



Deep Learning for Vision & Language

Post-Training for LLMs and VLMs: Thinking/Reasoning Models



Last Class

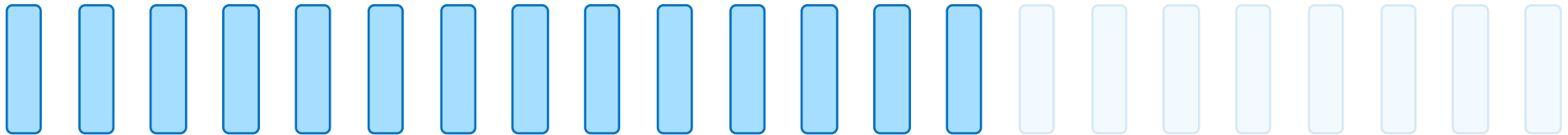
- Efficient Finetuning: LoRA
- Instruction Finetuning (IFT)
 - Supervised Finetuning
- Learning from Human Preferences
 - Reinforcement Learning from Human Feedback (RLHF)
 - Direct Preference Optimization (DPO)

Today

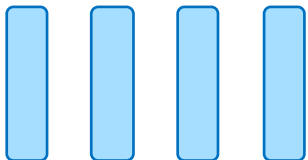
- Chain-of-Thought Reasoning
- Thinking/Reasoning Models
 - Reinforcement Learning with Verifiable Rewards (RLVR)
 - Group Reward Preference Optimization (GRPO)
 - DeepSeek-R1
- Paper Discussion: Budget Forcing

Pretrained: LLM or VLM: Next Token Prediction

Because these are two different constants and numerically ...



