Yandex Research

Parallel LLM Inference

George Yakushev

Motivation

- Harder problems may require long chains of reasoning

Question: Find all triples (x, y, z) of positive integers such that $x \leq y \leq z$ and $x^3(y^3 + z^3) = 2012(xyz + 2)$.

Solution: First note that x divides $2012 \cdot 2 = 2^3 \cdot 503$. If $503 \mid x$ then the right-hand side of the equation is divisible by 503^3, and it follows that $503^2 \mid xyz + 2$. This is false as $503 \mid x$. Hence $x = 2^m$ with $m \in \{0,1,2,3\}$. If $m \geq 2$ then $2^6 \mid 2012(xyz + 2)$. However the highest powers of 2 dividing 2012 and $xyz + 2 = 2^m yz + 2$ are 2^2 and 2^1 respectively. So $x = 1$ or $x = 2$, yielding the two equations

$$y^3 + z^3 = 2012(yz + 2),$$
$$y^3 + z^3 = 503(yz + 1)$$

In both cases …… It follows that $y \equiv -z (mod\ 503)$ as claimed. Therefore $y + z = 503k$ with $k \geq 1$. In view of $y^3 + z^3 = (y + z)((y - z)^2 + yz)$ the two equations take the form

$$k(y - z)^2 + (k - 4)yz = 8 \quad (1)$$
$$k(y - z)^2 + (k - 1)yz = 1 \quad (2)$$

In (1) we have $(k - 4)yz \leq 8$, which implies $k \leq 4$ ……
Therefore (1) has no integer solutions.
Equation (2) implies $0 \leq (k - 1)yz \leq 1$, so that $k = 1$ or $k = 2$.
Also $0 \leq k(y - z)^2 \leq 1$, hence $k = 2$ only if $y = z$. However then $y = z = 1$, which is false in view of $y + z \geq 503$.
Therefore $k = 1$ and (2) takes the form $(y - z)^2 = 1$, yielding $z - y = |y - z| = 1$. Combined with $k = 1$ and $y + z = 503k$, this leads to $y = 251, z = 252$.
In summary the triple $(2,251,252)$ is the only solution.

Final answer: $(2,251,252)$ **Subfield:** Number theory

Answer type: Tuple **Question type:** Open-ended

Motivation

- Harder problems may require long chains of reasoning
- Determining the best parallel solving strategies in advance is challenging

Question: Find all triples (x, y, z) of positive integers such that $x \leq y \leq z$ and $x^3(y^3 + z^3) = 2012(xyz + 2)$.

Solution: First note that x divides $2012 \cdot 2 = 2^3 \cdot 503$. If $503 \mid x$ then the right-hand side of the equation is divisible by 503^3, and it follows that $503^2 \mid xyz + 2$. This is false as $503 \mid x$. Hence $x = 2^m$ with $m \in \{0,1,2,3\}$. If $m \geq 2$ then $2^6 \mid 2012(xyz + 2)$. However the highest powers of 2 dividing 2012 and $xyz + 2 = 2^m yz + 2$ are 2^2 and 2^1 respectively. So $x = 1$ or $x = 2$, yielding the two equations

$$y^3 + z^3 = 2012(yz + 2),$$
$$y^3 + z^3 = 503(yz + 1)$$

In both cases …… It follows that $y \equiv -z (mod\ 503)$ as claimed. Therefore $y + z = 503k$ with $k \geq 1$. In view of $y^3 + z^3 = (y + z)((y - z)^2 + yz)$ the two equations take the form

$$k(y - z)^2 + (k - 4)yz = 8 \quad (1)$$
$$k(y - z)^2 + (k - 1)yz = 1 \quad (2)$$

In (1) we have $(k - 4)yz \leq 8$, which implies $k \leq 4$ ……
Therefore (1) has no integer solutions.
Equation (2) implies $0 \leq (k - 1)yz \leq 1$, so that $k = 1$ or $k = 2$.
Also $0 \leq k(y - z)^2 \leq 1$, hence $k = 2$ only if $y = z$. However then $y = z = 1$, which is false in view of $y + z \geq 503$.
Therefore $k = 1$ and (2) takes the form $(y - z)^2 = 1$, yielding $z - y = |y - z| = 1$. Combined with $k = 1$ and $y + z = 503k$, this leads to $y = 251, z = 252$.
In summary the triple $(2,251,252)$ is the only solution.

