Side-Channel Attacks and Fairness Concerns in Speculative Sampling LLMs

Yinuo Zhao
[1] Department of Computer Science, Faculty of Arts and Science
University of Toronto

January 2024 - April 2024

Abstract: This report outlines the ongoing research conducted as part of CSC494H1 S(Winter), supervised by Professor Gururaj Saileshwar. The primary objective is to investigate the timing differentials resulting from speculative execution (including speculative sampling or decoding) in Large Language Models (LLMs), and to assess whether the information leaked through these differentials exacerbates existing LLM attacks. Furthermore, we aim to explore the potential fairness issues in servicing LLM requests among differently represented groups. Initial experiments have revealed a 4.63% difference in average token generation time between male and female controlled groups, but the significance of which remains uncertain pending further progress in this study. ***Keywords*:** *large language model, speculative decoding, side channel attacks, fairness, safety.*

1 Introduction

In recent years, there has been a surge in proposals to implement Large Language Models (LLMs) utilizing **Speculative Decoding** [1], also known as Speculative Sampling[2], as a means to expedite the processing of requests. This approach involves initial inference using smaller LLMs followed by low-cost verification with larger ones, promising faster response times.

While this methodology has demonstrated efficacy in accelerating LLM performance, its implications for fairness and safety remain relatively uncharted territory. Of growing concern is the potential vulnerability to side-channel attacks and fairness biases inherent in employing multiple models, which may amplify existing safety concerns and biases.

In this study, we aim to address this gap by examining the impact of speculative sampling on LLM performance through experimentation with existing implementations. Our findings reveal timing variations in translation tasks based on textual context, such as gender indicators, attributable to speculative sampling. This underscores the significance of speculative sampling in shaping the safety and fairness of LLM performance, raising concerns about the trade-offs between speed and integrity. Given the increasing popularity of speculative sampling and its consequential risks, there is a pressing need for advanced and tailored strategies to effectively mitigate these issues.

2 Literature review

The study of related work on this field consists of two part, one of which focuses on the speculative sampling algorithm and the other focuses on fairness and social biases in LLMs.

2.1 Speculative Decoding (Sampling)

SpecInfer [1] is an LLM serving system designed to enhance latency through speculative inference and token tree verification. It employs boost-tuned small language models (SSMs) to predict LLM outputs, which are then verified against the LLM using parallel decoding. Challenges such as large search spaces and verifying speculated tokens are addressed through collective boost-tuning and multi-step sampling.

SPEED [3] introduces a method to accelerate inference by predicting multiple future tokens alongside the current one, leveraging early-layer hidden states. This speculative execution enables parallelism, thus reducing memory constraints and expediting inference. Furthermore, it demonstrates that this approach maintains model accuracy and enables deeper decoder training with parameter sharing, all while incurring minimal runtime overhead.

CS Drafting [4] presents an algorithm involving multiple draft models, beginning with a statistical language model as the baseline. These draft models review content and propose it to larger models or the target, a process termed Vertical Cascade, effectively eliminating autoregression. The study finds that later tokens have lower acceptance probability, allowing for the utilization of smaller, faster models for these tokens, a concept termed Horizontal Cascade.

Other publications, such as those by Chen et al. [2] and Kim et al. [5], introduce similar ideas centered around speculative sampling. The core concept involves generating multiple tokens within a single transformer call. This approach leverages a faster but less accurate draft model to generate short continuations, matching the parallel scoring latency of sampling a single token from the larger target model. Additionally, it utilizes a novel rejection sampling scheme to maintain the target model's distribution within hardware limitations.

2.2 Fairness and Social Bias in LLMs

The issue of fairness in large language models (LLMs) has been extensively researched in the field. Numerous studies have evaluated LLMs across various scenarios and use cases, and researchers have introduced datasets tailored for investigating these issues.