Multi-layer Self-Attention Causal Transformer Model

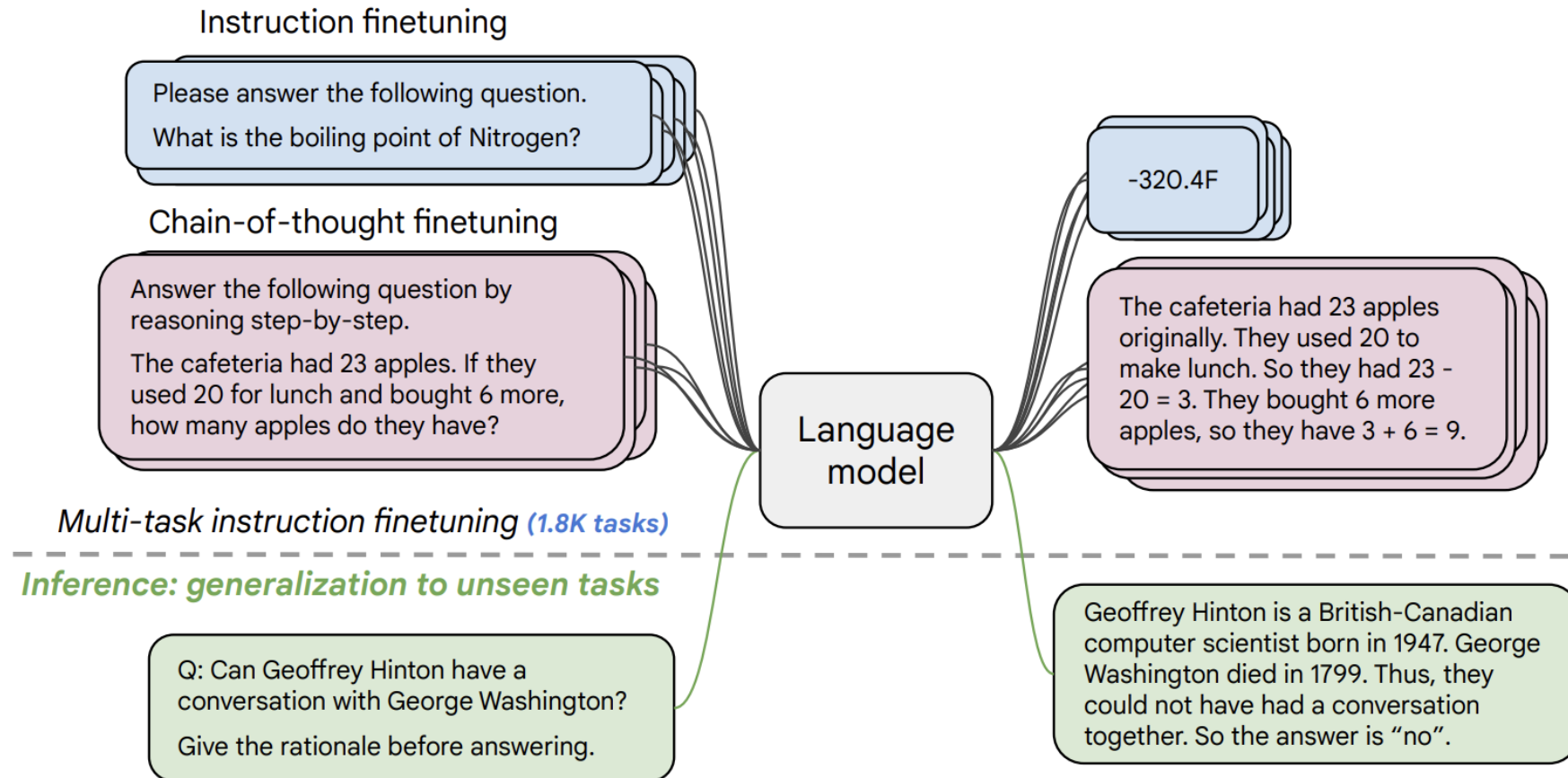


Why is $e > \pi$?

Next Word Prediction is limited

- Predicting the next word can lead to intelligent behavior however this still limited when trying to get answers out of LLMs.
- What makes some of the new LLMs special?
ChatGPT (GPT-3.5, 3.5 Turbo, 4, 4-turbo), FLAN-T5, OPT-IML
 - Instruction Tuning:
 - SFT: Supervised Finetuning (Curated Input/Output Instruction Sets)
 - Preference Tuning:
 - DPO: Direct Preference Optimization
 - PPO: Proximal Policy Optimization (Reinforcement Learning with Human Feedback)
 - Reinforcement Tuning for Reasoning:
 - GRPO: Group-relative Policy Optimization (DeepSeek)

Instruction Tuning (e.g. FLAN-T5 by Google)



InstructGPT (ChatGPT)

SFT (IFT) + Reinforcement Learning with Human Feedback (RLHF)

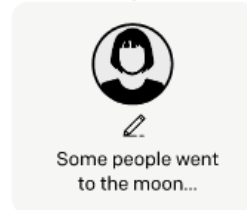
Step 1

Collect demonstration data, and train a supervised policy.

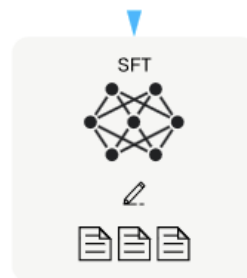
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



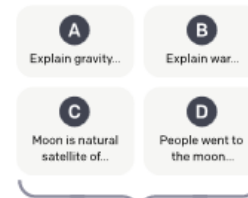
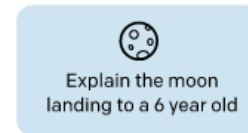
This data is used to fine-tune GPT-3 with supervised learning.



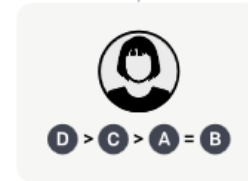
Step 2

Collect comparison data, and train a reward model.

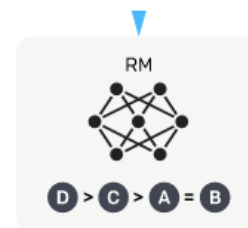
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



Chain-of-Thought Prompting

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

**Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou**

Google Research, Brain Team
{jasonwei, dennyzhou}@google.com

January 2023

Chain-of-Thought Prompting

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. ✅

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go?

