クラスタリングに基づく KV Cache 残差量子化による大規模言語モデル推論の高速化

渡邊梢平 [1] 田浦健次朗 [1]
[1] 東京大学大学院 情報理工学系研究科
{shohei,tau}@eidos.ic.i.u-tokyo.ac.jp

概要

大規模言語モデル (LLM) の急速な発展に伴い，長文生成や多数同時リクエストを伴う推論サービスが広く利用されている．一方で，LLM の推論は計算資源およびメモリ資源を大量に消費し，とくにオンライン推論環境においてはスループットが大きな制約となる．本研究では，精度劣化を最小限に抑えた KV Cache 量子化手法を提案する．提案手法では，事前にサンプルデータから取得した KV Cache をクラスタリングし，代表点との誤差を Key と Value の統計的特性の違いを考慮した量子化処理を用いて圧縮することで，推論スループットを最大で約 3.5 倍向上させながら，従来手法と比較して量子化誤差を RMSE で 40%以上削減することを確認した．

1 はじめに

近年，LLM の汎用性能を活かした応用が拡大している．一方でモデルの大規模化に伴い，推論時の計算・メモリ要求が増大し，効率的推論が重要な課題となっている．

多くの LLM は Transformer [1] をベースとしており，Attention の計算が必要となる．この計算を高速化するため，各層の Key/Value を保存して再利用する KV Cache が一般に用いられる．しかし KV Cache は層数・ヘッド数・隠れ次元・生成長に比例して増加し，深刻なメモリボトルネックとなる．特に GPU 環境では KV Cache がメモリを圧迫し，バッチサイズや同時処理数を制限してスループット/レイテンシや運用コストに影響する [2].

この課題に対し，低ビット量子化 [3, 4]，重要度の低いトークンの削除・要約 [5, 6]，Attention 構造や保持方式の改変 [7, 8, 9] などが提案されているが，タスク依存の性能劣化 [10]，既存実装との互換性といったトレードオフが残る．また Key と Value は役割・統計特性が異なり，一様な量子化は品質劣化の要因となり得る．さらにアルゴリズム上の工夫が必ずしも実スループット向上に直結せず，メモリアクセスや計算効率を含む検討が必要である．

本研究では KV Cache の統計的構造を活用した量子化を提案する．事前サンプルから KV Cache をクラスタリングして重心を得て，推論時は各 Key/Value を最近傍重心で近似し，残差のみを量子化して保持することで，高圧縮と低誤差を両立する．さらに Key には次元ごとのダイナミックレンジ差を考慮した次元単位量子化を適用し，Value にはランダム回転による分散均一化 [11] 後に Lloyd–Max 量子化 [12, 13] を行うことで誤差分散を低減する．加えて GPU 上で量子化・復号と Attention を統合した最適化 CUDA カーネルを実装し，メモリ削減に加えて推論スループット向上を達成した．実験ではメモリ使用量を大幅に削減しつつ，ベースライン比で最大約 3.5 倍のスループット向上を確認した．

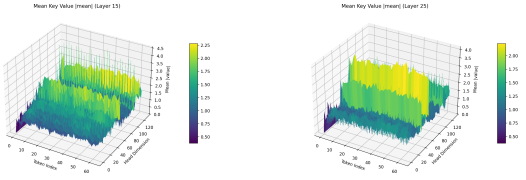
2 関連研究

2.1 KV Cache の量子化

2.1.1 Key 行列の分布特性

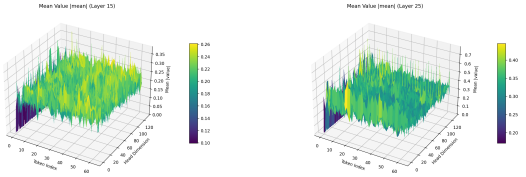
Transformer の Attention 機構において，Key 行列は Value 行列とは異なり，特定の次元に極端な外れ値が集中する傾向がある．予備実験として，Llama 3.1 8B Instruct[1)] を用いて Key 行列および Value 行列の値の分布を観測した結果を，図 1 および図 2 に示す．Key 行列では，層を問わず特定のチャネルが他の次元と比較して著しく大きな値を持ち，全体のダイナミックレンジを支配していることが確認できる．一方で，Value 行列にはそのような顕著な偏りは観測されなかった．既存の min-max 量子化は外れ値に弱

