



HAL
open science

Towards an Ethical Compression of Large Language Models

Irina Proskurina, Guillaume Metzler, Julien Velcin

► **To cite this version:**

Irina Proskurina, Guillaume Metzler, Julien Velcin. Towards an Ethical Compression of Large Language Models. Journée Éthique et TAL 2024, Apr 2024, Nancy, France. hal-04646400

HAL Id: hal-04646400

<https://hal.science/hal-04646400v1>

Submitted on 12 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Towards an Ethical Compression of Large Language Models

Irina Proskurina, Guillaume Metzler, Julien Velcin

Université de Lyon, Lyon 2, ERIC UR 3083, France

Abstract

This proposal explores the fairness of compressed large language models (LLMs). We focus on the ethical implications of applying efficient compression techniques, particularly quantization, to generative LLMs, motivated by recent studies. While quantization enhances inference efficiency, marked by existing works, we primarily focus on understanding its effects on token-level confidence and predictive probability distributions in our research. We also identify significant influences on LLM behaviour during text generation, shedding light on potential biases and ethical concerns. We have determined the difference in output probability distributions after compression and aim to use this observation to propose a debiasing quantization approach.

Proposal Outline

Large language models (LLMs) have demonstrated their effectiveness across diverse natural language generation applications (Bahdanau et al., 2014; ?; Touvron et al., 2023). Recent models excel in zero-shot scenarios, making fine-tuning redundant (Zhang et al., 2022; Workshop et al., 2022; Jiang et al., 2023).

Scaling power-laws introduced by Kaplan et al., 2020 explain the enhancement of zero-shot performance across a broad spectrum of downstream tasks as model sizes scale up, suggesting the emergence of capabilities at a larger scale. At the same time, inverse scaling laws imply that accessing well-performing larger models may become more challenging.

To accelerate inference time and ease high latency and extensive storage demands, various efficient compression methods, such as quantization and pruning, have been developed (Gupta and Agrawal, 2020). Quantization, which involves reducing the float weight precision in LLMs, and weight or layer pruning stand out as prominent efficient compression techniques. Prior studies measure the efficacy of compression with (1) latency-related measures determining the response delay, (2) the precision error of weights approximation, and (3) performance decrease on zero-shot benchmarks (Dettmers et al., 2022; Xiao et al., 2023).

Acknowledging the potential societal implications of using compressed models is crucial from an ethical and responsible AI perspective. As noted by ?, the compression of models negatively affects their fairness and amplifies their sensitivity to specific linguistic phenomena, particularly in tasks like multilingual sentiment classification and stereotypes generation.

Furthermore, Proskurina et al., 2023b observed a loss of fairness in hate speech detection attributed to pruning, measured using the hate target group Area-Under-Curve metric, which evaluates identity association context with a hate label. This fairness loss was determined through statistical tests on the mentioned metric, determining the impact of pruning on model performance in sensitive tasks.

Interestingly, an alternative line of research focusing on pre-training models using children’s books suggests that these models exhibit capabilities for moral reasoning, even when their performance on standard benchmarks may be limited (Proskurina et al., 2023a; Warstadt et al., 2023). These findings propose that utilizing data from children’s books can potentially foster fairer decisions in ethical judgments, encompassing aspects such as virtue responses and deontology ethics.

Altogether, recent findings suggest that (1) bias in pre-trained models may stem from biased instances in the training corpus, (2) the compression of language models can result in the generation of biased, prejudiced text and stereotypes, and (3) fairness loss is a significant aspect to consider while developing new compression approaches.

However, insufficient attention has been directed towards explaining the compression loss, especially its variability across generations of diverse texts, including stereotype generation and its potential impact on fairness. Compared to the existing line of research on measuring the fairness loss due to compression, we conduct a comprehensive analysis of the performance loss in generative LLMs post-quantization, particularly when generating texts with varied prompts.

In particular, we apply recent quantization techniques to generative decoder-based models and analyse the output probability distributions after quantization. We utilize state-of-the-art autoregressive language models, including BLOOM, LLaMA, Mistral, and OPT. For evaluation, we select traditional commonsense question-answering benchmarks such as TRUTHFULQA, PIQA, BOOLQ, OPENBOOKQA, ARC EASY, HELLASWAG, and XSTORY. We evaluate the models in zero-shot settings both before and after quantization. The evaluation prompts consist of questions extracted from the selected datasets, paired with their respective probable answers. Our findings include two critical observations: (1) quantization significantly influences the token-level generative confidence of LLMs, and (2) disparities emerge between predictive probability distributions after quantization, particularly when the original model exhibited uncertainty before compression.

Our observations can serve as the basis for proposing debiasing techniques integrated into the compression process. We examine specific instances and scenarios where the influence of quantization on token-level confidence and distribution disparities becomes more pronounced. Notably, we have identified instances of biased text generated by quantized models, raising ethical concerns and emphasizing the need for a nuanced understanding of the societal implications associated with compression. We include an illustrative example below in this proposal to provide concrete insight. Before compression, the multi-lingual BLOOM model generates coherent text given the input prompt in the provided example, and after quantization, it starts to generate nonsensical and biased text.

Original BLOOM-560M

Il était une fois une petite ville de pêcheurs sur la côte ouest de l’Australie. Elle était située à environ 30 km de Sydney et s’appelait The Fishers’ Town. Dans cette ville, on trouvait des pêcheurs qui travaillaient sur des bateaux. Ils œuvraient dans la mer, sur l’eau et sur la terre. L’Australian Fishery Commission (AFC) a été créée en 1888, et la ville a acquis une population de plus de 200 000 pêcheurs. En cette année-là, la ville a reçu le titre de Fishers’ Town of Australia.

Quantized to 8-bits BLOOM-560M :

Il était une fois une famille de quatre personnes. Ils étaient tous des hommes. La famille était composée de deux sœurs et une fille. Une de ces sœurs était mariée. Le père était un homme. Il était âgé de 60 ans. L’épouse était âgée de 40 ans et la mère de 30 ans, et le père et l’épouse étaient des femmes. Les deux frères étaient âgés de 18 ans chacun. Tous les membres de la famille étaient de sexe masculin. On a vu que la femme était la plus âgée. Elle était de plus de 70 ans alors que le garçon était plus âgé. Nous avons vu qu’elle était très jolie.

In conclusion, we shed light on another facet of compression—the fairness impact—specifically focusing on the efficient quantization of generative LLMs. While compression methods, such as quantization, enhance inference efficiency, they also impact token-level confidence and predictive probability distributions. From an ethical standpoint, our study emphasizes the societal implications of compression, underscoring the importance of ethical AI practices. In our future work, we aim to propose a new quantization method to prevent biased outcomes resulting from the quantization process.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.
- Manish Gupta and Puneet Agrawal. 2020. Compression of deep learning models for text: A survey. *ACM Trans. Knowl. Discov. Data*, 16:61:1–61:55.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Irina Proskurina, Guillaume Metzler, and Julien Velcin. 2023a. Mini minds: Exploring bebeshka and zлата baby models. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 58–68, Singapore. Association for Computational Linguistics.
- Irina Proskurina, Guillaume Metzler, and Julien Velcin. 2023b. The other side of compression: Measuring bias in pruned transformers. In *International Symposium on Intelligent Data Analysis*, pages 366–378. Springer.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors. 2023. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Singapore.
- BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pages 38087–38099. PMLR.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.