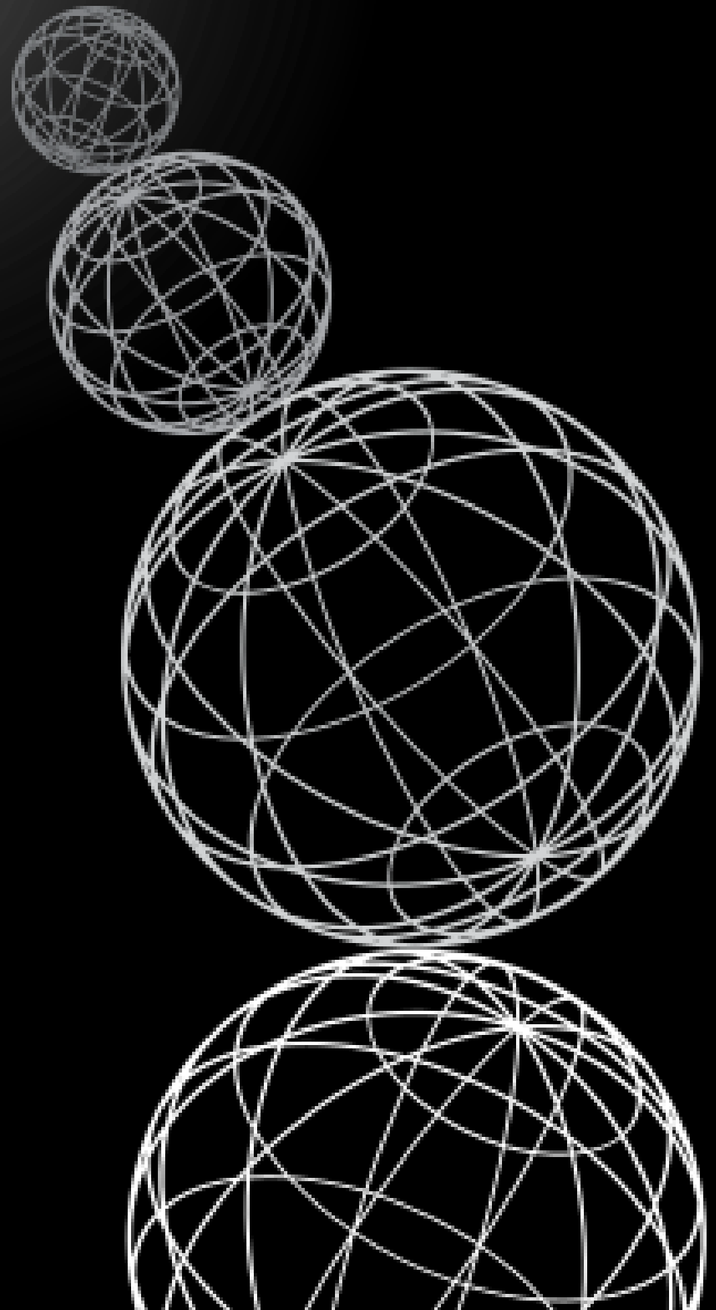
MULTIFACETED EVALUATION OF LARGE LANGUAGE MODELS' POTENTIAL FOR SCIENTIFIC DISCOVERY

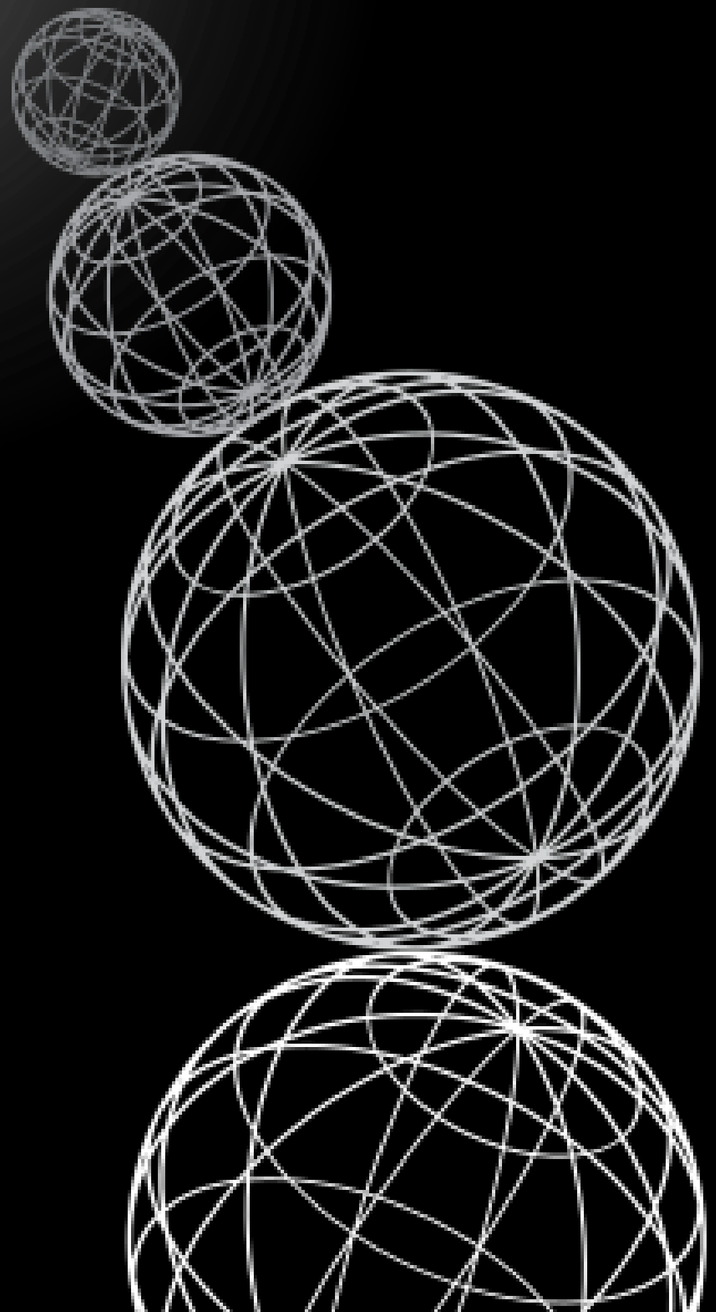
By: Olivia Barsemian, Sophie Chen, XiaoQing Cui, Sahasra Datla, Daniel Egusquiza Bronstein, Sreethan Gangavarapu, Justin Lambert, John Loatman, Nathaniel Nadler, Shivan Shrinivas, Michael Xu, Adhithya Yuvaraj

Advisor: Dr. Minjoon Kouh

Assistants: John McLaughlin, Giorgi Tsnobiladze



Can LLMs do Science?