1） https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct



(a) 15th layer (b) 25th layer

図 1: Key 行列の分布



(a) 15th layer (b) 25th layer

図 2: Value 行列の分布

く，Key 行列のような分布では有効な分解能が著しく低下するため，これらの特性を考慮した手法が必要となる．

2.1.2 KIVI

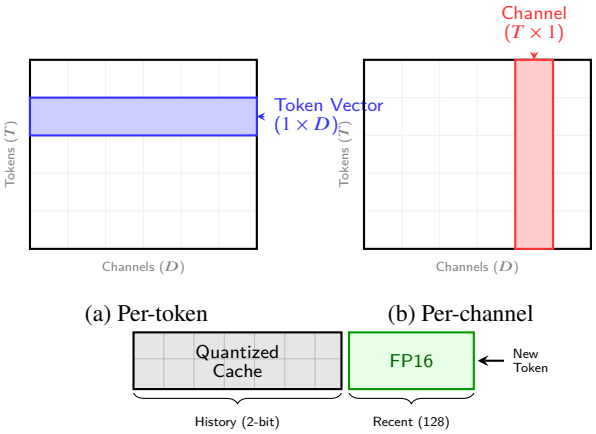
Liu らの KIVI [3] は，KV Cache の分布の非対称性に着目し，外れ値が特定チャネルに集中する Key は per-channel，トークン間干渉を防ぎたい Value は per-token と，異なる軸で量子化を行う (図 3)．また，推論時のストリーミング処理においては，直近のトークンを FP16 で保持し，一定長に達した段階で量子化済み領域へ結合することで精度劣化を抑制している．

2.1.3 QuaRot

QuaRot [11] は，量子化の妨げとなる外れ値を，Hadamard 変換による回転で平滑化する手法である．入力への回転を重み行列への逆回転で相殺することで，モデルの出力を変えずに内部表現の外れ値を抑制できる．KV Cache に対しては，RoPE との整合性を保つために Query と Key の双方にオンラインで回転を適用することで，4-bit の end-to-end 量子化を実現している．

2.1.4 Lloyd-Max 量子化

Lloyd-Max 量子化 [12, 13] は，確率密度関数 $p(x)$ に基づき，量子化誤差 $E[(X - Q(X))^2]$ を最小化する最適な非一様量子化手法である．最適な量子化器は，決定境界 $\mathcal{T} = \{t_i\}$ と代表値 $\mathcal{R} = \{r_i\}$ に関して以下の 2 つの条件を満たす必要がある．



(a) Per-token (b) Per-channel

(c) Streaming Inference

図 3: KIVI [3] における量子化軸の比較とストリーミング処理．

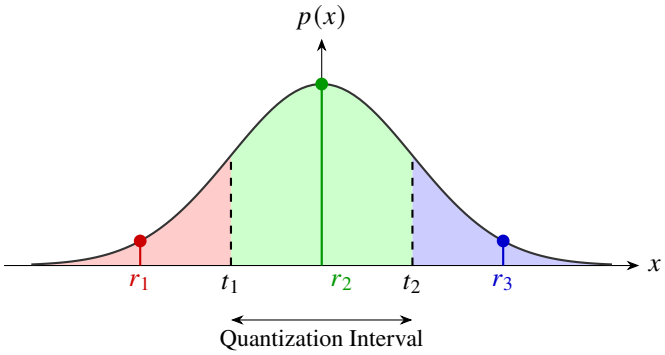


図 4: Lloyd-Max 量子化の概念図．決定境界 t_k と代表値 r_k が最適化される．

1. 最近傍条件: 決定境界 t_k は隣接する代表値の中点でなければならない．

$$t_k = \frac{r_k + r_{k+1}}{2} \tag{1}$$

2. 重心条件: 代表値 r_k は，その区間における確率分布の重心 (条件付き期待値) でなければならない．

$$r_k = \frac{\int_{t_{k-1}}^{t_k} x p(x) dx}{\int_{t_{k-1}}^{t_k} p(x) dx} \tag{2}$$

これらの条件は相互に依存するため，Lloyd のアルゴリズムでは，初期値から出発して上記 2 式を交互に更新する反復計算により，局所最適解へと収束させる．図 4 に，最適化された量子化器における各区間の配置の概念図を示す．

