Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024c. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9.

Liu, T., Xu, C., and McAuley, J. Repobench: Benchmarking repository-level code auto-completion systems. In *The Twelfth International Conference on Learning Representations*, 2024d. URL https://openreview.net/forum?id=pPjZIOuQuF.

Liu, Z., Desai, A., Liao, F., Wang, W., Xie, V., Xu, Z., Kyrillidis, A., and Shrivastava, A. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36, 2024e.

Luo, W., Ma, S., Liu, X., Guo, X., and Xiao, C. Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=GC4mXVfquq.

Meta. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/, 2024. Accessed: 2024-06-07.

Mohtashami, A. and Jaggi, M. Landmark attention: Random-access infinite context length for transformers. *ArXiv preprint*, abs/2305.16300, 2023. URL https://arxiv.org/abs/2305.16300.

OpenAI. Gpt-4o-mini: Advancing cost-efficient intelligence, 2023. Accessed: 2023-12-14.

Pan, R., Liu, X., Diao, S., Pi, R., Zhang, J., Han, C., and Zhang, T. Lisa: Layerwise importance sampling for memory-efficient large language model fine-tuning. *ArXiv preprint*, abs/2403.17919, 2024a. URL https://arxiv.org/abs/2403.17919.

Pan, R., Xing, S., Diao, S., Sun, W., Liu, X., Shum, K., Zhang, J., Pi, R., and Zhang, T. Plum: Prompt learning using metaheuristics. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 2177–2197, Bangkok, Thailand and virtual meeting, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.129. URL https://aclanthology.org/2024.findings-acl.129.

Pires, B. Á. and Szepesvári, C. Multiclass classification calibration functions. *arXiv preprint arXiv:1609.06385*, 2016.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv preprint*, abs/2403.05530, 2024. URL https://arxiv.org/abs/2403.05530.

Shaham, U., Ivgi, M., Efrat, A., Berant, J., and Levy, O. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7977–7989, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.536. URL https://aclanthology.org/2023.findings-emnlp.536.

Shi, W., Min, S., Lomeli, M., Zhou, C., Li, M., Lin, X. V., Smith, N. A., Zettlemoyer, L., Yih, W.-t., and Lewis, M. In-context pretraining: Language modeling beyond document boundaries. In *The Twelfth International Conference on Learning Representations*.

Smith, B. and Troynikov, A. Evaluating chunking strategies for retrieval. Technical report, Chroma, 2024. URL https://research.trychroma.com/evaluating-chunking.

Steinwart, I. How to compare different loss functions and their risks. *Constructive Approximation*, 26:225–287, 2007. URL https://api.semanticscholar.org/CorpusID:16660598.

Sun, Y., Dong, L., Zhu, Y., Huang, S., Wang, W., Ma, S., Zhang, Q., Wang, J., and Wei, F. You only cache once: Decoder-decoder architectures for language models. *arXiv preprint arXiv:2405.05254*, 2024.

Tang, J., Zhao, Y., Zhu, K., Xiao, G., Kasikci, B., and Han, S. Quest: Query-aware sparsity for efficient long-context llm inference. *ArXiv preprint*, abs/2406.10774, 2024. URL https://arxiv.org/abs/2406.10774.

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., and Metzler, D. Long range arena : A benchmark for efficient transformers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k.

Tay, Y., Dehghani, M., Tran, V. Q., Garcia, X., Bahri, D., Schuster, T., Zheng, H. S., Houlsby, N., and Metzler, D. Unifying language learning paradigms. *ArXiv preprint*, abs/2205.05131, 2022. URL https://arxiv.org/abs/2205.05131.