Self-Evaluation Guided Beam Search for Reasoning

Yuxi Xie[1*], Kenji Kawaguchi[1], Yiran Zhao[1], James Xu Zhao[1],
Min-Yen Kan[1#], Junxian He[2#], Michael Qizhe Xie[1#],

- 1. National University of Singapore
- 2. Hong Kong University of Science and Technology

[*]Correspondence [#]Equal Advising

Speaker: Yuxi Xie

twitter: @sigrid_xie

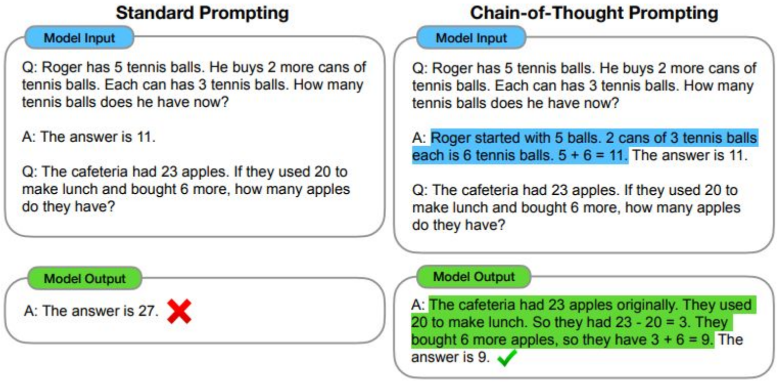
Recent Breakthroughs in Large Language Model Reasoning

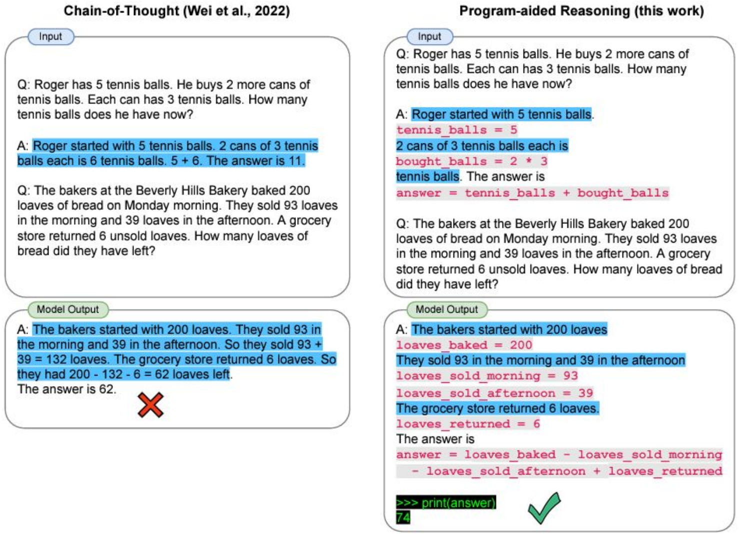
Breaking down a problem into intermediate steps facilitates reasoning

– Various prompting approaches have been proposed to define the intermediate reasoning chains, such as *chain-of-thought* (CoT), *program-aided language models* (PAL)