Lie et al. (2023) [6] investigate the performance of LLMs in tabular tasks and their susceptibility to social biases. They find that LLMs inherit biases from their training data, which affect fairness in predictions. Despite attempts to mitigate bias, LLMs still exhibit larger fairness gaps compared to traditional models. Label-flipping of examples can reduce biases, indicating inherent bias within LLMs. Kotek et al. (2023) [7] examine how recent LLMs perform regarding gender stereotypes, showing that they often align with biased assumptions about occupations and gender. The study finds that LLMs tend to choose stereotypical occupations, reflecting societal perceptions rather than official statistics. Moreover, the models exacerbate biases and frequently overlook sentence ambiguities, providing inaccurate explanations for their choices. These findings underscore the need for careful testing and addressing biases in LLMs to ensure fair treatment of marginalized groups.

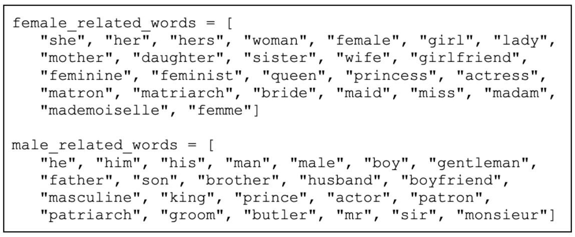
Tamkin et al. (2023) [8] present a dataset containing a diverse set of prompts covering 70 hypothetical decision scenarios, ranging from approving a loan to providing press credentials. Each prompt instructs the model to make a binary decision (yes/no) about a particular person described in the prompt. DELPHI [9] is a new dataset of controversial questions, built upon the Quora Question Pairs Dataset, to facilitate research in this area. The dataset addresses challenges such as knowledge recency, safety, fairness, and bias. The Crowdsourced Stereotype Pairs benchmark (CrowS-Pairs) [10] comprises 1508 examples covering stereotypes related to nine types of bias, such as race, religion, and age. In CrowS-Pairs, models are presented with pairs of sentences, one expressing a stereotype more strongly and the other less so. The focus is on stereotypes about historically disadvantaged groups compared to advantaged ones. StereoSet [11] is a comprehensive dataset designed to measure stereotypical biases in pretrained language models across domains like gender, profession, race, and religion. Unlike existing studies that use artificially constructed sentences, StereoSet is a large-scale natural dataset in English. These datasets provide valuable resources for investigating bias and fairness concerns in Large Language Models.

3 Experiment Setup

To address the research questions, we employed an existing open-source implementation of speculative coding LLMs models and curated datasets. The following sections detail the models, datasets, and evaluation metrics utilized in our experiments.

3.1 Models

We utilized the Big Little Decoder [5], an open-source implementation built on PyTorch and the Hugging Face Transformers library. Specifically, we employed the mT5-large and mT5-small models [12] for conducting machine translation tasks. These models vary in size by approximately a factor of 20. Our experiments were configured with a rollback value of 2 and a fallback value of 1, with further details provided in the associated literature.



```
female_related_words = [
   "she", "her", "hers", "woman", "female", "girl", "lady",
   "mother", "daughter", "sister", "wife", "girlfriend",
   "feminine", "feminist", "queen", "princess", "actress",
   "matron", "matriarch", "bride", "maid", "miss", "madam",
   "mademoiselle", "femme"]

male_related_words = [
   "he", "him", "his", "man", "male", "boy", "gentleman",
   "father", "son", "brother", "husband", "boyfriend",
   "masculine", "king", "prince", "actor", "patron",
   "patriarch", "groom", "butler", "mr", "sir", "monsieur"]
```

Figure 1: Filtering Keywords.

