

CSCE 689: Special Topics in Trustworthy NLP

Lecture 16: Long-Context Language Models

Kuan-Hao Huang
khhuang@tamu.edu



Course Project – Midterm Report

- Due: 11/6
- Page limit: 4 pages (excluding reference)
- Format: [ACL style](#)
- The report should include
 - Introduction to the topic you choose and problem definition
 - Related literature and overview of existing progress and challenges
 - Proposed solutions, novelty, and expected contributions
 - Evaluation metrics
 - Planned implementation details, including dataset, models, codebases, etc.
 - Current progress (preliminary results, baseline results, etc.)
 - Next steps and expected timeline

Long-Context Language Models

- Language models that can handle very long inputs
- Why we care about this?

Chain-of-Thought (CoT) Prompting

- Ask the model to explain its reasoning before making an answer

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

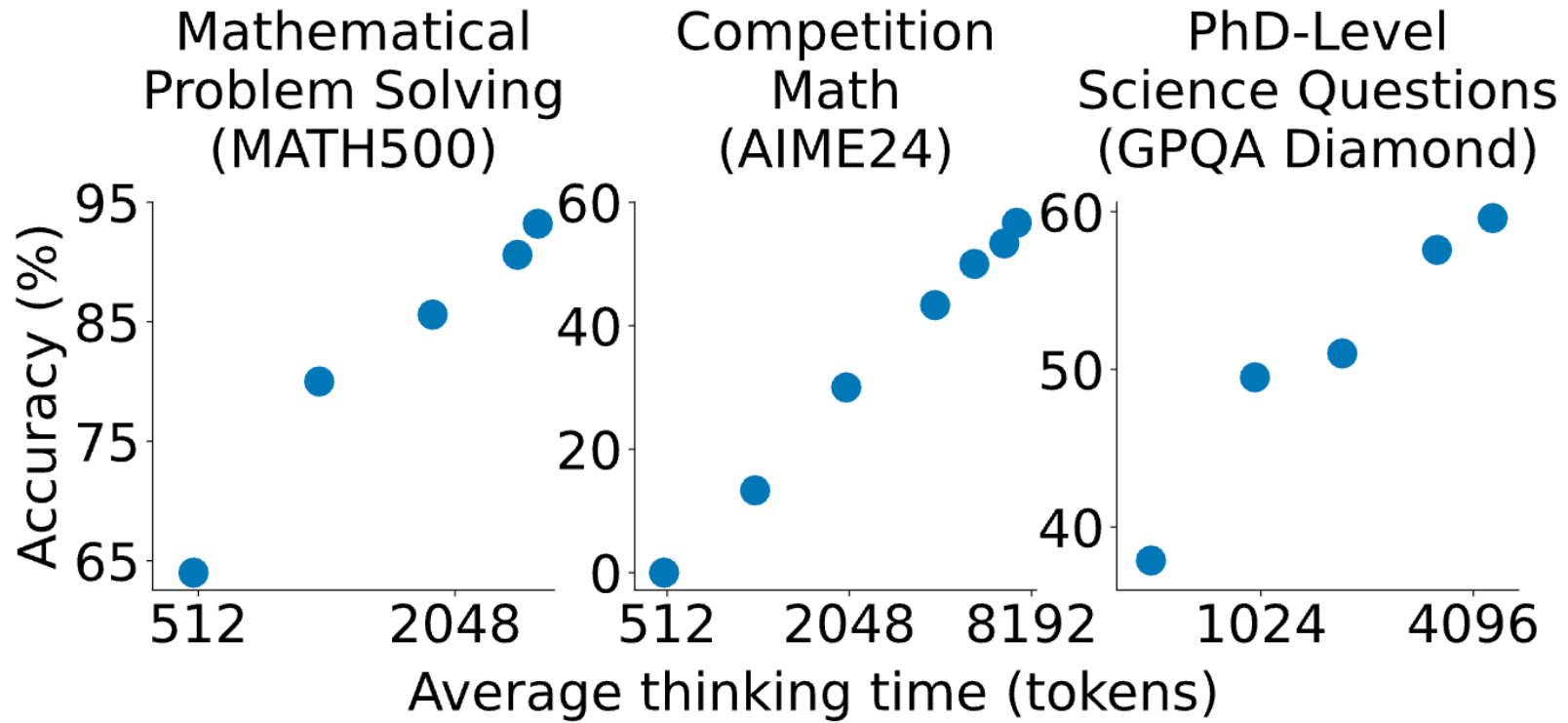
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