$$P(a \mid x) = \mathbb{E}_{R \sim P(R \mid x)} P(a \mid \boxed{R}, x)$$

decompose the answer generation process into a reasoning chain





Challenge in Multi-Step Reasoning

LLMs struggle with error accumulation across multiple steps

factorize the reasoning process in an autoregressive manner

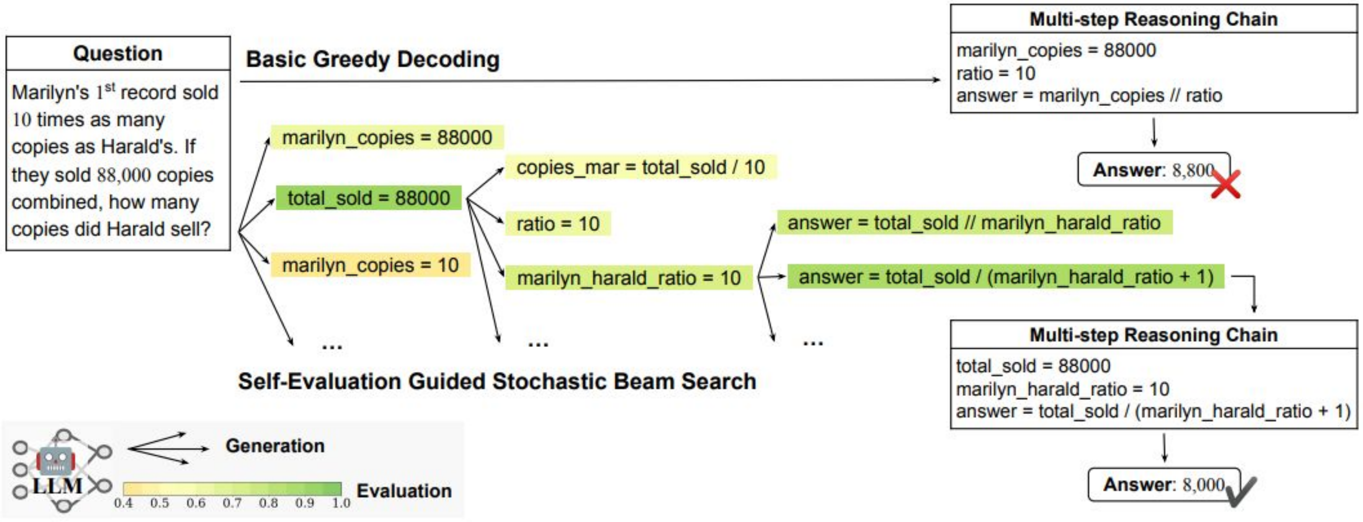
$$P(R = s^{1:T} \mid x) = \prod_t P(s^t \mid x, s^{1:t-1})$$

– As the complexity and length of reasoning chains increase with the difficulty of tasks, LLMs struggle with errors and imperfections that accumulate across multiple intermediate steps

– The growing number of steps leads to an exponential growth in the search space, making it exceedingly difficult to obtain accurate final outcomes

Leverage LLM Self-Evaluation to Guide Reasoning

We integrate stepwise self-evaluation to guide the reasoning process



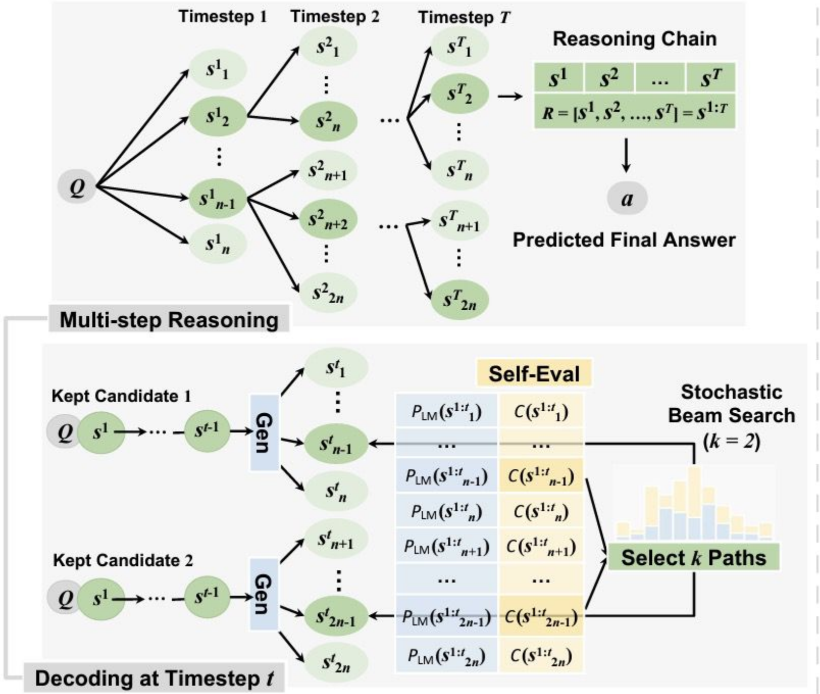
We propose Self-Evaluation Guided Stochastic Beam Search, a framework of stepwise reasoning.

