PREFIX AND OUTPUT LENGTH-AWARE SCHEDULING FOR EFFICIENT ONLINE LLM INFERENCE

Iñaki Arango & Ayush Noori
Harvard John A. Paulson School of Engineering and Applied Sciences
Allston, MA 02134, USA
{inakiarango,anoori}@college.harvard.edu

Yepeng Huang
Harvard Medical School
Boston, MA 02115, USA
yepeng@fas.harvard.edu

Rana Shahout & Minlan Yu
Harvard John A. Paulson School of Engineering and Applied Sciences
Allston, MA 02134, USA
{rana, minlanyu}@seas.harvard.edu

ABSTRACT

To meet high user demand, production LLM inference systems use data parallelism to allocate the request pool evenly across multiple GPUs. However, in modern AI applications like chatbots, code generation, search engines, or agents, prompt prefixes are often shared, allowing for improved performance if requests with shared prefixes are assigned to the same GPU and the intermediate KV cache is reused. Previous work like PREBLE (Srivatsa et al., 2024) has developed distributed LLM serving platforms that optimize for prompt sharing; however, we hypothesized that additional performance gains could be achieved by integrating prefix-aware and output length-aware scheduling. To that end, we extend the adaptive prefix-aware scheduler of PREBLE to account for output length, which can be estimated using a lightweight BERT model or other cheap predictor. To benchmark this modification to PREBLE, we also build on PREBLE's online LLM inference simulation to support overhead tracking, variable output lengths, experiment caching, and data analysis. This simulation platform allows us to demonstrate that including output length in the per-GPU load calculation improves the performance of PREBLE, with 14.31% and 28.89% reduced latency at 64 and 128 requests per second, respectively, on 8 GPUs. Thus, considering both output length and shared prefixes may enable improved efficiency of online LLM inference in high demand settings.

1 INTRODUCTION

LLM inference can benefit from several parallelization approaches, including data, tensor, and pipeline parallelism (Li et al., 2024a; Liu et al., 2024). Tensor and pipeline parallelism distribute individual operators or model layers across multiple GPU devices, respectively (Jiang et al., 2024). By contrast, the most popular approach, data parallelism, involves replicating model and optimizer states across GPU devices. Model replicas can together serve requests in parallel, thereby increasing the throughput of the inference system, and additional GPU devices can be added as needed to meet user demand. Indeed, providers like OpenAI, Google, Anthropic, and others have built large-scale production AI clusters with tens of thousands of GPUs (Gangidi et al., 2024; Zu et al., 2024) across which they deploy model replicas to concurrently process user requests.

Traditional data parallelism strategies divide requests evenly and randomly across the available GPU workers. However, in real-world LLM applications, requests exhibit patterns that can be exploited

to improve performance over naïve data parallelism. Specifically, two features of modern LLM use cases stand out (for a detailed discussion, see Srivatsa et al. (2024)):

High prompt-to-output ratio. LLM inference can be divided into two stages: prefill and decode. In the prefill phase, the (tokenized) input, known as the prompt, is processed by Transformer layers, and the generated key and value (KV) vectors are cached (Pope et al., 2023). During the subsequent decode phase, the LLM autoregressively generates new tokens based on the KV cache to produce output. In contemporary LLM applications, prompts are often orders of magnitude longer than outputs, including in document (Li et al., 2024b; Saad-Falcon et al., 2023) and video (Xiao et al., 2021; Rawal et al., 2024) question answering (QA), in-context learning (Dong et al., 2024), complex reasoning tasks (*e.g.*, using chain-of-thought (Wei et al., 2024) or tree-of-thought (Yao et al., 2023a) prompting), multi-agent systems (Wu et al., 2023b), and many other settings. In fact, state-of-the-art LLMs like Gemini are equipped with context lengths up to 10 million tokens (Team et al., 2024; Google, 2024). Since LLM inference is increasingly dominated by prompt processing rather than output generation, optimizing prefill efficiency is critical to improving performance (Srivatsa et al., 2024).