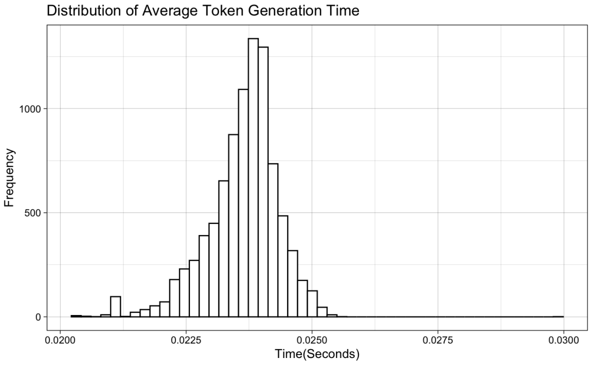


Figure 2: Average token generation time of all validating and testing samples, no filtering, with outlier.

3.2 Datasets

The raw dataset employed in our experiments is the IWSLT 2017 De-En dataset [13], containing text in both English and German. To investigate gender bias issues, we performed additional filtering on the testing data. Given the limited utility of other Natural Language Processing methods such as LDA, we opted to split the data based on the presence of words related to specific genders. The keywords utilized for filtering are depicted in Figure 1.

3.3 Evaluation Metrics

We adopted the evaluation metrics from the Big Little Decoder [5], encompassing the following: BLEU (BiLingual Evaluation Understudy) [14] score, generation length, prediction loss, runtime, samples per second, and steps per second during testing. Additionally, we measured the average token generation time of the samples. Inference evaluation was conducted with a batch size of 1 on a single NVIDIA RTX A6000 GPU.

4 Results and Discussion

4.1 Token Generation Time

The average token generation across all validation and testing samples(8967 observations) exhibits a distribution that closely resembles a normal distribution, with minimal variation, as depicted in Figure 2. Each data point represents the average generation time of tokens within one sample. However, it's noteworthy that an outlier with a generation time of 0.0297 seconds exists, while the remaining data points all fall below 0.026 seconds. Consequently, Figure 3 displays the same plot after excluding this outlier. The resulting distribution appears unimodal with a slight left tail.

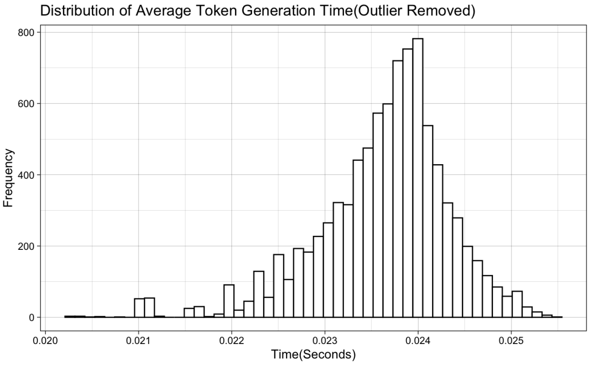


Figure 3: Average token generation time of all validating and testing samples, no filtering, without outlier.

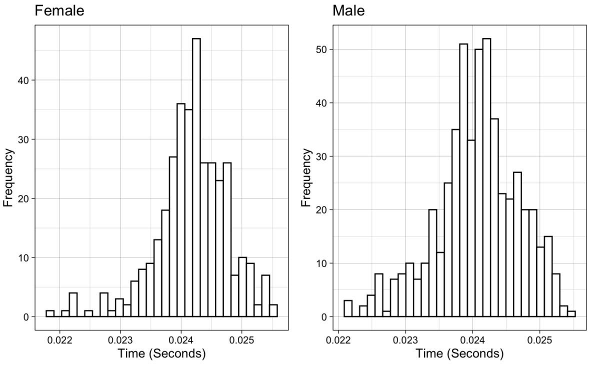


Figure 4: Average token generation time by gender.