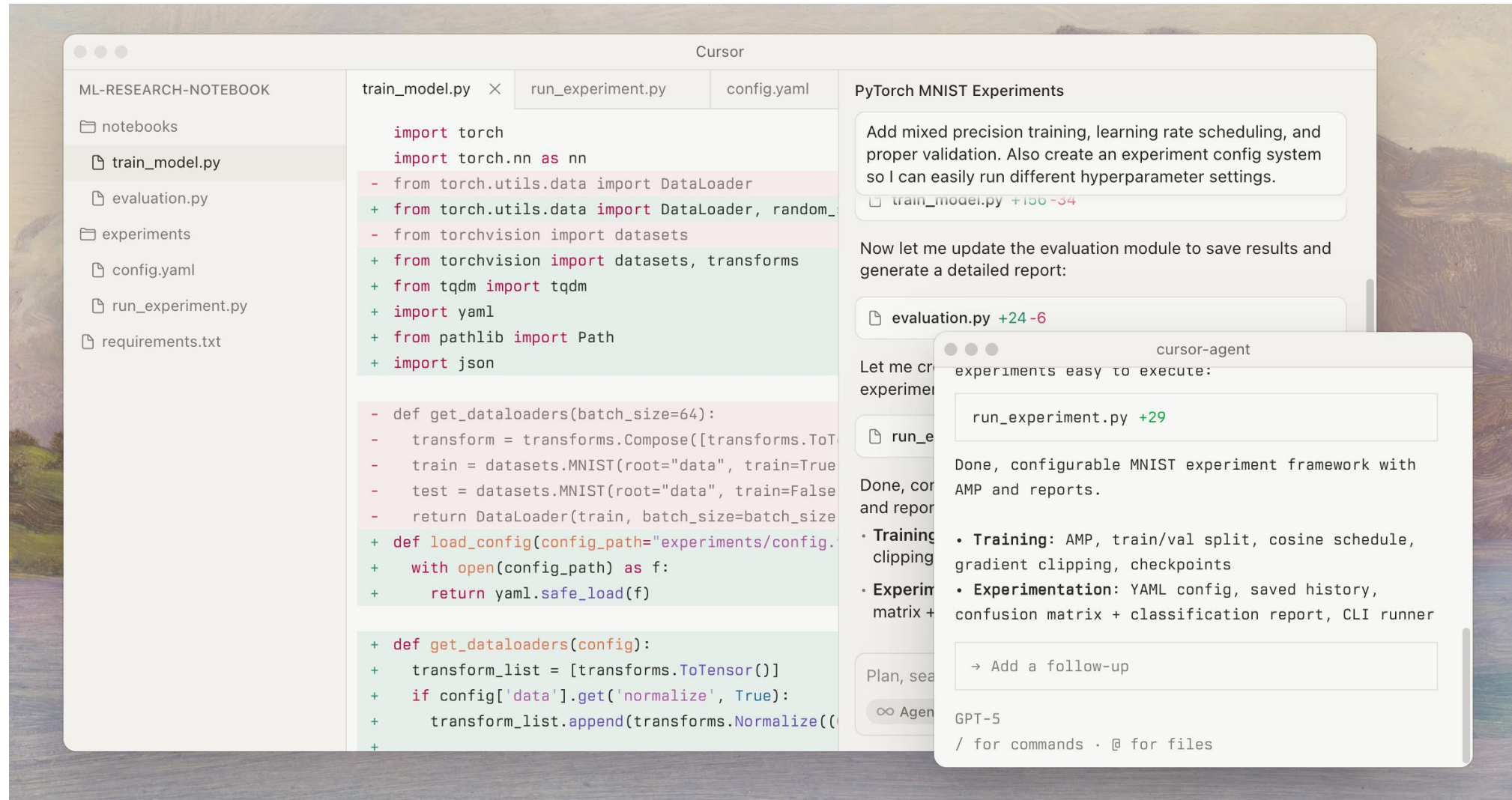
Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Test-Time Scaling