Options: (a) race track (b) populated areas
(c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

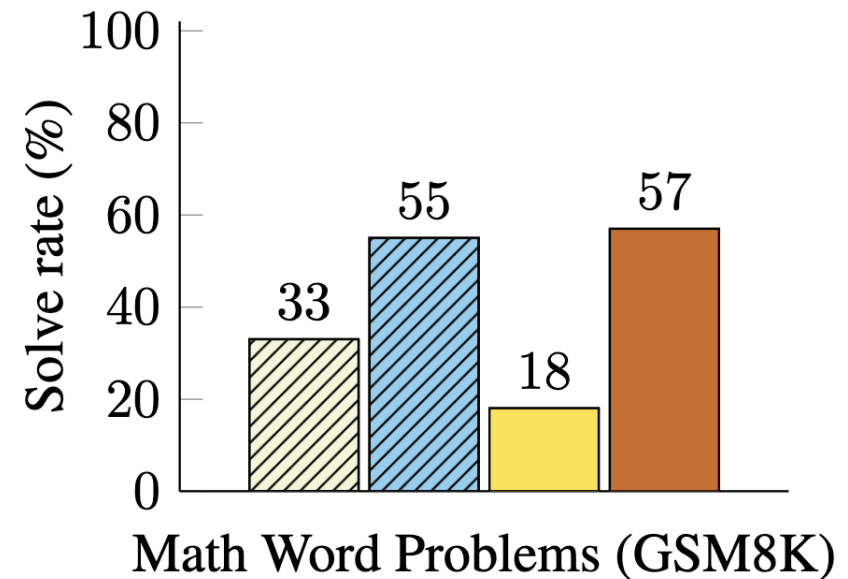
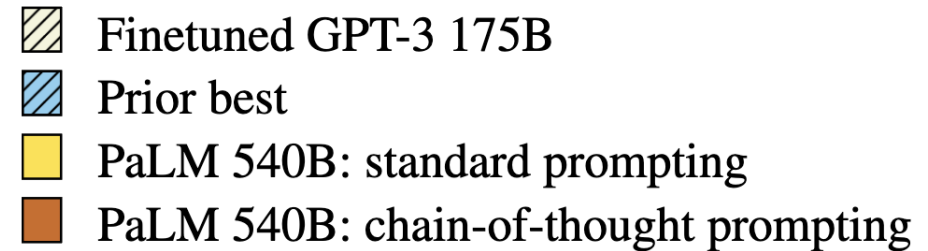
A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Chain-of-Thought Prompting

Weng earns \$12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?

Albert is wondering how much pizza he can eat in one day. He buys 2 large pizzas and 2 small pizzas. A large pizza has 16 slices and a small pizza has 8 slices. If he eats it all, how many pieces does he eat that day?

It takes Roque two hours to walk to work and one hour to ride his bike to work. Roque walks to and from work three times a week and rides his bike to and from work twice a week. How many hours in total does he take to get to and from work a week with walking and biking?

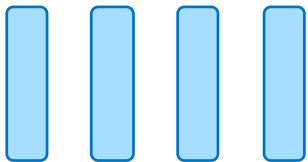


Reasoning Models

<think> Let us analyze each constant e = ... </think> <answer> The question is incorrect ...



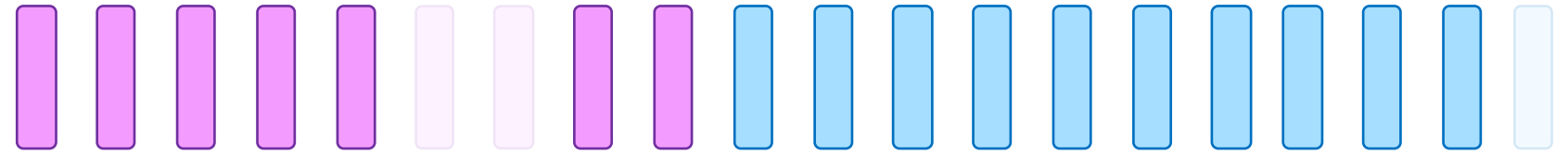
Multi-layer Self-Attention Causal Transformer Model