3 提案手法

3.1 概要

KV Cache を (i) 事前学習した重心と (ii) 低ビット残差の組み合わせで表現し，メモリ削減と精度の両立を図る．KV Cache として，重心インデックスと低ビット量子化された残差のみを保持することで，高い圧縮率を実現する．なお，Key に関しては，RoPE 適用前の $\boldsymbol{k}_{\text{pre-RoPE}}$ を対象にクラスタリングを行う．

$$\tilde{\boldsymbol{k}} \approx \mathrm{RoPE}\big[\mathrm{CodebookCentroid}\big(\boldsymbol{k}_{\text{pre-RoPE}}\big)\big] + Q(r_k), \quad (3)$$

$$\tilde{\boldsymbol{v}} \approx \mathrm{CodebookCentroid}(\boldsymbol{v}) + Q_{\mathrm{MSE}}(r_v). \quad (4)$$

3.2 Value ベクトルの量子化器

Value 残差 $\boldsymbol{r}_v$ の量子化には，ランダム回転とスカラー量子化を組み合わせる手法を採用する．Zandieh らによる TurboQuant [14] の知見に基づき，高次元ベクトルにランダム回転 $\boldsymbol{\Pi}$ を適用した後の各成分の分布 $f_X(x)$ は，以下のようにモデル化できる．

$$f_X(x) \propto (1-x^2)^{(d-3)/2}. \quad (5)$$

本手法では，この分布 f_X に対して Lloyd-Max 量子化器 [12, 13] を設計し，代表値集合 $\mathscr{C}$ を決定し，$\mathscr{C}$ 内の値を用いて量子化する．

$$Q_{\mathrm{MSE}}(\boldsymbol{r}) := \operatorname*{argmin}_{\boldsymbol{c}\in\mathscr{C}^d} \|\boldsymbol{\Pi}\boldsymbol{r} - \boldsymbol{c}\|_2 . \quad (6)$$

計算の効率を考慮し，QuaRot [11] と同様に，実装上は固定のランダム直交行列を用い，復元時は逆回転 $\boldsymbol{\Pi}^\top$ を適用する．

3.3 Key ベクトルの量子化器

Key の残差 r_k については，KIVI [3] と同様に次元ごとのスカラー量子化を採用する．これにより RoPE 適用後の空間における誤差を抑制する．

3.4 推論時における KV Cache の管理

計算コストと精度維持のトレードオフを考慮し，KV Cache を図 5 のように 2 領域に分割管理する．

1. 直近ウィンドウ: 直近トークンは重要度が高いため，量子化せず FP16 で保持する．
2. 量子化済み領域: ウィンドウから溢れたトークンは，pre-RoPE 重心インデックスと 2bit 残差コード等に圧縮して保存する．