The distribution of average token generation times be-

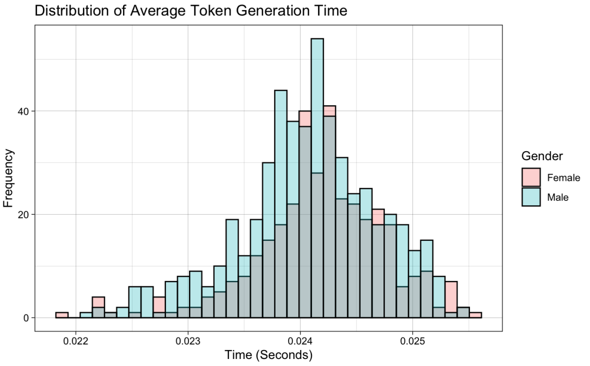


Figure 5: Combined histogram of average token generation time by gender.

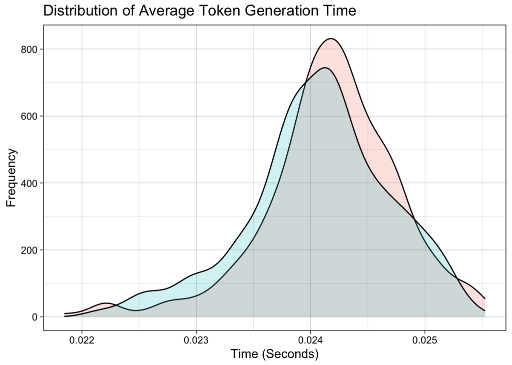


Figure 6: Density plot of average token generation time by gender.

tween male and female datasets exhibits largely identical shapes, as illustrated in Figure 4. However, there is a slight distinction in average token generation times, with the male text demonstrating slightly lower values compared to the female text, although this difference is minimal. Figure 5 provides a visual representation of this distinction, with the blue bars skewed towards the left, indicating shorter generation times for male text, while the red bars are shifted towards the right, suggesting slightly longer generation times for female text. For a clearer depiction of this observation, Figure 6 offers a refined illustration of the same graph.

On average, the time it takes to generate a token with the female dataset is 0.02418 seconds, compared to 0.02407 seconds for the male dataset. This reflects a difference of 4.63%.

4.2 Other Evaluation Metrics

In terms of translation quality, the BLEU score for the male data is 34.5083, whereas for the female data it is slightly higher at 34.7681, indicating a 0.75% improvement in translation quality for the female dataset. Additionally, the average generation length for the male data is 30.8061, compared to 30.4661 for the female data, resulting in a 1.11% difference, with male-generated text being slightly longer. Regarding prediction loss, the male data records a loss of 1.0629, while the female data incurs a slightly higher loss of 1.0986, indicating a 3.30% increase in loss for female translation.

However, in terms of speed, the female data outperforms the male data in both samples per second and steps per second, showing a 1.62% increase in speed. Although these differences are present, they remain relatively minor, all below 2%. This complexity makes it challenging to conclusively attribute the disparities to model bias rather than chance variation without further investigation using more sophisticated experimental designs. A full summary of the statistics is shown in Table 1.

	Male	Female	%
BLEU	34.5083	34.7681	0.75
gen_length	30.8061	30.4661	1.11
loss	1.0629	1.0986	3.30
runtime	0:07:06.00	0:04:47.03	N/A
samples	526	354	N/A
sample/second	1.235	1.233	1.62
steps/second	1.235	1.233	1.62

Table 1: Caption

5 Limitations and Future Works

The current experiment and its results are constrained by the data utilized for evaluating model performance. Despite efforts to filter data by gender using keyword searches, variations in the content of the text being translated across datasets may impact timing performance to varying degrees. For instance, sentences with more complex structures are generally harder to translate as opposed to simple sentences. A more robust approach would involve employing datasets in which translation texts are identical except for gender-indicating vocabularies. By controlling for additional factors, we can more accurately assess gender bias in large language models, mitigating potential confounders. Moreover, the filtered dataset has relatively small sample size which limits its reliability.

Another avenue for exploration involves examining timing differences in translation among languages with varying degrees of popularity. This approach offers insight into comparing more and less represented groups, providing a nuanced understanding of translation performance across diverse linguistic contexts. Furthermore, we can extend this analysis to investigate biases related to age, race, and nationality, leveraging appropriate datasets to discern patterns and disparities.