– Stochastic beam search balances exploitation and exploration with sampling temperatures.

– Self-Evaluation can calibrate the decoding direction step by step.

Stochastic Beam Search

Stochastic beam search balances exploitation and exploration with sampling temperatures



We can set annealing sampling temperatures through reasoning – to have higher diversity for the initial steps and focus more on the quality at the end.

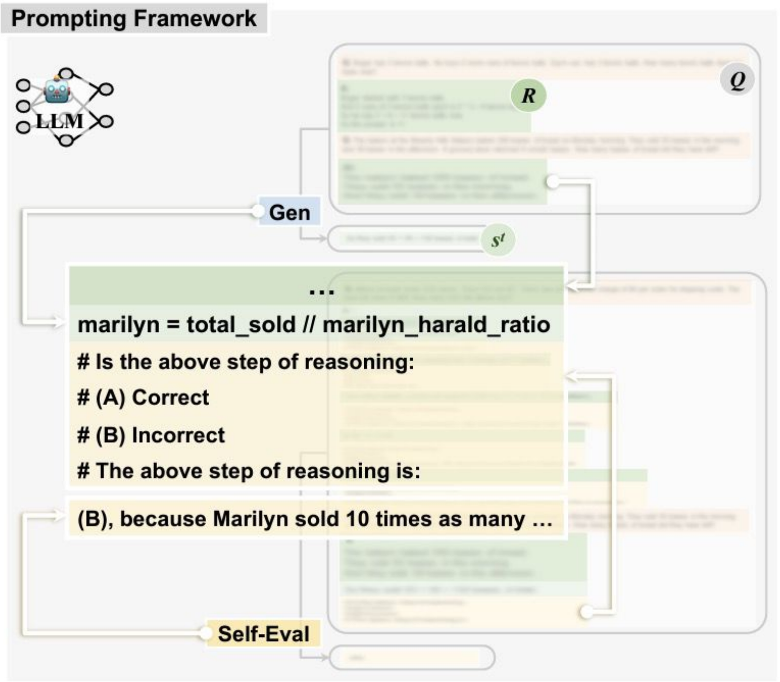
With the beam obtained through reasoning, our approach can inherently be integrated with majority voting (on the result beam) to elicit better final results.

Self-Evaluation Score

Self-Evaluation can calibrate the decoding direction step by step

We combine the LM probability and the correctness confidence as a calibrated objective for our controlled decoding:

$$\mathcal{E}(s^{1:T}) = \prod_t P^{\lambda}_{\mathrm{LM}_G}(s^t \mid x, s^{1:t-1})\, \mathcal{C}^{1-\lambda}(s^t)$$



Experimental Results

Self-Evaluation Guided Beam Search can outperform Self-Consistency (of equal cost), especially on multi-step reasoning where the reasoning chain is particularly lengthy

Table 1: Result comparison (accuracy %) on arithmetic and symbolic reasoning tasks. The best result is in **bold** and the lowest cost is in green. We report methods all with Codex backbone for a fair comparison. Similar to Huang et al. (2022), Diverse (Li et al., 2022) fine-tune task-specific verifiers to apply weights on samples in self-consistency (SC). Other fine-tuning methods include reward-based supervision (Uesato et al., 2022) and content-specific training (Lewkowycz et al., 2022). We also report the number of tokens (# Tokens) on GSM8K to compare the costs of different methods.

