Lecture 20: LLM Test-Time Inference and Reasoning
CS486/686 Intro to Artificial Intelligence

2026-3-19

Pascal Poupart
David R. Cheriton School of Computer Science
CIFAR AI Chair at Vector Institute

UNIVERSITY OF WATERLOO

Outline

Improvements to RL from Human Feedback:

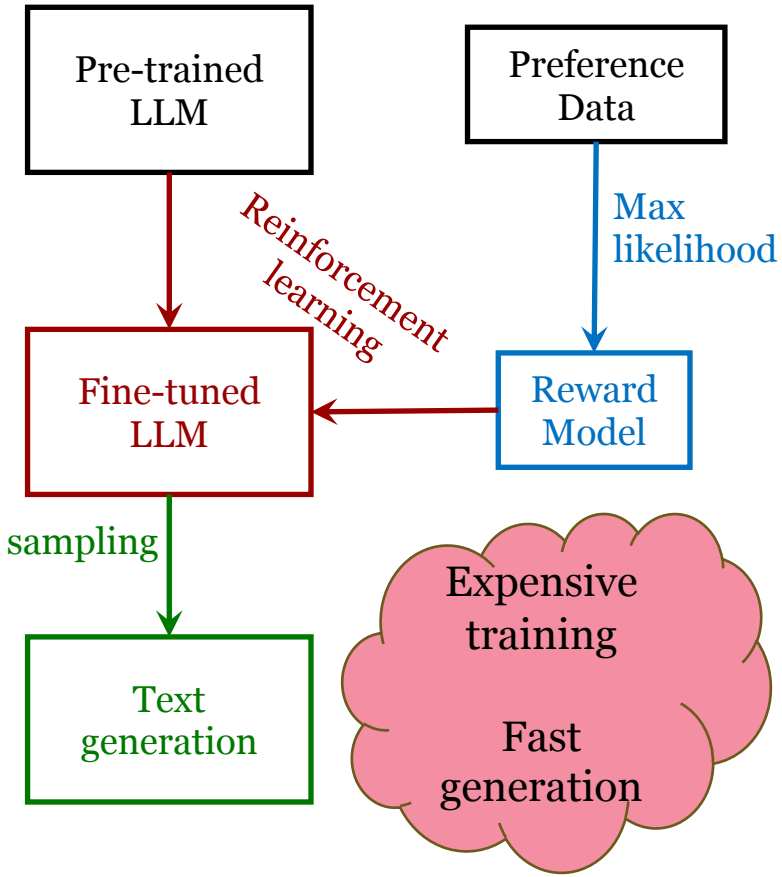
- Direct Preference Optimization
- Reward Guided Text Generation

LLM Reasoning (test time inference):

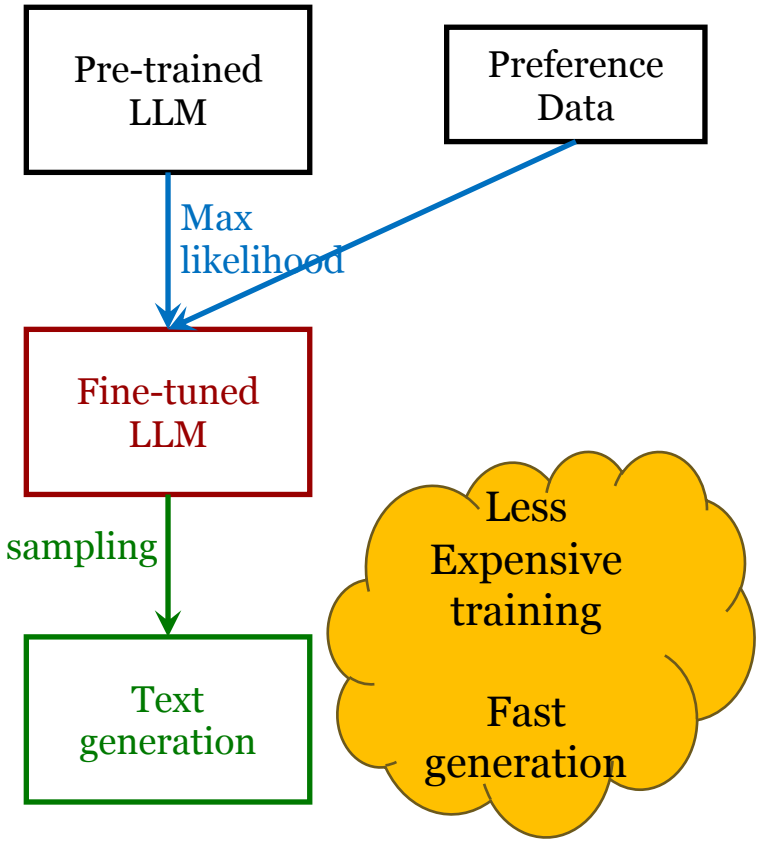
- Chain of thought reasoning
- Reasoning by searching
- Self-reflection

RLHF Improvements

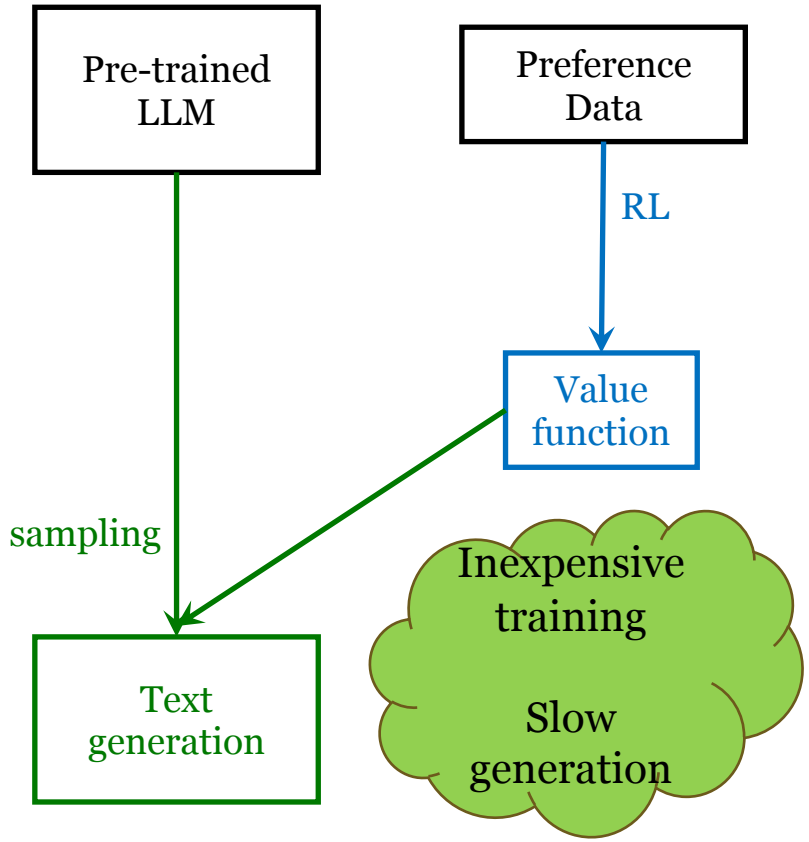
Proximal Policy Optimization (PPO)
Ouyang et al., 2022