Shared prompts. Furthermore, in many LLM applications, prompts often overlap across user requests and may share certain prefixes. In fact, Srivatsa et al. (2024) report that, in some scenarios, 85% to 97% of tokens in a prompt are shared with other prompts. Examples of settings with frequent prompt sharing include conversational agents (Anthropic, 2024), tool use (Schick et al., 2023; Qin et al., 2023), code generation, question answering (Wang et al., 2024; Li et al., 2024b; Saad-Falcon et al., 2023; Jia et al., 2024; Xiao et al., 2021; Rawal et al., 2024), complex reasoning (Zhang et al., 2022; Wei et al., 2024; Yao et al., 2023a; Besta et al., 2024), batch inference (Li et al., 2022), in-context learning (Dong et al., 2024), and agent systems (Chen et al., 2024; Chan et al., 2023; Wu et al., 2023b; Gao et al., 2024b; Guo et al., 2024; Xi et al., 2023).

Given the diverse situations in which prompts may be shared for LLM inference, we and others (Srivatsa et al., 2024; Juravsky et al., 2024) have explored how to leverage shared prompts to improve the efficiency of online LLM inference. Notably, the first approach to benefit from prompt sharing under a distributed LLM serving system with data parallelism across multiple GPUs was PREBLE (Srivatsa et al., 2024). However, PREBLE sets the expected decode output length equal to the average output length of requests during scheduling, thereby assuming low-variance response lengths. In real-world scenarios, such as chatbot applications, output lengths are highly variable and distinct. Therefore, PREBLE may induce imbalanced workloads on different GPUs, causing performance degradation in certain settings. Here, we extend PREBLE by integrating its prefix-aware scheduling with the output length-aware scheduling of S3 (Jin et al., 2024) to achieve improved performance.

2 RELATED WORK

Recently, many solutions have been proposed to leverage prompt sharing for LLM inference (Li et al., 2024a; Wu et al., 2023a; Kwon et al., 2023; Gao et al., 2024a) across multiple levels of abstraction: from hardware-proximal modifications of attention computation – both lossless, like HYDRAGEN (Juravsky et al., 2024), and lossy, like PROMPTCACHE (Gim et al., 2024) – to optimizations of inter-GPU scheduling (Srivatsa et al., 2024) that are built on top of existing LLM serving systems (Zheng et al., 2024; Kwon et al., 2023). Prior work, including SGLANG (Zheng et al., 2024), HYDRAGEN (Juravsky et al., 2024), PROMPTCACHE (Gim et al., 2024) and CACHEBLEND (Yao et al., 2024), reuse the cached KV state when prompt prefixes are shared on a single GPU. PREBLE optimizes for KV state reuse while balancing the computational load across GPUs. The PREBLE scheduler maintains a global prefix tree with references to all KV caches stored in each GPU, enabling rapid matching of new requests. This tree is combined with a least recently used (LRU) cache, which determines when to evict unused KV states and load new requests into memory. Follow-on research after PREBLE, like MEMSERVE (Hu et al., 2024) and MOONCAKE (Qin et al., 2024), also leverage prompt sharing in multi-GPU settings. Nonetheless, PREBLE remains state-of-the-art at the time of writing, *e.g.*, on the LooGLE dataset with 0.7 requests per second, MEMSERVE is outperformed by PREBLE by a factor of $6.25\times$ for 2 GPUs and $18.18\times$ for 4 GPUs. Separately, previous work like S3 (Jin et al., 2024) has demonstrated the potential for performance improvements by incorporating a decoding output length prediction module into the scheduler. The S3 authors fine-tune a DistilBERT model to predict output sequence lengths given an input prompt, achieving $6.49\times$ improved throughput over worst case benchmarks. Similarly, TRAIL (Shahout