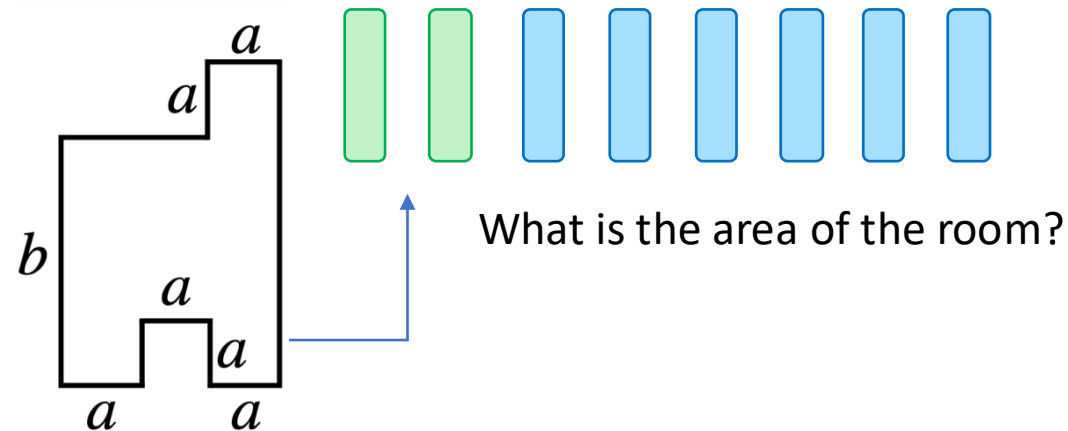
Why is $e > \pi$?

Reasoning + (Visual) Grounding Models

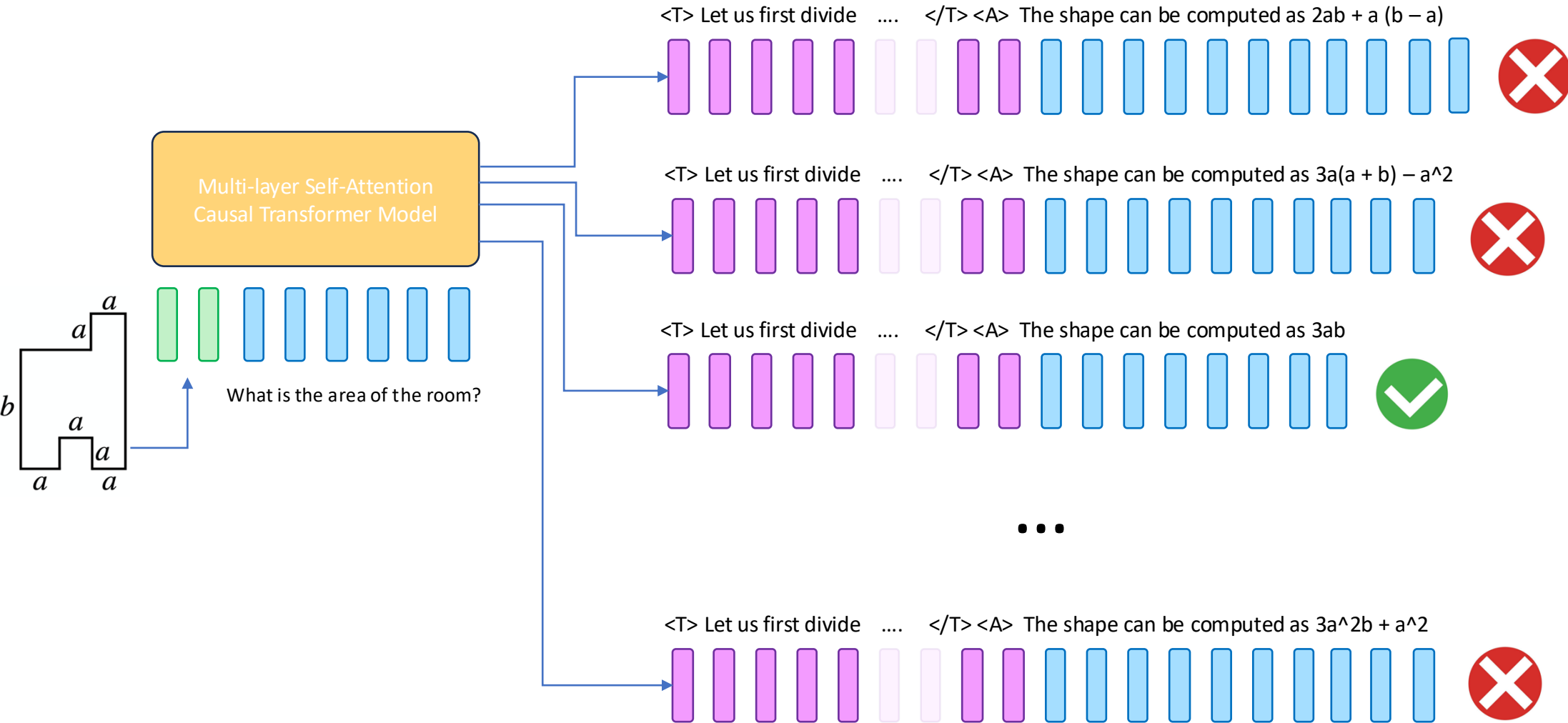
<think> Let us first divide ... </think> <answer> The shape can be computed as...