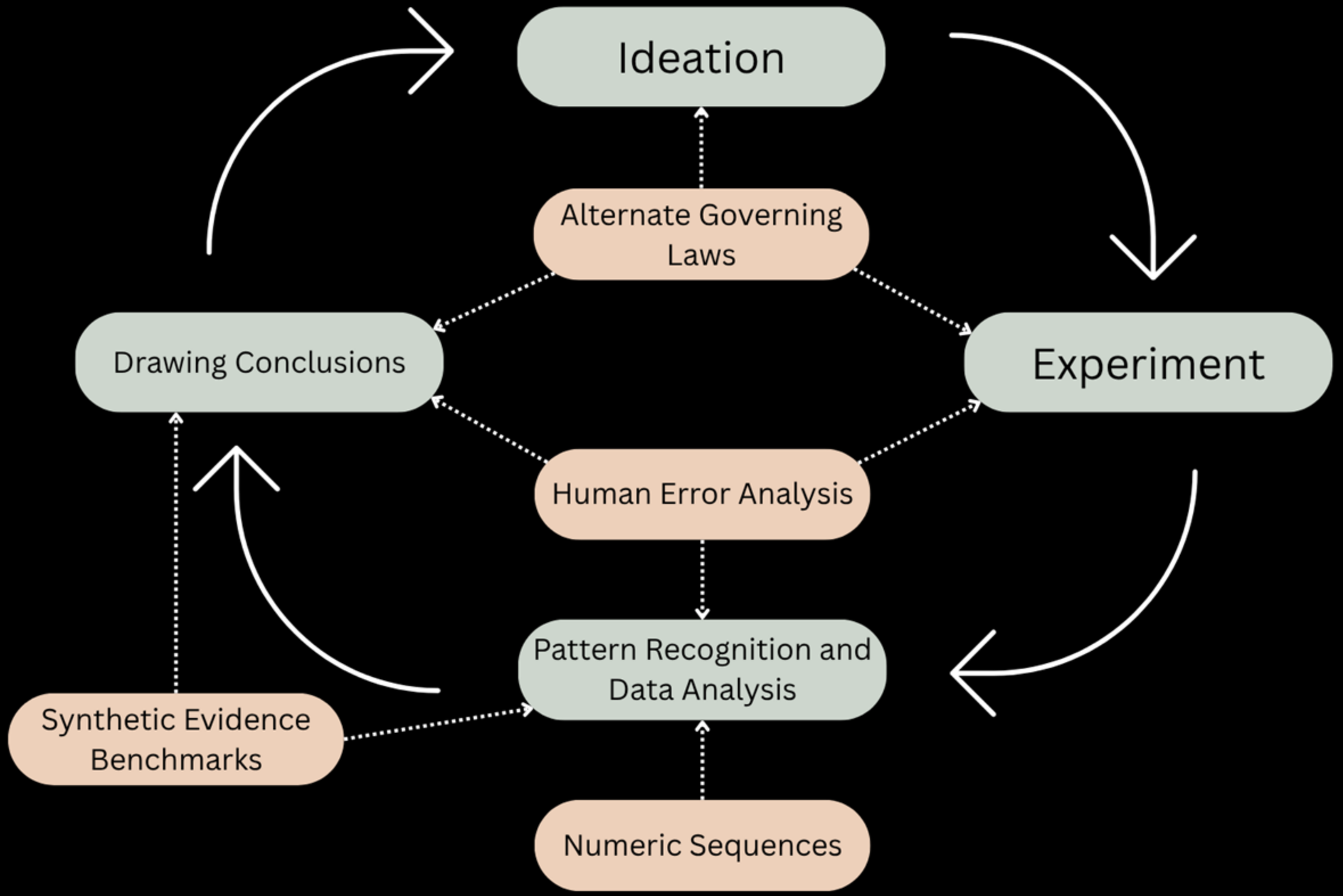


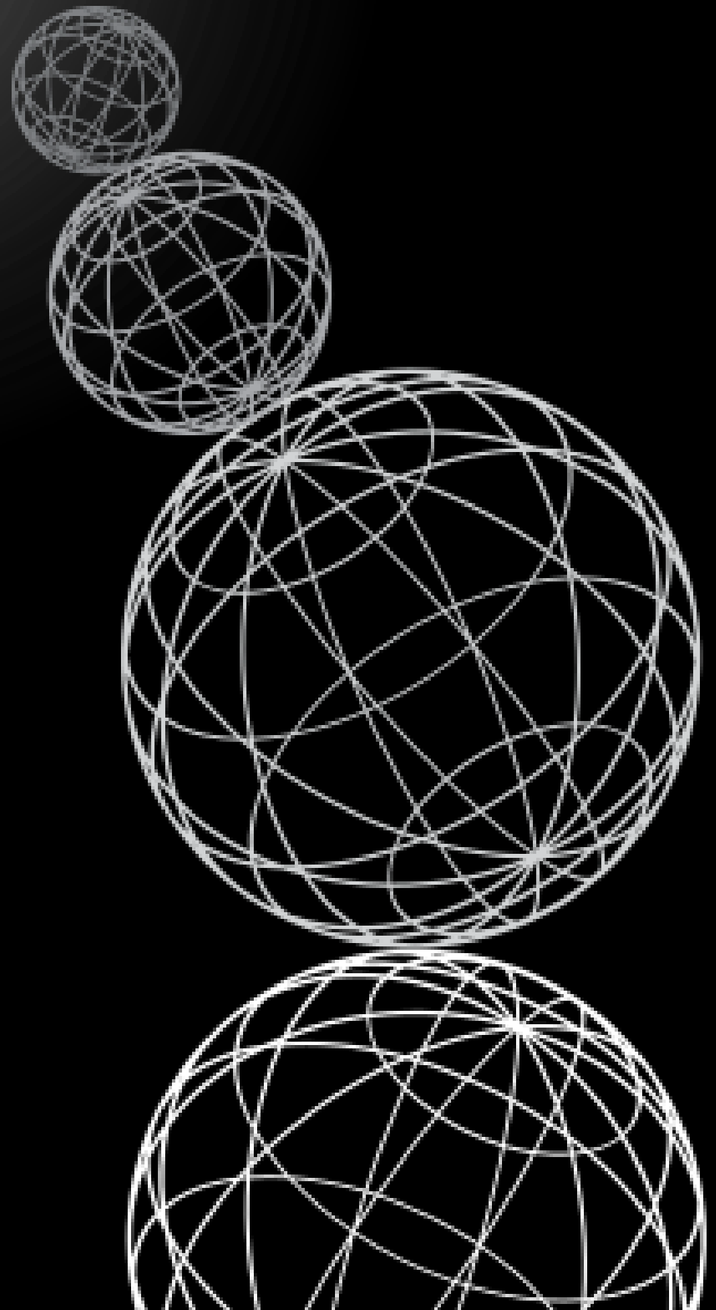
Can LLMs do Science?

Many benchmarks already exist

“Doing science” is very broad

Evaluate specific skills in the scientific method


Ideation
Alternate Governing Laws
Drawing Conclusions
Experiment
Human Error Analysis
Pattern Recognition and Data Analysis
Synthetic Evidence Benchmarks
Numeric Sequences



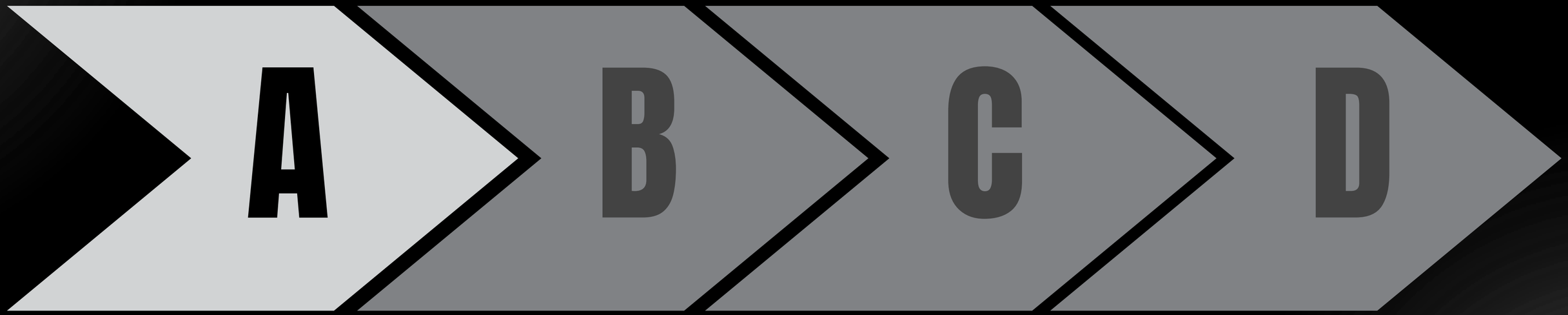
Can LLMs do Science?

Models: ChatGPT, Claude, Gemini

Limited Background Knowledge

Process

PART A: Can LLMs find numerical patterns?





Guess the Pattern!

3, 4, 5, 2, 5, 5

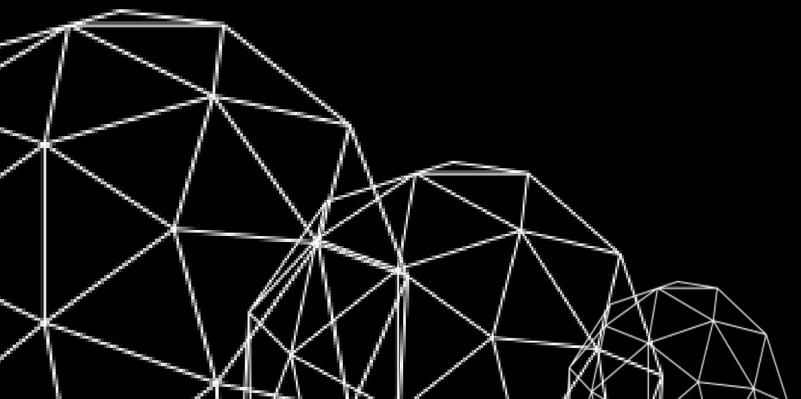
→ π (every other digit!)

1, 4, 27, 256, 3125, 46656

→ n^n

Why are we testing these sequences?

- **Constant**: knowledge retrieval
- **Closed form**: mathematical relationships
- **Others:** recursive, number-theoretic, combinatorial, specialized



0 1 5 13 26 50 **65 85**

Babylonian Spiral

[71, 105]

$a(n) = n(n^2 + 6n - 1)/6$

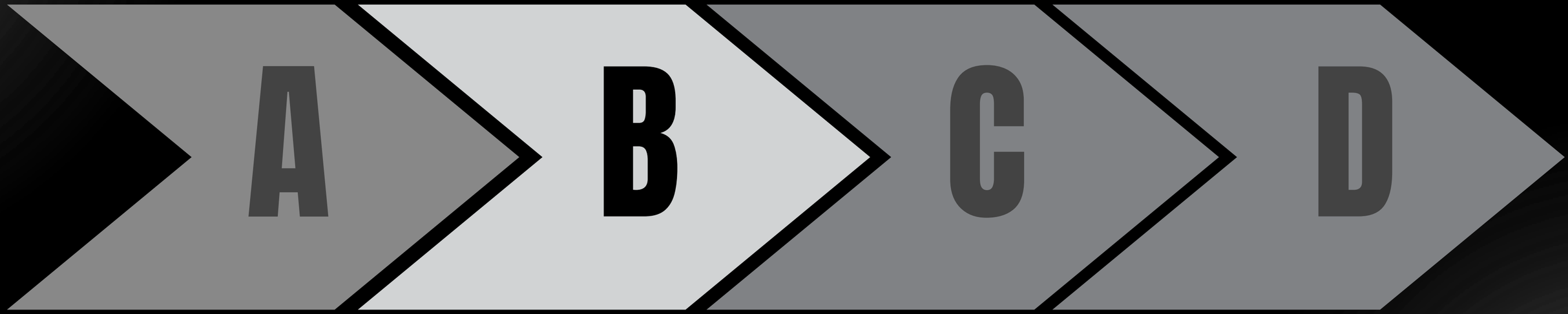
“ Treating 50 as a likely typo for 45, the continuation is 71, 105. If 50 is taken as literally correct, no simple polynomial or linear recurrence fits all six terms cleanly.”