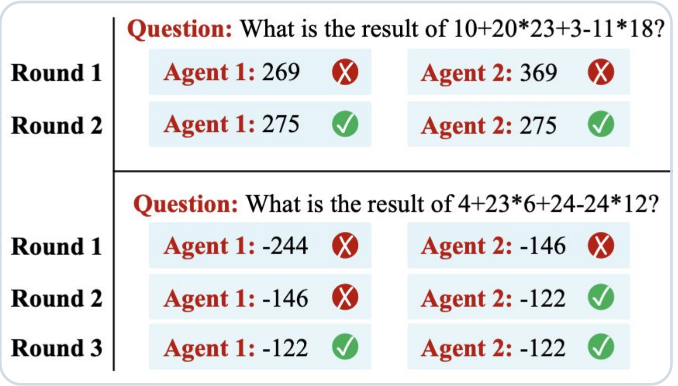
Final answer: $(2,251,252)$ **Subfield:** Number theory
Answer type: Tuple **Question type:** Open-ended

Existing Methods for Parallel LLM Generation

Multiagent Debate

Multiple LLM instances reason independently, then vote on the final solution

- can be offset with better single-agent prompting
- no acceleration: 1+ agent has to solve the entire problem sequentially



Improving Factuality and Reasoning in Language Models through Multiagent Debate, Du et al.

Existing Methods for Parallel LLM Generation

Multiagent Debate

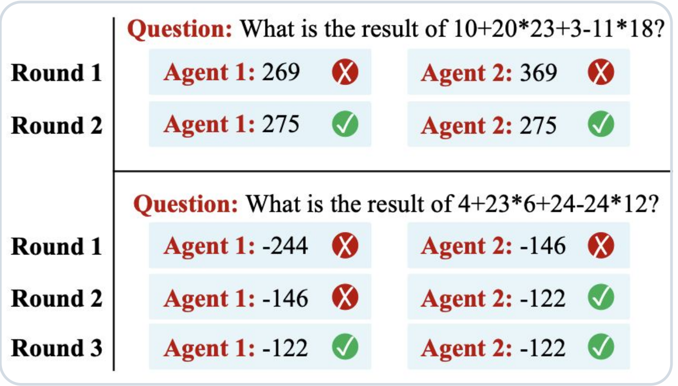
Multiple LLM instances reason independently, then vote on the final solution

- can be offset with better single-agent prompting
- no acceleration: 1+ agent has to solve the entire problem sequentially

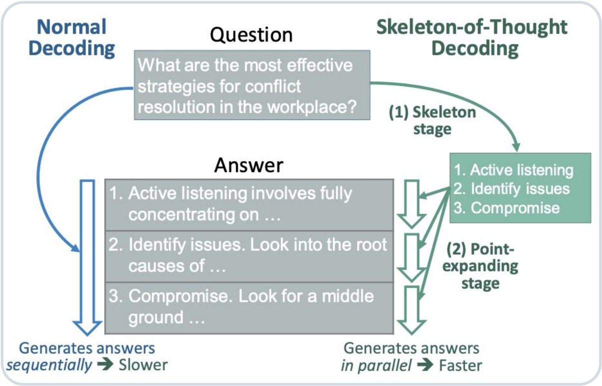
Skeleton-of-Thought

LLM creates a plan with parallel sub-tasks, then launches parallel LLM instances

- can accelerate inference
- harm reasoning for problems that do not fit their framework



Improving Factuality and Reasoning in Language Models through Multiagent Debate, Du et al.



Skeleton-of-Thought: Prompting LLMs for Efficient Parallel Generation, Ning et al.

Existing Methods for Parallel LLM Generation

Multiagent Debate

Multiple LLM instances reason independently, then vote on the final solution

- can be offset with better single-agent prompting
- no acceleration: 1+ agent has to solve the entire problem sequentially

Skeleton-of-Thought

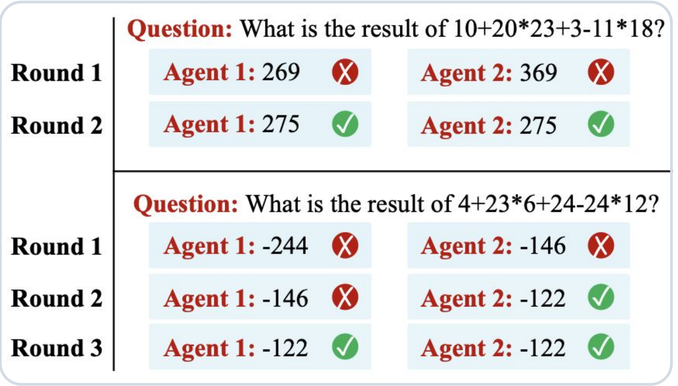
LLM creates a plan with parallel sub-tasks, then launches parallel LLM instances

- can accelerate inference
- harm reasoning for problems that do not fit their framework

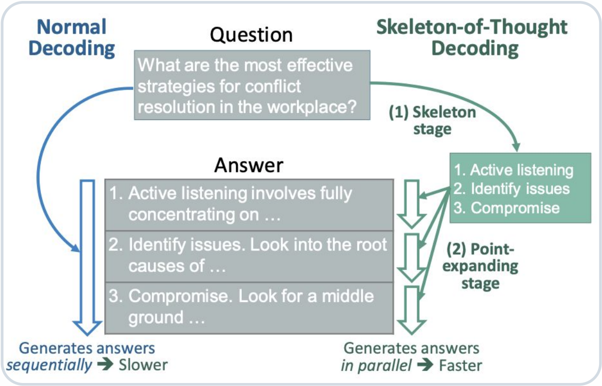
PASTA

LLM creates a plan with parallel sub-tasks in specific language, then launches parallel LLM instances, collects all results and gives the final answer