Moreover, addressing fairness and safety concerns may be better achieved through tasks such as question answering rather than translation. Question answering tasks yield richer outputs from large language models, offering clearer insights into their behavior compared to translation tasks, which involve relatively less cognitive processing and therefore provide less information on such issues.

6 Acknowledgements

This research project is under the supervision of Professor Gururaj Saileshwar in the Department of Computer Science, Faculty of Arts and Science at the University of Toronto, in collaboration with Sophia Abolore.

References

[1] X. Miao, G. Oliaro, Z. Zhang, X. Cheng, Z. Wang, Z. Zhang, R. Y. Y. Wong, A. Zhu, L. Yang, X. Shi, C. Shi, Z. Chen, D. Arfeen, R. Abhyankar, and Z. Jia, "Specinfer: Accelerating generative large language model serving with tree-based speculative inference and verification," 2024.

[2] C. Chen, S. Borgeaud, G. Irving, J.-B. Lespiau, L. Sifre, and J. Jumper, "Accelerating large language model decoding with speculative sampling," 2023.

[3] C. Hooper, S. Kim, H. Mohammadzadeh, H. Genc, K. Keutzer, A. Gholami, and S. Shao, "Speed: Speculative pipelined execution for efficient decoding," 2024.

[4] Z. Chen, X. Yang, J. Lin, C. Sun, K. C.-C. Chang, and J. Huang, "Cascade speculative drafting for even faster llm inference," 2024.

[5] S. Kim, K. Mangalam, S. Moon, J. Malik, M. W. Mahoney, A. Gholami, and K. Keutzer, "Speculative decoding with big little decoder," 2023.

[6] Y. Liu, S. Gautam, J. Ma, and H. Lakkaraju, "Investigating the fairness of large language models for predictions on tabular data," 2024. [Online]. Available: https://openreview.net/forum?id=6jJFmwAlen

[7] H. Kotek, R. Dockum, and D. Sun, "Gender bias and stereotypes in large language models," in *Proceedings of The ACM Collective Intelligence Conference*, ser. CI '23. ACM, Nov. 2023. [Online]. Available: http://dx.doi.org/10.1145/3582269.3615599

[8] A. Tamkin, A. Askell, L. Lovitt, E. Durmus, N. Joseph, S. Kravec, K. Nguyen, J. Kaplan, and D. Ganguli, "Evaluating and mitigating discrimination in language model decisions," 2023.

[9] D. Q. Sun, A. Abzaliev, H. Kotek, Z. Xiu, C. Klein, and J. D. Williams, "Delphi: Data for evaluating llms' performance in handling controversial issues," 2023.

[10] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, "CrowS-pairs: A challenge dataset for measuring social biases in masked language models," B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1953–1967. [Online]. Available: https://aclanthology.org/2020.emnlp-main.154

[11] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," 2020.

[12] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," 2021.

[13] M. Cettolo, M. Federico, L. Bentivogli, J. Niehues, S. Stüker, K. Sudoh, K. Yoshino, and C. Federmann, "Overview of the IWSLT 2017 evaluation campaign," S. Sakti and M. Utiyama, Eds. Tokyo, Japan: International Workshop on Spoken Language Translation, Dec. 14-15 2017, pp. 2–14. [Online]. Available: https://aclanthology.org/2017.iwslt-1.1

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

7 Artifact Appendix

The code base for running the analysis in this report is available at the GitHub repository here. To run the translation tasks, navigate to `src` and run the following command:

```
CUDA_VISIBLE_DEVICES=0 python3 \
run_bild_translation.py \
--model bild \
--small kssteven/mT5-small-wmt2014-de-en \
--large kssteven/mT5-large-wmt2014-de-en \
--dataset_name iwslt2017 \
--dataset_config iwslt2017-de-en \
--source_lang de --target_lang en \
--bild_rollback 2 --bild_fallback 1
```