ARB: Advanced Reasoning Benchmark for Large Language Models

Tomohiro Sawada[1,2], Daniel Paleka[1,3], Alexander Havrilla[1,2],
Pranav Tadepalli[1,2], Paula Vidas[1,3], Alexander Kranias[1,2], John J. Nay[4], Kshitij Gupta[1,5], Aran Komatsuzaki[1,2]

[1]DuckAI, [2]Georgia Tech, [3]ETH Zurich
[4]Stanford Center for Legal Informatics [5]MILA

Contributions:

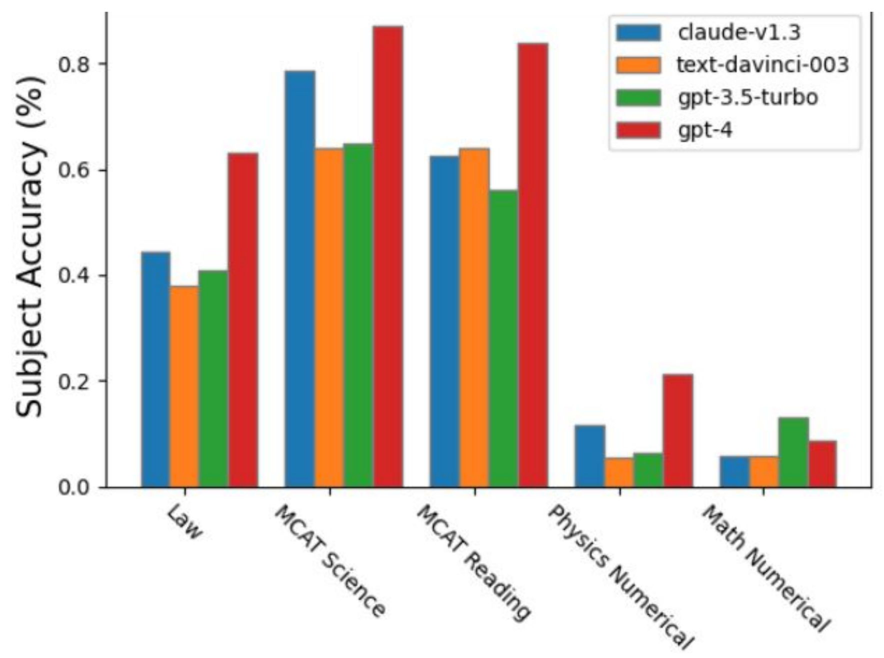
1. LLM benchmark with challenging math/physics problems with numerical, symbolic, and proof-like answers
2. introduce a rubric-based eval approach that uses GPT-4 to score intermediate reasoning steps
3. novel framework for classifying reasoning errors made by LMs

1. Hard LLM benchmark

- 1300 problems in multiple areas, 25% in math and physics
- Current models score badly!

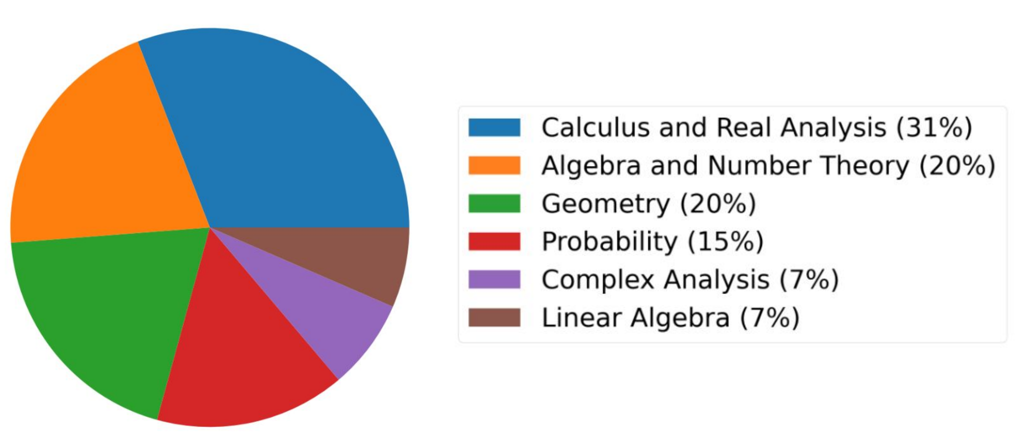
Table 1: Types of problems in the benchmark by subject area.