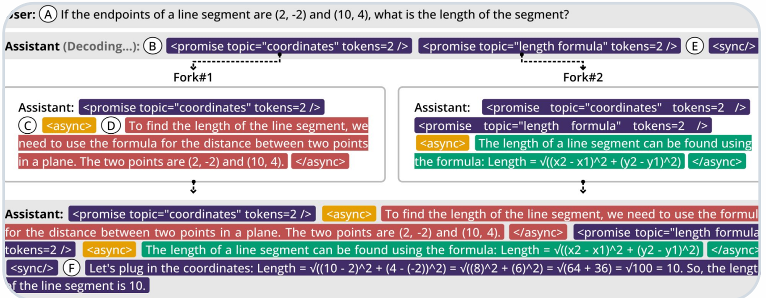
- can accelerate inference
- long subtasks force other instances to idle, wasting resources



Improving Factuality and Reasoning in Language Models through Multiagent Debate, Du et al.



Skeleton-of-Thought: Prompting LLMs for Efficient Parallel Generation, Ning et al.



Learning to Keep a Promise: Scaling Language Model Decoding Parallelism with Learned Asynchronous Decoding, Jin et al.

Motivation

- Harder problems may require long chains of reasoning
- Determining the best parallel solving strategies in advance is challenging
- Humans may interact dynamically:
  - re-planning on the fly
  - abandoning tasks half-way
  - switching to another approach
  - debating strategy

Question: Find all triples (x, y, z) of positive integers such that $x \leq y \leq z$ and $x^3(y^3 + z^3) = 2012(xyz + 2)$.

Solution: First note that x divides $2012 \cdot 2 = 2^3 \cdot 503$. If $503 \mid x$ then the right-hand side of the equation is divisible by 503^3, and it follows that $503^2 \mid xyz + 2$. This is false as $503 \mid x$. Hence $x = 2^m$ with $m \in \{0,1,2,3\}$. If $m \geq 2$ then $2^6 \mid 2012(xyz + 2)$. However the highest powers of 2 dividing 2012 and $xyz + 2 = 2^m yz + 2$ are 2^2 and 2^1 respectively. So $x = 1$ or $x = 2$, yielding the two equations

$$y^3 + z^3 = 2012(yz + 2),$$
$$y^3 + z^3 = 503(yz + 1)$$

In both cases ...... It follows that $y \equiv -z (mod\ 503)$ as claimed. Therefore $y + z = 503k$ with $k \geq 1$. In view of $y^3 + z^3 = (y + z)((y - z)^2 + yz)$ the two equations take the form

$$k(y - z)^2 + (k - 4)yz = 8 \quad (1)$$
$$k(y - z)^2 + (k - 1)yz = 1 \quad (2)$$

In (1) we have $(k - 4)yz \leq 8$, which implies $k \leq 4$ ......
Therefore (1) has no integer solutions.
Equation (2) implies $0 \leq (k - 1)yz \leq 1$, so that $k = 1$ or $k = 2$.
Also $0 \leq k(y - z)^2 \leq 1$, hence $k = 2$ only if $y = z$. However then $y = z = 1$, which is false in view of $y + z \geq 503$.
Therefore $k = 1$ and (2) takes the form $(y - z)^2 = 1$, yielding $z - y = |y - z| = 1$. Combined with $k = 1$ and $y + z = 503k$, this leads to $y = 251, z = 252$.
In summary the triple $(2,251,252)$ is the only solution.

Final answer: $(2,251,252)$ **Subfield:** Number theory

Answer type: Tuple **Question type:** Open-ended

Hogwild! Inference: Parallel LLM generation via Concurrent Attention

Gleb Rodionov
Yandex

Roman Garipov
HSE University, Yandex
ITMO University

Alina Shutova
HSE University, Yandex

George Yakushev
HSE University, Yandex

Erik Schultheis
IST Austria

Vage Egizarian
IST Austria

Anton Sinitsin
Yandex

Denis Kuznedelev
Yandex

Dan Alistarh
IST Austria

Hogwild! Inference Scheme

Common Cache

Task: compute $x^2 + x^4$ for x in {1, 2, 3}.