Direct Preference Optimization (DPO)
Rafailov et al., 2023

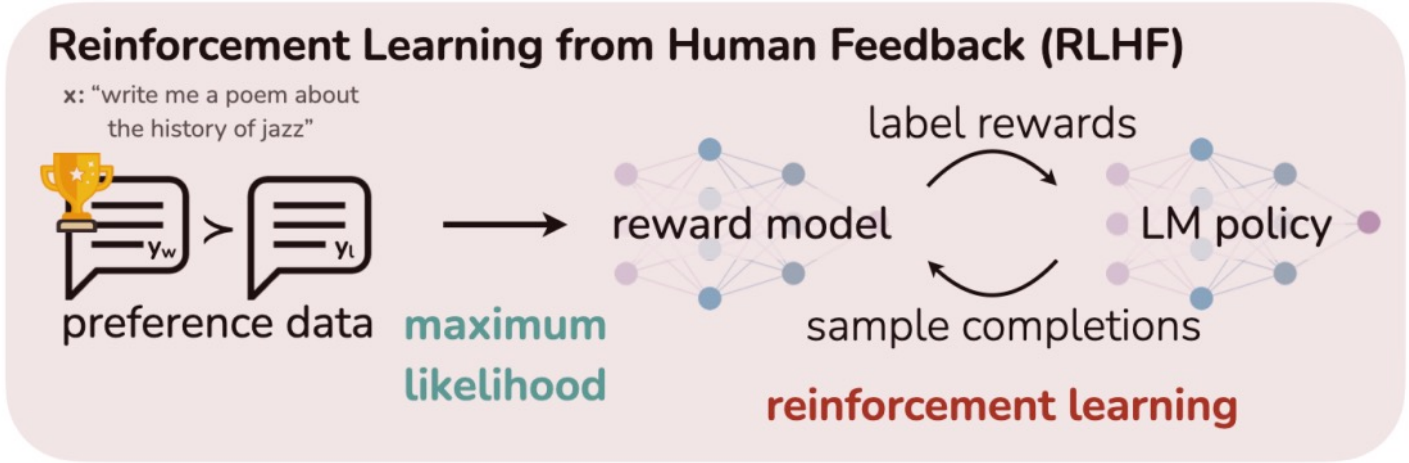


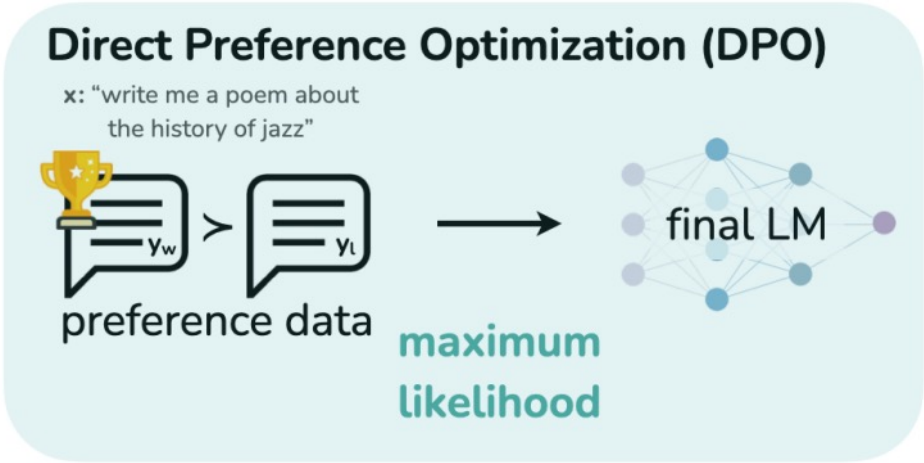
Reward Guided Text Generation (RGTG)
Khanov et al, 2024
Rashid et al., 2025



Direct Preference Optimization

Rafailov, Sharma, Mitchell, Ermon, Manning, Finn (2023) **Direct Preference Optimization: Your Language Model is Secretly a Reward Model**, *NeurIPS*.





Bypassing RL

- Recall RL objective:

$$\max_{\phi} E_{s \in Dataset}\left[E_{a \sim \pi_\phi(a|s)}[r_\theta(s,a)] - \beta\, KL\left(\pi_\phi(\cdot\,|s)\middle|\pi_{ref}(\cdot\,|s)\right)\right]$$

- Closed form solution (proof on next slide):

$$\pi_\phi(a|s) = \frac{1}{Z(s)}\pi_{ref}(a|s)\exp\left(\frac{r_\theta(s,a)}{\beta}\right)$$

- Isolate reward: $r_\theta(s,a) = \beta \log \frac{\pi_\phi(a|s)}{\pi_{ref}(a|s)} + \beta \log Z(s)$

- Plug into preference objective:

$$Loss(\theta) = -\frac{1}{\binom{k}{2}} E_{(s,a_i,a_j) \in Dataset} \log \sigma\left(r_\theta(s,a_i) - r_\theta\left(s,a_j\right)\right)$$

$$= -\frac{1}{\binom{k}{2}} E_{(s,a_i,a_j) \in Dataset} \log \sigma\left(\beta \log \frac{\pi_\phi(a_i|s)}{\pi_{ref}(a_i|s)} - \beta \log \frac{\pi_\phi\left(a_j|s\right)}{\pi_{ref}\left(a_j|s\right)}\right)$$

Optimal Policy Derivation

$$\arg\max_{\phi} E_{s\in Dataset}\left[E_{a\sim\pi_\phi(a|s)}[r_\theta(s,a)] - \beta\, KL\left(\pi_\phi(\cdot\,|s)\middle|\pi_{ref}(\cdot\,|s)\right)\right]$$

$$= \arg\max_{\phi} E_{s\in Dataset}\left[E_{a\sim\pi_\phi(a|s)}\left[r_\theta(s,a) - \beta\log\frac{\pi_\phi(a|s)}{\pi_{ref}(a|s)}\right]\right] \quad \text{by KL definition}$$

$$= \arg\min_{\phi} E_{s\in Dataset}\left[E_{a\sim\pi_\phi(a|s)}\left[\log\frac{\pi_\phi(a|s)}{\pi_{ref}(a|s)} - \frac{1}{\beta}r_\theta(s,a)\right]\right] \quad \text{since max = - min}$$

$$= \arg\min_{\phi} E_{s\in Dataset}\left[E_{a\sim\pi_\phi(a|s)}\left[\log\frac{\pi_\phi(a|s)}{\frac{1}{Z(s)}\pi_{ref}(a|s)\exp\left(\frac{r_\theta(s,a)}{\beta}\right)} - \log Z(s)\right]\right] \quad \text{where } Z(s) = \sum_a \pi_{ref}(a|s)\exp\left(\tfrac{r_\theta(s,a)}{\beta}\right)$$

$$= \arg\min_{\phi} E_{s\in Dataset}\left[E_{a\sim\pi_\phi(a|s)}\left[\log\frac{\pi_\phi(a|s)}{\frac{1}{Z(s)}\pi_{ref}(a|s)\exp\left(\frac{r_\theta(s,a)}{\beta}\right)}\right]\right] \quad \text{since } \log Z(s) \text{ is independent of } \phi$$

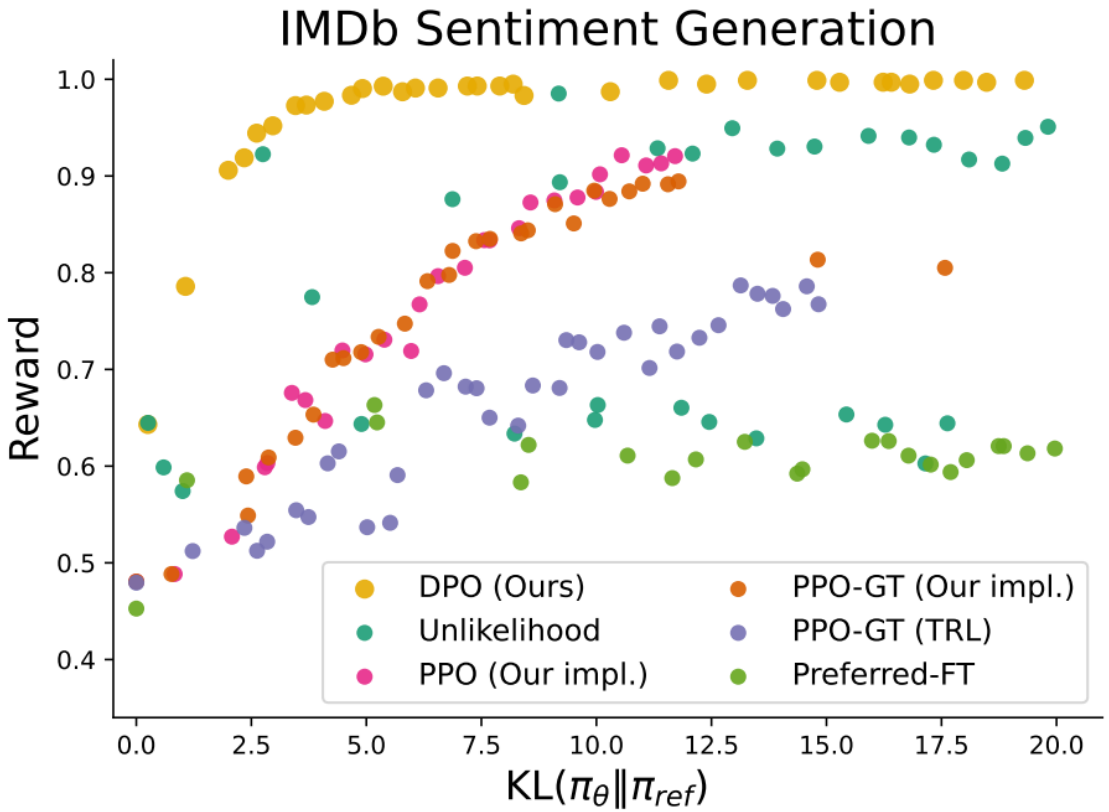
$$= \arg\min_{\phi} E_{s\in Dataset}\left[E_{a\sim\pi_\phi(a|s)}\left[\log\frac{\pi_\phi(a|s)}{\pi_{\phi^*}(a|s)}\right]\right] \quad \text{where } \pi_{\phi^*}(a|s) = \frac{1}{Z(s)}\pi_{ref}(a|s)\exp\left(\frac{r_\theta(s,a)}{\beta}\right)$$