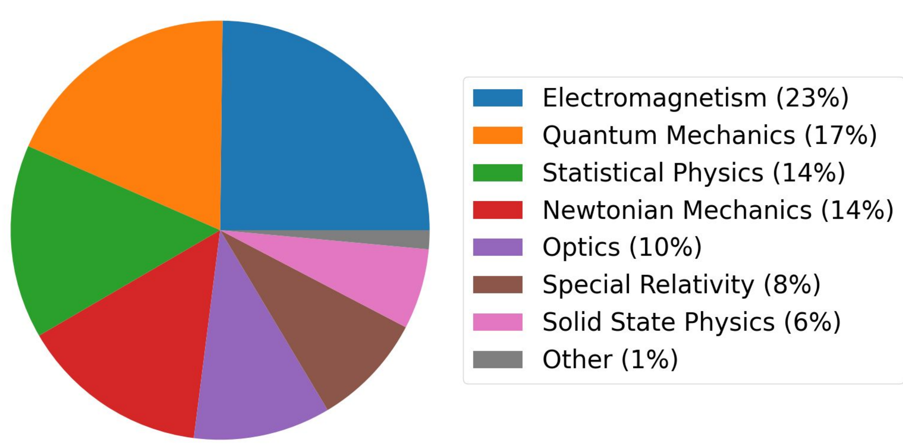
Subject	Answer Type	Number
Physics	Numerical	113
	Numerical (w/ image)	18
	Symbolic	51
	Symbolic (w/ image)	13
Mathematics	Numerical	69
	Symbolic	52
	Proof-like	19



- **Example problem:** Given a classical model of the tritium atom with a nucleus of charge +1 and a single electron in a circular orbit of radius r_0, suddenly the nucleus emits a negatron and changes to charge +2. (The emitted negatron escapes rapidly and we can forget about it.) The electron in orbit suddenly has a new situation. Find the ratio of the electron's energy after to before the emission of the negatron (taking the zero of energy, as usual, to be for zero kinetic energy at infinite distance).

- Diverse coverage:





2. Rubric-based auto-evaluation

- Problem: Self-evaluation with LLMs is unreliable
- Proposal: Multi-step rubric-based grading
  - 1. Get LLM-generated solution
  - 2. Generate rubric from reference solution
  - 3. Grade with GPT-4, following the rubric
- Example of the procedure: =>
- Results: surprisingly large correlation with human graders!

	Physics Symbolic	Math Symbolic	Proof-like
Human eval score	5.00	3.13	2.65
Model eval score	5.05	3.37	3.8
Correlation	0.91	0.78	0.82

Table 7: Example model generated rubric and evaluation. Evaluation with a rubric reveals that the model begins with the model makes partial progress, but is unable to arrive at the correct final answer.

Problem Statement

Find the maximum area of all triangles that can be inscribed in an ellipse with semiaxes a and b, and describe the triangles that have maximum area.

Ground Truth Solution

Using the parameterization

$$x = a\cos t, y = b\sin t,$$

a triple of points on the ellipse is given by

$$(a\cos t_i, b\sin t_i), \quad i = 1,2,3.$$

So the area of an inscribed triangle is given by

$$\frac{1}{2}\begin{vmatrix} 1 & a\cos t_1 & b\sin t_1 \\ 1 & a\cos t_2 & b\sin t_2 \\ 1 & a\cos t_3 & b\sin t_3 \end{vmatrix} = \frac{ab}{2}\begin{vmatrix} 1 & \cos t_1 & \sin t_1 \\ 1 & \cos t_2 & \sin t_2 \\ 1 & \cos t_3 & \sin t_3 \end{vmatrix}$$

which is ab times the area of a triangle inscribed in the unit circle. In the case of the circle, among all inscribed triangles with a given base $2w(0 < w \leqslant 1)$, the one of maximum area is an isosceles triangle whose area equals

$$g(w) = w\left(1+\sqrt{1-w^2}\right).$$

Using elementary calculus one finds that the maximum of g on the interval $0 \leqslant w \leqslant 1$ occurs at $w = \sqrt{3}/2$, corresponding to an equilateral triangle, and equals $3\sqrt{3}/4$. Alternatively, fixing one side of the triangle as the basis, we easily see that among all the inscribed triangles the one with the greatest area is isosceles because of the maximum height, showing that the angle at the basis is the same. Fixing another side we see that the triangle is indeed equilateral. Hence, the area is maximal when

$$t_2 = t_1 + \frac{2\pi}{3} \text{ and } t_3 = t_2 + \frac{2\pi}{3}$$

that is, when the corresponding triangle inscribed in the unit circle is regular.
For the ellipse with semiaxes a, b, this corresponds to an inscribed triangle with maximum area equals $3ab\sqrt{3}/4$.

