LLMSteer: Improving Long-Context LLM Inference by Steering Attention on Reused Contexts

Zhuohan Gu, Jiayi Yao,
Kuntai Du, Junchen Jiang

THE UNIVERSITY OF CHICAGO

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction and Background

Large Language Models (LLMs) excel at tasks like question answering, summarization, and reasoning but require augmentation for domain-specific or user-specific knowledge.

- **Challenge:** Incorporating supplemental contexts (thousands of tokens) is challenging due to the *lost-in-the-middle problem* and high runtime costs.
- **Prefix Caching:** Systems reuse Key-Value (KV) caches for frequently used contexts to avoid repeated prefill processing, improving efficiency.
- **Persistent Limitations:** Despite prefix caching, models often lose track of critical information because KV pairs remain static.
- **We tackle just one question**: Can we achieve both high efficiency and high quality *without* fine-tuning models?

Previous Work

- **Attention Steering:** PASTA [1] and AutoPASTA [2] reweight highlighted input tokens. AutoPASTA automates this via iterative prompting.
- **Limitations:** AutoPASTA increases runtime (2 LLM calls) and breaks prefix caching, making KV caches non-reusable.

Our Approach

Insights

- Single-pass processing may overlook important tokens in the context.
- Multi-pass prompting with varied prefixes identifies tokens needing more attention through consistently high attention scores.

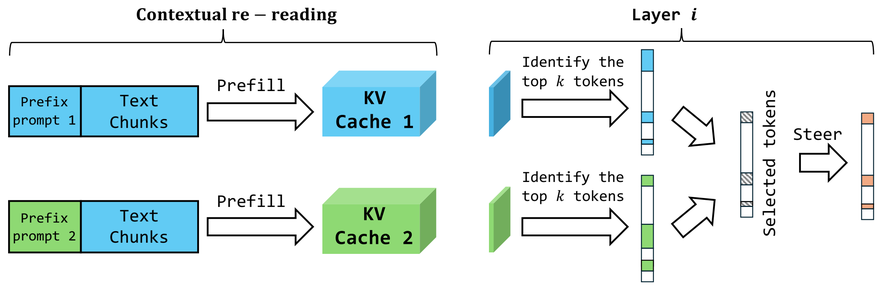


Figure 1: LLMSTEER processes the same context twice with different prefix prompts, generates KV caches, and reweights selected tokens.

Design: LLMSTEER consists of THREE steps

- **Contextual re-reading:** Process the context twice with different prefix prompts.
- **Token Selection:** Compute cumulative attention scores, select top-k tokens.
- **Steering**: Scale up the attention weight of selected tokens.

Example

GSM8K

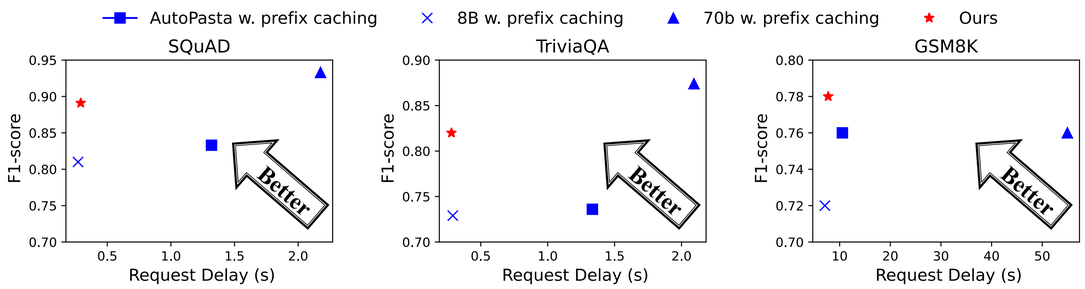
Q: A bag of chips has 250 calories per serving. If a 300g bag has 5 servings, how many grams can you eat if your daily calorie target is 2000 and you've already consumed 1800 calories?
A: 48g

W. Prefix Prompt 1:
"... Since there are 5 servings in the bag, the number of grams that can be eaten is 0.8 x 300g = 240g ..."
(*Incorrect: Miscalculated by using the total bag weight instead of grams per serving.*)

W. Prefix Prompt 2:
"... 0.8 servings is approximately equal to 1 serving. The bag has 5 servings, and it weighs 300g. So, each serving weighs: 300g ÷ 5 servings = 60g. Therefore, you can eat approximately: 60g"
(*Incorrect: Rounded 0.8 servings up to 1 serving, ignoring the fraction of grams.*)

Steered:
"Since the bag is 300g and there are 5 servings, each serving is 300 / 5 = 60g. Therefore, the number of grams that can be eaten is 0.8 x 60 = 48g."
(*Correct: Combines Prefix 1's fraction handling and Prefix 2's grams-per-serving calculation, producing the correct process and answer.*)

Evaluations



References

[1] Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for llms, 2023.
[2] Anonymous. Model tells itself where to attend: Faithfulness meets automatic attention steering. In *Submitted to ACL Rolling Review - June 2024*, 2024.

Paper Link

Homepage