Approach	Arithmetic						Symbolic	
	GSM8K	# Tokens	AQuA	SVAMP	ASDiv	TabMWP	DATE	OBJECT
single reasoning chain								
CoT	65.6	0.2k	45.3	74.8	76.9	65.2	64.8	73.0
PoT	71.6	–	54.1	85.2	–	73.2	–	–
PAL	72.0	0.3k	–	79.4	79.6	–	76.2	96.7
Ours-PAL	80.2	27.7k	55.9	89.6	84.9	79.1	**78.6**	**96.8**
multiple reasoning chains								
CoT, SC	78.0	5.3k	52.0	86.8	–	75.4	–	–
CoT, Diverse	82.3	–	–	87.0	**88.7**	–	–	–
PoT, SC	80.0	–	58.6	89.1	–	**81.8**	–	–
PAL, SC	80.4	7.4k	–	–	–	–	–	–
Ours-PAL	**85.5**	550.0k	**64.2**	**90.3**	85.8	80.9	–	–

computational cost overhead

Table 3: Cost (# Tokens) and result (accuracy %) comparison on arithmetic and commonsense reasoning tasks. We base our experiments on Llama-2 (13B) since Codex is not available. We show the results of the baseline and our method both in the multiple-chain scenario for a fair comparison. Here we use PAL and CoT prompting for arithmetic and commonsense reasoning, respectively.

Approach	Arithmetic (PAL)					Commonsense (CoT)	
	GSM8K	AQuA	SVAMP	ASDiv	TabMWP	StrategyQA	CommonsenseQA
Baseline	41.8	30.7	71.2	66.2	43.7	**71.0**	**74.4**
# Tokens	13.9k	6.6k	5.9k	2.7k	1.9k	2.7k	1.2k
Ours	**46.1**	**31.5**	**74.6**	**67.7**	**49.6**	70.6	74.0
# Tokens	12.6k	6.0k	5.0k	2.5k	1.2k	2.6k	1.2k

shorter reasoning chains but even higher uncertainty

Table 4: Absolute accuracy (in %) increases on instances of different complexity determined by the length of reasoning chains (represented as # Steps).

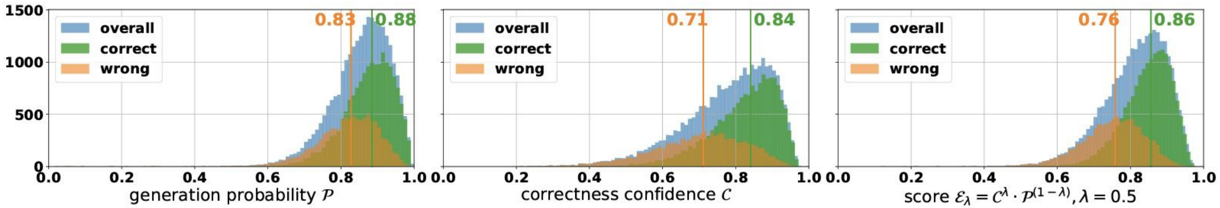
GSM8K				
# Steps	# Ins.	PAL	Ours	ΔAccu.
< 7	437	85.8	91.3	+5.49
$\in (7, 9]$	524	74.8	82.6	+7.82
≥ 9	358	72.9	82.6	+9.78

StrategyQA				
# Steps	# Ins.	CoT	Ours	ΔAccu.
< 4	637	84.6	84.9	+0.31
$\in [4, 5)$	1,301	78.6	79.1	+0.46
≥ 5	351	68.4	71.8	+3.42

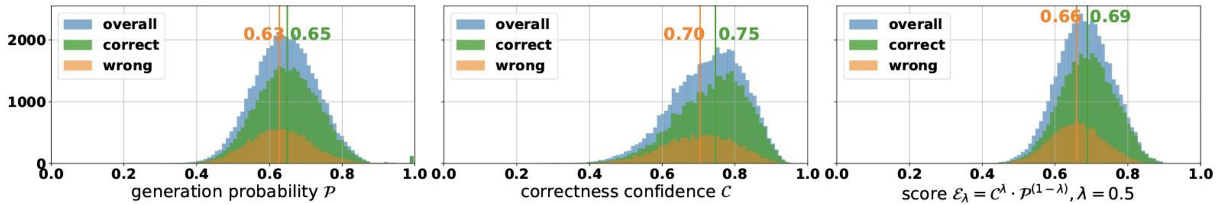
performance gain mainly comes f longer reasoning chains

Analysis: Self-Evaluation Score

Our Self-Evaluation Score can better determine the correctness of reasoning steps, especially on arithmetic reasoning tasks.