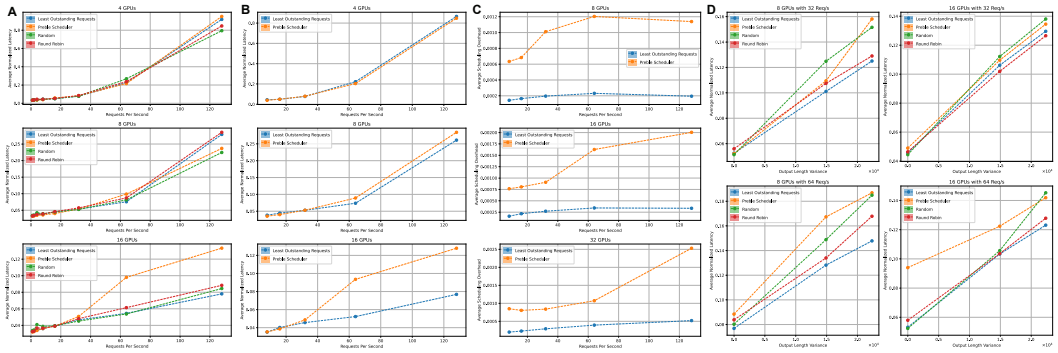

Figure 1: We evaluate performance of PREBLE and other baseline naïve schedulers across **(A)** request rate, **(B)** GPU count, **(C)** scheduling overhead, and **(D)** variance in output length.

et al., 2024) demonstrated that output length predictions can be cheaply obtained from the layer embeddings of the target LLM itself, achieving $2.66\times$ lower error as compared to BERT predictions.

3 RESULTS

3.1 CHARACTERIZATION OF PREBLE SCALABILITY

First, we extend the simulator of LLM inference and scheduling created by Srivatsa et al. (2024) to evaluate and benchmark PREBLE and our proposed improvements. We added support for overhead tracking, variable output lengths, and experiment caching. This environment also facilitated running entire experiment suites and performing detailed data analysis, serving as the primary testbed for our experiments. Using this simulator, we evaluated PREBLE and other baseline naïve schedulers under multiple configurations to identify opportunities for performance improvement by integrating output length-aware scheduling. We focused on the scalability of PREBLE as the number of GPUs and requests increased. Tests were conducted on 4, 8, and 16 GPUs wtih request rates in the range 2^n for $n \in \{0, 1, \ldots, 7\}$. Unless otherwise stated, all experiments were performed on our modified simulator with A6000 GPUs, FlashInfer (FlashInfer Contributors, 2024) enabled, and with artificial datasets constructed from the ReAct training data (Yao et al., 2023b) configured with 10 workloads and 4 in-context examples.

We find that, at high request rates, PREBLE is outperformed by other non-prefix-aware schedulers (Figure 1A). This effect is exacerbated by an increase in the number of GPUs available for inference. In the 4-GPU setup, PREBLE performs comparably to baseline naïve schedulers, including random, round robin, or least outstanding requests (LOR), as measured by average normalized latency (*i.e.*, the average latency per token). However, PREBLE demonstrates a disadvantage at the highest request rate of 128 requests per second (RPS). Further, as the rate of requests inbound to the server grows, the latency increases faster for PREBLE than for the naïve scheduling methods. In the 8-GPU configuration, PREBLE's falls behind all naïve methods at 32 RPS, and is still outperformed by the random scheduler at the highest rate of 128 RPS. Under the 16-GPU configuration, PREBLE falls behind the baselines with a $1.6\times$ higher latency at both 64 RPS and 128 RPS. We also investigated how PREBLE performance changes based on the number of available GPUs (Figure 1B). We observe that latency growth for PREBLE accelerates with the number of GPUs faster than for the LOR scheduler. With 4 GPUs, PREBLE achieves comparable latency to LOR across all tested request rates. However, with both 8 and 16 GPUs, PREBLE's latency exceeds that of LOR at 64 RPS and 128 RPS.