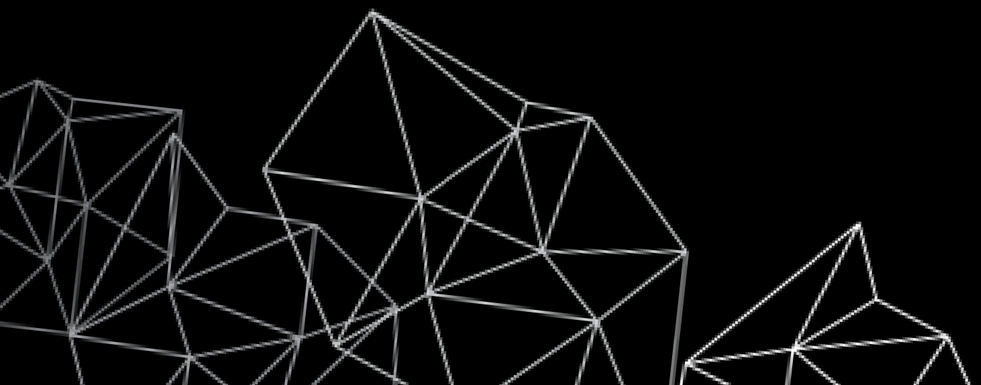
Results

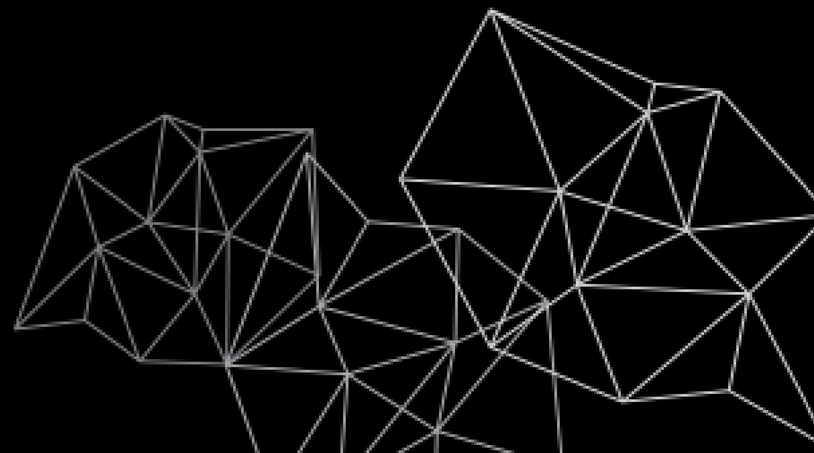
Metric	ChatGPT	Claude
Intended	43.8%	62.5%
Alternative	25.0%	22.9%
Incorrect	31.3%	14.6%
Valid Reasoning Rate	68.8%	85.4%

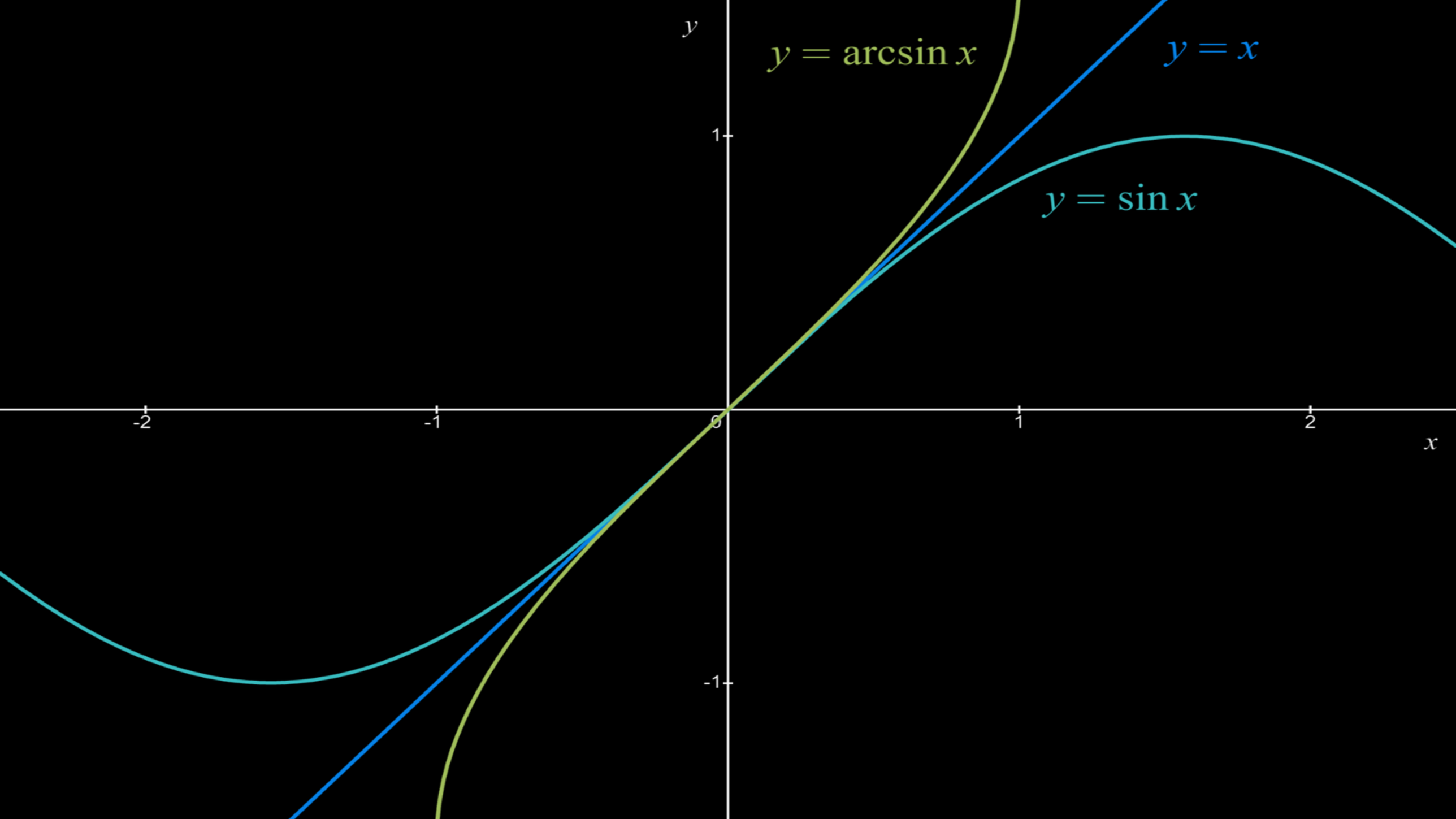
PART B: Do LLMs know when they know enough?