Model Generated Rubric

- Parameterize the ellipse using $x = a\cos t, y = b\sin t$ (2 points).
- Write the expression for the area of an inscribed triangle using the determinant formula (2 points):

$$\frac{ab}{2}\begin{vmatrix} 1 & \cos t_1 & \sin t_1 \\ 1 & \cos t_2 & \sin t_2 \\ 1 & \cos t_3 & \sin t_3 \end{vmatrix} \quad (1)$$

- Connect the problem to the maximum area of an inscribed triangle in the unit circle (2 points).
- Correctly identify that the maximum area occurs when the inscribed triangle is equilateral (2 points).
- Find the maximum area of an inscribed triangle in the ellipse, and provide the correct answer (2 points):

$$\frac{3ab\sqrt{3}}{4} \quad (2)$$

Model Response

To maximize the area of the inscribed triangle, we can use the following approach:
1. Parameterize the ellipse: The equation of an ellipse with semiaxes a and b is given by $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$. We can parameterize the ellipse using the parameter t as follows: $x = a\cos(t)$ and $y = b\sin(t)$, where $0 \le t \le 2\pi$.
2. Determine the vertices of the triangle: Let $A(a\cos(t_1), b\sin(t_1))$, $B(a\cos(t_2), b\sin(t_2))$, and $C(a\cos(t_3), b\sin(t_3))$ be the vertices of the inscribed triangle, where $0 \le t_1, t_2, t_3 \le 2\pi$.
3. Calculate the area of the triangle: Using the Shoelace Theorem, the area of the triangle is given by:

$$\text{Area} = \frac{1}{2}\left|\sum_{i=1}^{3} x_i y_{i+1} - \sum_{i=1}^{3} x_{i+1} y_i\right|$$

where $(x_4, y_4) = (x_1, y_1)$.
Substituting the parameterized coordinates, we get:

$$\text{Area} = \frac{1}{2}|a^2\cos(t_1)\sin(t_2) + a^2\cos(t_2)\sin(t_3) + a^2\cos(t_3)\sin(t_1) - a^2\cos(t_2)\sin(t_1) - a^2\cos(t_3)\sin(t_2) - a^2\cos(t_1)\sin(t_3)|$$

Simplifying the expression, we get:

$$\text{Area} = \frac{a^2}{2}|\cos(t_1)\sin(t_2) - \cos(t_2)\sin(t_1) + \cos(t_2)\sin(t_3) - \cos(t_3)\sin(t_2) + \cos(t_3)\sin(t_1) - \cos(t_1)\sin(t_3)|$$

Using the sine difference identity, $\sin(A-B) = \sin(A)\cos(B) - \cos(A)\sin(B)$, we can rewrite the area as:

$$\text{Area} = \frac{a^2}{2}|\sin(t_1 - t_2) + \sin(t_2 - t_3) + \sin(t_3 - t_1)|$$

4. Maximize the area: Notice that the maximum value of each sine term is 1, and this occurs when the angles t_1, t_2, t_3 are equally spaced around the ellipse, i.e., when the triangle is equilateral. In this case, the area becomes:

$$\text{Area} = \frac{a^2}{2}|1+1+1| = \frac{3a^2}{2}$$

5. Relate the maximum area to the semiaxes: Since the triangle is equilateral, the ratio of the area of the triangle to the area of the ellipse is constant. The area of the ellipse is given by πab. Thus, we can write:

$$\frac{\text{Area}_{\text{max}}}{\pi ab} = \frac{\frac{3a^2}{2}}{\pi ab}$$

Solving for the maximum area, we get:

$$\text{Area}_{\text{max}} = \frac{3a^2}{2} \cdot \pi ab = \frac{3\pi a^3 b}{2}$$

Thus, the maximum area of an inscribed triangle in an ellipse with semiaxes a and b is $\frac{3\pi a^3 b}{2}$, and the triangles with maximum area are equilateral.
ANSWER: $\frac{3\pi a^3 b}{2}$

Rubric Based Evaluation

1. The student parameterized the ellipse correctly using $x = a\cos(t)$ and $y = b\sin(t)$. (2 points)
2. The student used the Shoelace Theorem to find the area of the triangle, but did not use the determinant formula as specified in the rubric. (0 points)
3. The student did not explicitly connect the problem to the maximum area of an inscribed triangle in the unit circle. (0 points)
4. The student correctly identified that the maximum area occurs when the inscribed triangle is equilateral. (2 points)
5. The student found the maximum area of the inscribed triangle in the ellipse, but their answer is incorrect. The correct answer is $\frac{3ab\sqrt{3}}{4}$, not $\frac{3\pi a^3 b}{2}$. (0 points)
Rubric Score: 4 points

3. Error analysis (GPT-4)

Table 3: Mistakes on mathematics and physics problems in ARB, GPT-4.

	Misread problem	Wrong approach	Logical error or hallucination	Arithmetic mistake	Correct answer	Correct reasoning
Math Numerical	0%	25%	88%	48%	3%	3%
Math Symbolic	16%	50%	29%	4%	16%	16%
Math Proof-like	5%	50%	72%	16%	n/a	5%
Physics Numerical	0%	80%	53%	6%	6%	6%
Physics Symbolic	0%	37%	68%	31%	28%	12%

- Access the paper and the dataset here:

arb.duckai.org

DuckAI