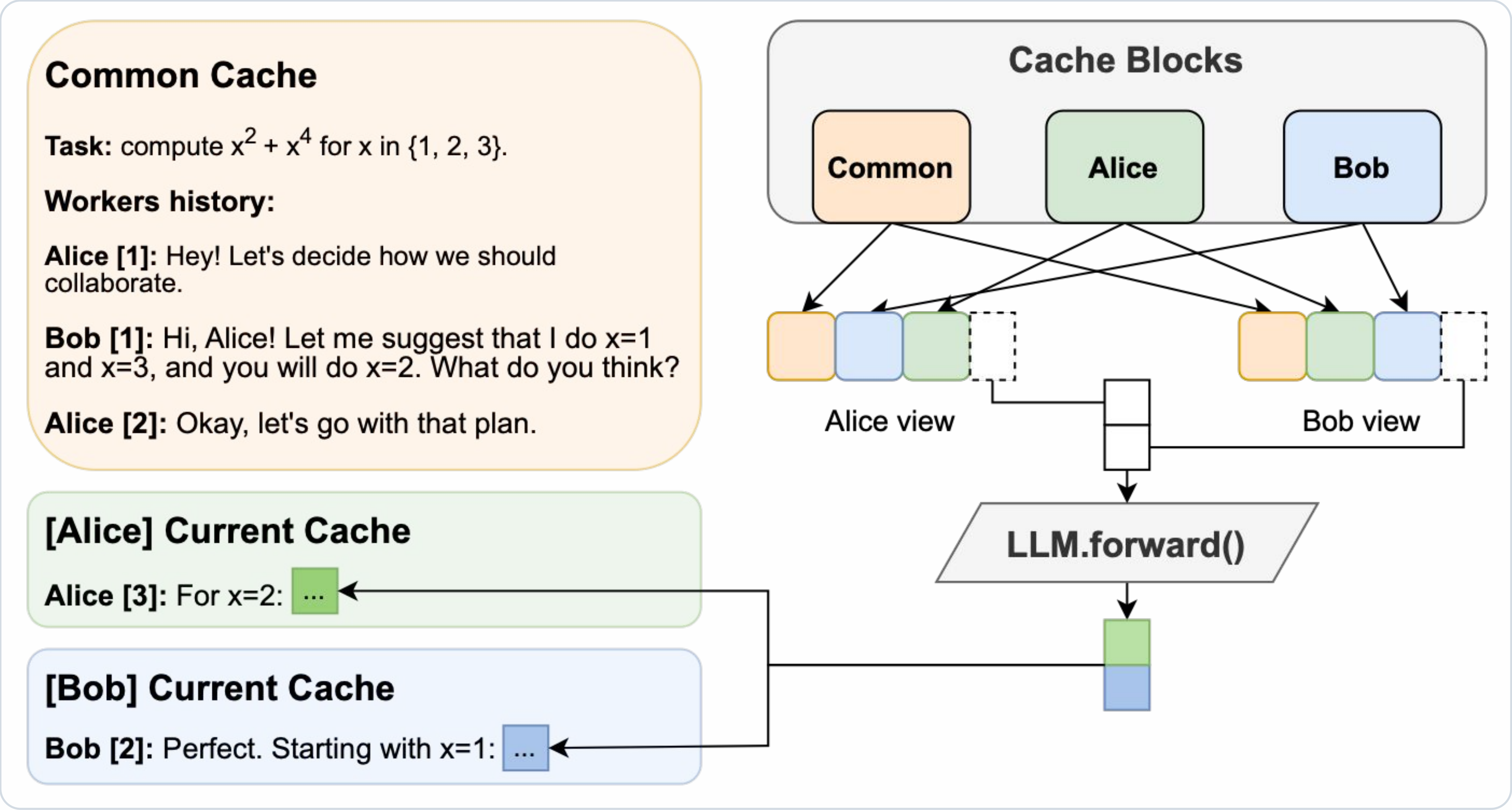
Workers history:

Alice [1]: Hey! Let's decide how we should collaborate.