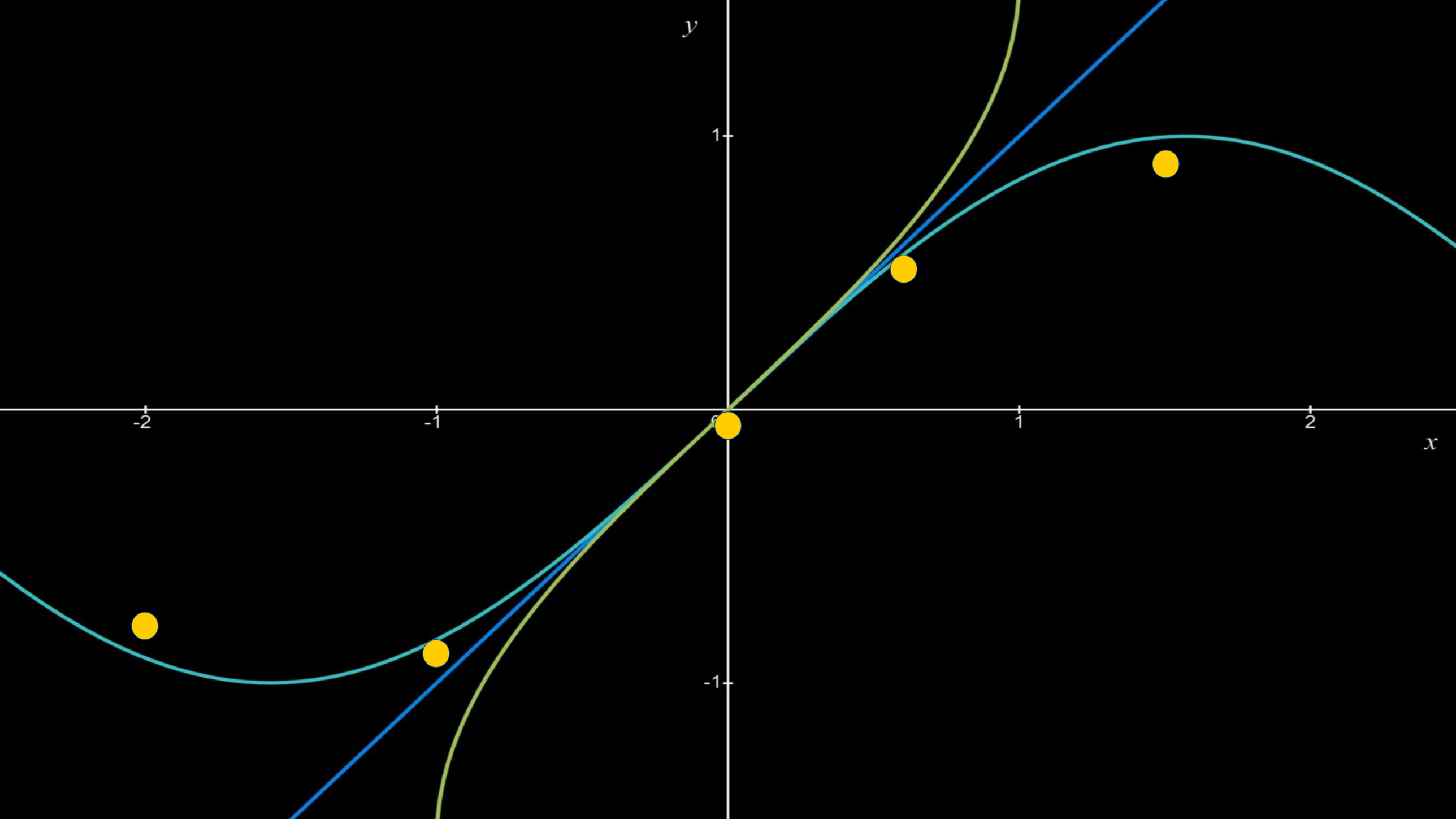


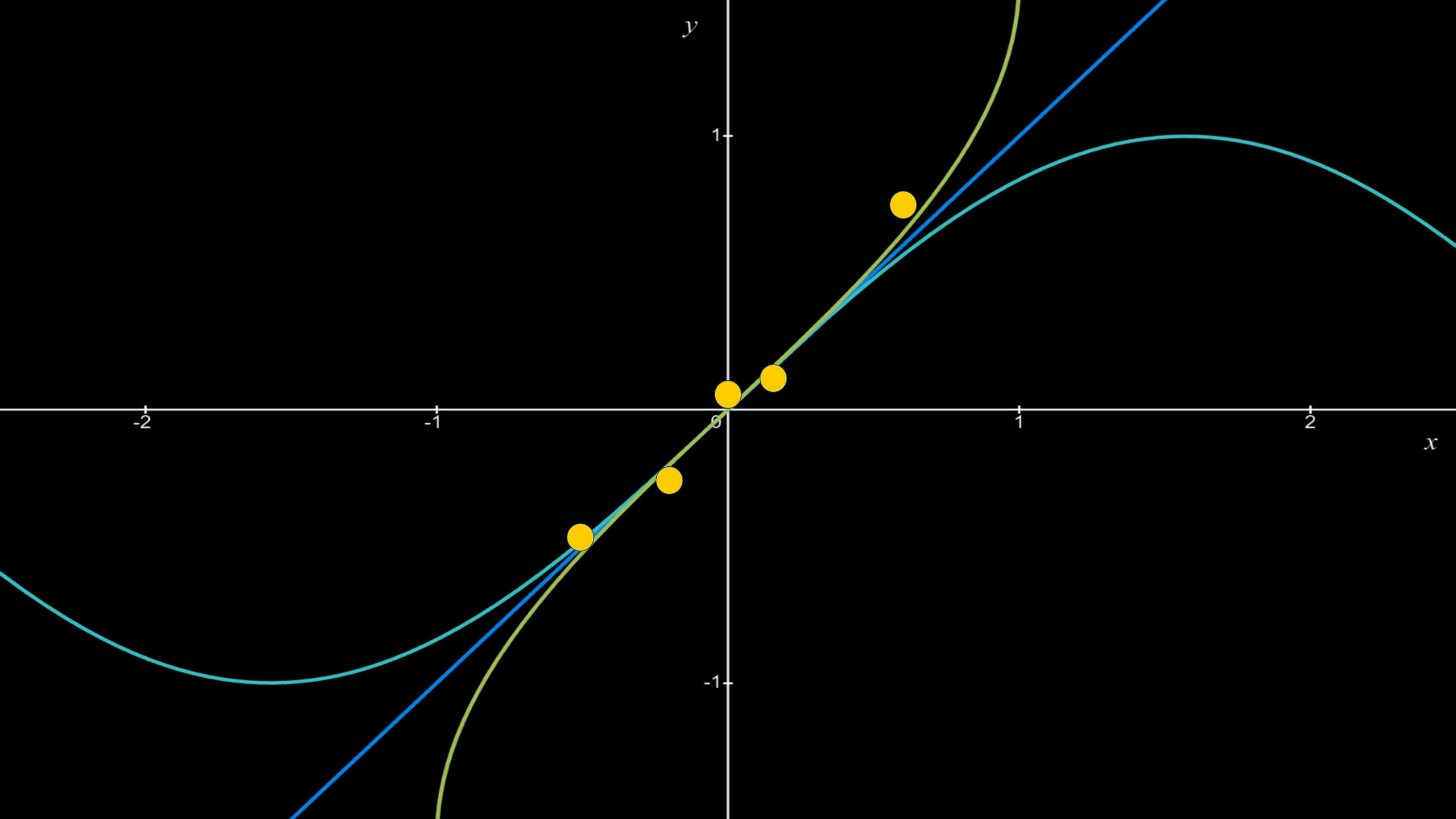
Can LLMs judge when evidence is sufficient?