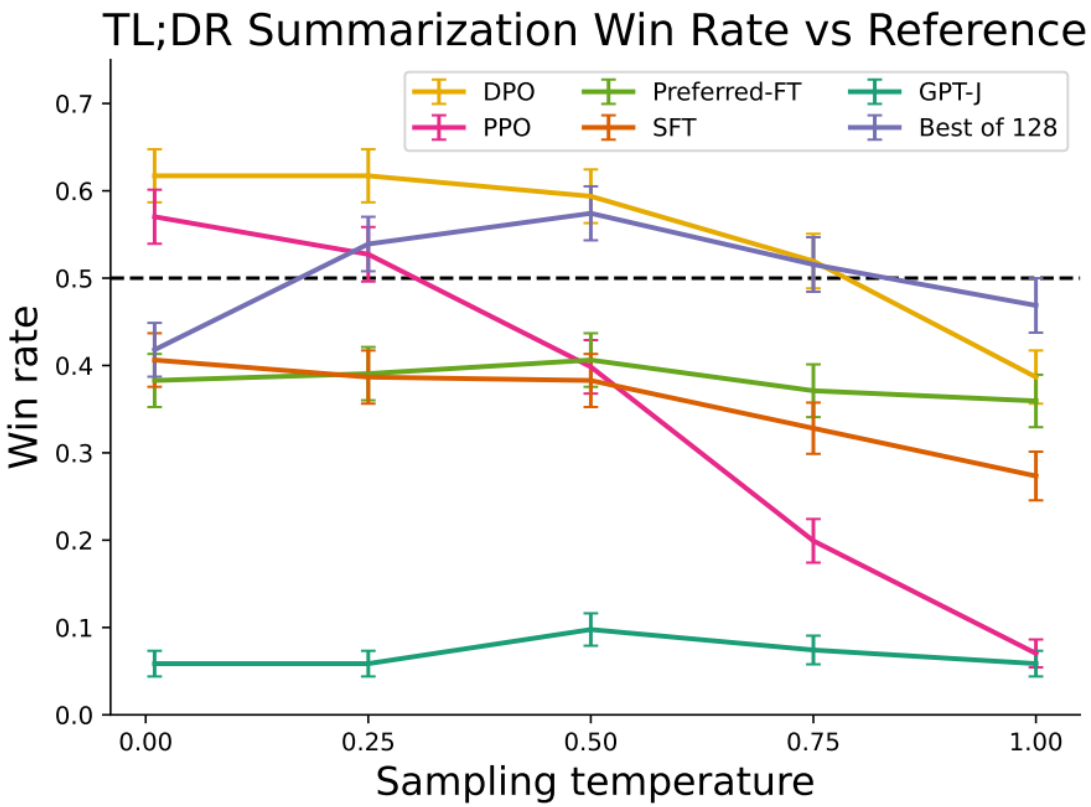
$$= \arg\min_{\phi} E_{s\in Dataset}\left[KL(\pi_\phi(\cdot\,|s)||\pi_{\phi^*}(\cdot\,|s))\right] \quad \text{by KL definition}$$