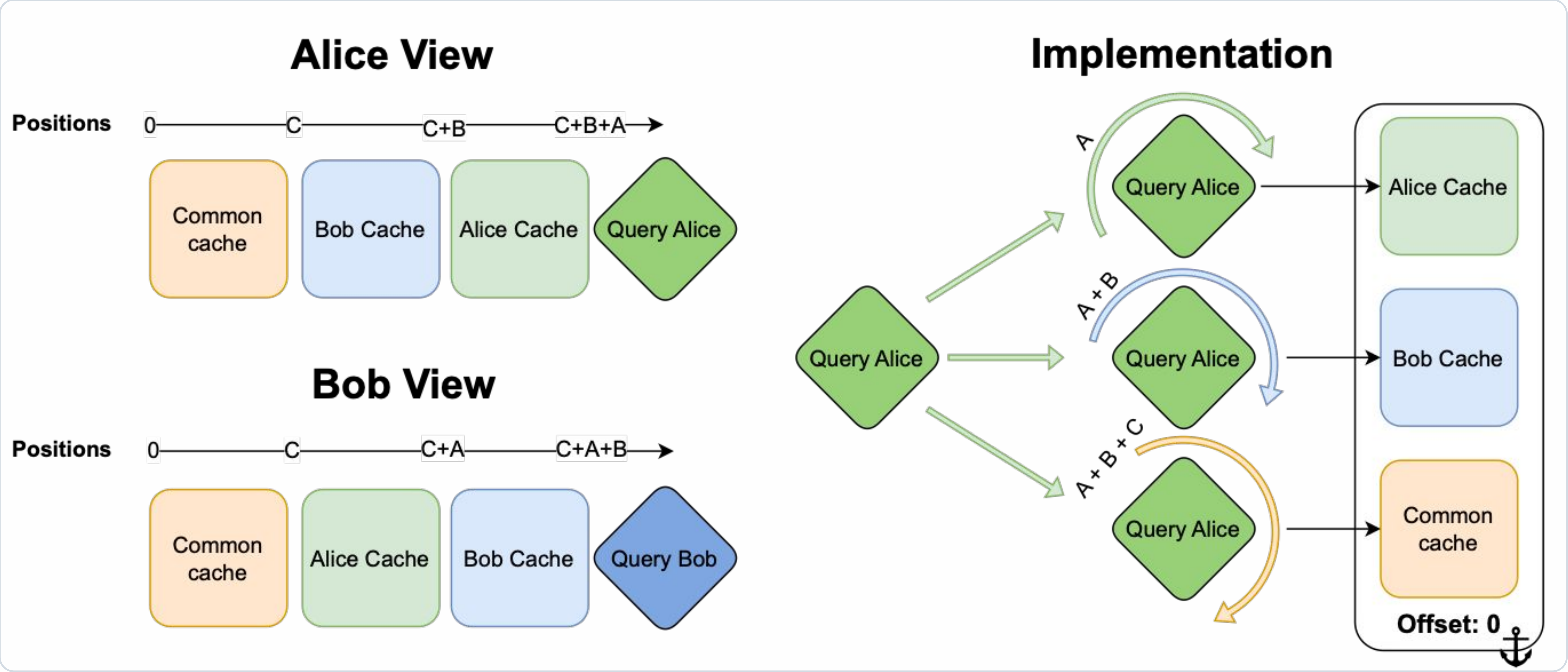
Bob [1]: Hi, Alice! Let me suggest that I do x=1 and x=3, and you will do x=2. What do you think?

Alice [2]: Okay, let's go with that plan.