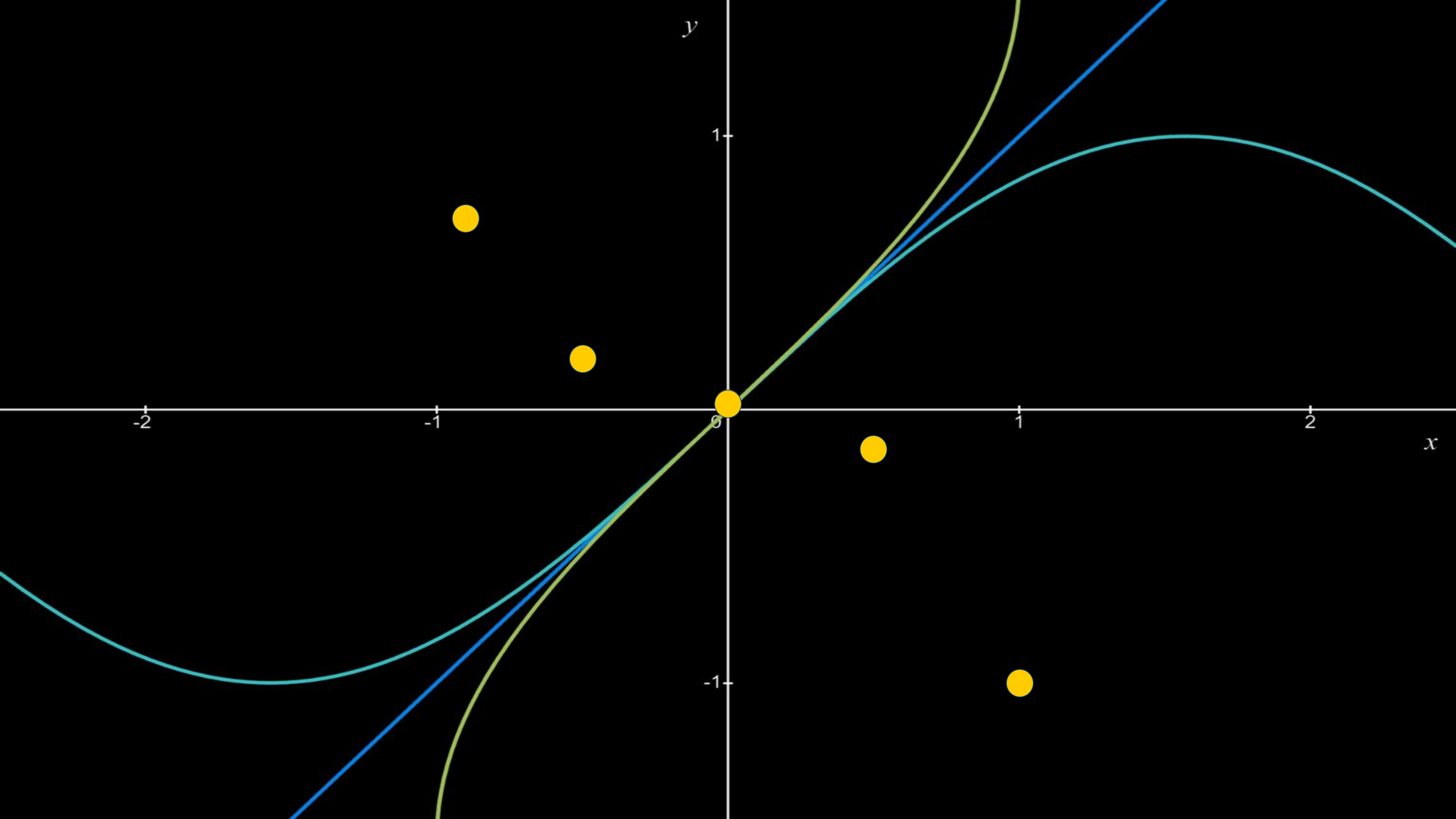




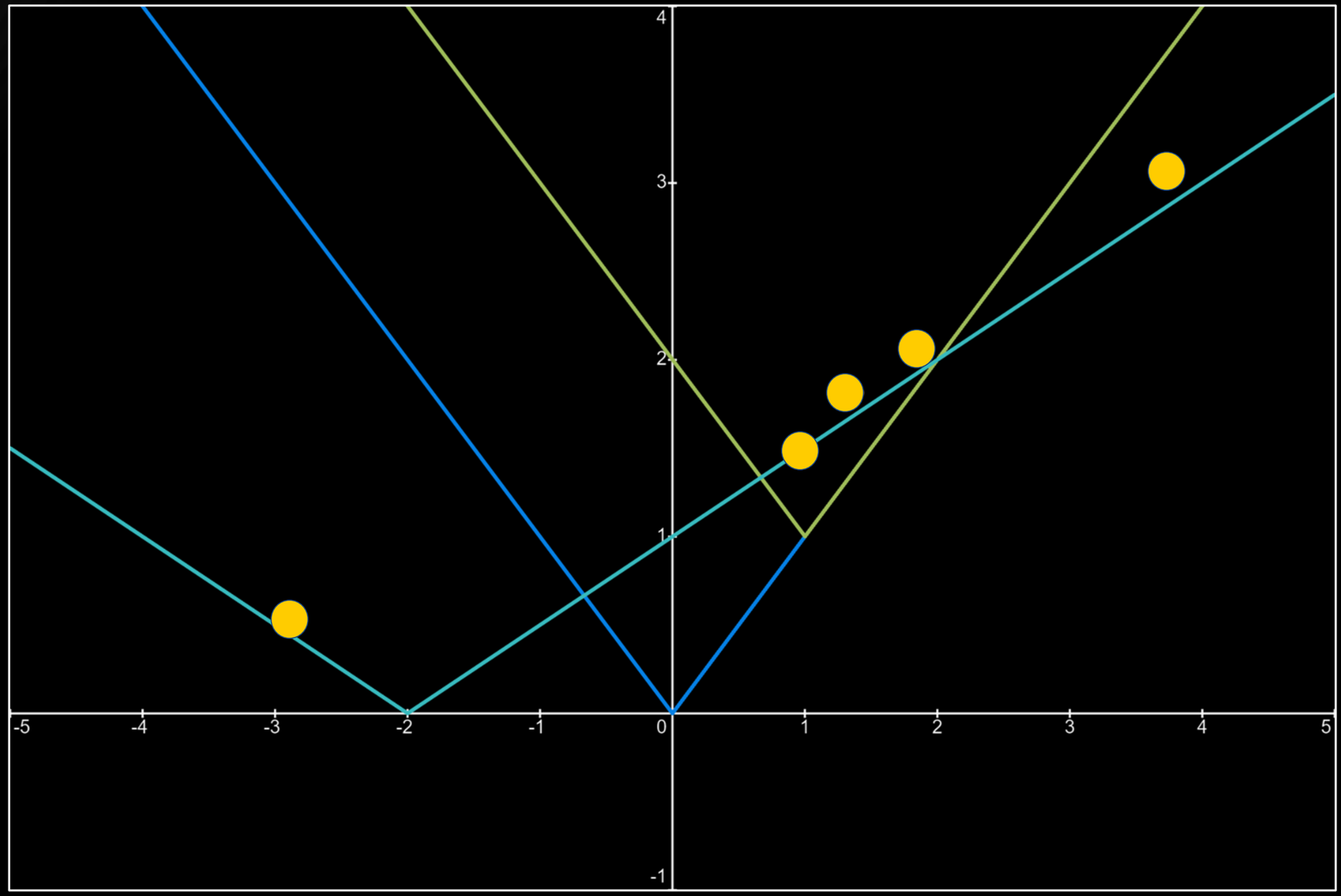

y
x
y = arcsin x
y = x
y = sin x
-2
-1
0
1
2
1
-1


y
1
-2
-1
1
2
x
-1


y
1
-2
-1
0
1
2
x
-1


y
1
-2
-1
0
1
2
x
-1

LLM Examples

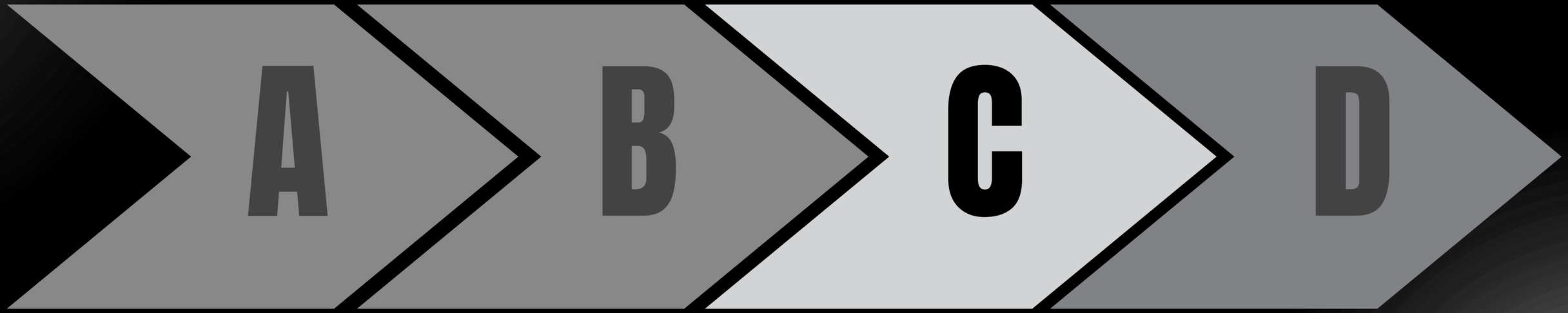