$$= \phi^* \quad \text{since KL is minimized when both arguments are equal}$$

Empirical Results

Rafailov et al. 2023

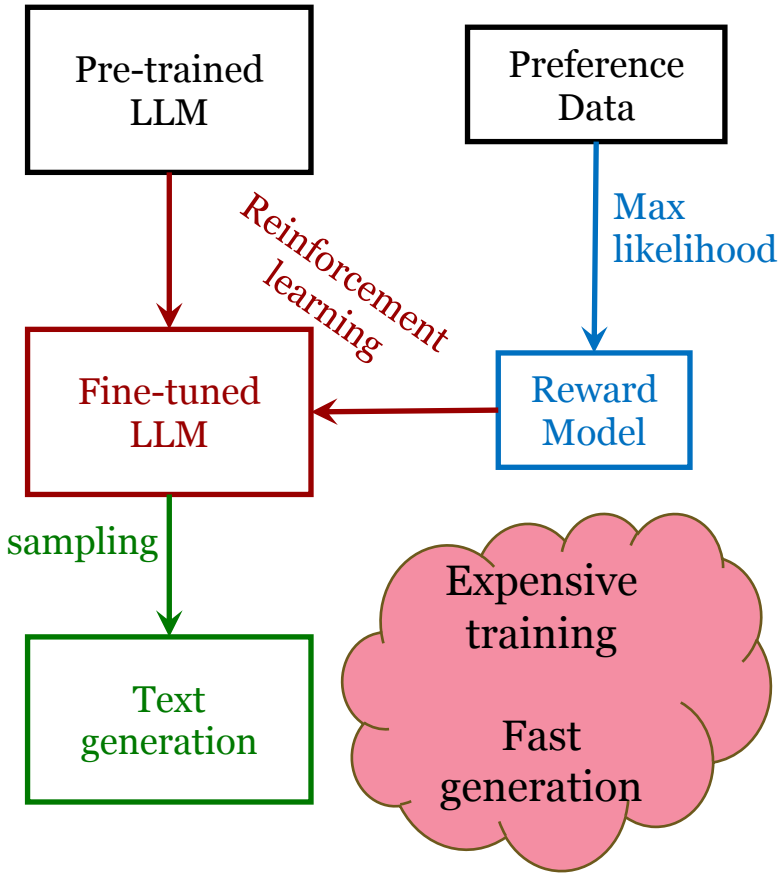




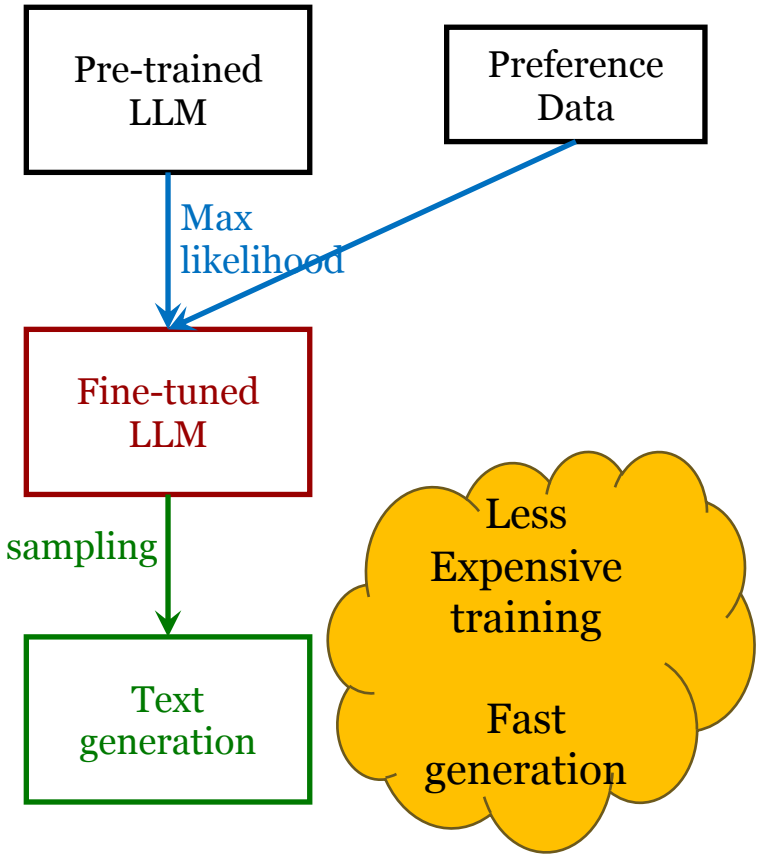
UNIVERSITY OF WATERLOO

RLHF Improvements

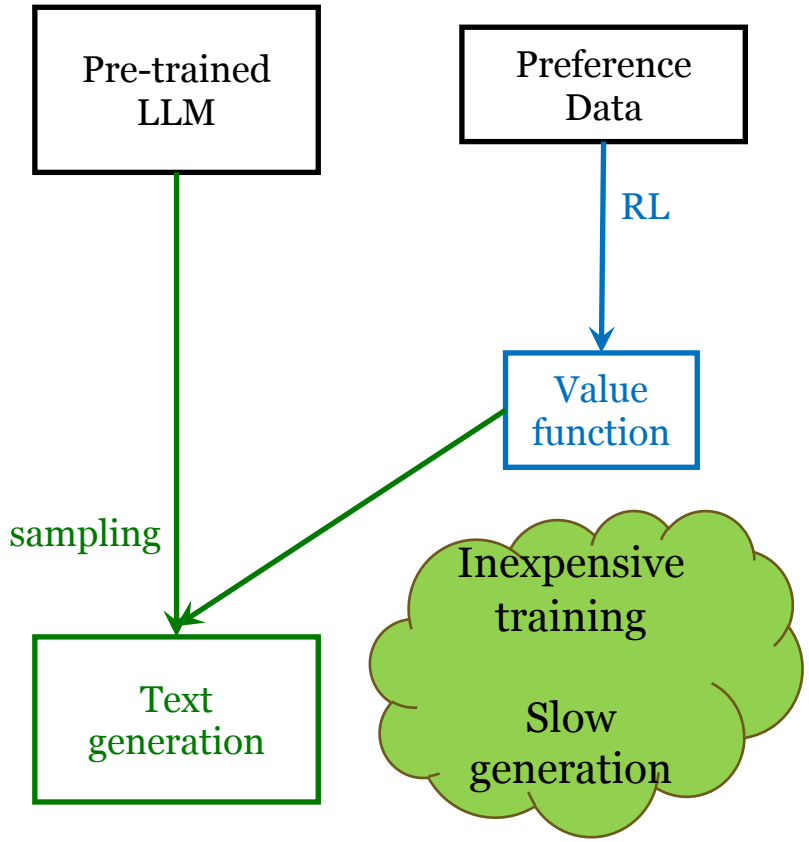
Proximal Policy Optimization (PPO)
Ouyang et al., 2022



Direct Preference Optimization (DPO)
Rafailov et al., 2023



Reward Guided Text Generation (RGTG)
Khanov et al, 2024
Rashid et al., 2025



Text Generation

- Utterance-level RL (learn reward function)
  - State: prompt
  - Action: response
  - Reward: score complete response
  - Horizon: 1 time step (no transition function)
- Token-level RL (learn Q-function)
  - State: [prompt, partial response]
  - Action: next token
  - Reward: score complete response (0 for partial response)
  - Horizon: length of response
    - Deterministic transition function:
      T([prompt, partial response], next token) →[prompt, partial response, next token]
    - Q(s,a): score next token

Sequence Generation

- Recall closed form solution

$$\pi_\phi(\boldsymbol{a}|\boldsymbol{s}) = \frac{1}{Z(\boldsymbol{s})}\pi_{ref}(\boldsymbol{a}|\boldsymbol{s})\exp\left(\frac{r_\theta(\boldsymbol{s},\boldsymbol{a})}{\beta}\right)$$

$$= softmax\left(\log \pi_{ref}(\boldsymbol{a}|\boldsymbol{s}) + \frac{r_\theta(\boldsymbol{s},\boldsymbol{a})}{\beta}\right)$$

- Text generation:

$$\boldsymbol{a} \sim softmax\left(\log\begin{pmatrix}\pi_{ref}(\boldsymbol{a}_1|\boldsymbol{s})\\ \pi_{ref}(\boldsymbol{a}_2|\boldsymbol{s})\\ \pi_{ref}(\boldsymbol{a}_3|\boldsymbol{s})\\ \dots\\ \pi_{ref}(\boldsymbol{a}_n|\boldsymbol{s})\end{pmatrix} + \begin{pmatrix}r_\theta(\boldsymbol{s},\boldsymbol{a}_1)\\ r_\theta(\boldsymbol{s},\boldsymbol{a}_2)\\ r_\theta(\boldsymbol{s},\boldsymbol{a}_3)\\ \dots\\ r_\theta(\boldsymbol{s},\boldsymbol{a}_n)\end{pmatrix}/\beta\right)$$

Token Generation

- Token-wise LLM modeling

$$\pi_\phi\left(a^i|\boldsymbol{s},\boldsymbol{a}^{1:i-1}\right)=\frac{1}{Z(\boldsymbol{s})}\pi_{ref}\left(a^i|\boldsymbol{s},\boldsymbol{a}^{1:i-1}\right)\exp\left(\frac{Q_\theta\left(\boldsymbol{s},\boldsymbol{a}^{1:i}\right)}{\beta}\right)$$

$$=softmax\left(\log\pi_{ref}\left(a^i|\boldsymbol{s},\boldsymbol{a}^{1:i-1}\right)+\frac{Q_\theta\left(\boldsymbol{s},\boldsymbol{a}^{1:i}\right)}{\beta}\right)$$

- Token generation:

$$a^i\sim softmax\left(\log\begin{pmatrix}\pi_{ref}\left(a_1^i|\boldsymbol{s},\boldsymbol{a}^{1:i-1}\right)\\ \pi_{ref}\left(a_2^i|\boldsymbol{s},\boldsymbol{a}^{1:i-1}\right)\\ \pi_{ref}\left(a_3^i|\boldsymbol{s},\boldsymbol{a}^{1:i-1}\right)\\ \dots\\ \pi_{ref}\left(a_n^i|\boldsymbol{s},\boldsymbol{a}^{1:i-1}\right)\end{pmatrix}+\begin{pmatrix}Q_\theta\left(\boldsymbol{s},\boldsymbol{a}^{1:i-1},a_1^i\right)\\ Q_\theta\left(\boldsymbol{s},\boldsymbol{a}^{1:i-1},a_2^i\right)\\ Q_\theta\left(\boldsymbol{s},\boldsymbol{a}^{1:i-1},a_3^i\right)\\ \dots\\ Q_\theta\left(\boldsymbol{s},\boldsymbol{a}^{1:i-1},a_n^i\right)\end{pmatrix}/\beta\right)$$

FaRMA: Faster Reward Model for Alignment

- Rashid, Wu, Fan, Li, Kristiadi, Poupart (2025) **Towards Cost-Effective Reward Guided Text Generation**, *ICML*.
- Optimization problem:

$$\max_{\theta} E_{(\boldsymbol{s},\boldsymbol{a}_+,\boldsymbol{a}_-)\in Dataset} \log \sigma\big(Q_\theta(\boldsymbol{s},\boldsymbol{a}_+) - Q_\theta(\boldsymbol{s},\boldsymbol{a}_-)\big)$$
$$\text{Subject to } Q_\theta\big(\boldsymbol{s},\boldsymbol{a}^{1:i}\big) = \max_{\boldsymbol{a}^{i+1:|\boldsymbol{a}|}} Q_\theta\big(\boldsymbol{s},[\boldsymbol{a}^{1:i},\boldsymbol{a}^{i+1:|\boldsymbol{a}|}]\big)\ \forall \boldsymbol{s},\boldsymbol{a},i$$