Further investigation revealed that PREBLE performance is hampered by overhead introduced by its E2 scheduler (Figure 1C). This scheduling overhead scales with both the number of GPUs and the request rate, disproportionately impacting PREBLE's ability to maintain low latency. Concretely, in the 4-GPU setup, PREBLE's average scheduling overhead increases from 6×10^{-4} to 12×10^{-4} as the request rate rises from 8 RPS to 64 RPS, and remains above 11×10^{-4} at 128 RPS. By contrast, the LOR baseline shows only a slight increase, from 2×10^{-4} to below 3×10^{-4} over the same range. Similar trends are observed in the 8-GPU setup, where PREBLE's overhead grows from 8×10^{-4} to 20×10^{-4} from 8 RPS to 128 RPS, while the LOR baseline rises only marginally from 2×10^{-4}

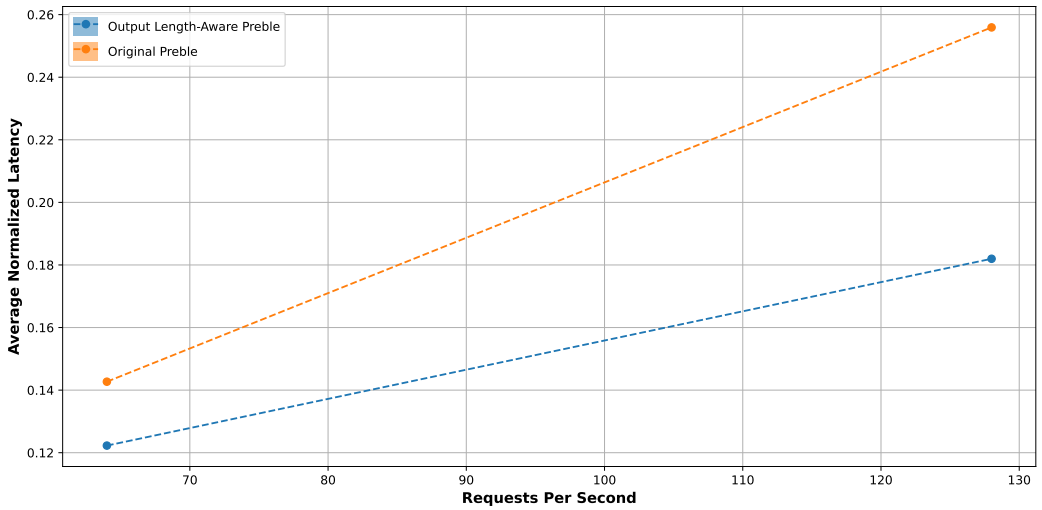

Figure 2: Output length-aware scheduling improves performance of PREBLE on 8 GPUs.

to 3×10^{-4}. The disparity is more pronounced in the 32-GPU setup, where PREBLE's overhead increases from 8×10^{-4} to 25×10^{-4} over the same request rate range, compared to a more modest increase from 2×10^{-4} to 5×10^{-4} for the LOR baseline.

Additionally, as we initially hypothesized, decode length heterogeneity adversely impacts the performance of PREBLE (Figure 1D). Despite the fact that PREBLE incorporates a scheduler that estimates future GPU load based on historical output lengths, both naïve schedulers and PREBLE schedulers suffer from increased request latency when there is high variance in the length of decoded request responses (Appendix A.2 and Figure 3), supporting the need for an output length-aware scheduler.

3.2 EXTENDING PREBLE TO INCLUDE OUTPUT LENGTH-AWARE SCHEDULING

Having carefully characterized the scalability challenges associated with PREBLE, we sought to improve its performance by leveraging both prefix-aware and output length-aware scheduling. We build on the E2 scheduler of PREBLE by considering prefix sharing, fairness, and output length. As a proof-of-concept, we provide PREBLE with access to a perfect oracle of true output length rather than predicted output length. Output length could also be predicted; for example, in Appendix A.1, we show that a lightweight BERT-based language model fine-tuned on `Alpaca-52K` achieves $4.8\times$ greater performance than random at this task.

To incorporate output length, we modify the global prefix tree maintained by the PREBLE E2 scheduler. In the original E2 prefix tree, each tree node represents a prefix, and is associated with three properties: the number of tokens in the node (*i.e.*, maximum number of new tokens), the GPUs caching the node KV values, and the per-GPU number of requests sharing the node prefix over a history window H. We add additional functionality to store the true output length of each tree node by replacing the average maximum number of new tokens over all requests in H with the exact true output length of each request. We compare this modified version of PREBLE on a dataset with high variance in token lengths, created using our ReAct-based variance-customizable dataset generator (Appendix A.2). When using output length for per-GPU load calculation, we improve the performance of PREBLE in high-demand settings, with 14.31% reduced latency at 64 RPS (0.1223 vs. 0.1427) and 28.89% reduced latency at 128 RPS (0.1820 vs. 0.2559).