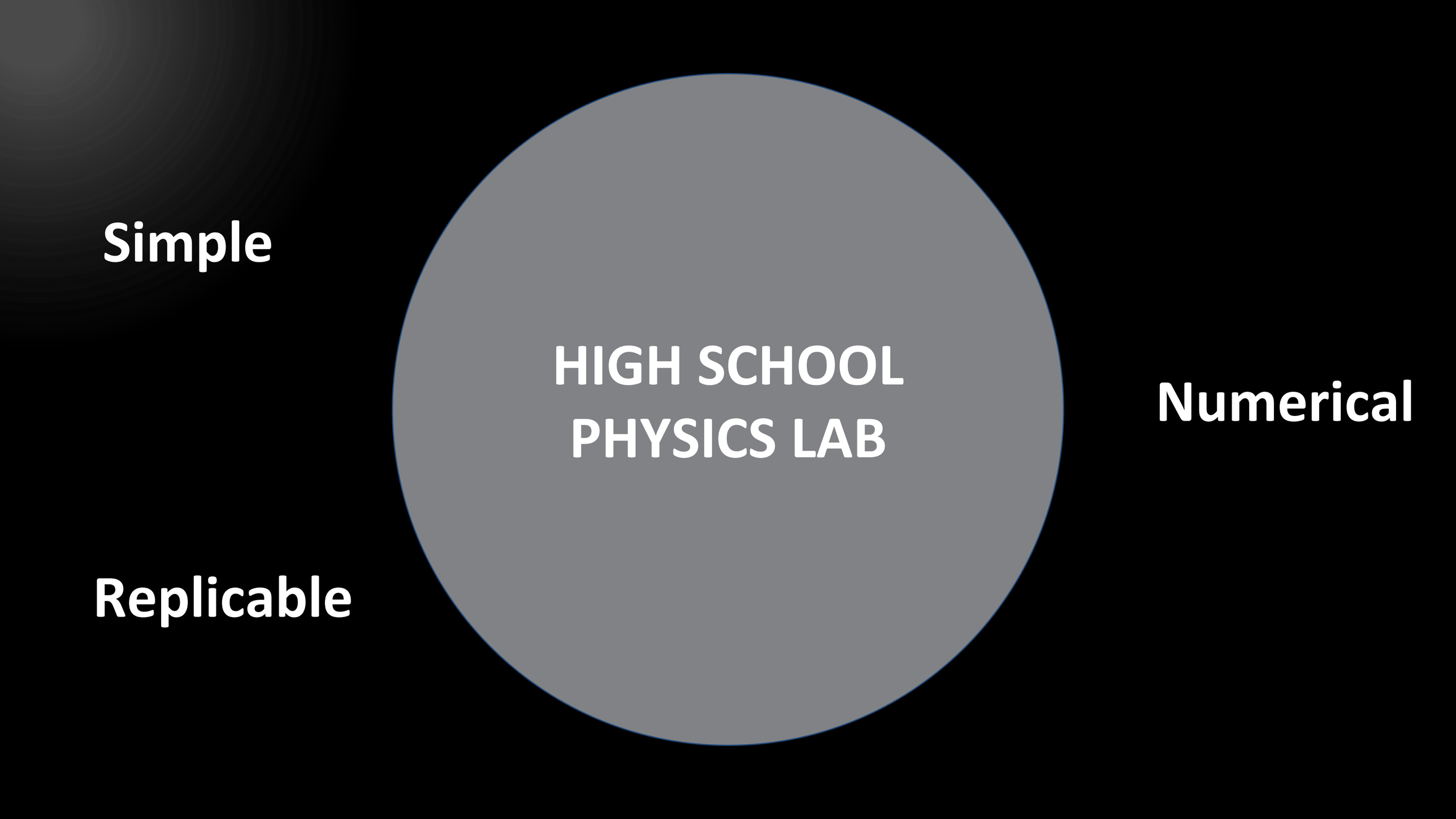
Error Patterns and Implications

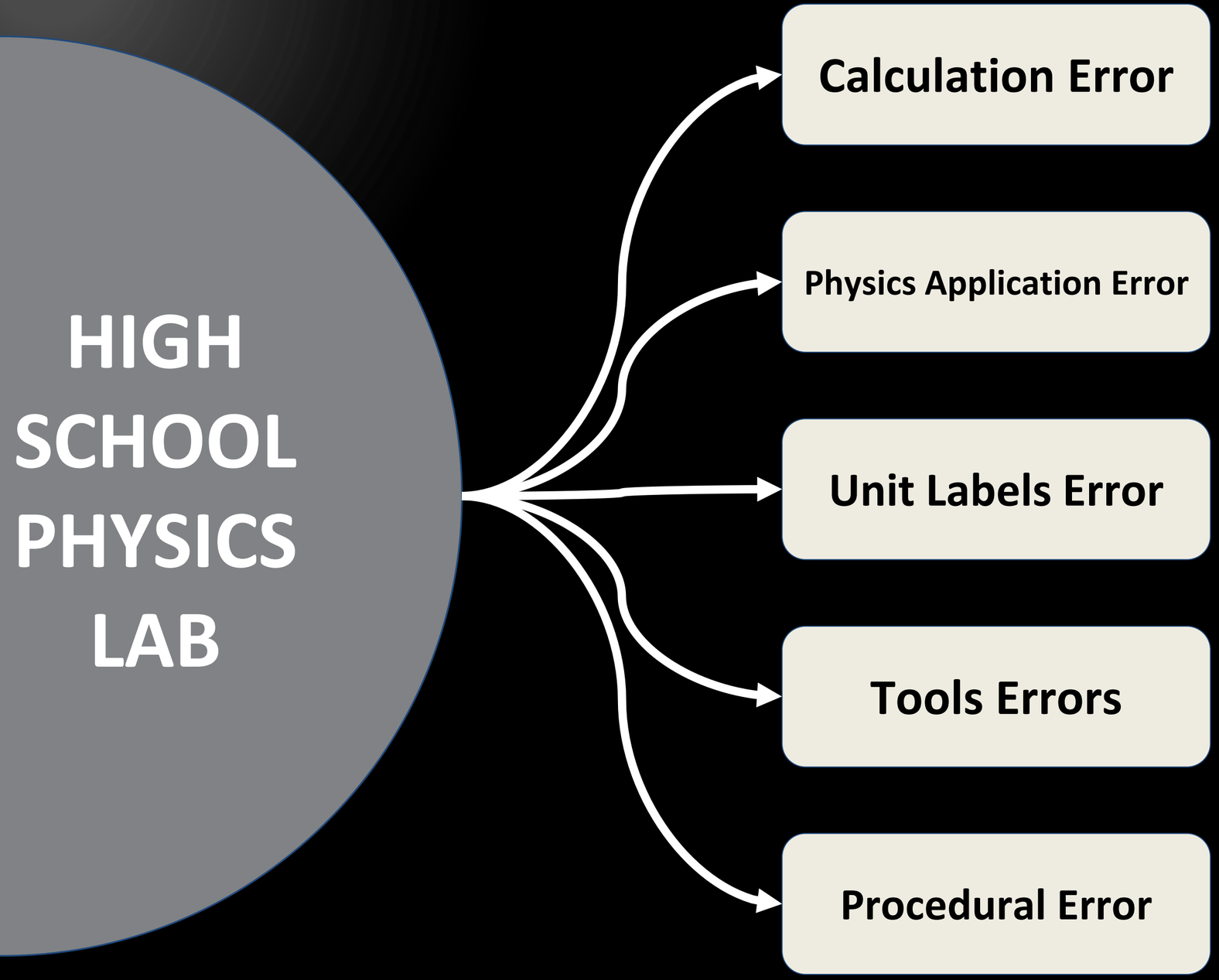
LLMs were overly cautious

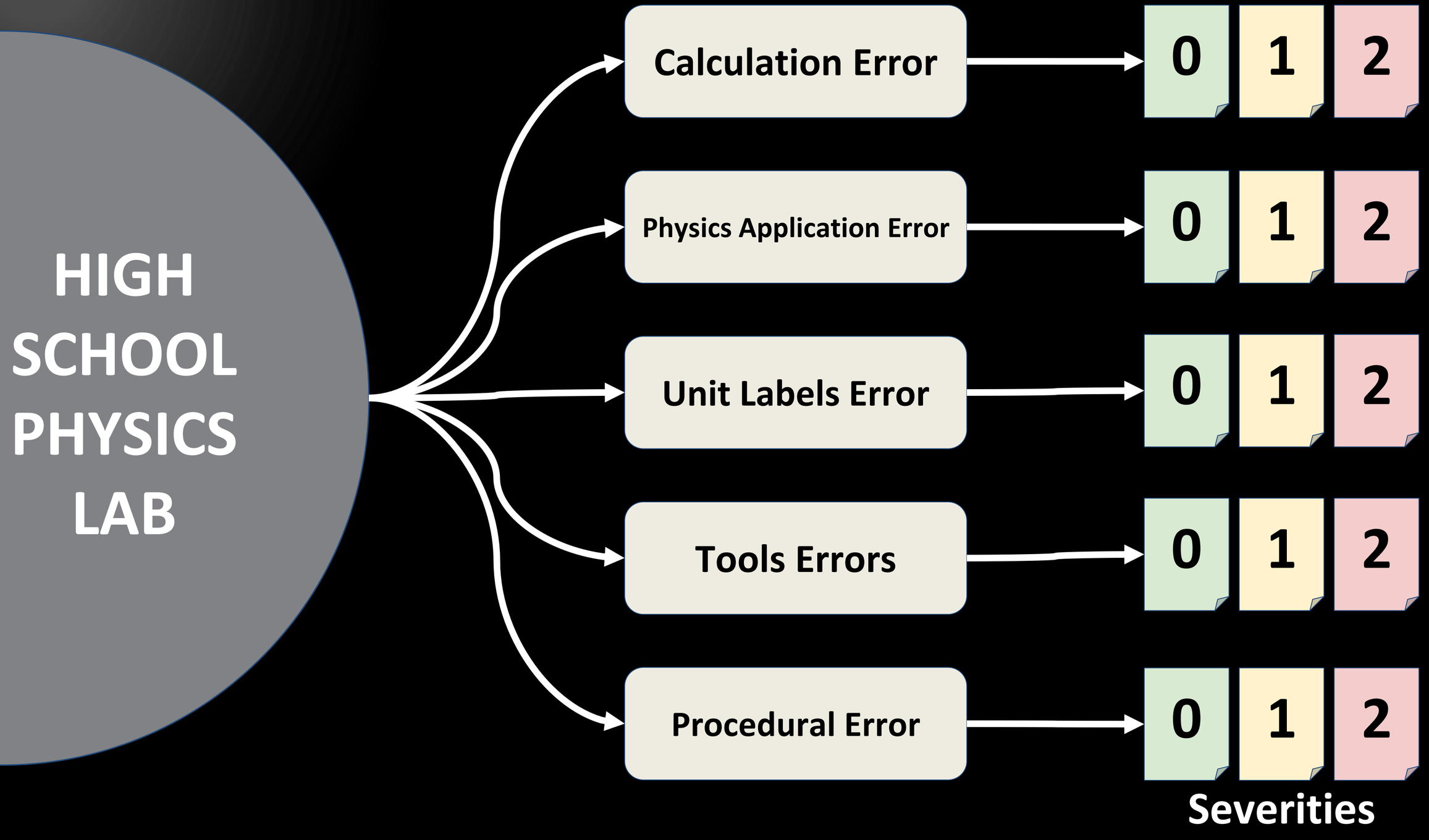
Manners of Failure > Accuracy

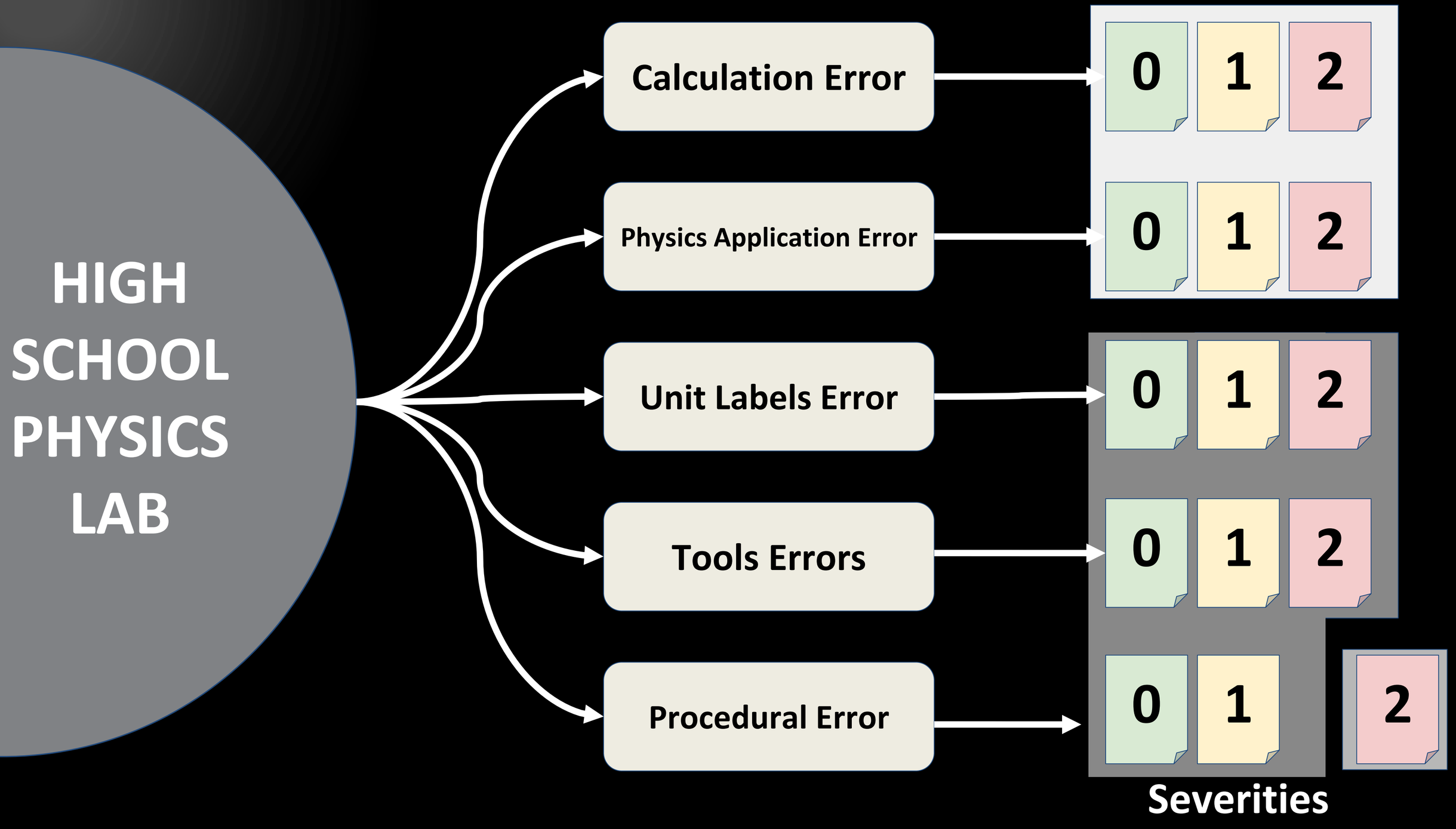
PART C: Can LLMs find and prove errors?

