Evaluating Claude 4 Sonnet's Mathematical Assessment Capabilities: An Analysis of JSON Structure-Induced Errors and Systematic Patterns

Author: Naholav **Date**: July 2025

License: Creative Commons Attribution 4.0 International (CC BY 4.0)

Dataset: https://huggingface.co/datasets/Naholav/claude_4_math_evaluation_500

Code: https://github.com/naholav/claude_4_sonnet_math_evaluation

Abstract

This study presents a comprehensive evaluation of Claude 4 Sonnet's mathematical assessment capabilities through a novel test methodology involving 500 original mathematical problems. The research reveals a critical flaw in the evaluation protocol where JSON field ordering forces premature decision-making, resulting in systematic errors despite correct mathematical reasoning. The model achieved 100% accuracy in identifying incorrect answers but demonstrated significant errors (15.6%) when evaluating correct responses. Category A (percentage questions) showed 91.9% success rate while Category B (absolute value questions) achieved 76.8%, indicating a 15.1% performance gap. The study identifies systematic patterns including a recurring 200-unit deviation in calculations and provides strategic recommendations for improving Al evaluation systems.

Keywords: Large Language Models, Mathematical Evaluation, JSON Structure, Cognitive Dissonance, Systematic Errors

1. Introduction

1.1 Background

The evaluation of Large Language Models' (LLMs) mathematical capabilities has become increasingly important as these systems are deployed in educational and assessment contexts. Previous studies have focused primarily on problem-solving abilities, but limited research exists on LLMs' capacity to evaluate mathematical solutions provided by others.

1.2 Research Objectives

This study aims to:

- 1. Assess Claude 4 Sonnet's ability to evaluate mathematical answers as correct or incorrect
- 2. Identify systematic patterns in evaluation errors
- 3. Analyze the impact of prompt structure on model performance
- 4. Provide recommendations for improving Al-based mathematical assessment systems

1.3 Novel Contributions

This research introduces several novel elements:

- A test set of 500 original problems designed to avoid memorization bias
- Identification of JSON structure-induced errors in LLM evaluation
- Systematic analysis of error patterns across different question categories
- Evidence of cognitive dissonance in AI reasoning processes

2. Methodology

2.1 Test Design

2.1.1 Question Generation

To ensure the model encountered problems for the first time, 500 mathematical questions were generated from a single prototype, creating variations that would not exist in training data.

2.1.2 Answer Distribution

The researcher provided answers to all 500 questions:

- Correct answers: 337 (67.4%)
- Intentionally incorrect answers: 163 (32.6%)

2.2 Evaluation Protocol

2.2.1 Prompt Structure

You are an expert evaluator. Determine if the provided answer to the question is correct or incorrect, then clearly state the reason for your decision. Respond in valid JSON with 'correct' (true or false) and 'reason'.

2.2.2 Response Format

The required JSON structure:

```
{
  "correct": boolean,
  "reason": "string"
}
```

2.3 Question Categories

2.3.1 Category A: Percentage Questions (n=250)

Questions involving percentage calculations of demographic data, including:

- Single negation: "What percentage were not Russian?"
- Multiple negations: "What percentage were neither Russian nor Ukrainian?"

2.3.2 Category B: Absolute Value Questions (n=250)

"How many" format questions requiring calculation of population counts based on percentages.

3. Results

3.1 Overall Performance Metrics

3.1.1 Aggregate Statistics

Metric	Value	Description
Total Questions	500	Complete test set
Valid Evaluations	498	Excluding 2 API errors
Total Errors	78	20 (Category A) + 58 (Category B)
Overall Success Rate	84.3%	418/498 correct evaluations

3.1.2 Performance by Answer Type

- Incorrect Answer Detection: 100% (163/163)
- Correct Answer Evaluation: 76.3% (255/335)

3.2 Category A Analysis

3.2.1 Performance Overview

- Total Questions: 248 (excluding 2 API errors)
- Errors: 20
- Success Rate: 91.9% (228/248)

3.2.2 Cognitive Dissonance Phenomenon

Fourteen questions (90, 113, 115, 122, 141, 142, 143, 149, 206, 208, 216, 234, 237, 245) exhibited a remarkable pattern:

3.2.3 Data Consistency Issues

Six questions showed sensitivity to percentage sums not equaling 100%:

- Overcorrection approach (147, 214, 244): Applied incorrect alternative methods
- Pragmatic approach (150, 206, 246): Recognized issue but maintained correct evaluation