- In practice: alternate between minimizing two loss functions
  - $L_1(\theta) = -E_{(\boldsymbol{s},\boldsymbol{a}_+,\boldsymbol{a}_-)\in Dataset} \log \sigma\big(Q_\theta(\boldsymbol{s},\boldsymbol{a}_+) - Q_\theta(\boldsymbol{s},\boldsymbol{a}_-)\big)$
  - $L_2(\theta) = \frac{1}{2} E_{(\boldsymbol{s},\boldsymbol{a})\in Dataset, i\leq|\boldsymbol{a}|} \left(Q_\theta(\boldsymbol{s},\boldsymbol{a}^{1:i}) - \max_{\boldsymbol{a}^{i+1:|\boldsymbol{a}|}} Q_\theta\big(\boldsymbol{s},[\boldsymbol{a}^{1:i},\boldsymbol{a}^{i+1:|\boldsymbol{a}|}]\big)\right)^2$

FaRMA Pseudocode

Repeat

Repeat for each $(s, \boldsymbol{a}_+, \boldsymbol{a}_-)$ in minibatch

$$L_1(\theta) = \log \sigma\big(Q_\theta(\boldsymbol{s}, \boldsymbol{a}_+) - Q_\theta(\boldsymbol{s}, \boldsymbol{a}_-)\big)$$

$$\theta \leftarrow \theta \ - \alpha \nabla L_1(\theta)$$

Repeat for each $(\boldsymbol{s}, \boldsymbol{a}, i)$ in minibatch

$$L_2(\theta) = \frac{1}{2}\left(Q_\theta\left(\boldsymbol{s}, \boldsymbol{a}^{1:i}\right) - \max_{a^{i+1}} Q_\theta\left(\boldsymbol{s}, \boldsymbol{a}^{1:i+1}\right)\right)^2$$

$$\theta \leftarrow \theta \ - \alpha \nabla L_2(\theta)$$

Empirical Results

TL;DR Summarization			
Method	**LLM**	$r \pm$ SE	**Time(min)**
π_{ref}	frozen	0.98±0.18	2
ARGS	frozen	1.46±0.16	32
PARGS	frozen	1.56±0.19	31
CD	frozen	1.15±0.16	29
FaRMA	frozen	2.05±0.15	5
CARDS	frozen	1.73±0.16	17
DPO	trained	2.08±0.18	2
PPO	trained	2.05±0.14	2

Table 2. Avg. reward (over 100 samples) ± standard error total generation time for the TL;DR summarization task.

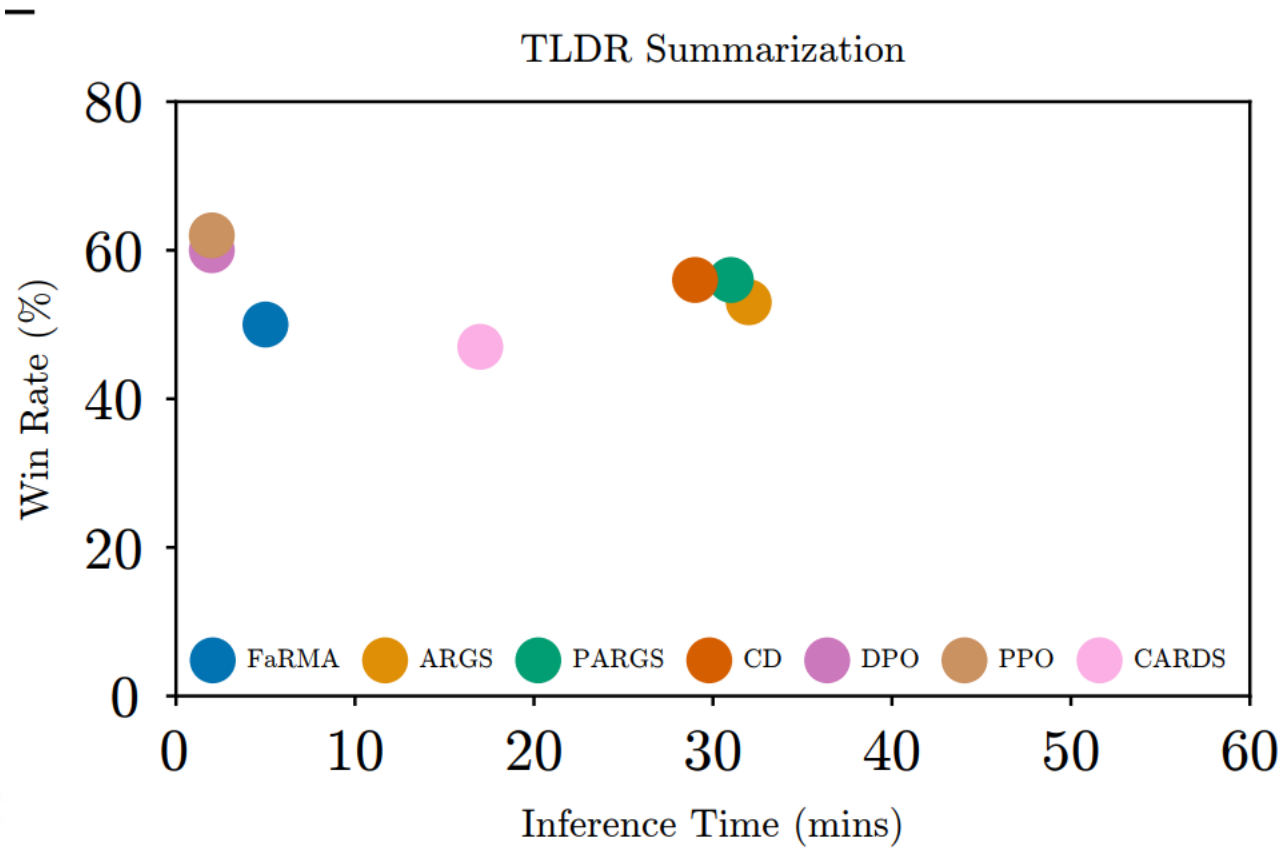


Figure 2. GPT4 evaluation on TLDR

Reasoning LLMs

Source: Pan, Ji et al. (2025) A Survey of Slow Thinking-based Reasoning LLMs using Reinforced Learning and Inference-time Scaling Law