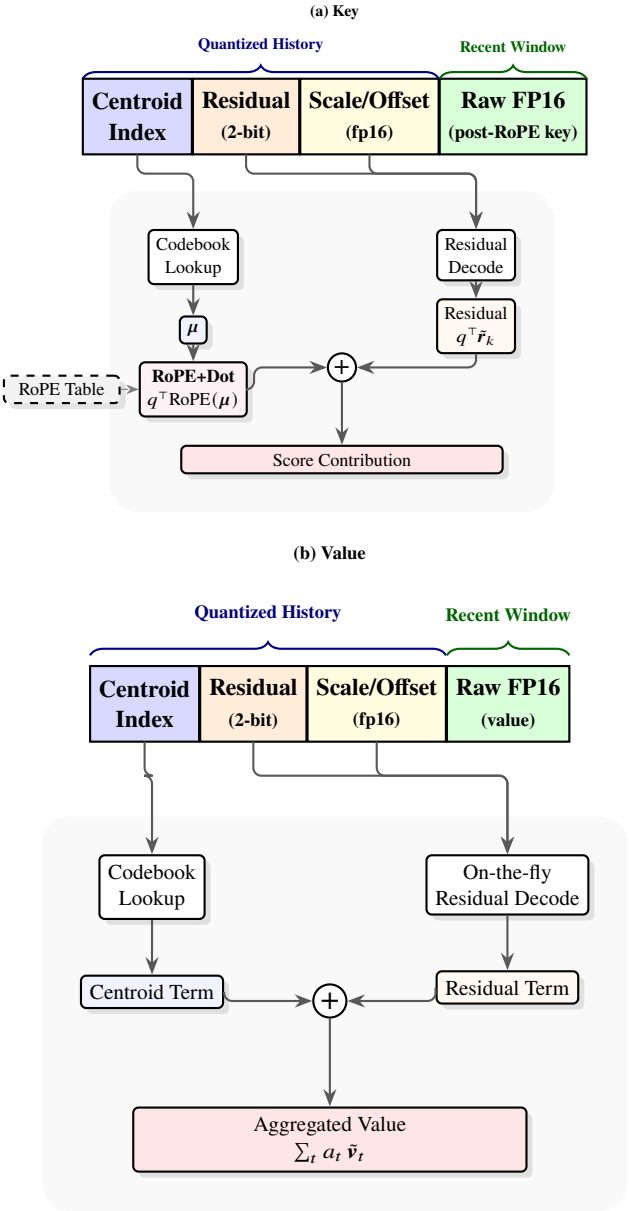


図 5: 推論時の KV Cache の利用の流れ．

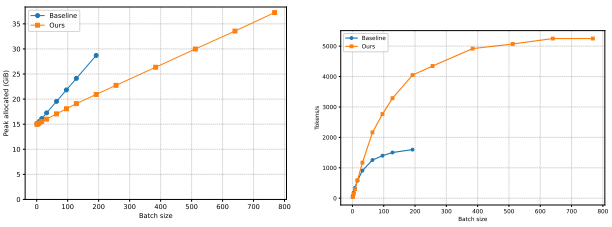
推論時，量子化済み領域のデータはメモリから読み出された後，カーネル内で復号される．Key/Value ともに明示的な復元を行わないことで，メモリ帯域を節約しつつ高速な Attention 計算を実現する．

4 評価・実験

本章では，提案手法の有効性を (i) メモリ・スループット，(ii) タスク性能，(iii) 量子化誤差の観点から検証する．

実験環境として，VRAM 40GB の NVIDIA A100 Tensor Core GPU[2] を搭載した mdx [15] 上の仮想マシンを使用した．モデルは Llama 3.1 8B Instruct を使用

2) https://www.nvidia.com/ja-jp/data-center/a100/



(a) メモリ使用量　　(b) スループット

図 6: メモリ使用量とスループット

表 1: LM-Eval における性能比較

Task	Baseline (16bit)	KIVI (2bit)	Ours (2bit)
CoQA	0.6847	**0.6800**	0.6773
TruthfulQA	0.5402	**0.5403**	**0.5403**
GSM8K	0.7779	0.7460	**0.7710**
Average	0.6676	0.6554	**0.6629**

し，比較対象として，KIVI [3] と Hugging Face のデフォルト実装 (Baseline) を使用した．

提案手法の実装は最適化された CUDA を用いて行った．

4.1　メモリ使用量とスループット

Baseline (16bit) と提案手法を比較した．図 6a に示す通り，提案手法はバッチサイズ増加に伴うメモリ増分を約 2.5 倍削減した．これにより，Baseline では OOM となる領域でも推論が可能である．

また，図 6b に示す通り，最大で約 3.5 倍のスループット向上を確認した．これは最大バッチサイズの増加と，メモリ操作オーバーヘッドの低減に起因する．

4.2　タスク性能

LM-Eval [16] を用いた複数タスクにおける性能を表 1 に示す．提案手法は，既存手法 KIVI を平均スコアで上回り，特に推論能力を要する GSM8K において高い性能を示した．

LongBench [17] による評価結果を表 2 に示す．提案手法は全 21 タスク中 13 タスクで KIVI を上回った．特に Retrieval や QA タスクでの改善が顕著であり，Baseline に近い精度を維持している．一部の要約タスクでは KIVI が優れるものの，平均スコアでは提案手法が同等以上の性能を示している．

表 2: LongBench 評価結果

Task	Baseline	Ours (2bit)	KIVI (2bit)
2WikiMQA	0.1220	**0.1089**	0.1045
DuReader	0.2848	**0.2607**	0.2558
GovReport	0.3338	0.3174	**0.3304**
HotpotQA	0.1255	**0.1188**	0.1176
LCC	0.6339	0.6186	**0.6248**
LSHT	0.4650	**0.4550**	0.4500
MultiNews	0.2686	0.2631	**0.2694**
MultiFieldQA (EN)	0.2837	0.2816	**0.2880**
MultiFieldQA (ZH)	0.1887	**0.1899**	0.1764
MuSiQue	0.0671	**0.0685**	0.0673
NarrativeQA	0.1319	**0.1408**	0.1312
Passage Count	0.0646	0.0623	**0.0689**
Passage Retrieval (EN)	0.9078	**0.9804**	0.9729
Passage Retrieval (ZH)	0.9663	0.8980	**0.9154**
Qasper	0.1551	**0.1624**	0.1406
QMSum	0.2338	0.2360	**0.2433**
RepoBench-P	0.5229	**0.5079**	0.5055
SAMSum	0.4434	0.4337	**0.4447**
TREC	0.7300	**0.7300**	0.7250
TriviaQA	0.7894	**0.7946**	0.7944
VCSum	0.1276	**0.1492**	0.1248
Average	0.3736	**0.3704**	0.3691