LLMs for Coding



LLM Agent History

User starts interaction

Interaction Turn #1

User

Task: Together Lily, David, and Bodhi collected 43 insects. Lily found 7 more than David. David found half of what Bodhi found. How many insects did Lily find? Solution output format: an integer.

LLM

Thought: To solve this problem, let's start by setting up equations [...] assume that David found x insects. Based on the given information, Lily found 7 more than David, so Lily found x + 7 insects [...]

Execute:

```
from sympy import symbols, Eq, solve
x, b = symbols('x b')
# set up the equation
equation = Eq(x + (x + 7) + (1/2) * b, 43)
# solve the equation
solution = solve(equation, x)
solution[0]
```

Interaction Turn #2

Python Interpreter

Observation: Out[1]: 18.0 - 0.25*b
You have 4 steps left and 2 chances to propose solution left.

User (optional, LLM-simulated)

This is NOT GOOD. [...] You should have used the information that David found half of what Bodhi found [...]. Then, you can express the total insects in terms of Bodhi's insects and solve for it [...]

LLM

Thought: [Apologize ...] Bodhi found b insects. David found half of what Bodhi found, so David found (1/2) * b insects [...]

Execute:

```
from sympy import symbols, Eq, solve
b = symbols('b')
# set up the equation
equation = Eq((1/2) * b + 7 + (1/2) * b + b, 43)
# solve the equation
solution = solve(equation, b)
solution[0]
```

Interaction Turn #3

Python Interpreter

Observation: Out[2]: 18.000000000000000
You have 3 steps left and 2 chances to propose solution left.

Interaction Turn #4

Python Interpreter

Observation: Out[3]: 25.000000000000000
You have 2 steps left and 2 chances to propose solution left.

User (optional, LLM-simulated)

This is NOT GOOD. You made a mistake in calculating the number of insects Lily found. You should have calculated the number of insects David found first, which is half of what Bodhi found [...]

LLM

Thought: [Apologize ...] Bodhi found 18 insects [...] David found half of what Bodhi found, so David found (1/2) * 18 insects. Lily found 7 more than David, so Lily found (1/2) * 18 + 7 insects.

Execute:

```
david_insects = (1/2) * solution[0]
lily_insects = david_insects + 7
lily_insects
```

Interaction Turn #5

Python Interpreter

Observation: Out[4]: 16.000000000000000
You have 1 steps left and 2 chances to propose solution left. You should take the last step to propose a solution.

User (optional, LLM-simulated)

This is GOOD. You have correctly calculated the number of insects Lily found. Now you can propose the solution.

LLM

Propose Solution: The number of insects Lily found is 16.

User

Good job! You have successfully solved the task! [stops interaction]

User (optional, LLM-simulated)

This is GOOD. [...] Now you can use this information to find out how many insects Lily found.

LLM

Thought: Now that we know that Bodhi found 18 insects [...] Lily found 7 more than David, so Lily found 18 + 7 insects.

Execute:

```
lily_insects = solution[0] + 7
lily_insects
```

LLMs Context Length

- Llama 2: ~4K tokens
- GPT-4: ~8K tokens
- Llama 3.1: 128K tokens
- GPT-4o: ~128K tokens
- Qwen 2.5: ~128K tokens
- Claude 3.5 Sonnet: ~200K tokens
- GPT-4.1: ~1M tokens
- Gemini 1.5 Pro: ~1M tokens

LLMs Context Length

- Can we extend the context length during the **inference time**?
 - Positional encoding?
 - Understanding ability?