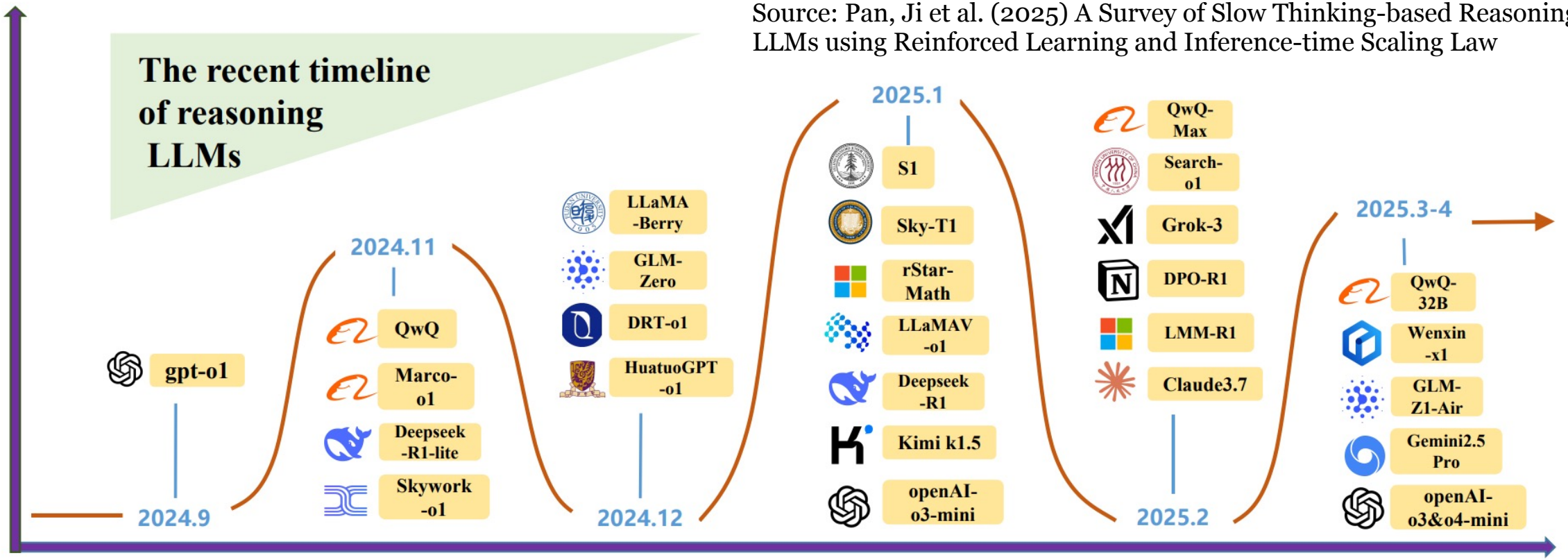
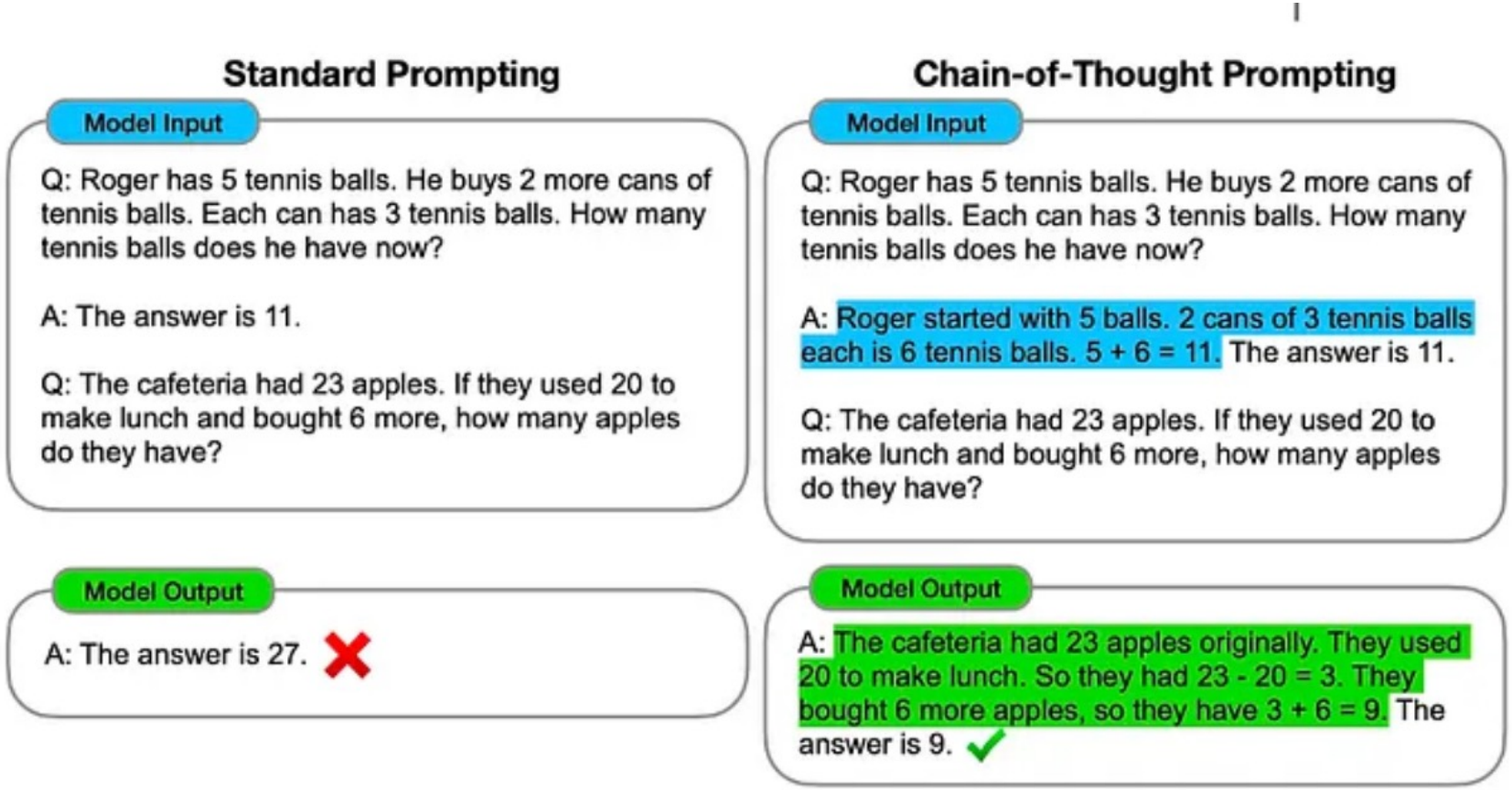


Fig. 1. The timeline of main reasoning LLMs.

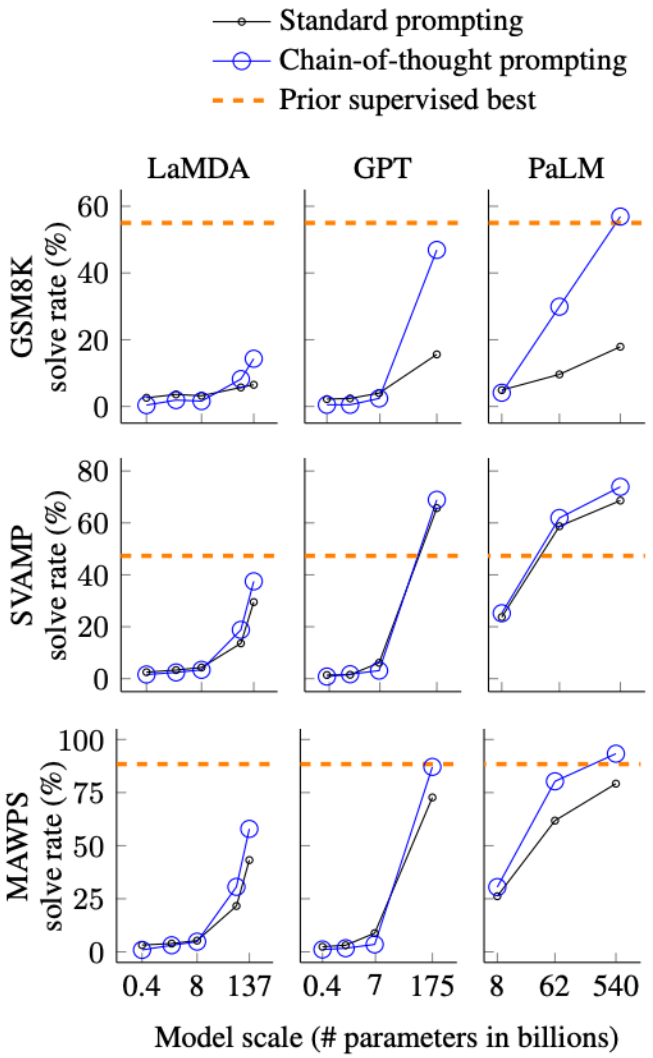
Chain of Thought (CoT) Prompting



Credit: https://cameronrwolfe.substack.com/p/chain-of-thought-prompting-for-llms

Impact of CoT

Source: Wei et al. (2022) Chain of thought prompting elicits reasoning in large language models.



