A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners*

Bowen Jiang[1,2], Yangxinyu Xie[1,2], Zhuoqun Hao[1], Xiaomeng Wang[1], Tanwi Mallick[2], Weijie J. Su[1], Camillo J. Taylor[1], Dan Roth[1]

University of Pennsylvania[1] Philadelphia, PA, 19104, USA

Argonne National Laboratory[2] Lemont, IL, 60439, USA

{bwjiang@seas, xinyux@wharton, zhuoqunh@sas, xwang1@wharton}.upenn.edu,

tmallick@anl.gov, {suw@wharton, cjtaylor@seas, danroth@seas}.upenn.edu

1 Introduction

Large language models (LLMs) have achieved remarkable progress in understanding and generating human-like text, triggering growing interest in their reasoning capabilities and response rationality (Lyu et al., 2023; Xu et al., 2024; Jiang et al., 2024; Zhang et al., 2024; Yang et al., 2024; Chen et al., 2024; Jin et al., 2024; Wu et al., 2023; Liang et al., 2022; McCoy et al., 2023). There exists a variety of evaluation benchmarks, focusing different reasoning topics like arithmetic (Patel et al., 2021; Mishra et al., 2022), commonsense (Geva et al., 2021; Bisk et al., 2020), and logical problems (Han et al., 2022; Morishita et al., 2023).

However, existing works emphasize the overall accuracy of LLMs in performing benchmark tasks, ignoring one thing hidden in the problem statement that could cause brittle generalization capabilities: the **token bias**. In this extended abstract, we define that an LLM is subject to token bias in a reasoning task if systematic changes to some or all tokens in the task descriptions - while keeping the underlying logic intact - lead to predictable shifts of the model's output. A strong token bias suggests that the model is relying on superficial patterns in the input rather than truly understanding the underlying reasoning task, making it fail to generalize well to novel examples and phrasings encountered in the wild that differ from the spurious patterns it has learned from the training data.

2 Examples of the Token Bias

It's likely that most LLMs have been trained to recognize classic examples that frequently appear in literature, but those examples usually have iconic narratives with protagonists having fixed names. As a result, the question remains whether they acquire genuine reasoning skills or merely learn to falsely associate frequently appearing names with the correct reasoning outcomes they should have.

The following example shows a token bias in the Linda Problem, i.e., the conjunction fallacy, in psychology (Tversky and Kahneman, 1983). We alter the name "Linda" to other names like "Luna" and rephrase the story telling, while maintaining the same logical structure.

The Linda Problem in Psychology

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations. Which is more probable?

(a) Linda is a bank teller.
(b) Linda is a bank teller and is active in the feminist movement.

→ Luna is 29 years old, married, deeply passionate about environmental conservation, and volunteers their weekends at local park clean-ups. They studied physics and applied math in college, and held several campaigns to reduce the campus's carbon footprint. Which is more probable?

(a) Luna is an assistant professor in aerospace engineering and is an active member of an environmental advocacy group.
(b) Luna is an assistant professor in aerospace engineering.

Similarly, we perturb the famous "Twenty-Five Horses" problem in graph theory by replacing "horses" with "bunnies", a change that shouldn't affect the logic. If a token bias toward "horses" exists, a systematic drop in the LLM's performance on the altered problem should be observed.

The Twenty-Five Horses Problem in Mathematics

You want to find the fastest 3 horses→ bunnies in a group of 25→ 36 horses→ bunnies. You can only race 5→ 6 horses→ bunnies at a time. You don't have a stopwatch, so you can only know the ranking of each horse→ bunny within each race. How many races do you need?

[0]This is an extended abstract of the full paper accepted at EMNLP 2024 https://arxiv.org/abs/2406.11050.

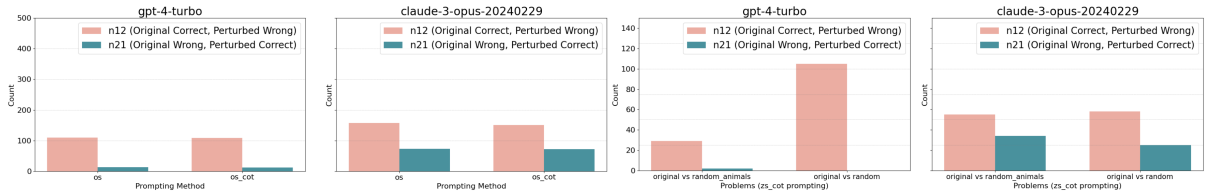

Figure 1: Hypothesis testing results ($n = 200$). In the left two sub-figures, we compare the original Linda problem with the one rephrased under a different name. "os_cot" and "os" mean one-shot learning with and without chain-of-thought (CoT) prompting "let's think step by step" (Wei et al., 2022). In the right two sub-figures, "original" means we query the LLM on the original "twenty-five horses" problem using zero-shot CoT prompting directly, "random_animals" means we perturb "horses" to another random animal name, and "random" means we perturb both the animal names and the numerical values "25" (and "5"). To reject the null, we expect $n12 > n21$.

