

CAN SPECULATIVE SAMPLING ACCELERATE REACT WITHOUT COMPROMISING REASONING QUALITY?

Han Xu, Jingyang Ye, Yutong Li
University of Illinois Urbana-Champaign
{hanxu8, ye30, yutong23}@illinois.edu

Haipeng Chen
William & Mary
{hchen23}@wm.edu

ABSTRACT

Large language models (LLMs) are increasingly used as agents for interaction with external environments. These interplays are commonly facilitated through various prompting paradigms. However, such paradigms require extended interaction traces between the LLMs and the environment, resulting in low task-solving efficiency. In this work, we integrate speculative sampling (SpS) into the novel ReAct paradigm. In particular, we investigate speculative sampling’s impact on the efficiency of ReAct and the quality of reasoning tasks. Our evaluations using HotPotQA and FEVER datasets demonstrate that implementing speculative sampling alongside ReAct results in a 2.18x-2.62x acceleration compared to using ReAct alone, while only introducing a negligible impact on the reasoning abilities¹.

1 INTRODUCTION

ReAct (Yao et al., 2023) has demonstrated exceptional performance across various tasks, including question answering (Yang et al., 2018), fact verification (Thorne et al., 2018), text game (Shridhar et al., 2021), and webpage navigation (Yao et al., 2022). Despite these successes, the prompt-based paradigm, which synergizes reasoning and acting in language models, leads to prolonged task-solving traces. This characteristic ultimately compromises the efficiency of solving tasks.

Recently, speculative sampling (SpS) has been introduced to accelerate the inference process of a target language model M_p , with an approximate model M_q (Leviathan et al., 2023; Chen et al., 2023). This method involves three steps: (1) the more efficient model M_q generates approximations for k tokens, where $k \in \mathbb{Z}^+$; (2) M_p evaluates all the guesses, accepting those that lead to an identical distribution; and (3) if all guesses are accepted, M_p samples an additional token from the adjusted distribution, or, in the event of any rejected guesses, M_p corrects the rejected token. This procedure ensures the generation of 1 to $k + 1$ tokens in one run. The acceleration is attributed to the lower computational and memory bandwidth costs of making these guesses, coupled with a high likelihood of their acceptance. For example, for a basic text generation task, SpS demonstrates a 2.19x increase in throughput and latency compared to the Llama2 baseline model.

Model	Throughput (token/sec)	Latency (sec/token)	Throughput/Latency Improvement
Llama2-70B	1.28	0.780	-
SpS (k=4)	2.80	0.356	2.19

Table 1: Comparison between Llama2-70B and SpS

While SpS has shown promise in accelerating translation and summarization tasks in the original literature, its impact on reasoning tasks remains unexplored. In this paper, we investigate whether SpS can expedite the ReAct paradigm without affecting reasoning quality. This involves incorporating SpS into ReAct (Figure 1) and evaluating its impact using the HotPotQA and FEVER datasets. Our findings reveal that SpS, using a combination of Llama2-7B and Llama2-70B models (Touvron et al., 2023), significantly speeds up the question-answering (QA) process within ReAct by an average factor of 2.18x-2.62x, without compromising the accuracy of the answers.

¹Code and data are available at <https://github.com/wmd3i/ReAct-SpS>

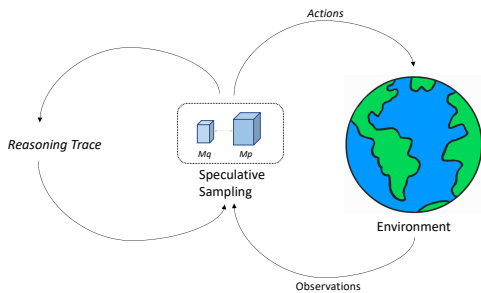


Figure 1: The integration of SpS within the ReAct paradigm

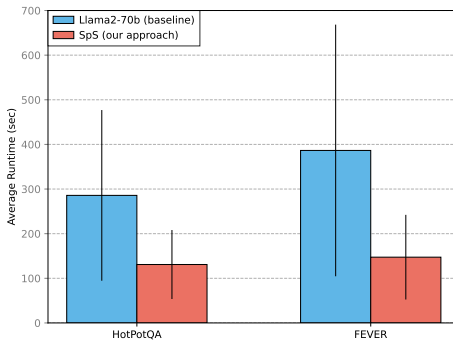


Figure 2: Average runtime of ReAct alone vs. ReAct with SpS (lower is better)

