5 | 80.5 | 65.5 | 81.2 | 66.3 | 75.6 | 45.6 (+22.3%) |
| + $SFT_{Syn.}$ | 34.2 | **32.1** | 17.3 | **31.7** | 20.2 | 11.9 | 78.8 | 70.8 | 80.0 | 65.1 | 74.7 | 47.0 (+23.7%) |
| + $SFT_{Mix}$ | **36.9** | 31.7 | **19.2** | 30.2 | **22.1** | **12.7** | **81.4** | **72.3** | **82.7** | **69.0** | **78.1** | **48.8** (**+25.5%**) |
| Qwen3-8B-Base | 13.3 | 7.5 | 2.9 | 10.2 | 6.0 | 1.2 | 58.5 | 46.2 | 61.1 | 38.9 | 54.4 | 27.3 (+ 0.0%) |
| + $SFT_{Ori.}$ | 44.4 | 35.4 | 20.0 | 32.7 | 21.2 | 15.0 | 85.7 | 74.0 | **86.9** | 73.6 | 83.9 | 52.1 (+24.8%) |
| + $SFT_{Syn.}$ | **47.7** | **44.8** | **30.6** | **37.5** | 25.8 | **20.6** | 84.4 | 76.4 | 85.6 | 72.8 | 83.3 | 55.4 (+27.9%) |
| + $SFT_{Mix}$ | 46.2 | 42.4 | 29.5 | 35.9 | **26.8** | 20.5 | **85.9** | **76.7** | **86.9** | **74.4** | **86.6** | **55.6** (**+28.3%**) |
| Qwen3-14B-Base | 14.8 | 11.0 | 4.8 | 10.2 | 7.1 | 1.2 | 60.8 | 50.2 | 63.5 | 43.9 | 55.9 | 29.4 (+ 0.0%) |
| + $SFT_{Ori.}$ | 50.4 | 39.0 | 25.2 | 41.7 | 25.2 | 20.4 | 85.9 | 73.4 | 85.5 | 75.0 | 89.1 | 55.5 (+26.1%) |
| + $SFT_{Syn.}$ | **58.3** | 46.4 | **38.8** | **50.0** | 30.3 | 23.9 | 85.5 | 77.5 | **88.7** | **76.3** | 88.9 | 60.4 (+31.0%) |
| + $SFT_{Mix}$ | 56.5 | **52.5** | **38.8** | 47.2 | **34.6** | **26.5** | **86.7** | **77.8** | 88.2 | 76.1 | **89.8** | **61.3** (**+31.9%**) |

# Improving Contextual Mathematical Reasoning

- **Path 2: Decoupling with a Dedicated Formulation Model**
  - **Setup**:
    - **Models:** Qwen3-8B and Qwen3-14B used as both formulators and reasoners.
    - **Pipeline**
      - A "Formulation Model" is fine-tuned to translate scenarios into abstract math.
      - Its output is fed to a “Reasoning Model” to solve.

# Improving Contextual Mathematical Reasoning

- **Path 2: Decoupling with a Dedicated Formulation Model**
  - **Untuned**: Adding an untuned formulation stage introduces extra errors that propagate to reasoning.
  - **Tuned**: Formulation is difficult to learn effectively from scenario–original pairs alone.

| Reasoning Model | Formulation Model (Qwen3-8/14B) | | | | |
|---|---|---|---|---|---|
| | w/o | Untuned | | Tuned | |
| | | 8B | 14B | 8B | 14B |
| Qwen3-8B | 53.9 | 48.9 | 53.4 | 20.8 | 22.3 |
| Qwen3-14B | 57.7 | 51.8 | 56.2 | 21.8 | 24.6 |

# Conclusion & The Path Forward

- **Contextual mathematical reasoning** is a major, unsolved challenge for LLMs.
- The primary weakness is **problem formulation**, not calculation.
- Success is constrained by a **dual bottleneck**: models must both formulate the problem and execute the reasoning correctly.
- Future work should focus on methods that jointly improve formulation and reasoning skills.