4 DISCUSSION

Here, we show that considering both output length and shared prefixes may enable improved efficiency of online LLM inference in high-demand settings more representative of real-world LLM inference needs. Future work will use a BERT-based language model for output length prediction (Appendix A.1) instead of a perfect oracle to output length, and will also benchmark PREBLE and

other scheduling solutions under multiple hardware and load configurations to further characterize scalability and reduce the variance in performance results.

To reduce scalability challenges due to overhead, we also propose a new, alternative strategy for scheduling. Motivated by SPLITWISE (Patel et al., 2024) rather than the chunked prefill approach (Agrawal et al., 2024) currently used by PREBLE, VLLM, and SGLANG, we propose splitting prefill and decode phases across different GPU workers and scheduling accordingly in a prefix and output length-aware manner. Specifically, by leveraging single-token activations to predict output lengths for load-aware SPLITWISE scheduling, such a scheduler would be capable of dynamically allocating resources between prefill and decode tasks. This approach has the potential to significantly minimize scaling overhead on GPUs by aligning computational resources more closely with actual workload demands. Future work can consider the development of this scheduling system for efficient online LLM serving.

ACKNOWLEDGEMENTS

Rana Shahout and Minlan Yu are partially supported by NSF CNS NeTS 2107078. This work was supported in part by ACE, one of the seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

REFERENCES

Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, Alexey Tumanov, and Ramachandran Ramjee. Taming Throughput-Latency Tradeoff in LLM Inference with Sarathi-Serve, June 2024. URL http://arxiv.org/abs/2403.02310. arXiv:2403.02310 [cs].

Anthropic. Prompt caching with Claude, August 2024. URL https://www.anthropic.com/news/prompt-caching.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, March 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i16.29720. URL http://arxiv.org/abs/2308.09687. arXiv:2308.09687 [cs].

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate, August 2023. URL http://arxiv.org/abs/2308.07201. arXiv:2308.07201 [cs].

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs, June 2024. URL http://arxiv.org/abs/2309.13007. arXiv:2309.13007 [cs].

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, October 2024. URL http://arxiv.org/abs/2301.00234. arXiv:2301.00234 [cs].

FlashInfer Contributors. FlashInfer: Kernel Library for LLM Serving, December 2024. URL https://github.com/flashinfer-ai/flashinfer. original-date: 2023-07-22T00:29:04Z.

Adithya Gangidi, Rui Miao, Shengbao Zheng, Sai Jayesh Bondu, Guilherme Goes, Hany Morsy, Rohit Puri, Mohammad Riftadi, Ashmitha Jeevaraj Shetty, Jingyi Yang, Shuqiang Zhang, Mikel Jimenez Fernandez, Shashidhar Gandham, and Hongyi Zeng. RDMA over Ethernet for Distributed Training at Meta Scale. In *Proceedings of the ACM SIGCOMM 2024 Conference*, ACM SIGCOMM '24, pp. 57–70, New York, NY, USA, August 2024. Association for Computing Machinery. ISBN 9798400706141. doi: 10.1145/3651890.3672233. URL https://doi.org/10.1145/3651890.3672233.

Bin Gao, Zhuomin He, Puru Sharma, Qingxuan Kang, Djordje Jevdjic, Junbo Deng, Xingkun Yang, Zhou Yu, and Pengfei Zuo. Cost-Efficient Large Language Model Serving for Multi-turn Conversations with CachedAttention, June 2024a. URL http://arxiv.org/abs/2403.19708. arXiv:2403.19708 [cs].

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering Biomedical Discovery with AI Agents, July 2024b. URL http://arxiv.org/abs/2404.02831. arXiv:2404.02831 [cs].