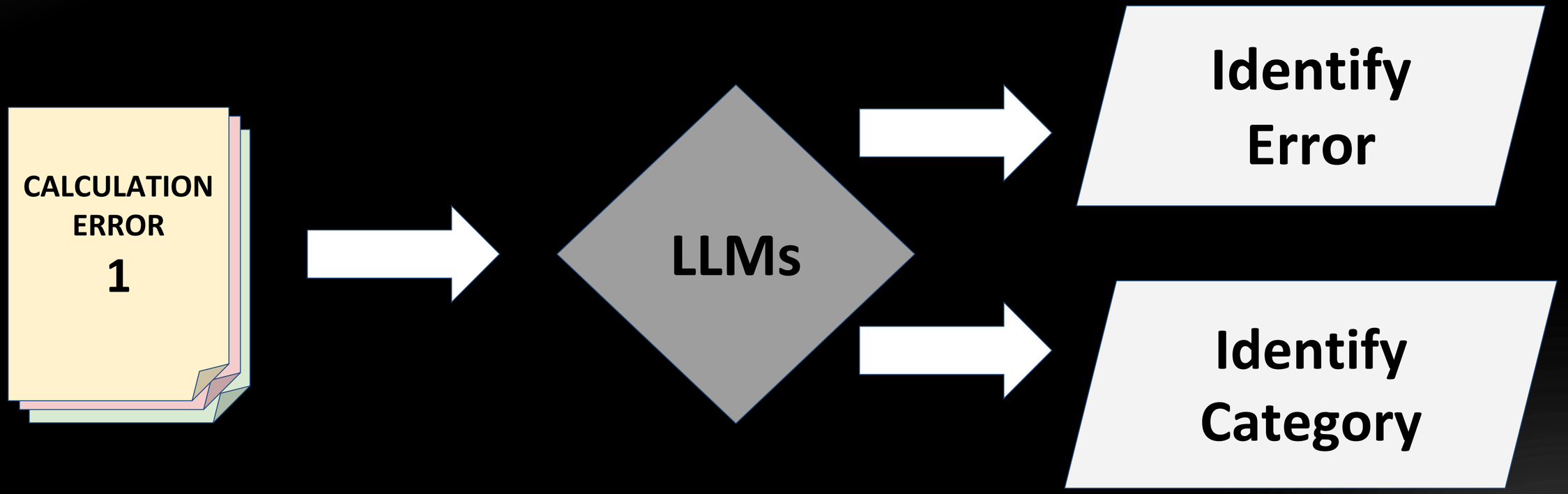



Simple
HIGH SCHOOL
PHYSICS LAB
Numerical
Replicable


HIGH SCHOOL PHYSICS LAB
Calculation Error
Physics Application Error
Unit Labels Error
Tools Errors
Procedural Error


HIGH SCHOOL PHYSICS LAB
Calculation Error
Physics Application Error
Unit Labels Error
Tools Errors
Procedural Error
0 1 2
0 1 2
0 1 2
0 1 2
0 1 2
Severities

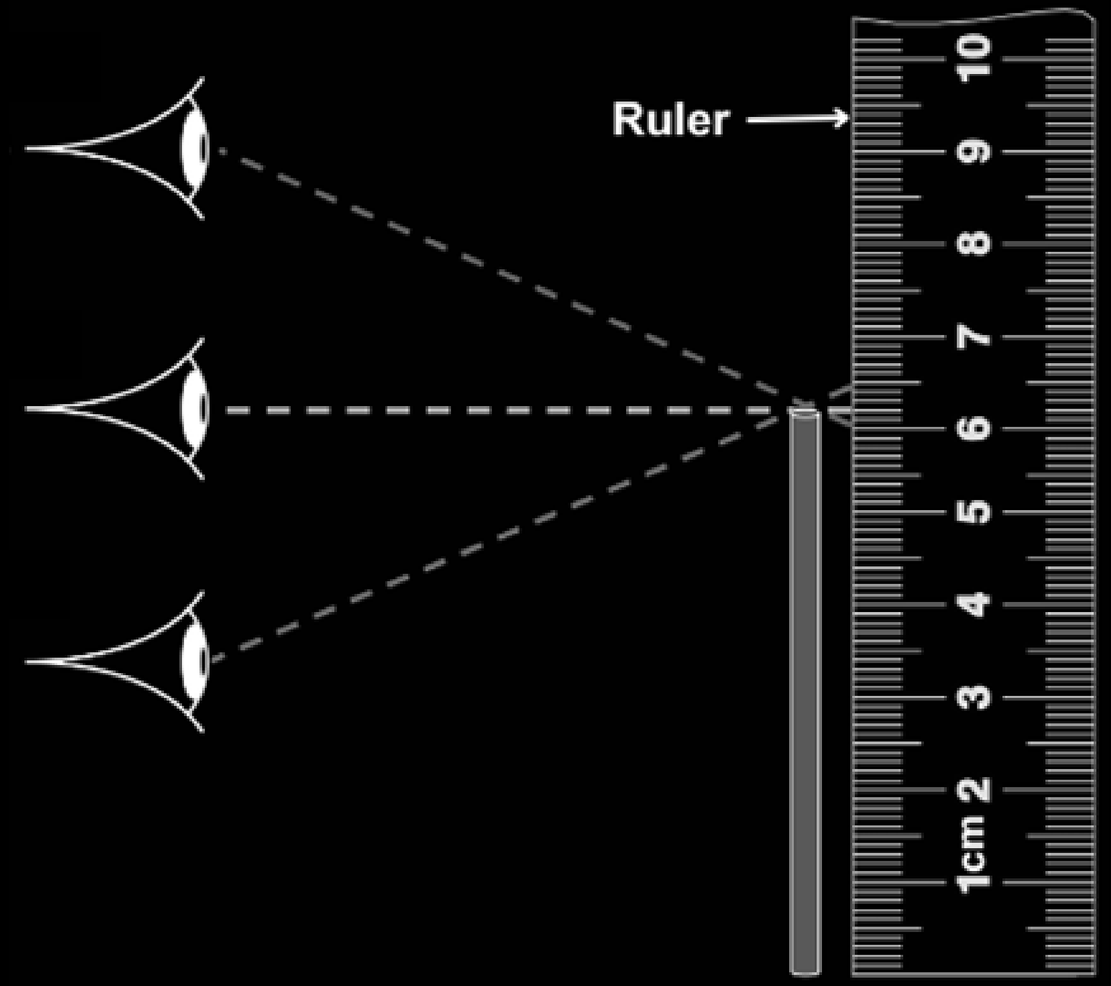

HIGH SCHOOL PHYSICS LAB
Calculation Error
Physics Application Error
Unit Labels Error
Tools Errors
Procedural Error
0 1 2
0 1 2
0 1 2
0 1 2
0 1
2
Severities


CALCULATION
ERROR
1
LLMs
Identify
Error
Identify
Category

Strong Identification

BUT

Textbook Fallback



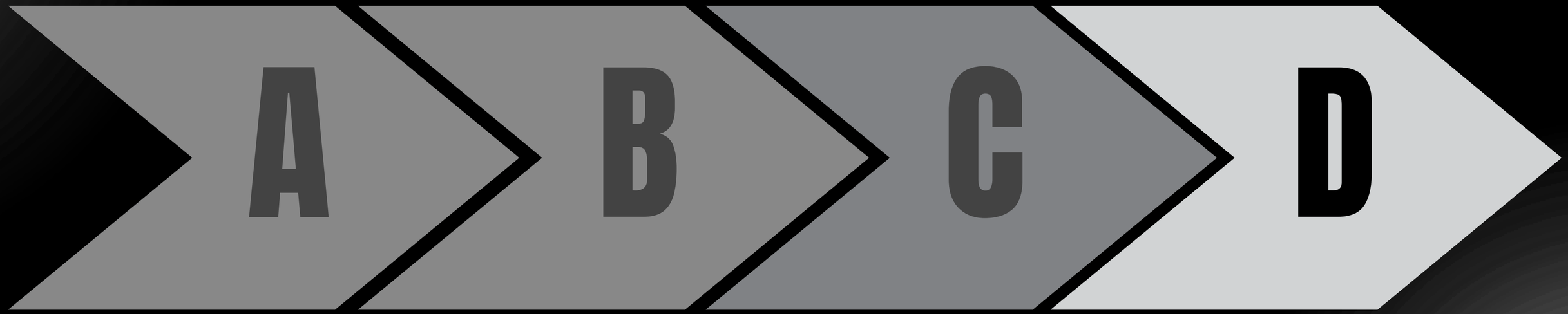
The measurements alone do not identify the single human error.
However, these possibilities could be distinguished [...]. Therefore, the uncertainty is **resolvable**, not fundamentally non-identifiable.