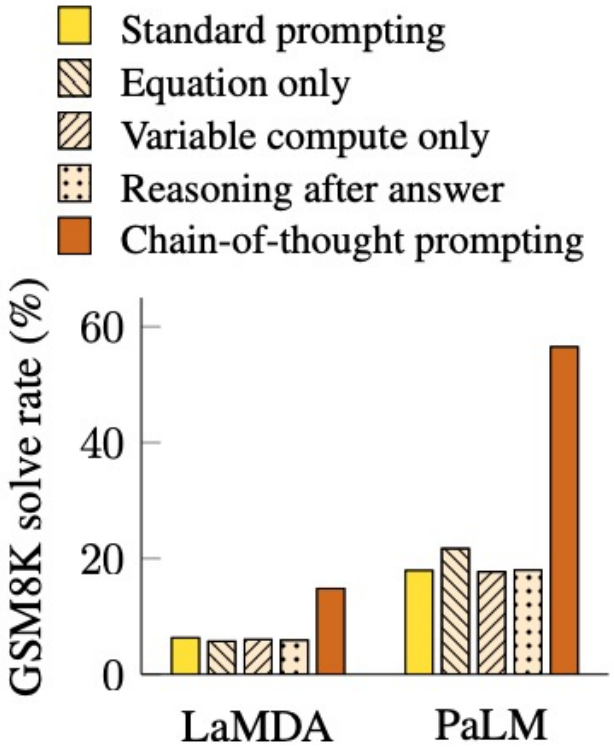
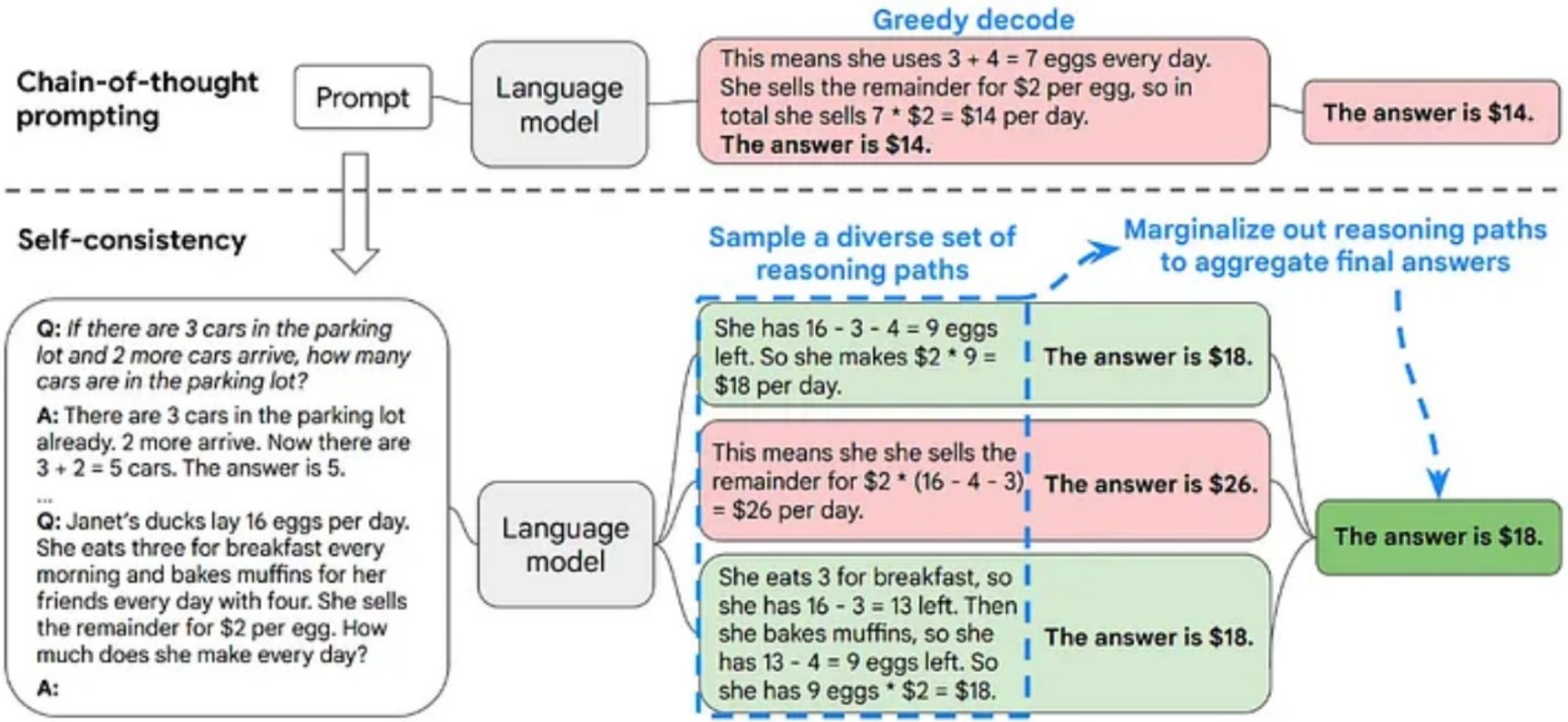


Figure 5: Ablation study for different variations of prompting using LaMDA 137B and PaLM 540B. Results for other datasets are given in Appendix Table 6 and Table 7.

CoT with Self-Consistency



Credit: https://cameronrwolfe.substack.com/p/chain-of-thought-prompting-for-llms

Reasoning by Searching

Source: Pan, Ji et al. (2025) A Survey of Slow Thinking-based Reasoning LLMs using Reinforced Learning and Inference-time Scaling Law

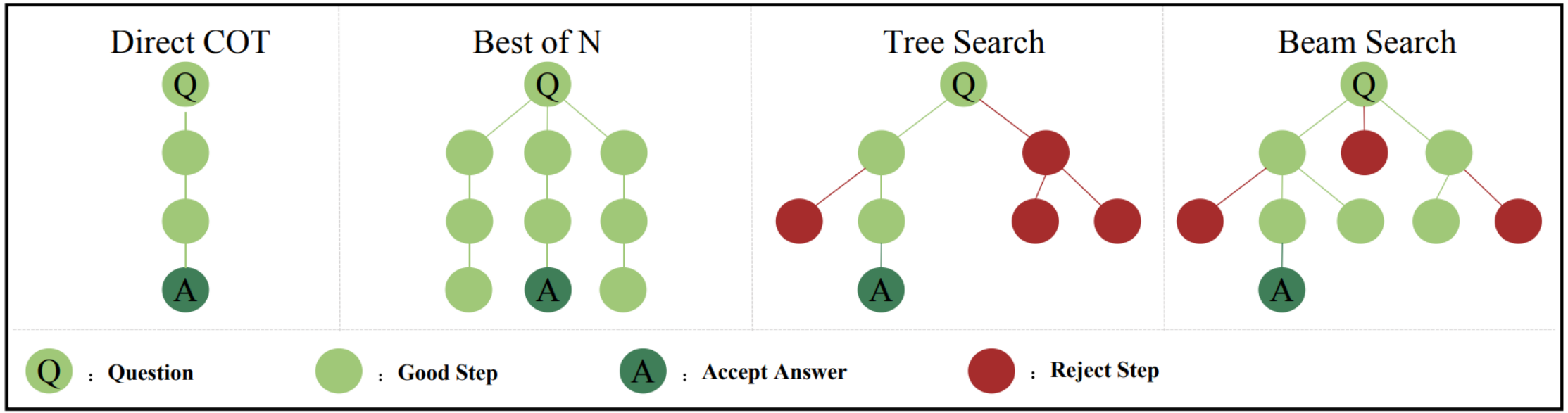


Fig. 3. The search algorithms for test-time scaling

Learning to Reason

Source: Pan, Ji et al. (2025) A Survey of Slow Thinking-based Reasoning LLMs using Reinforced Learning and Inference-time Scaling Law