In Gim, Guojun Chen, Seung-seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. Prompt Cache: Modular Attention Reuse for Low-Latency Inference, April 2024. URL http://arxiv.org/abs/2311.04934. arXiv:2311.04934.

Google. Our next-generation model: Gemini 1.5, February 2024. URL https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large Language Model based Multi-Agents: A Survey of Progress and Challenges, April 2024. URL http://arxiv.org/abs/2402.01680. arXiv:2402.01680 [cs].

Cunchen Hu, Heyang Huang, Junhao Hu, Jiang Xu, Xusheng Chen, Tao Xie, Chenxi Wang, Sa Wang, Yungang Bao, Ninghui Sun, and Yizhou Shan. MemServe: Context Caching for Disaggregated LLM Serving with Elastic Memory Pool, June 2024. URL http://arxiv.org/abs/2406.17565. arXiv:2406.17565 [cs].

Shuyue Jia, Subhrangshu Bit, Edward Searls, Meagan V. Lauber, Lindsey A. Claus, Pengrui Fan, Varuna H. Jasodanand, Divya Veerapaneni, William M. Wang, Rhoda Au, and Vijaya B. Kolachalama. PodGPT: An audio-augmented large language model for research and education, November 2024. URL https://www.medrxiv.org/content/10.1101/2024.07.11.24310304v2. Pages: 2024.07.11.24310304.

Ziheng Jiang, Haibin Lin, Yinmin Zhong, Qi Huang, Yangrui Chen, Zhi Zhang, Yanghua Peng, Xiang Li, Cong Xie, Shibiao Nong, Yulu Jia, Sun He, Hongmin Chen, Zhihao Bai, Qi Hou, Shipeng Yan, Ding Zhou, Yiyao Sheng, Zhuo Jiang, Haohan Xu, Haoran Wei, Zhang Zhang, Pengfei Nie, Leqi Zou, Sida Zhao, Liang Xiang, Zherui Liu, Zhe Li, Xiaoying Jia, Jianxi Ye, Xin Jin, and Xin Liu. MegaScale: Scaling Large Language Model Training to More Than 10,000 GPUs, February 2024. URL http://arxiv.org/abs/2402.15627. arXiv:2402.15627 [cs].

Yunho Jin, Chun-Feng Wu, David Brooks, and Gu-Yeon Wei. S3: increasing GPU utilization during generative inference for higher throughput. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pp. 18015–18027, Red Hook, NY, USA, May 2024. Curran Associates Inc.

Jordan Juravsky, Bradley Brown, Ryan Ehrlich, Daniel Y. Fu, Christopher Ré, and Azalia Mirhoseini. Hydragen: High-Throughput LLM Inference with Shared Prefixes, May 2024. URL http://arxiv.org/abs/2402.05099. arXiv:2402.05099.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention, September 2023. URL http://arxiv.org/abs/2309.06180. arXiv:2309.06180.

Baolin Li, Yankai Jiang, Vijay Gadepally, and Devesh Tiwari. LLM Inference Serving: Survey of Recent Advances and Opportunities, July 2024a. URL http://arxiv.org/abs/2407.12391. arXiv:2407.12391 [cs].

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. LooGLE: Can Long-Context Language Models Understand Long Contexts?, September 2024b. URL http://arxiv.org/abs/2311.04939. arXiv:2311.04939 [cs].

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-Level Code Generation with AlphaCode, February 2022. URL http://arxiv.org/abs/2203.07814. arXiv:2203.07814.

Yiheng Liu, Hao He, Tianle Han, Xu Zhang, Mengyuan Liu, Jiaming Tian, Yutong Zhang, Jiaqi Wang, Xiaohui Gao, Tianyang Zhong, Yi Pan, Shaochen Xu, Zihao Wu, Zhengliang Liu, Xin Zhang, Shu Zhang, Xintao Hu, Tuo Zhang, Ning Qiang, Tianming Liu, and Bao Ge. Understanding LLMs: A Comprehensive Overview from Training to Inference, January 2024. URL http://arxiv.org/abs/2401.02038. arXiv:2401.02038 [cs].

Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. Splitwise: Efficient generative LLM inference using phase splitting, May 2024. URL http://arxiv.org/abs/2311.18677. arXiv:2311.18677.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently Scaling Transformer Inference. *Proceedings of Machine Learning and Systems*, 5:606–624, March 2023. URL https://proceedings.mlsys.org/paper_files/paper/2023/hash/c4be71ab8d24cdfb45e3d06dbfca2780-Abstract-mlsys2023.html.

Ruoyu Qin, Zheming Li, Weiran He, Mingxing Zhang, Yongwei Wu, Weimin Zheng, and Xinran Xu. Mooncake: A KVCache-centric Disaggregated Architecture for LLM Serving, July 2024. URL http://arxiv.org/abs/2407.00079. arXiv:2407.00079 [cs].

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs, October 2023. URL http://arxiv.org/abs/2307.16789. arXiv:2307.16789 [cs].

Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. CinePile: A Long Video Question Answering Dataset and Benchmark, October 2024. URL http://arxiv.org/abs/2405.08813. arXiv:2405.08813 [cs].

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, David Seunghyun Yoon, Ryan A. Rossi, and Franck Dernoncourt. PDFTriage: Question Answering over Long, Structured Documents, November 2023. URL http://arxiv.org/abs/2309.08872. arXiv:2309.08872 [cs].

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

Rana Shahout, Eran Malach, Chunwei Liu, Weifan Jiang, Minlan Yu, and Michael Mitzenmacher. Don't Stop Me Now: Embedding Based Scheduling for LLMs, October 2024. URL http://arxiv.org/abs/2410.01035. arXiv:2410.01035 [cs].

Vikranth Srivatsa, Zijian He, Reyna Abhyankar, Dongming Li, and Yiying Zhang. Preble: Efficient Distributed Prompt Scheduling for LLM Serving. October 2024. URL https://openreview.net/forum?id=meKEKDhdnx.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo-yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose

Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kepa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G. Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, Z. J. Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang,

Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Be-

nigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, T. J. Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, M. K. Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M, Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A Family of Highly Capable Multimodal Models, June 2024. URL http://arxiv.org/abs/2312.11805. arXiv:2312.11805 [cs].

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. (arXiv:2012.15828), June 2021. doi: 10.48550/arXiv.2012.15828. URL http://arxiv.org/abs/2012.15828. arXiv:2012.15828.

Yubo Wang, Xueguang Ma, and Wenhu Chen. Augmenting Black-box LLMs with Medical Textbooks for Biomedical Question Answering. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1754–1770, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.95. URL https://aclanthology.org/2024.findings-emnlp.95.

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 *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 24824–24837, Red Hook, NY, USA, April 2024. Curran Associates Inc. ISBN 978-1-71387-108-8.

Bingyang Wu, Yinmin Zhong, Zili Zhang, Gang Huang, Xuanzhe Liu, and Xin Jin. Fast Distributed Inference Serving for Large Language Models, May 2023a. URL http://arxiv.org/abs/2305.05920. arXiv:2305.05920 [cs].

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, October 2023b. URL http://arxiv.org/abs/2308.08155. arXiv:2308.08155 [cs].

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The Rise and Potential of Large Language Model Based Agents: A Survey, September 2023. URL http://arxiv.org/abs/2309.07864. arXiv:2309.07864 [cs].

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. pp. 9777–9786, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Xiao_NExT-QA_Next_Phase_of_Question-Answering_to_Explaining_Temporal_Actions_CVPR_2021_paper.html.

Jiayi Yao, Hanchen Li, Yuhan Liu, Siddhant Ray, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. CacheBlend: Fast Large Language Model Serving for RAG with Cached Knowledge Fusion, June 2024. URL http://arxiv.org/abs/2405.16444. arXiv:2405.16444 [cs].

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models, December 2023a. URL http://arxiv.org/abs/2305.10601. arXiv:2305.10601 [cs].

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models, March 2023b. URL http://arxiv.org/abs/2210.03629. arXiv:2210.03629 [cs].

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic Chain of Thought Prompting in Large Language Models, October 2022. URL http://arxiv.org/abs/2210.03493. arXiv:2210.03493 [cs].