3 Hypothesis Testing as an Evaluation Framework with Statistical Guarentee

We reconceptualize the evaluation of reasoning capabilities into a general and rigorous statistical testing framework beyond accuracy. Figure 2 shows that given a dataset of n problems with potential token biases, the process begins with perturbations trying to remove token biases. It generates n matched pairs, enabling us to create a contingency table for hypothesis testing. We use McNemar's Test (McNemar, 1947) with p-values < 0.05.

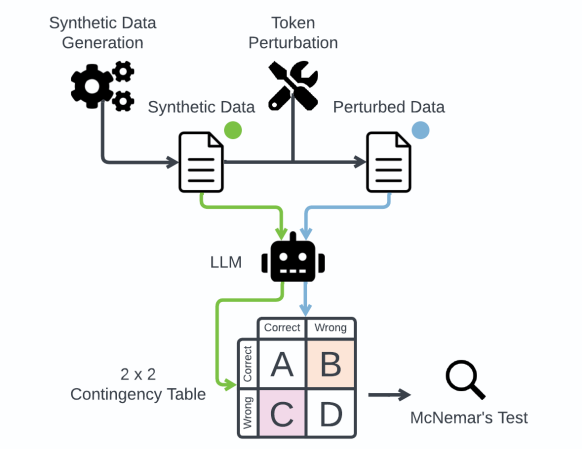

Figure 2: An illustration of the overall framework. We generate synthetic data, perform systematic token perturbations, and evaluate an LLM for comparative studies. The resulting contingency table, where A-D are integer values of counts, allows for subsequent statistical tests.

Hypothesis Testing and Results We use $\boldsymbol{H_0}$ to denote the null hypothesis and $\boldsymbol{H_a}$ the alternative hypothesis. $\boldsymbol{H_0}$ assumes LLMs have genuine reasoning capabilities, and we either accept or reject $\boldsymbol{H_0}$. We define π_{12} as the probability of answering the original problem correctly and the perturbed problem wrong, with π_{21} representing the reverse scenario. n_{12} and n_{21} are the respective counts.

Hypothesis 1 Genuine reasoning LLM should withstand surface-level alterations to the one-shot exemplar in the problem statements.

P is the original problem while P' is the perturbed problem with tokens perturbations irrelevant to the underlying logic.

$\boldsymbol{H_0}$: $\pi_{12} = \pi_{21}$.

$\boldsymbol{H_a}$: $\pi_{12} > \pi_{21}$ or $\pi_{12} < \pi_{21}$ (it depends).

An LLM might do well when presented with the original Linda problem or the twenty-five horses problem as a one-time example and asked to solve a similar problem. However, as shown in Figure 1, there are token biases related to specific words like "Linda", "25", and "horses". Substituting them with other logically equivalent words systematically degrades the LLM's performance, surprisingly, resulting in $\pi_{12} > \pi_{21}$. *Such changes should not influence outcomes for genuine reasoners, as those names are irrelevant to the logical process.*

In terms of accuracy, we find that changing "Linda" to other names in one-shot learning with CoT reduces accuracy from 95.0% to 24.0% on GPT-4 (Achiam et al., 2023) and from 40.5% to 30.0% on Claude-3-opus (Anthropic, 2024). For the "twenty-five horses" problem, replacing "horses" with other animals reduces 98.5% to 85.0% on GPT-4 and 40.5% to 30.0% on Claude-3-opus. Further changing "25" to other values decreases accuracy to 46.0% and 24.0%, respectively.

4 Discussions

LLMs do not consistently apply genuine reasoning in their decision-making process, but primarily rely on token bias for response generation. Therefore, any robust evaluation of the LLM's generalization should account for the fundamental impact of token bias hidden in the current benchmark problems.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. Models overview. Software available from Anthropic. Accessed: 2024-05-20.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Xinyun Chen, Ryan A Chi, Xuezhi Wang, and Denny Zhou. 2024. Premise order matters in reasoning with large language models. *arXiv preprint arXiv:2402.08939*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Weijie J Su, Camillo J Taylor, and Tanwi Mallick. 2024. Multi-modal and multi-agent systems meet rationality: A survey. *arXiv preprint arXiv:2406.00252*.

Mingyu Jin, Qinkai Yu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, Mengnan Du, et al. 2024. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. 2023. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, et al. 2022. Lila: A unified benchmark for mathematical reasoning. *arXiv preprint arXiv:2210.17517*.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2023. Learning deductive reasoning from synthetic corpus based on formal logic. In *International Conference on Machine Learning*, pages 25254–25274. PMLR.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.

Amos Tversky and Daniel Kahneman. 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4):293.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*.

Han Xu, Jingyang Ye, Yutong Li, and Haipeng Chen. 2024. Can speculative sampling accelerate react without compromising reasoning quality? In *The Second Tiny Papers Track at ICLR 2024*.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? *arXiv preprint arXiv:2402.16837*.

Yizhe Zhang, He Bai, Ruixiang Zhang, Jiatao Gu, Shuangfei Zhai, Josh Susskind, and Navdeep Jaitly. 2024. How far are we from intelligent visual deductive reasoning? *arXiv preprint arXiv:2403.04732*.