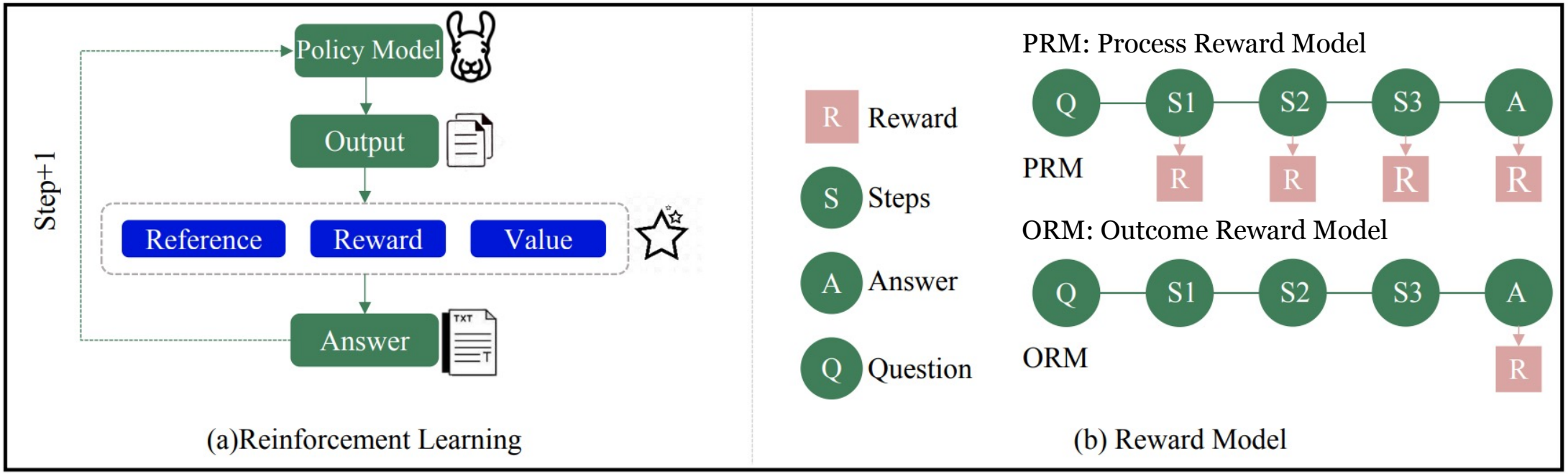
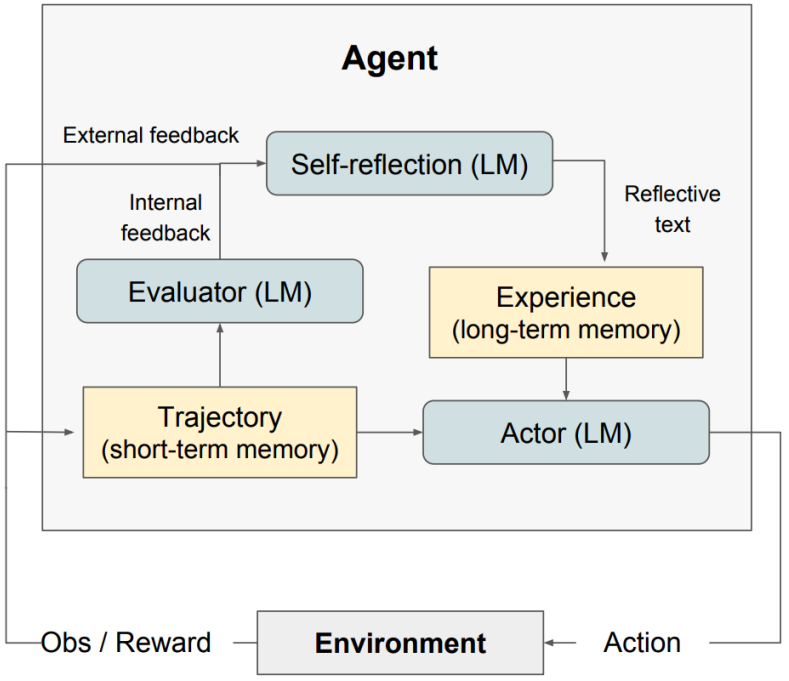


Fig. 4. The reinforcement learning framework and reward model

Reflexion: Verbalized Reinforcement Learning

Source: Shinn, Cassano et al. (2023) Reflexion: Language Agents with Verbal Reinforcement Learning



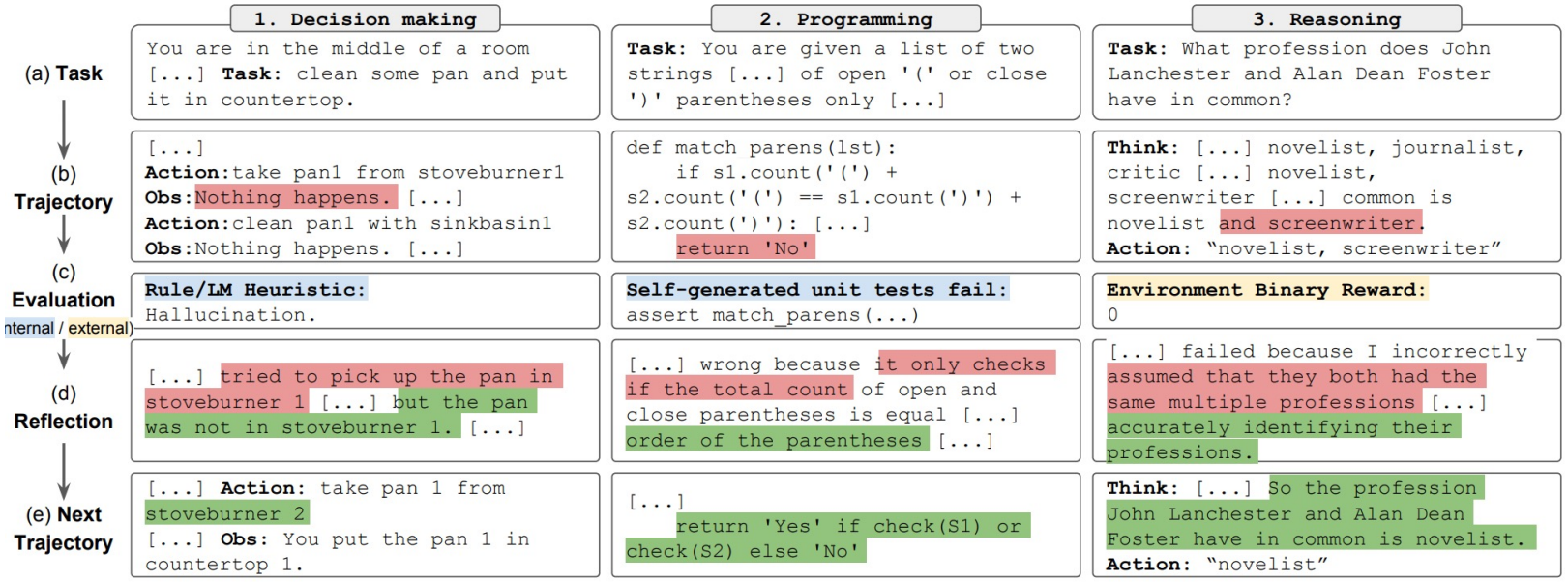


Figure 1: Reflexion works on decision-making 4.1, programming 4.3, and reasoning 4.2 tasks.

Improved Reasoning by Self-Reflection

Source: Shinn, Cassano et al. (2023) Reflexion: Language Agents with Verbal Reinforcement Learning

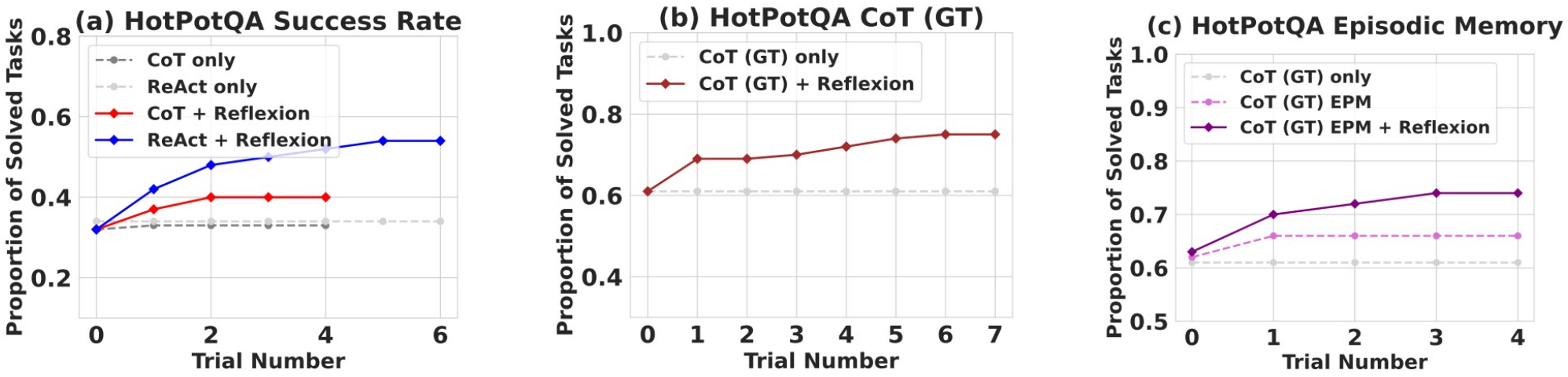


Figure 4: Chain-of-Thought (CoT) and ReAct. Reflexion improves search, information retrieval, and reasoning capabilities on 100 HotPotQA questions. (a) Reflexion ReAct vs Reflexion CoT (b) Reflexion CoT (GT) for reasoning only (c) Reflexion vs episodic memory ablation.

Inference Time Reasoning

Source: Ke, Jiao et al. (2025) A Survey of Frontiers in LLM Reasoning