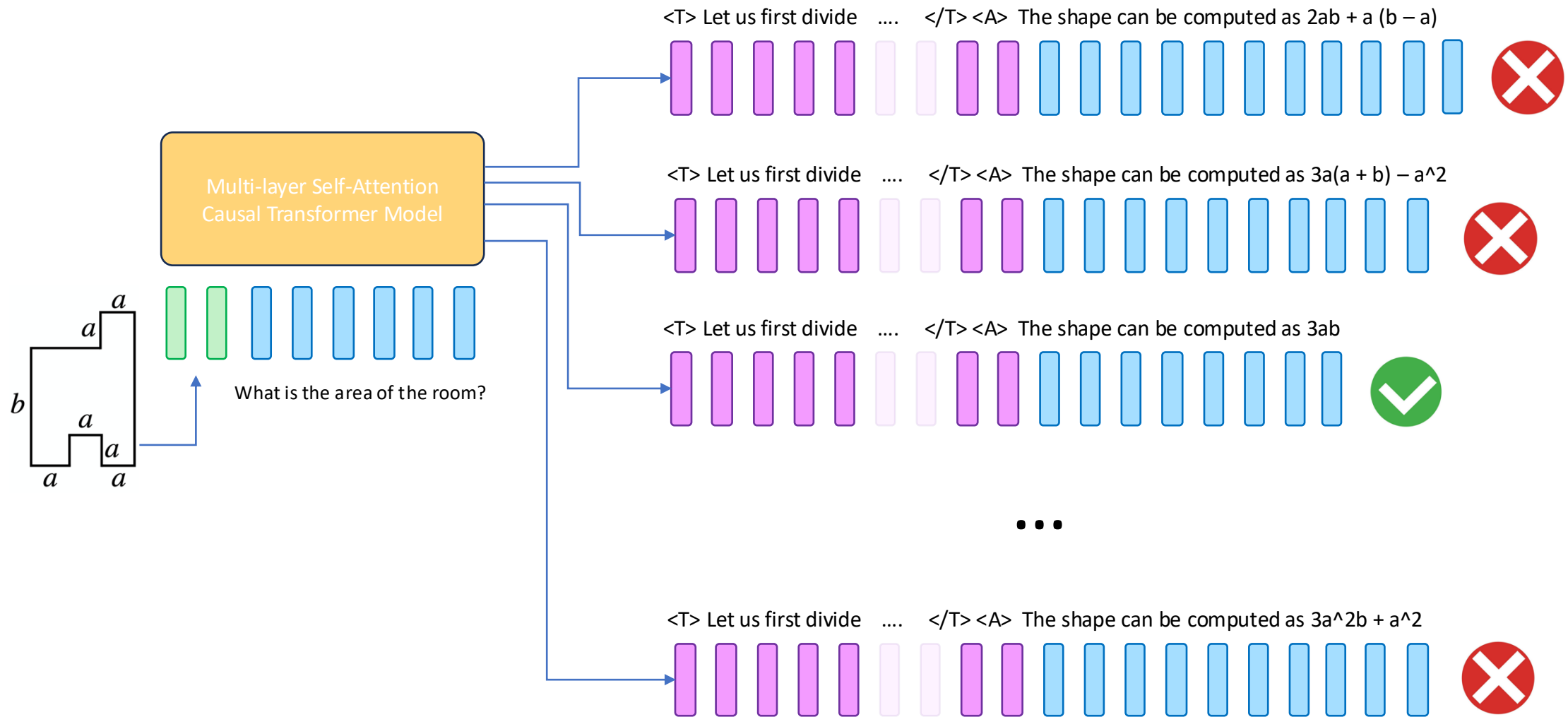
Multi-layer Self-Attention Causal Transformer Model



Reinforcement Tuning (Rollouts!)



Reinforcement Learning with Verifiable Rewards (RLVR)



What problems have verifiable rewards?

- When can it be applied?
- What types of problems?

How does it work?