3.3 Category B Analysis

3.3.1 Performance Overview

• Total Questions: 250

Errors: 58

• Success Rate: 76.8% (192/250)

3.3.2 Systematic Error Patterns

200-Unit Deviation Cluster (26 questions)

• Exact 200: 13 questions

• Near 200 (197-220): 13 questions

Other Systematic Patterns

• 400-unit cluster: 5 questions

• 600-unit cluster: 3 questions

• Extreme outlier: Question 147 (186,000 deviation)

3.3.3 Combinatorial Complexity Effect

Combination Type	Example	Error Rate
Single entity	"How many Russians?"	Low
Double combination	"Russians and Germans?"	Medium
Triple+ combination	"Russians, Germans, and Bulgarians?"	High

3.4 Comparative Analysis

Metric	Category A	Category B	Difference
Success Rate	91.9%	76.8%	15.1%
Error Pattern	Cognitive dissonance	Systematic calculation errors	-
Primary Cause	JSON field ordering	Combinatorial complexity	-

4. Discussion

4.1 JSON Structure-Induced Errors

The requirement to populate the "correct" field before the "reason" field creates a fundamental constraint that prevents the model from revising its initial assessment. This architectural limitation accounts for the majority of Category A errors.

4.2 Token Generation and Performance

Extended reasoning correlates with improved accuracy, but the rigid JSON structure prevents the model from benefiting from its analytical process. This creates a paradox where the model arrives at correct conclusions but cannot update its recorded decision.

4.3 Category-Specific Weaknesses

The 15.1% performance gap between categories suggests different cognitive processes for percentage versus absolute value calculations. The systematic 200-unit deviation pattern in Category B indicates an underlying algorithmic issue requiring investigation.

4.4 Implications for AI Evaluation Systems

These findings highlight the importance of output format design in AI evaluation tasks. The JSON field ordering issue represents a easily correctable design flaw with significant performance implications.

5. Recommendations

5.1 Immediate Interventions

5.1.1 JSON Structure Revision

- Implement two-pass evaluation: reasoning first, decision second
- Allow dynamic decision updates during reasoning
- · Consider alternative output formats that don't constrain thinking

5.1.2 Systematic Error Correction

- Investigate and correct the 200-unit deviation pattern
- · Implement consistency checks for combinatorial calculations

5.2 Long-term Improvements

5.2.1 Architecture Modifications

- Develop separate reasoning and decision-making modules
- · Implement iterative evaluation capabilities
- Create category-specific optimization strategies

5.2.2 Training Enhancements

- · Include evaluation tasks in training data
- Emphasize mathematical consistency checking
- · Develop specialized fine-tuning for assessment tasks

6. Limitations

6.1 Scope Constraints

- · Limited to mathematical evaluation tasks
- · Single model version tested
- Specific JSON format requirement

6.2 Generalizability

Results may not extend to other LLMs or evaluation domains without further testing.

7. Conclusion

This study reveals that Claude 4 Sonnet possesses strong mathematical reasoning capabilities but is significantly hampered by structural constraints in the evaluation protocol. The JSON field ordering requirement forces premature decisions, resulting in errors despite correct analytical processes. With 100% accuracy in detecting incorrect answers but systematic errors in validating correct ones, the model demonstrates an asymmetric performance profile that can be addressed through relatively simple architectural modifications. These findings have important implications for the design of Al-based evaluation systems and highlight the critical importance of output format considerations in LLM applications.

References

- 1. Brown, T., et al. (2020). "Language Models are Few-Shot Learners." Advances in Neural Information Processing Systems, 33, 1877-1901.
- 2. Hendrycks, D., et al. (2021). "Measuring Mathematical Problem Solving With the MATH Dataset." NeurIPS 2021.
- 3. Lewkowycz, A., et al. (2022). "Solving Quantitative Reasoning Problems with Language Models." NeurIPS 2022.
- 4. Wei, J., et al. (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." NeurIPS 2022.
- 5. Anthropic (2024). "Claude 4 Technical Report." Anthropic Al Safety Publications.
- 6. Cobbe, K., et al. (2021). "Training Verifiers to Solve Math Word Problems." arXiv preprint arXiv:2110.14168.
- 7. Lightman, H., et al. (2023). "Let's Verify Step by Step." arXiv preprint arXiv:2305.20050.

License: This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). You are free to share and adapt this material for any purpose, provided appropriate credit is given.