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient Execution of Structured Language Model Programs, June 2024. URL http://arxiv.org/abs/2312.07104. arXiv:2312.07104.

Yazhou Zu, Alireza Ghaffarkhah, Hoang-Vu Dang, Brian Towles, Steven Hand, Safeen Huda, Adekunle Bello, Alexander Kolbasov, Arash Rezaei, Dayou Du, Steve Lacy, Hang Wang, Aaron Wisner, Chris Lewis, and Henri Bahini. Resiliency at Scale: Managing {Google's} {TPUv4} Machine Learning Supercomputer. pp. 761–774, 2024. ISBN 978-1-939133-39-7. URL https://www.usenix.org/conference/nsdi24/presentation/zu.

A Appendix

A.1 Predicting Decode Output Length

We trained a request output length predictor to enable more accurate and efficient scheduling. To limit the overhead incurred by the scheduler, we utilized the lightweight 6-layer BERT-based `all-MiniLM-L6-v2` language model (Wang et al., 2021), which is a `sentence-transformers` model Reimers & Gurevych (2019) with 80 million parameters contrastively trained on over a billion sentences to map text to vectors in $\mathbb{R}^{384}$. We fine-tuned this language model on paired `(request, output length)` data from the training split of the `Alpaca-52K` dataset curated by Stanford (Taori et al., 2023). As suggested by prior work Jin et al. (2024), for BERT models, framing output length prediction as a binned classification task yields better results than regression. We used 20 bins for output length (*i.e.*, a 20-class classification problem) and partitioned the data into a 60% training set, 20% validation set, and 20% test set. On the independent test set, the model achieved $4.8\times$ greater performance than random. Model performance metrics are reported in Table 1.

Table 1: Performance metrics for the fine-tuned BERT-based `all-MiniLM-L6-v2` language model on the independent `Alpaca-52K` test set.

Performance Metric	**Value**
Accuracy	0.24
F1 Score	0.22
Precision	0.23
Recall	0.24
AUROC	0.85
MCC	0.20

A.2 Generating Variance-Maximizing Benchmarks

To understand how decode length heterogeneity impacts the performance of PREBLE, we generated artificial benchmarking datasets from ReAct whose output lengths can follow any user-specified discrete distribution, then used this to create a variance-maximizing benchmark with high variance in token lengths (Figure 3). For this dataset, all requests were created with a maximum new token length of 300 tokens. With 50% probability, requests had a true output length of 1 token, and with 50% probability, they had a true output length of 300 tokens. We used the same number of workloads and in-context examples as in previous evaluations (10 and 4, respectively).

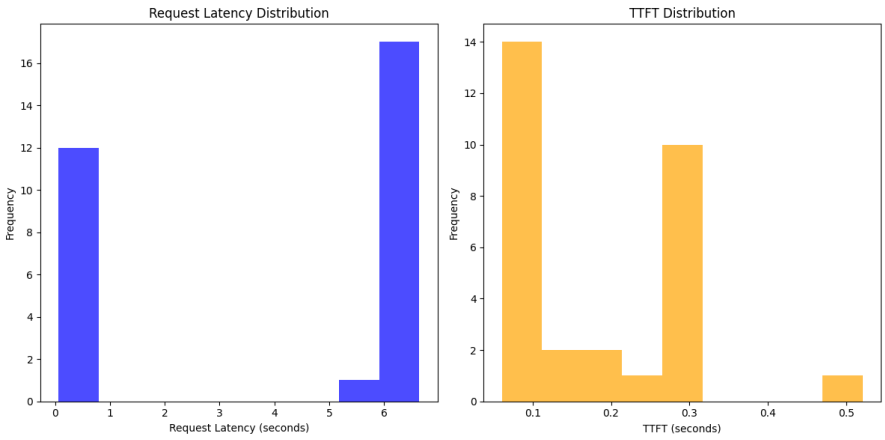

Figure 3: We generated artificial benchmarking datasets from ReAct whose output lengths can follow any user-specified discrete distribution; for example, here we generate a variance-maximizing benchmark.