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

`research@deepseek.com`

January 2025

Group Relative Policy Optimization (GRPO)

We ask the model to attempt the problem 5 times. It must structure its output using XML-style tags.

INPUT PROMPT (x)

"Find x if $3x + 7 = 22$ "

y_1 : <think> $3x=15 \Rightarrow x=5$ </think> <answer>5</answer>

(Correct)

y_2 : <think> $3x=29$ </think> <answer>9.6</answer>

(Logic Error)

y_3 : <think> $3x=15$ </think> <answer>5</answer>

(Correct)

y_4 : <think> $3x=15$ </think> <answer>5</answer>

(Correct)

y_5 : <think> $3x=15$ </think> <answer>15</answer>

(Logic Error)

GRPO – Reward Computation

y_1 : `<think> 3x=15 \Rightarrow x=5 </think> <answer>5</answer>`

(Correct)

y_2 : `<think> 3x=29 </think> <answer>9.6</answer>`

(Logic Error)

y_3 : `<think> 3x=15 </think> <answer>5</answer>`

(Correct)

y_4 : `<think> 3x=15 </think> <answer>5</answer>`

(Correct)

y_5 : `<think> 3x=15 </think> <answer>15</answer>`

(Logic Error)

A Rule-Based Verifier ignores the `<think>` content and only extracts the value between the `<answer>` tags to assign a reward (r_i).

Path	Extracted <code><answer></code>	Reward (r_i)	Status
y_1	5	1.0	Correct
y_2	9.6	0.0	Error
y_3, y_4	5	1.0	Correct
y_5	15	0.0	Error

GRPO – Relative Group Advantage Computation

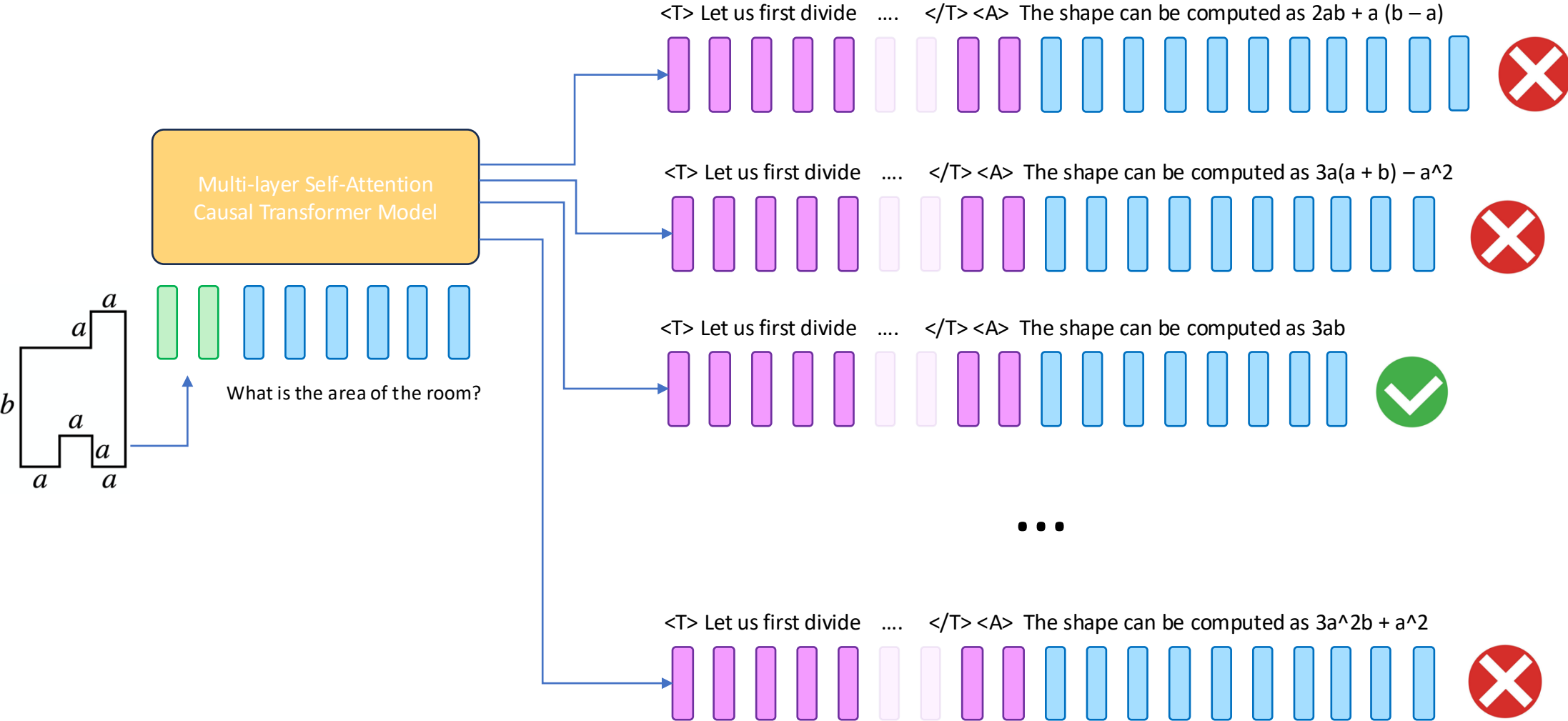
<code>y₁: <think> 3x=15 ⇒ x=5 </think> <answer>5</answer></code>	(Correct)
<code>y₂: <think> 3x=29 </think> <answer>9.6</answer></code>	(Logic Error)
<code>y₃: <think> 3x=15 </think> <answer>5</answer></code>	(Correct)
<code>y₄: <think> 3x=15 </think> <answer>5</answer></code>	(Correct)
<code>y₅: <think> 3x=15 </think> <answer>15</answer></code>	(Logic Error)