2 EXPERIMENTS AND RESULTS

2.1 EXPERIMENT SETUP

Dataset Following (Yao et al., 2023), we employ HotpotQA and FEVER to evaluate the reasoning capabilities. Specifically, we randomly select 150 questions from each dataset. The prompts used in the QA process comprise multiple thought-action-observation steps (Appendix A.1).

Baseline We combine SpS with ReAct using Llama2-7B (M_p) and Llama2-70B (M_q), and set $k = 4$ for a balance between speedup and acceptance rate. The Llama2-70B model is chosen as the baseline. All models are quantized with normalized float 4 and run on 8 NVIDIA A100 80G GPUs.

Measurements We first measure the throughput (the number of tokens generated per second) and latency (the time taken per token) by comparing the SpS setup against the baseline using short sentence prompting. Next, we evaluate the average speed and standard deviation of the duration for each ReAct QA process, as well as the accuracy rate of the answers. The correctness is determined by exact match (EM). Due to hardware constraints, we impose a 2048 token limit for the QA process, and mark a process as failed if the limit is reached. Additionally, we limit ReAct to 7 steps for each QA process, as further steps do not necessarily yield improved results.

2.2 RESULTS AND OBSERVATIONS

Consistent with the 2.19x increase in speed outlined in Table 1, SpS also significantly enhances the QA process. Here, we observe a considerable acceleration in the average QA duration – 2.18x for HotPotQA and 2.62x for FEVER. Additionally, there is a notable reduction in the standard deviation by a factor of 2.47x for HotPotQA and 2.97x for FEVER (Figure 2). These improvements collectively contribute to more stable and predictable QA durations.

As Appendix A.2 shows, both methods achieve the same accuracy - 0.25 for HotPotQA and 0.53 for FEVER. Despite the uniformity in FEVER evaluations, a close examination of HotPotQA traces reveals that the only differences occur when handling Q7 and Q14. As detailed in Appendix A.3, the ReAct method correctly answers Q7, while SpS does not. Conversely, SpS successfully answers Q14, which the ReAct method fails to do.

3 CONCLUSION

By integrating SpS into the ReAct paradigm, we achieve a significant acceleration in QA processing, ranging from 2.18x to 2.62x. Despite this efficiency gain, we observe only a minor impact on reasoning quality, suggesting a promising balance between speed and accuracy in LLM performance. Our approach is model-agnostic, and as future work, we plan to extend to other prompting paradigms such as Chain-of-Thought (Wei et al., 2022), Reflexion (Shinn et al., 2023) and ComposerX (Deng et al., 2024). Another avenue of research will explore incorporating mechanisms like PID control (Chen et al., 2024) to improve the robustness of SpS.

URM STATEMENT

All authors of this work meet the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

- Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling, 2023.
- Zhuotong Chen, Zihu Wang, Yifan Yang, Qianxiao Li, and Zheng Zhang. PID control-based self-healing to improve the robustness of large language models. *Transactions on Machine Learning Research*, 2024.
- Qixin Deng, Qikai Yang, Ruibin Yuan, Yipeng Huang, Yi Wang, Xubo Liu, Zeyue Tian, Jiahao Pan, Ge Zhang, Hanfeng Lin, Yizhi Li, Yinghao Ma, Jie Fu, Chenghua Lin, Emmanouil Benetos, Wenwu Wang, Guangyu Xia, Wei Xue, and Yike Guo. Composerx: Multi-agent symbolic music composition with llms, 2024.
- Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org, 2023.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Aleworld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757. Curran Associates, Inc., 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