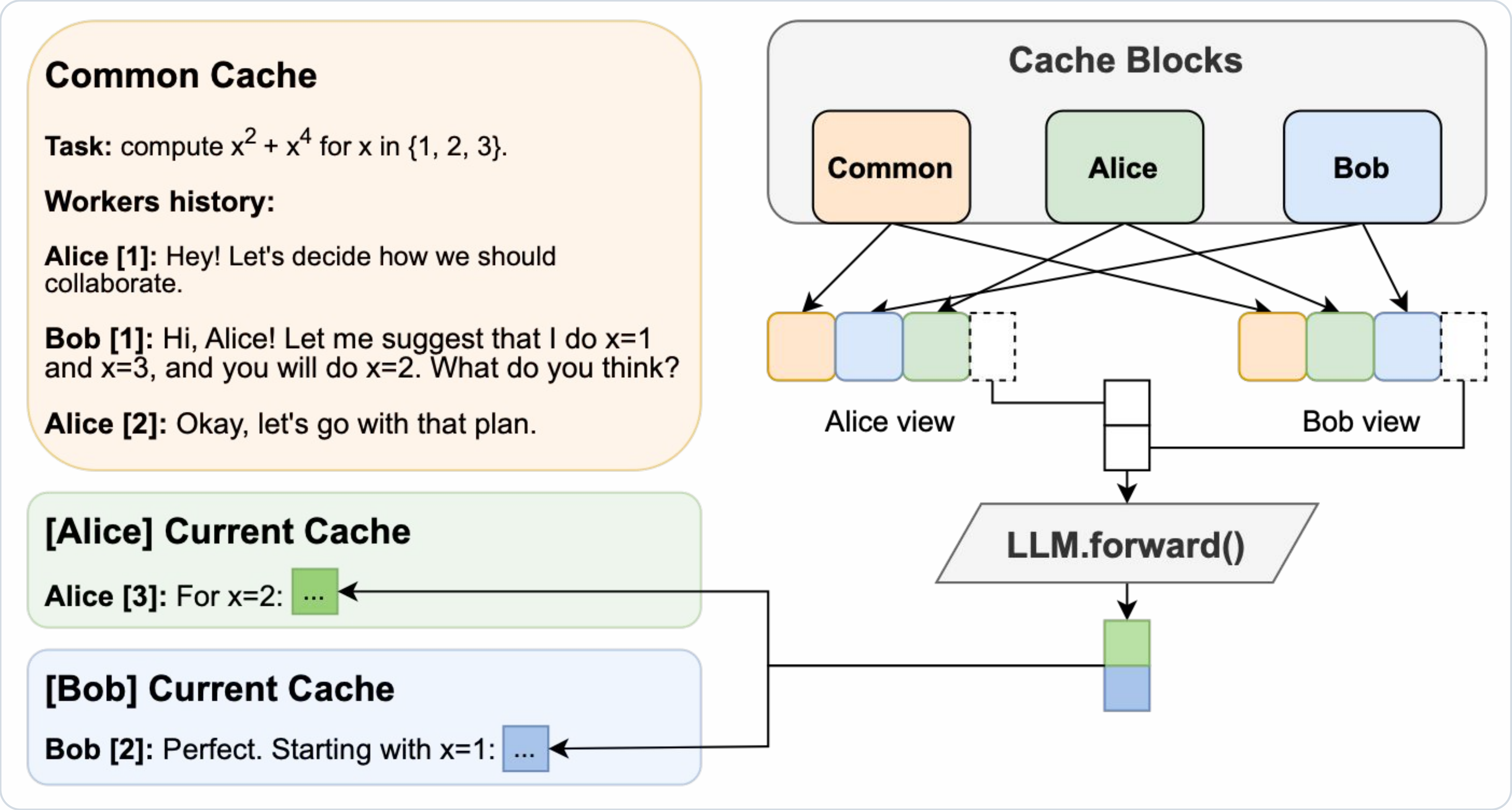
Hogwild! Inference Scheme



Hogwild! Inference Scheme



Hogwild! Inference Scheme



Hogwild! Inference Scheme

- Alice: Quick check: am I doing a redundant work? (yes/no): no. Bob, I see you’re working on the same product. Let me think of another approach. Let’s consider writing out the first few terms to see if there’s a pattern.

Hogwild! Inference Scheme

- Alice: Quick check: am I doing a redundant work? (yes/no): no. Bob, I see you’re working on the same product. Let me think of another approach. Let’s consider writing out the first few terms to see if there’s a pattern.