Correct paths are rewarded for being above average. Failed paths are penalized for being below it.



$$A_i = \frac{r_i - \mu}{\sigma} = \frac{r_i - 0.60}{0.49}$$

Reinforcement Tuning (Rollouts!)



Simplest Loss Function to Use (REINFORCE)

$$\sum A_i \cdot \log \pi_{\theta}(y_i | x)$$

Standard policy gradients update weights directly by the log probability. This is **volatile**. One extremely lucky or weird sample can cause a massive gradient step that destroys the model.

Importance Sampling Ratio (r_i)

To safely reuse data for multiple gradient steps, we calculate the ratio between the policy currently being trained (π_θ) and the policy that originally generated the sample ($\pi_{\theta_{old}}$).

$$r_i(\theta) = \frac{\pi_\theta(y_i|x)}{\pi_{\theta_{old}}(y_i|x)}$$

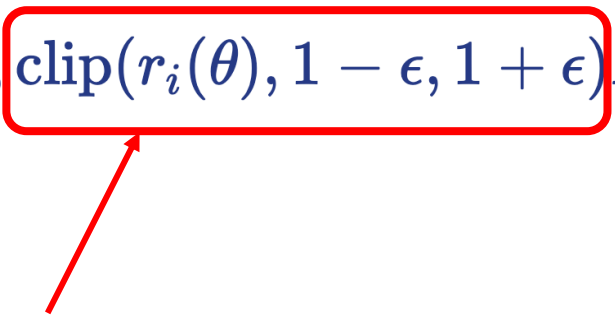
If $r_i(\theta) > 1$, the new model is **more likely** to output this path than the old model.

If $r_i(\theta) < 1$, the new model is **less likely** to output this path.

The GRPO Objective Function

$$J(\theta) = \frac{1}{G} \sum_{i=1}^G [\min(r_i(\theta)A_i, \text{clip}(r_i(\theta), 1 - \epsilon, 1 + \epsilon)A_i) - \beta D_{KL}(\pi_{\theta} || \pi_{ref})]$$

The GRPO Objective Function

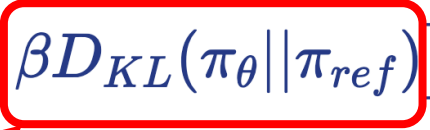
$$J(\theta) = \frac{1}{G} \sum_{i=1}^G [\min(r_i(\theta)A_i, \text{clip}(r_i(\theta), 1 - \epsilon, 1 + \epsilon)A_i) - \beta D_{KL}(\pi_\theta || \pi_{ref})]$$


We restrict how much the ratio can change in a single update using a clipping hyperparameter (usually $\epsilon = 0.2$).

Prevents Collapse

Without clipping, a massive advantage ($A_i = +5.0$) might cause the model to update so aggressively it forgets grammar just to force that specific output.