A APPENDIX

A.1 PROMPTS

HotpotQA Prompts

Original	
Question	Musician and satirist Allie Goertz wrote a song about the “The Simpsons” character Milhouse, who Matt Groening named after who?
Answer	Richard Nixon
Question	Which magazine was started first Arthur’s Magazine or First for Women?
Answer	Arthur’s Magazine
ReAct	
Question	Musician and satirist Allie Goertz wrote a song about the “The Simpsons” character Milhouse, who Matt Groening named after who?
Thought 1	The question simplifies to “The Simpsons” character Milhouse is named after who. I only need to search Milhouse and find who it is named after.
Action 1	Search[Milhouse]
Observation 1	Milhouse Mussolini Van Houten is a recurring character in the Fox animated television series The Simpsons voiced by Pamela Hayden and created by Matt Groening.
Thought 2	The paragraph does not tell who Milhouse is named after, maybe I can look up “named after”.
Action 2	Lookup[named after]
Observation 2	(Result 1 / 1) Milhouse was named after U.S. president Richard Nixon, whose middle name was Milhous.
Thought 3	Milhouse was named after U.S. president Richard Nixon, so the answer is Richard Nixon.
Action 3	Finish[Richard Nixon]
Question	Which magazine was started first Arthur’s Magazine or First for Women?
Thought 1	I need to search Arthur’s Magazine and First for Women, and find which was started first.
Action 1	Search[Arthur’s Magazine]
Observation 1	Arthur’s Magazine (1844-1846) was an American literary periodical published in Philadelphia in the 19th century.
Thought 2	Arthur’s Magazine was started in 1844. I need to search First for Women next.
Action 2	Search[First for Women]
Observation 2	First for Women is a woman’s magazine published by Bauer Media Group in the USA.[1] The magazine was started in 1989.
Thought 3	First for Women was started in 1989. 1844 (Arthur’s Magazine) ; 1989 (First for Women), so Arthur’s Magazine was started first.
Action 3	Finish[Arthur’s Magazine]

FEVER Prompts

Original	
Claim	Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.

FEVER Prompts - Continued from the previous page	
Label	SUPPORTS
Claim	Stranger Things is set in Bloomington, Indiana.
Label	REFUTES
ReAct	
Claim	Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.
Thought 1	I need to search Nikolaj Coster-Waldau and find if he has worked with the Fox Broadcasting Company.
Action 1	Search[Nikolaj Coster-Waldau]
Observation 1	Nikolaj William Coster-Waldau (born 27 July 1970) is a Danish actor and producer. He graduated from the Danish National School of Performing Arts in Copenhagen in 1993,[1] and had his breakthrough role in Denmark with the film Nightwatch (1994). He played Jaime Lannister in the HBO fantasy drama series Game of Thrones, for which he received two Primetime Emmy Award nominations for Outstanding Supporting Actor in a Drama Series.. Coster-Waldau has appeared in numerous films in his native Denmark and Scandinavia, including Headhunters (2011) and A Thousand Times Good Night (2013). In the U.S, his debut film role was in the war film Black Hawk Down (2001), playing Medal of Honor recipient Gary Gordon.[2] He then played a detective in the short-lived Fox television series New Amsterdam (2008), and appeared in the 2009 Fox television film Virtuality, originally intended as a pilot.
Thought 2	Because he “appeared in the 2009 Fox television film Virtuality”, he should have worked with the Fox Broadcasting Company.
Action 2	Finish[SUPPORTS]
Claim	Stranger Things is set in Bloomington, Indiana.
Thought 1	I should search for Stranger Things, and see if it is set in Bloomington, Indiana.
Action 1	Search[Stranger Things]
Observation 1	Stranger Things is an American science fiction horror drama television series created by the Duffer Brothers. Set in the 1980s, primarily in the fictional town of Hawkins, Indiana, the series centers on a number of mysteries and supernatural events occurring around the town and their impact on an ensemble of child and adult characters.
Thought 2	The observation says that it is set in a “fictional town of Hawkins, Indiana”, so it is not set in Bloomington.
Action 2	Finish[REFUTES]

A.2 QA ACCURACY RATE

Method	HotPotQA	FEVER
ReAct	38/150 (0.25)	80/150 (0.53)
ReAct (SpS, k=4)	38/150 (0.25)	80/150 (0.53)

Table 2: Proportion of correctly answered questions from HotPotQA and FEVER

A.3 BASELINE AND SPS HOTPOTQA TRACES

Question 7 Traces	
Llama70B (baseline)	
Question	Jaelyn Stapp is married to the former frontman of a band that disbanded in what year?
Thought 1	I need to search Jaelyn Stapp and find who she is married to, and when the band disbanded.
Action 1	Search[Jaelyn Stapp]

Question 7 Traces - Continued from the previous page	
Observation 1	Jaclyn Nesheiwat Stapp (born July 29, 1980) is a beauty queen and fashion model. She is married to Scott Stapp, the singer of the band Creed and current solo touring artist. Her titles include Mrs. Florida America 2008, Miss New York USA 2004. In 2010 she wrote a children’s book, Wacky Jacky: The True Story of an Unlikely Beauty Queen. She is executive director of The Scott Stapp With Arms Wide Open Foundation, and founder of CHARM (Children Are Magical) by Jaclyn Stapp, which was formed to raise awareness of issues with which children deal and to provide underprivileged youth help with education.
Thought 2	Jaclyn Stapp is married to Scott Stapp, the singer of the band Creed. I need to search Creed and find when they disbanded.
Action 2	Search[Creed]
Observation 2	Could not find [Creed]. Similar: [‘Creed’, ‘Assassin’s Creed’, ‘Nicene Creed’, ‘Apostles’ Creed’, ‘Creed (band)’, ‘Creed (film)’, ‘Creed Brothers’, ‘Creed III’, ‘Assassin’s Creed (video game)’, ‘Assassin’s Creed Syndicate’]
Thought 3	Creed (band) was an American rock band formed in 1995 and disbanded in 2004.
Action 3	Finish[2004]
SpS (our approach)	
Question	Jaclyn Stapp is married to the former frontman of a band that disbanded in what year?
Thought 1	I need to search Jaclyn Stapp and find who she is married to, and when the band disbanded.
Action 1	Search[Jaclyn Stapp]
Observation 1	Jaclyn Nesheiwat Stapp (born July 29, 1980) is a beauty queen and fashion model. She is married to Scott Stapp, the singer of the band Creed and current solo touring artist. Her titles include Mrs. Florida America 2008, Miss New York USA 2004. In 2010 she wrote a children’s book, Wacky Jacky: The True Story of an Unlikely Beauty Queen. She is executive director of The Scott Stapp With Arms Wide Open Foundation, and founder of CHARM (Children Are Magical) by Jaclyn Stapp, which was formed to raise awareness of issues with which children deal and to provide underprivileged youth help with education.
Thought 2	Jaclyn Stapp is married to Scott Stapp, the singer of the band Creed. I need to find when Creed disbanded.
Action 2	Search[Creed disbanded]
Observation 2	Could not find [Creed disbanded]. Similar: [‘Creed (band)’, ‘Mark Tremonti’, ‘Scott Stapp’, ‘Creedence Clearwater Revival’, ‘Brett Hestla’, ‘Full Circle (Creed album)’, ‘Xavier Woods’, ‘List of Assassin’s Creed characters’, ‘Doug Clifford’, ‘Weathered’]
Thought 3	I cannot find when Creed disbanded, so I cannot answer the question.
Action 3	Finish[Unable to answer]

Note: The answer key is [2004]. The baseline correctly answers Q7 but SpS does not.