- Bob: Quick check: am I doing a redundant work? (yes/no): Yes, Alice is already calculating the x-coordinate of A. I’ll do something else. Let me find the coordinates of L instead.

Hogwild! Inference Scheme

- Alice: Quick check: am I doing a redundant work? (yes/no): no. Bob, I see you’re working on the same product. Let me think of another approach. Let’s consider writing out the first few terms to see if there’s a pattern.

- Bob: Quick check: am I doing a redundant work? (yes/no): Yes, Alice is already calculating the x-coordinate of A. I’ll do something else. Let me find the coordinates of L instead.

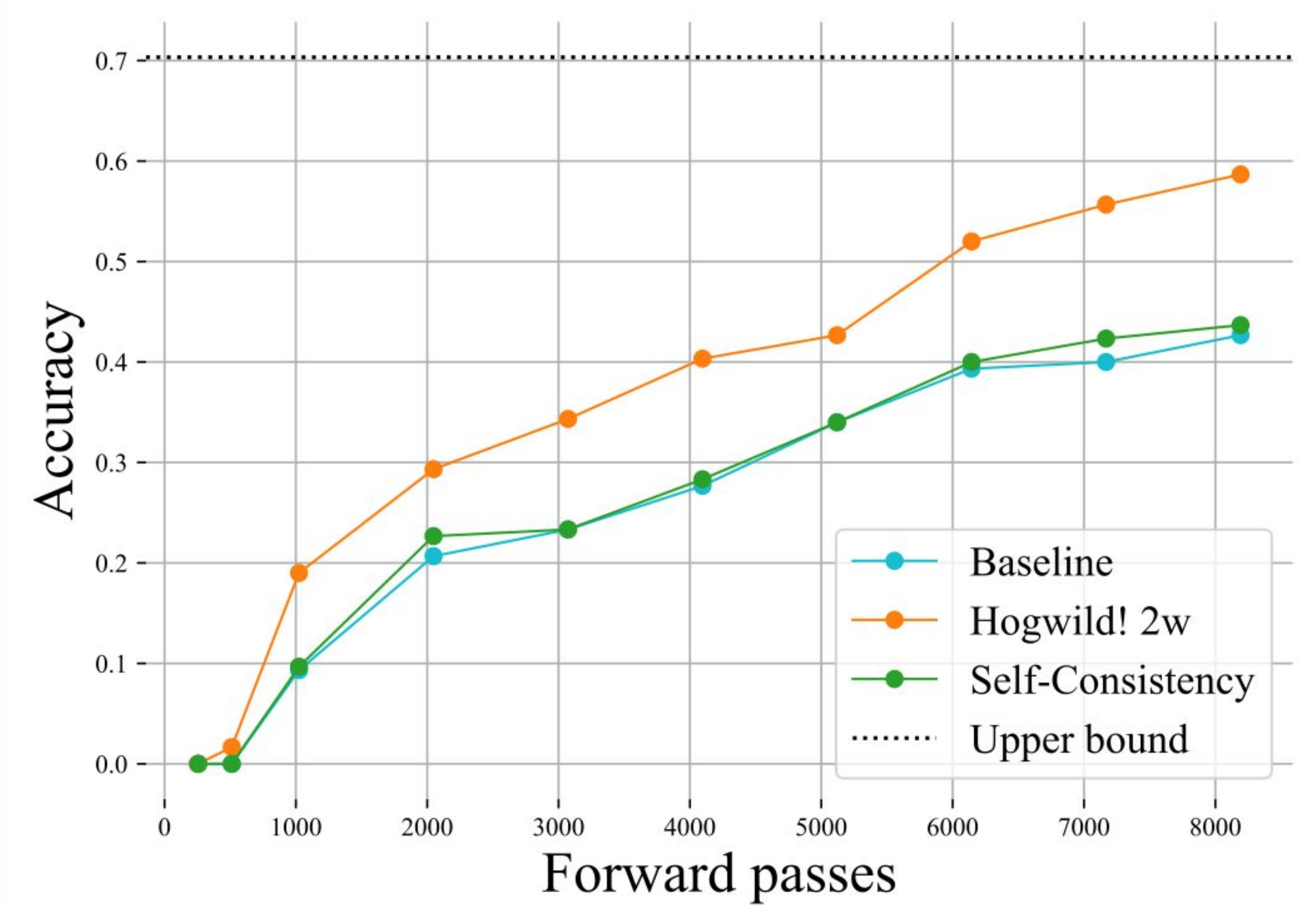
- Bob: Quick check: am I doing a redundant work? (yes/no): no. I’m finding coordinates of D, and Alice is finding coordinates of G.