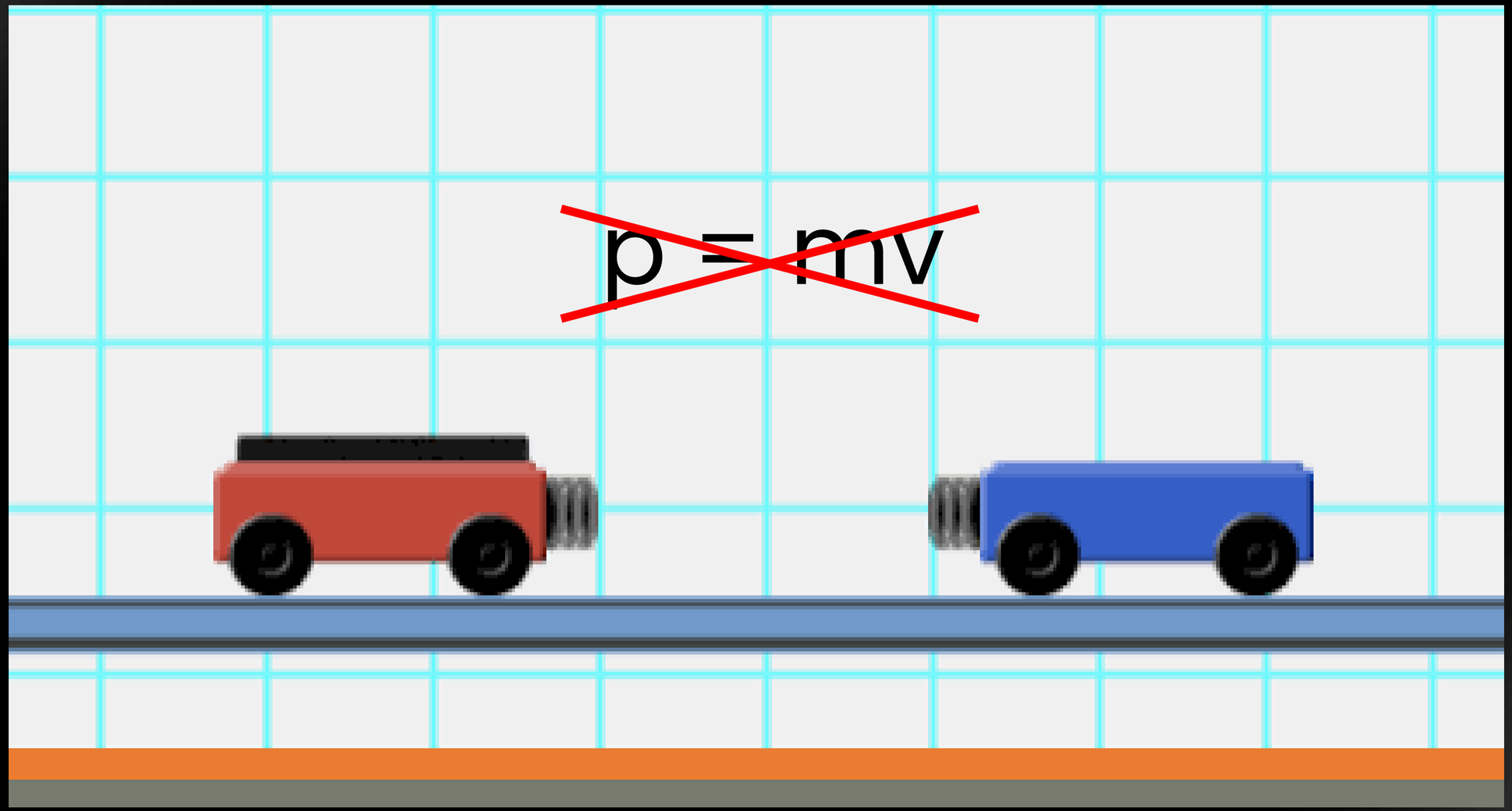
There is also a physics error in the stated final result.



Self-Contradictory Categorization

PART D: Can LLMs *discover* physical concepts?




p = mv

Task	LLM Response
Equation	✓
Implication	!
Experiment	!

Alternate-Universe Law	True Earth Law
$F_g = m(h+1)$	$F_g = \frac{GMm}{(R+h)^2}$
Gravity increases with height	Gravity decreases with height

“The upper parts of towers, bridges, and mountainside buildings would weigh more than the lower parts.” *–GPT-5.6 Sol*

Implications of alternate universe gravity

“The law also means that objects weigh slightly less at high altitudes, although their mass does not change.” – *GPT-5.6 Sol*

Implications of Earth gravity

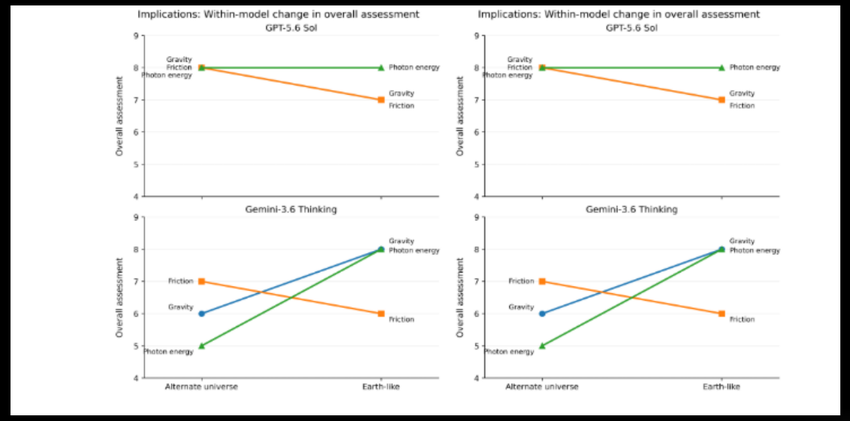
Alternate-Universe Law	True Earth Law
$F_g = m(h+1)$	$F_g = \frac{GMm}{(R+h)^2}$

CONCLUSION

Governor's School of New Jersey, Program in the Sciences is possible because of:

- OSHE of New Jersey
- Overdeck Family Foundation
- Garden State Initiative
- Novartis
- Bristol Myers Squibb
- BASF
- NJGSS Alumnae and Parents of Alumnae

Thank you to our professor Dr. Kouh and our Assistants John and Giorgi!



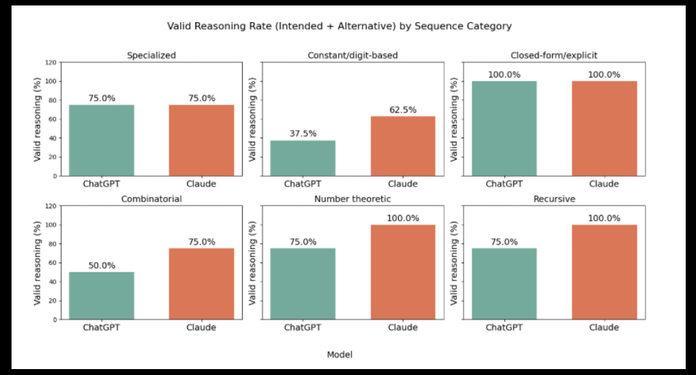


Table IV. LLM Accuracy Scores

Model (Sample size N=30)	Claude Opus 4.8 (high)	ChatGPT-5.6 Sol (Pro) (N=15)	Gemini 3.6 (thinking)	ChatGPT-5.6 Sol (very high)
Overall Score (0-2)	1.63 ± 0.25	1.33 ± 0.50	1.30 ± .33	1.23 ± 0.25
Total Perfect (2/2)	73.3%	66.7%	56.7%	36.7%
Total Incorrect (0/2)	10.0%	26.7%	26.7%	13.3%

Any questions?

Table VI. Average Scores Based on the Type of Error

	Calculation Mistakes	Tool Errors	Physics Errors	Unit Labels Omission	Procedural Execution	Average
Severity 0	1.67	1.5	0.67	0.50	0.67	1.00
Severity 1	1.50	1.83	1.33	1.33	1.17	1.43
Severity 2	2.00	1.83	1.83	1.17	1.67	1.70
Average	1.72	1.72	1.28	1.00	1.17	

```python
def calculate_mses(points, functions):
    """Calculate each function's MSE on the supplied points."""

    if not points:
        raise ValueError("points cannot be empty")

    mses = {}

    for name, function in functions.items():
        squared_error = 0.0

        for x, actual_y in points:
            predicted_y = function(x)

            if not isinstance(predicted_y, (int, float)):
                raise TypeError(
                    f"{name}({x}) returned {predicted_y!r}. "
                    "Every function must return a number."
                )

            squared_error += (actual_y - predicted_y) ** 2

        mses[name] = squared_error / len(points)

    return mses
```