(a) Score distribution of PAL baseline predictions on GSM8K.

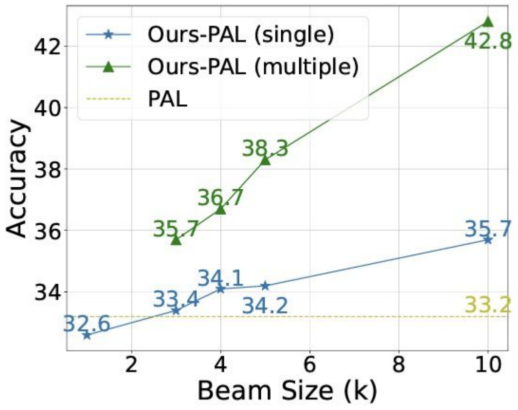


(b) Score distribution of CoT baseline predictions on StrategyQA.

Figure 5: Distributions of the self-evaluation score and its components (*i.e.*, generation confidence $\mathcal{P}$ and correctness confidence $\mathcal{C}$) on correct/incorrect baseline predictions. We highlight the median scores of the positive and negative cases using lines of the same colors respectively.

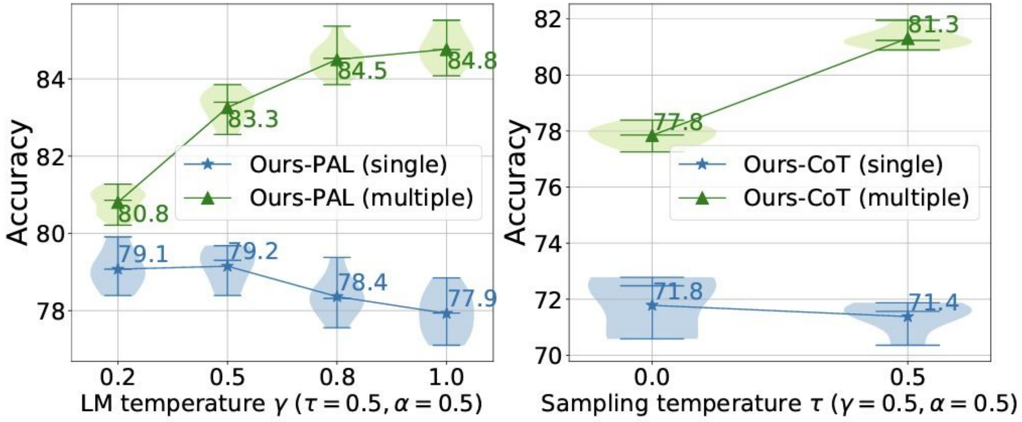
Analysis: Hyperparameters in Stochastic Beam Search

Our Beam Search algorithm can inherently be integrated with majority voting to achieve better performance.
Our step-specific sampling temperatures enable flexible control to balance the diversity and quality throughout the reasoning process.



(a) effect of beam size

performance vs. cost (beam size)



(b) effect of generation and sampling diversity

diversity vs. (single-chain) quality

Conclusion

To tackle the challenge of uncertainty in multi-step reasoning, we introduce a stepwise self-evaluation mechanism to guide and calibrate the reasoning process of LLMs. We propose a decoding algorithm integrating the self-evaluation guidance via stochastic beam search.

- Our beam search algorithm inherently enable majority voting on the result beam, leading to better performance compared with the self-consistency baseline of equal computational cost.
- Self-Evaluation demonstrates an efficient way to calibrate Generation.

However, model performance is constrained by the accessible search space within its own knowledge. Future works may explore more about how to integrate external feedback (e.g., tools, humans) for better guidance and calibration.

NEURAL INFORMATION PROCESSING SYSTEMS

Thank you

NUS National University of Singapore

HKUST

twitter: @sigrid_xie