Hogwild! Inference Scheme

- Alice: Quick check: am I doing a redundant work? (yes/no): no. Bob, I see you’re working on the same product. Let me think of another approach. Let’s consider writing out the first few terms to see if there’s a pattern.
- Bob: Quick check: am I doing a redundant work? (yes/no): Yes, Alice is already calculating the x-coordinate of A. I’ll do something else. Let me find the coordinates of L instead.
- Bob: Quick check: am I doing a redundant work? (yes/no): no. I’m finding coordinates of D, and Alice is finding coordinates of G.
- Alice: Quick check: am I doing a redundant work? (yes/no): yes, but that’s okay.

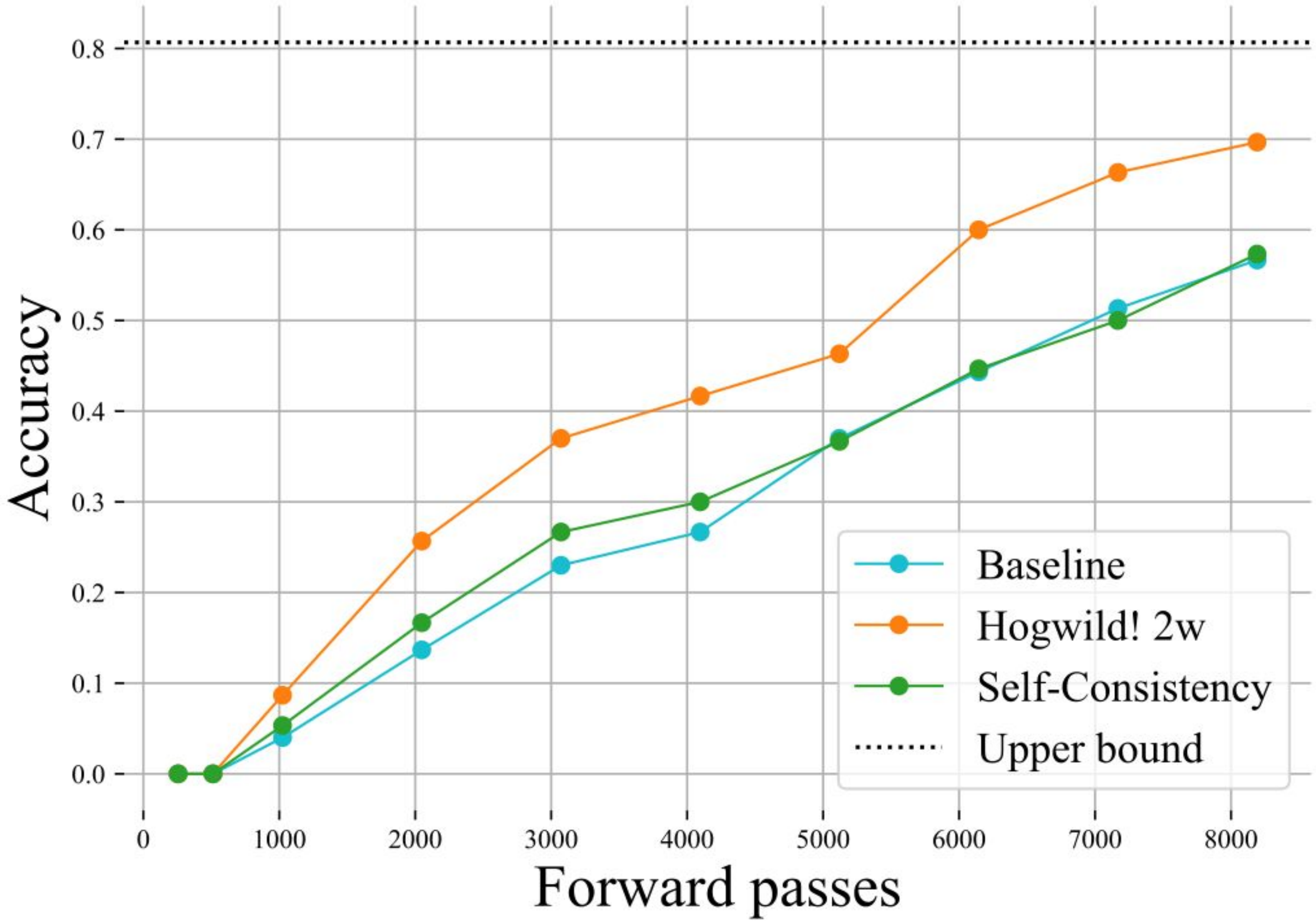
Experiments

DeepSeek-R1, AIME 2025



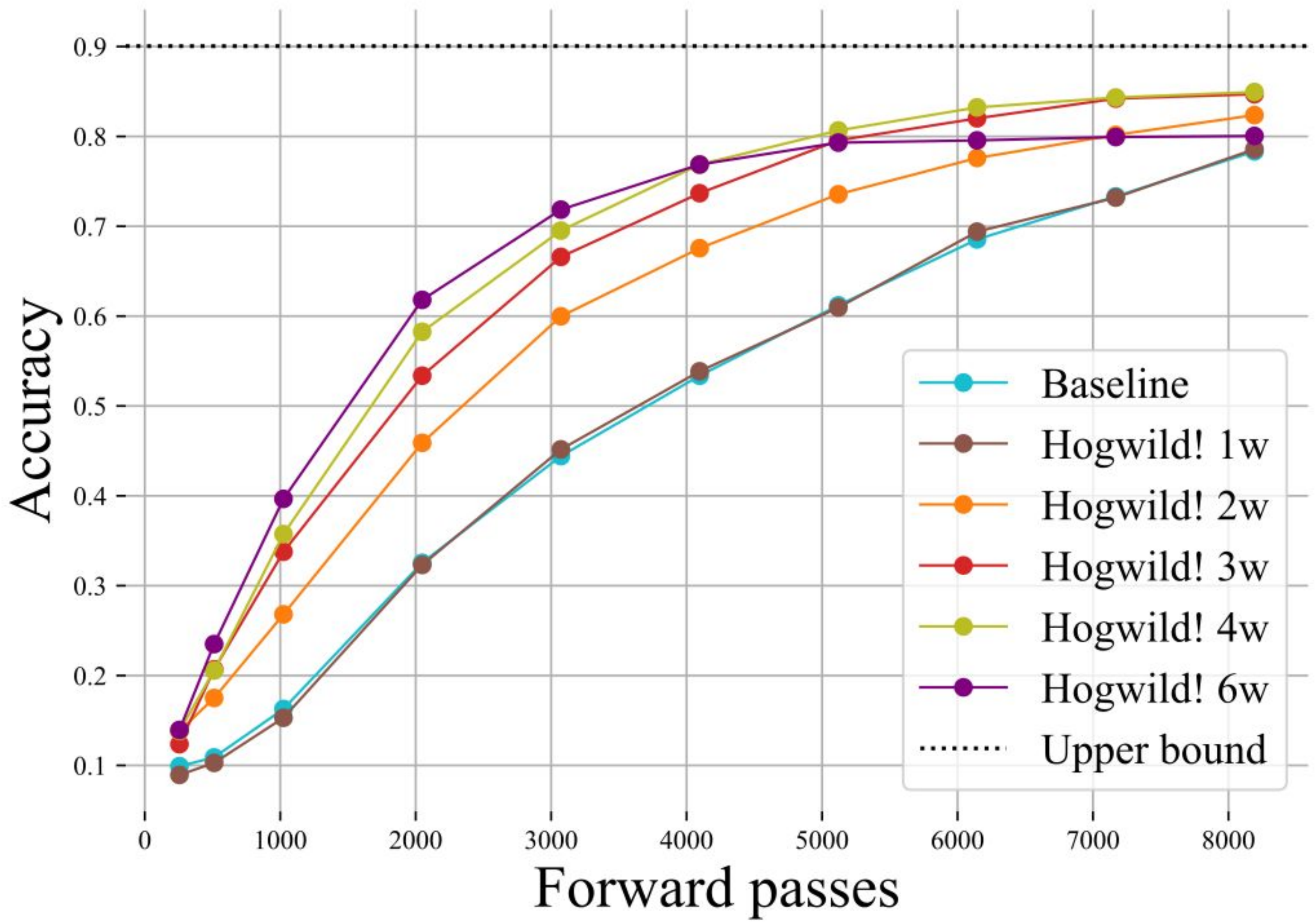
Experiments

Qwen3-235B-A22B, AIME 2025



Experiments

QwQ-32B, LIMO tasks



Demo

Paper, source code and
demo are available at

https://github.com/eqimp/hogwild_llm

Thank you!