# CS6795 – Term Project – Project Pitch

Yash Sarda, *Student, Georgia Institute of Technology*, Atlanta GA, USA, ysarda@gatech.edu

## I. Topic

The broad topic I have chosen for this project is a mix of Artificial Intelligence and Linguistics, specifically investigating symbolic reasoning and mathematical understanding of large language models (LLMs).

## II. Research Question

Through this project I hope to answer the following specific research question:

"When AI models encounter novel mathematical notation with minimal examples, do they rely primarily on symbol translation strategies or demonstrate genuine mathematical reasoning capabilities, and what does this reveal about the depth of their mathematical understanding?"

## III. Interest and Importance

The ability of LLMs to mimic human-level intelligence and prose was widely seen as just that – mimicry. With the emergence of reasoning models over the last two years, however, LLMs have been able to compete at the highest levels of mathematical competition, and replicate some of mankind's most difficult proofs with ease, even expected to soon surpass us and start solving unsolved problems. Because of this, I am supremely interested in testing whether LLMs' mathematical capabilities represent genuine understanding or sophisticated pattern matching by examining how they perform when mathematical concepts are presented in completely novel symbolic notations with minimal translation support.

This question is important because it addresses several topical scientific issues:

### A. Translation vs. Reasoning Distinction

When models encounter novel mathematical notation, do they rely primarily on symbol-to-symbol translation strategies or demonstrate genuine mathematical reasoning? This distinction is crucial for understanding the depth of AI mathematical cognition.

### B. Symbol Grounding

How do artificially intelligent systems connect abstract symbols to mathematical concepts? Do they develop notation-independent understanding of mathematical relationships [6], or are they bound to the specific symbols they were trained on?

### C. Reliability

How reliable are LLM-based mathematical systems in novel scenarios if they primarily learn surface-level pattern matching rather than deep mathematical understanding? This has critical implications for deploying AI in mathematical and scientific contexts.

### D. Mathematical Principle Transfer

Can models apply fundamental mathematical principles (like commutativity, distributivity, or algebraic laws) to novel symbols without explicit instruction? This tests whether they understand mathematical structure independently of familiar notation.

### E. Systematic Generalization in Mathematics

Which mathematical reasoning types (arithmetic operations, algebraic manipulation, logical inference) are most robust to notation changes? Can models combine learned operations in novel ways not explicitly demonstrated in examples? Language generalizability for models has been shown [5], but does this apply to mathematics as a whole?

## IV. Research Design and Outcomes

### A. Inputs

- Minimal notation examples inserted into the prompt, controlled for number of examples and simplicity of the problem. An example is as follows:

  *"You will be given mathematical problems using a novel notation system. Here are the basic symbol mappings:*
  ⨋ *is* 2, ∴ *is* 3, ⊕ *is* +, ≙ *is* =
  *Using only these examples, solve the following problems"*

- Complex mathematical problems requiring multi-step reasoning and novel operation combinations (normal and randomized notation) as seen in (1) and (2).

$$⨋ \oplus \therefore \triangleq \tag{1}$$

$$\text{If } ⨋ \oplus \therefore \triangleq \complement, \text{ then } \therefore \oplus ⨋ \triangleq \tag{2}$$

### B. Outputs

- Overall accuracy scores per model, for normal and randomized notation
- Classification of reasoning vs translation of model strategies
- Mathematical principle transfer scores (commutativity, distributivity, etc.)

- Performance degradation patterns across different reasoning types
- Comparative analysis of translation-dependent vs. reasoning-dependent problem solving

### C. Cognitive Science Concepts

- Symbol Grounding Theory: How abstract symbols acquire meaning through experience and context [2].
- Transfer Learning in Cognition: Research on how mathematical knowledge transfers across different representational formats [3].
- Analogical Reasoning: Mapping theory applied to mathematical concept transfers [4].

These concepts will be drawn from the lecture materials and literature review, like the papers already referenced and others on mathematical cognition and systematic generalization.

### D. Computational Model/Tool

The outcome of this project will be a multi-part tool used to evaluate LLMs on symbolic reasoning. The first part will be a controlled problem generator, that can provide either normal or randomized notation mathematical problems. These problems can focus on certain domains of mathematics to test for fundamental understanding of certain laws, or simply be translations of a math dataset. It will also control for the amount and type of notional examples provided to the model, testing if the models can generalize beyond explicitly provided documentation. This part of the tool will be constructed in Python with the assistance of the Google Deepmind Mathematics Dataset [7], and will be available publicly on GitHub.

The second part will be an evaluation framework of LLMs on the generated mathematical problems, including metrics like overall accuracy, impact of randomization (how much worse the model did after switching notation), specific accuracy on how well the model was able to apply mathematical laws, and whether its approach was translational or reasoning-based. This will be accomplished mostly via python, with an external LLM review of answers for the approach classification. This tool will also be available on GitHub.

The third and last part will be an accompanying dashboard of model scores from the above metrics, evaluated for the current state-of-the-art LLM models, hosted on a website. The code behind the dashboard will also be available on GitHub.

## V. Work Plan

The project is estimated to take a sum total of 102 hours, distributed over 12 weeks, as seen in Table I.

TABLE I

| Week | Task | Estimated Time (Hours) |
|---|---|---|
| 3 | Literature Review Foundation | 3 |
| 3 | Project Setup | 2 |
| 4 | Advanced Literature Review | 5 |
| 4 | Technical Exploration | 3 |
| 5 | Project Pitch Development | 4 |
| 5 | Symbol System Design | 3 |
| 6 | Problem Generator Development | 6 |
| 6 | Dataset Integration | 3 |
| 7 | Symbol System Implementation | 5 |
| 7 | Problem Generator Enhancement | 4 |
| 8 | LLM Evaluation Framework | 6 |
| 8 | Infrastructure Development | 3 |
| 9 | Evaluation System Completion | 5 |
| 9 | Midpoint Check-in | 3 |
| 10 | Comprehensive Model Testing | 6 |
| 10 | Data Collection | 2 |
| 11 | Advanced Analysis Implementation | 5 |
| 11 | Dashboard Development | 4 |
| 12 | Results Analysis | 5 |
| 12 | Dashboard Completion | 3 |
| 13 | Report Writing - Methods and Results | 5 |
| 13 | Code Documentation | 2 |
| 14 | Final Report Completion | 6 |
| 14 | Final Report Submission | 4 |
| 15 | Presentation Preparation | 3 |
| 15 | Final Deliverables | 2 |

## References

Reference [1] was used to generate the following citations, to assist with finding current literature on this topic, and to assist with finding literature on symbolic reasoning.


[1] Anthropic, "Claude AI assistant," 2024. [Online]. Available: https://www.anthropic.com/claude

[2] S. Harnad, "The symbol grounding problem," Mind & Machine, vol. 2, no. 3, pp. 335-378, 1990.

[3] M. L. Gick and K. J. Holyoak, "The cognitive basis of knowledge transfer," in Transfer of Learning: Contemporary Research and Applications, Academic Press, 1987, pp. 9-46.

[4] D. Gentner, "Structure-mapping: A theoretical framework for analogy," Cognitive Science, vol. 7, no. 2, pp. 155-170, 1983.

[5] B. M. Lake and M. Baroni, "Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks," in Proc. Int. Conf. Machine Learning (ICML), 2018, pp. 2879-2888.

[6] A. Lewkowycz et al., "Solving quantitative reasoning problems with language models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022, pp. 3843-3857.

[7] D. Saxton, E. Grefenstette, F. Hill, and P. Kohli, "Analysing mathematical reasoning abilities of neural models," arXiv preprint arXiv:1904.01557, 2019.