The GRPO Objective Function

$$J(\theta) = \frac{1}{G} \sum_{i=1}^G [\min(r_i(\theta)A_i, \text{clip}(r_i(\theta), 1 - \epsilon, 1 + \epsilon)A_i) - \beta D_{KL}(\pi_{\theta} || \pi_{ref})]$$


Same as PPO. We want to further prevent the model from deviating too much from the base “reference” model. We still want coherent language to be generated.



what are the best resources to learn a new language?

Thinking



what are the best resources to learn a new language?

 Assessing Contextual Relevance ▾

Answer now



what are the best resources to learn a new language?

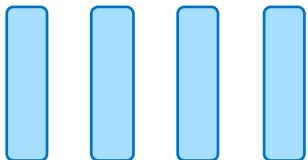
 Thinking about evaluating language learning resources and methods

Reasoning Models

<think> Let us analyze each constant e = ... </think> <answer> The question is incorrect ...



Multi-layer Self-Attention Causal Transformer Model



Why is $e > \pi$?

Paper Discussion

s1: Simple test-time scaling

**Niklas Muennighoff^{*134} Zitong Yang^{*1} Weijia Shi^{*23} Xiang Lisa Li^{*1} Li Fei-Fei¹ Hannaneh Hajishirzi²³
Luke Zettlemoyer² Percy Liang¹ Emmanuel Candès¹ Tatsunori Hashimoto¹**

Can someone explain this figure?

How many r in raspberry?

Question

Let's break down the process of counting the letter 'r' in the word "raspberry" ...

* First letter: 'r' - This is an 'r', count = 1.

* Second letter: 'a' - Not an 'r', count remains 1 ...

* Sixth letter: 'e' - Not an 'r', count remains 1.

* Seventh letter: 'r' - This is an 'r', count = 2.

* Eighth letter: 'y' - Not an 'r', count remains 2 ...

The number of 'r's in "raspberry" is 2.

Wait, let's re-read the question carefully. It asks "How many r in raspberry?" ... * r - a - s - p - b - e - r - r - y ... * First 'r' ... *

Second 'r' ... * Third 'r' ... Count = 3 ...

Reasoning trace

My initial answer of 2 was incorrect due to a quick reading of the word. **Final Answer:** The final answer is **3**

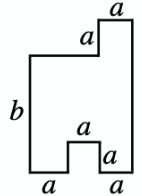
Response

Paper Discussion

PROXYTHINKER: TEST-TIME GUIDANCE THROUGH SMALL VISUAL REASONERS

Zilin Xiao¹, Jaywon Koo¹, Siru Ouyang², Jefferson Hernandez¹, Yu Meng³, Vicente Ordonez¹
¹Rice University ²University of Illinois Urbana-Champaign ³University of Virginia
{zilin,vicenteor}@rice.edu

Can someone explain this Figure?

Question:  **What is the area of the room?**
 A. $2ab + a(b - a)$ B. $3a(a + b) - a^2$
 C. $3a^2b$ D. $3a(b - a) + a^2$ E. $3ab$

Current Thinking Process:
 ...
 5. **Simplify the expression:**
 $4ab - 4a^2 = 4a(b - a)$
 6. **Rewrite the expression in a form that matches one of the given choices:**
 $4a(b - a) = 3a(a + b) - a^2$

Three models are shown:

- large, Base model:** Outputs "Final Thus Answer" with a bar chart. Final answer: **B**. ❌
- small, reasoning Expert:** Outputs "Wait However But" with a bar chart. Wait, does it seem right? Let's double check the choices: (A) $2ab + a(b - a)$ (B) $3a(a + b) - a^2$ (C) $3a^2b$ (D) $3a(b - a) + a^2$ (E) $3ab$. The correct choice is **B**. ❌
- small, Amateur model:** Outputs "Final Thus Answer" with a bar chart. Answer: **B**. ❌

Wait, let's double check. The room can be divided into a large rectangle and two smaller rectangles. Simplify the expression:
 Total area = $2ab + a^2 + ab - a^2 = 3ab$
 The expression $3ab$ matches choice (E).
 Thus, the area of the room is: **E** ✅

Diagram showing the combination of model outputs: α (Base model output) + α (Expert model output) - α (Amateur model output) = Final result (Expert model output).

Questions?