Question 14 Traces

Llama70B (baseline)	
Question	Lollipop Chainsaw featured Juliet Starling, who was voiced by a Canadian-American actress who has done voice roles for what Teen Titans spinoff series?
Thought 1	I need to search Juliet Starling and find who voiced her.
Action 1	Search[Juliet Starling]

Question 14 Traces - Continued from the previous page	
Observation 1	Could not find [Juliet Starling]. Similar: ['Lollipop Chainsaw', 'Jessica Nigri', 'Tara Strong', 'Juliet (disambiguation)', 'Highschool of the Dead', 'Chainsaws in popular culture', 'Tara Strong filmography', 'Starling (disambiguation)', 'List of cosplayers', 'List of Highschool of the Dead characters']
Thought 2	I cannot find Juliet Starling, maybe I can search the voice actress.
Action 2	Search[Tara Strong]
Observation 2	Tara Lyn Strong (née Charendoff; born February 12, 1973) is a Canadian and American actress. She is known for her voice work in animation, websites, and video games. Strong's voice roles include animated series such as The New Batman Adventures, Teen Titans, Teen Titans Go!, Rugrats, The Powerpuff Girls, The Fairly OddParents, The Proud Family, Xiaolin Showdown, Ben 10, Chowder, Wow! Wow! Wubbzy!, My Little Pony: Friendship Is Magic, Unikitty!, and DC Super Hero Girls. She has also voiced characters in the video games Mortal Kombat X, Ultimate Marvel vs. Capcom 3, Jak and Daxter, Final Fantasy X, X-2, Blue Dragon, and Batman: Arkham. She has earned Annie Award and Daytime Emmy nominations and won an award from the Academy of Interactive Arts & Sciences.
Thought 3	Tara Strong has done voice roles for Teen Titans and Teen Titans Go!.
Action 3	Finish[Teen Titans]
SpS (our approach)	
Question	Lollipop Chainsaw featured Juliet Starling, who was voiced by a Canadian-American actress who has done voice roles for what Teen Titans spinoff series?
Thought 1	I need to search Juliet Starling and find out who voiced her.
Action 1	Search[Juliet Starling]
Observation 1	Could not find [Juliet Starling]. Similar: ['Lollipop Chainsaw', 'Jessica Nigri', 'Tara Strong', 'Highschool of the Dead', 'Juliet (disambiguation)', 'Tara Strong filmography', 'Chainsaws in popular culture', 'Starling (disambiguation)', 'List of cosplayers', 'List of Highschool of the Dead characters']
Thought 2	I cannot find Juliet Starling, maybe I can search the voice actress.
Action 2	Search[Tara Strong]
Observation 2	Tara Lyn Strong (née Charendoff; born February 12, 1973) is a Canadian and American actress. She is known for her voice work in animation, websites, and video games. Strong's voice roles include animated series such as The New Batman Adventures, Teen Titans, Teen Titans Go!, Rugrats, The Powerpuff Girls, The Fairly OddParents, The Proud Family, Xiaolin Showdown, Ben 10, Chowder, Wow! Wow! Wubbzy!, My Little Pony: Friendship Is Magic, Unikitty!, and DC Super Hero Girls. She has also voiced characters in the video games Mortal Kombat X, Ultimate Marvel vs. Capcom 3, Jak and Daxter, Final Fantasy X, X-2, Blue Dragon, and Batman: Arkham. She has earned Annie Award and Daytime Emmy nominations and won an award from the Academy of Interactive Arts & Sciences.
Thought 3	She has earned Annie Award and Daytime Emmy nominations and won an award from the Academy of Interactive Arts & Sciences.
Action 3	Finish[Teen Titans Go!]

Note: The answer key is [Teen Titans Go!]. SpS correctly answers Q14 but the baseline does not.