表 3: 量子化誤差

Method	Component	Mean Error	RMSE
Ours	Key	2.583×10^{-3}	$\mathbf{8.968 \times 10^{-3}}$
Ours	Value	4.195×10^{-3}	$\mathbf{6.993 \times 10^{-3}}$
KIVI	Key	4.965×10^{-2}	1.248×10^{-1}
KIVI	Value	6.908×10^{-3}	1.229×10^{-2}

4.3　量子化誤差の定量評価

Key/Value の量子化誤差を表 3 に示す．提案手法は KIVI と比較して，RMSE において 40%以上改善した．

5　おわりに

本研究では，KV Cache の統計的構造を活用した量子化手法を提案した．今後の課題として，提案手法を様々なモデルで検証し，モデルサイズなどへの依存性を調査することや，既存の推論フレームワークとの統合を行うことが挙げられる．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[2] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In **Proceedings of the 29th Symposium on Operating Systems Principles**, SOSP '23, p. 611–626, New York, NY, USA, 2023. Association for Computing Machinery.

[3] Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen (Henry) Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: a tuning-free asymmetric 2bit quantization for kv cache. In **Proceedings of the 41st International Conference on Machine Learning**, ICML'24. JMLR.org, 2024.

[4] Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. Kvquant: towards 10 million context length llm inference with kv cache quantization. In **Proceedings of the 38th International Conference on Neural Information Processing Systems**, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc.

[5] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In **International Conference on Learning Representations**, 2023.

[6] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2o: heavy-hitter oracle for efficient generative inference of large language models. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[7] Utkarsh Saxena, Gobinda Saha, Sakshi Choudhary, and Kaushik Roy. Eigen attention: Attention in low-rank space for KV cache compression. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 15332–15344, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[8] Rongzhi Zhang, Kuang Wang, Liyuan Liu, Shuohang Wang, Hao Cheng, Chao Zhang, and Yelong Shen. Lorc: Low-rank compression for llms kv cache with a progressive compression strategy, 2024.

[9] Yifan Zhang, Yifeng Liu, Huizhuo Yuan, Zhen Qin, Yang Yuan, Quanquan Gu, and Andrew Chi-Chih Yao. Tensor product attention is all you need, 2025.

[10] Yixuan Wang, Shiyu Ji, Yijun Liu, Yuzhuang Xu, Yang Xu, Qingfu Zhu, and Wanxiang Che. Lookahead Q-cache: Achieving more consistent KV cache eviction via pseudo query. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 34146–34162, Suzhou, China, November 2025. Association for Computational Linguistics.

[11] Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: outlier-free 4-bit inference in rotated llms. In **Proceedings of the 38th International Conference on Neural Information Processing Systems**, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc.

[12] J. Max. Quantizing for minimum distortion. **IRE Transactions on Information Theory**, Vol. 6, No. 1, pp. 7–12, 1960.

[13] S. Lloyd. Least squares quantization in pcm. **IEEE Transactions on Information Theory**, Vol. 28, No. 2, pp. 129–137, 1982.

[14] Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. Turboquant: Online vector quantization with near-optimal distortion rate, 2025.

[15] Toyotaro Suzumura, Akiyoshi Sugiki, Hiroyuki Takizawa, Akira Imakura, Hiroshi Nakamura, Kenjiro Taura, Tomohiro Kudoh, Toshihiro Hanawa, Yuji Sekiya, Hiroki Kobayashi, Yohei Kuga, Ryo Nakamura, Renhe Jiang, Junya Kawase, Masatoshi Hanai, Hiroshi Miyazaki, Tsutomu Ishizaki, Daisuke Shimotoku, Daisuke Miyamoto, Kento Aida, Atsuko Takefusa, Takashi Kurimoto, Koji Sasayama, Naoya Kitagawa, Ikki Fujiwara, Yusuke Tanimura, Takayuki Aoki, Toshio Endo, Satoshi Ohshima, Keiichiro Fukazawa, Susumu Date, and Toshihiro Uchibayashi. mdx: A cloud platform for supporting data science and cross-disciplinary research collaborations. In **2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)**, pp. 1–7, 2022.

[16] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024.

[17] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics.