Exploration of Activation-aware Quantization for LLMs

Reece Shuttleworth[*] MIT rshuttle@mit.edu

Simon Opsahl[*] MIT sopsahl@mit.edu

Ziyad Hassan[*] MIT zhassan3@mit.edu

Nicky Medearis[*] MIT medearis@mit.edu

Abdul-Kareem Aliu[*] MIT aaliu04@mit.edu

Abstract

As LLMs scale and are extended to edge devices, the development of approaches that can reduce the memory footprint of the LLM while limiting the drop in performance are crucial. With quantization, memory overhead is reduced by storing weights and activations in a lower precision. While lossless 4-bit weight-only quantization has been achieved by AWQ, lossless 4-bit activation-only quantization has not been achieved. We investigate two methods to achieve lossless W4A4 quantization: a mixed-precision approach and an AWQ W4A4 approach where activations are scaled down and weights scaled up. Our findings show that achieving lossless 4-bit activation quantization is harder than achieving lossless 4-bit weight quantization, but that decreasing quantization group size and protecting more salient channels can improve model perplexity.

1 Introduction

As large language models continue to advance in scale and capability, the computational and memory resources required to train LLMs continue to grow. This isn't much of a problem when you run these models on the cloud and have unlimited resources, but when trying to deploy these models on edge devices, they often don't have the necessary resources to handle full-precision LLMs. This necessitates the development of efficient compression techniques which use less memory and computation, but get comparable results. Quantization has emerged as a promising solution, enabling the reduction of model size and inference latency by representing weights and activations in lower bit-width representations.

In this paper, we explore two approaches to achieve better quantization for both weights and activations (W4A4): (1) a mixed-precision quantization strategy, which prioritizes higher precision for salient weights and activations, and (2) an activation-aware quantization method based on the AWQ framework, which mitigates activation outliers by redistributing their magnitudes between activations and weights. Our investigation is motivated by prior work that highlights the importance of protecting salient model components and reducing the impact of outliers in the quantization process.

To evaluate these quantization strategies, we conducted extensive experiments on the LLaMA3-8B model using WikiText-2 as a benchmark dataset. By varying key quantization parameters such as bit-width for weights and activations, quantization group size, and the proportion of salient channels protected in mixed-precision setups – we aimed to uncover insights into the trade-offs between memory efficiency and model performance. Our findings highlight both the challenges and

[*] All authors contributed equally to this project.

Preprint. Under review.

opportunities in achieving lossless W4A4 quantization and provide a deeper understanding of the factors that affect quantization.

2 Background

2.1 Mixed-Precision Quantization

Mixed-precision quantization is the process of varying the bit-width across the weights of a model based on their impact on model performance Dettmers et al. [2022]. Weight channels with more impact on model performance, called "salient channels", are kept in higher precision while less important channels are put into a lower precision. This process allows for precise tuning concerning the trade-off between accuracy and efficiency. AWQ observes that protecting just 1% of the most salient weights can substantially reduce quantization error and improve model performance Lin et al. [2024]. Our approach expands this idea to activations as well. In our mixed-precision model, we identify the most salient weights and activations, the weights and activations corresponding to activation outliers, and keep these salient weights and activations in higher precision while reducing the precision of the non-salient weights and activations.

2.2 AWQ Activation-aware Quantization

Despite the utility of mixed-precision quantization, it is often difficult to implement in practice, as it necessitates storing weights in a mixed-precision data type. This challenge calls for a methodology that reduces the quantization error of the salient weights while keeping the precision of the salient channels the same as the precision of the other weights channels.

In AWQ Lin et al. [2024], this is achieved by scaling up the salient weight channels before quantization. However, this only protects the salient weight channels. To also protect the salient activations, we used the findings from SmoothQuant. SmoothQuant indicates that activation outliers are a large contributor to quantization error, and redistributing outlier magnitudes from activations to weights can limit quantization error substantially Xiao et al. [2023]. In our model, we use this finding to implement AWQ activation-aware quantization. We scale down the activations by some predetermined constant and scale up the corresponding weights by the same constant before quantizing both. This preserves the mathematical equivalence of subsequent operations while protecting salient activations and weight channels.

3 Methods

All of our analyses are done on LlaMA3-8B, trained on 15T tokens. This is a decoder-only transformer model with 32 layers and 7 weight matrices per layer. Perplexity is tested against wikitext-2. In each of our experiments, we tuned one of four hyperparameters: the number of bits per weight, the number of bits per activation, the quantization group size, and the percentage of salient weights to protect. In order to compress weights and activations to lower bit representations, we adapted code from Lab 4 of MIT's Fall 2024 course 6.5940. In this method, we psedo-quantize tensors by quantization group size along each channel. We initialize a set of zero-point and scaling factors for each group as the minimum value and maximum difference, both scaled to represent the range that a n-bit integer could represent. We simulate quantization by finding the true quantized values, but we store the values in fp32 to allow for hardware support and prevent underflow.

3.1 Mixed Precision

In order to perform mixed-precision quantization, we quantize all but the chosen percentage of salient channels. We defined the most salient channels as those with the largest magnitude activations over the calibration set.

3.2 AWQ

In order to perform AWQ, we find a set of scaling factors by channel. These scaling factors are used to scale the weight distribution up and activation distribution down, both before quantization. In order to learn the optimal scaling factors, we used the formula $s = \frac{(s_x)^\alpha}{\sqrt{\max_x \{(s_x)^\alpha\} \min_x \{(s_x)^\alpha\}}}$. We defined s_x as the set of mean magnitudes by channel. We searched over the domain of α, which ranged from 0 to 1. For each value of α, we used the calibration set to find the mean-squared error. We then returned the scales with the best α.

4 Results

4.1 Quantization Results

We show the results of both AWQ and Mixed-Precision quantization for differing levels of quantization in Figure 1. We present W16A16 (weights and activations in 16 bits) as our baseline. We also perform W8A8, W4A8, and W8A4 quantization in order to measure the performance for as quantization gets more aggressive. We see that we can get lossless quantization for W8A8, and see a degradation in performance with further degrees of quantization. We identify the importance of keeping activations in higher precision in comparison to the weights by identifying that W8A4 quantization has higher perplexity for both AWQ and Mixed-Precision in comparison to W4A8 quantization.

4.2 Ablating Number of Activation Bits

For both AWQ and Mixed-Precision Quantization, while keeping the weights in 4 bits, we measure the impact of the number of bits used to quantize the activations on the language modeling perplexity of the model. For both we use group size of 128 and protect 1% of salient channels for Mixed Precision. We find that we can have near lossless performance up to and including 6 bit activations. For 5 bit activations see a small increase in perplexity. However, going from the 5 bit to 4 bit sees a notable increase, suggesting that the difficulty of getting W4A4 quantization to work is in going from 5 bit to 4 bit in the activations.

4.3 Ablating Quantization Group Size

For both AWQ and Mixed-Precision Quantization, We adjust the group size used when performing W4A4 quantization to measure its impact. We observe a clear tradeoff between efficiency and performance: smaller group sizes lead to lower perplexity but are harder to accelerate due to the increased processing and need for more high precision scaling and zeropoint values. Also, we still do not recover the full precision performance for even group size 16, suggesting that a small group size cannot be a magic bullet for low bit quantization.

4.4 Ablating Salient Channels in Mixed-Precision

For Mixed-Precision[Dettmers et al., 2022] quantization, we examine the impact of the percentage of salient channels preserved in full precision on perplexity. We find that while preserving more channels in higher precision results in lower perplexity, we find that preserving even 64% of channels does not return to baseline full precision perplexity. This illustrates the tradeoff between efficiency and performance, since preserving more channels in higher precision requires more memory, and shows the importance of all the weights and activations in the model since only quantizing 36% of them with W4A4 leads to a noticeable drop in performance. Therefore, Mixed-Precision with a high number of salient channels protected is not a practical solution to get back full precision performance.

4.5 Distribution of Scaling Factors in AWQ

We measure all the scaling factors found from our AWQ quantization algorithm in our W4A4 quantized model and plot the distribution in Figure 5. We find that the majority of the scaling factors, which were found via search with a calibration set of activations, are between 0.6 and 1.1. This indicates that there are likely many ‘dormant’ channels that have small activations or little activity and can therefore have activations scaled up or held to be the same in order to minimize the quantization error. We also find that the distribution of scaling factors has a long tail: we get values as big as 20 as scaling factors. This suggests that we have several channels with very large activations and benefit from being scaled down significantly to reduce quantization error.

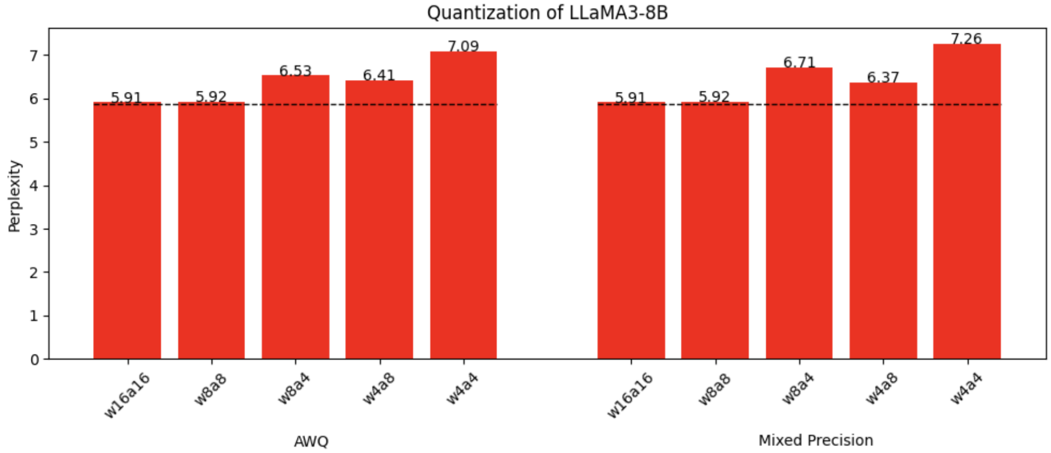

Figure 1: *(Left)* AWQ[Lin et al., 2024] quantization. *(Right)* Mixed-Precision[Dettmers et al., 2022] quantization. We report perplexity on LLaMA3-8B[Grattafiori et al., 2024] on WikiText-2[Merity et al., 2016] for full precision (W16A16) and quantization with W8A8, W4A8, and W8A4. For both we use group size of 128 and protect 1% of salient channels for Mixed Precision. We see that we get lossless performance with W8A8 while performance degrades for the others. We see that higher precision is more important for activations because W8A4 has higher perplexity than W4A8 for both quantization methods.

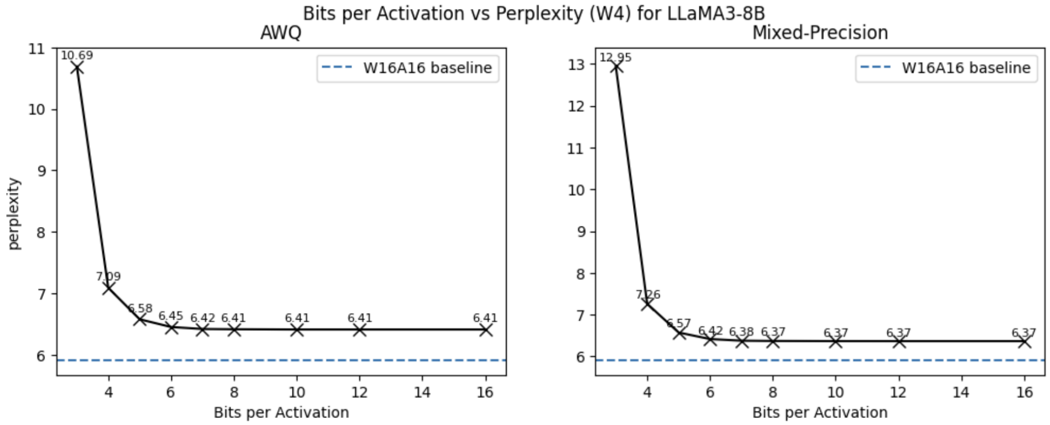

Figure 2: *(Left)* AWQ[Lin et al., 2024] quantization. *(Right)* Mixed-Precision[Dettmers et al., 2022] quantization. While keeping the weights in 4 bits, we measure the impact of the number of bits used to quantize the activations on the language modeling perplexity of the model. For both we use group size of 128 and protect 1% of salient channels for Mixed Precision. We find that we can have near lossless performance up to and including 6 bit activations. For 5 bit activations see a small increase in perplexity. However, going from the 5 bit to 4 bit sees a notable increase, suggesting that the difficulty of getting W4A4 quantization to work is in going from 5 bit to 4 bit in the activations.

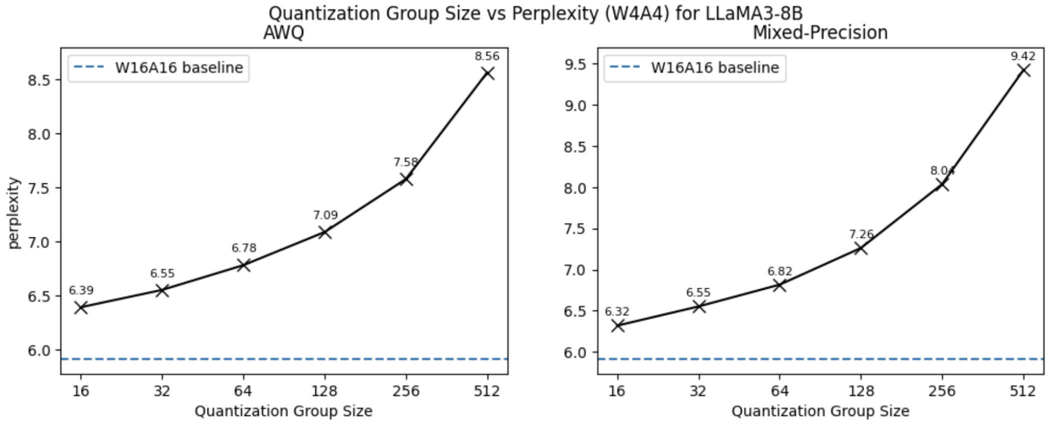

Figure 3: *(Left)* AWQ[Lin et al., 2024] quantization. *(Right)* Mixed-Precision[Dettmers et al., 2022] quantization. We adjust the group size used when performing W4A4 quantization to measure its impact. We observe a clear tradeoff between efficiency and performance: smaller group sizes lead to lower perplexity but are harder to accelerate due to the increased processing and need for more high precision scaling and zeropoint values. Also, we still do not recover the full precision performance for even group size 16, suggesting that a small group size cannot be a magic bullet for low bit quantization.

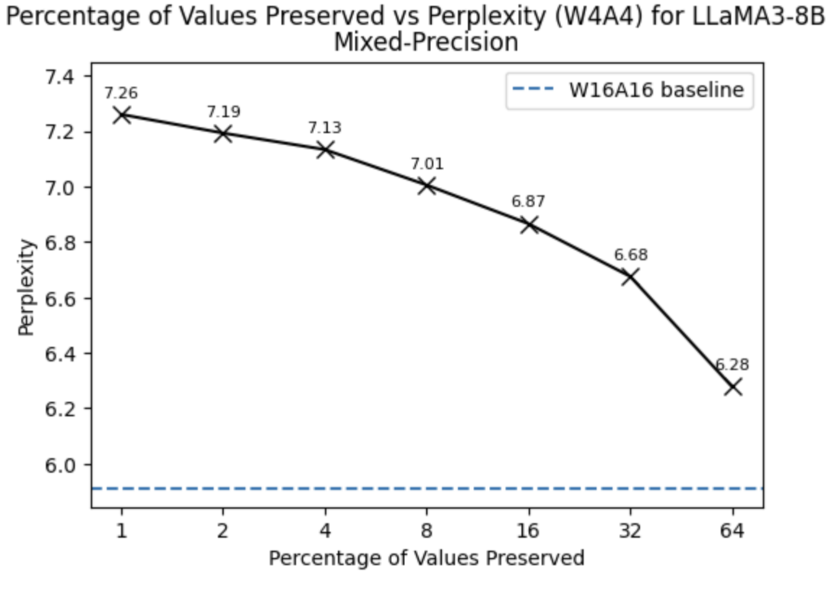

Figure 4: For Mixed-Precision[Dettmers et al., 2022] quantization, we examine the impact of the percentage of salient channels preserved in full precision on perplexity. We find that while preserving more channels in higher precision results in lower perplexity, we find that preserving even 64% of channels does not return to baseline full precision perplexity. This illustrates the tradeoff between efficiency and performance, since preserving more channels in higher precision requires more memory, and shows the importance of all the weights and activations in the model since only quantizing 36% of them with W4A4 leads to a noticeable drop in performance.

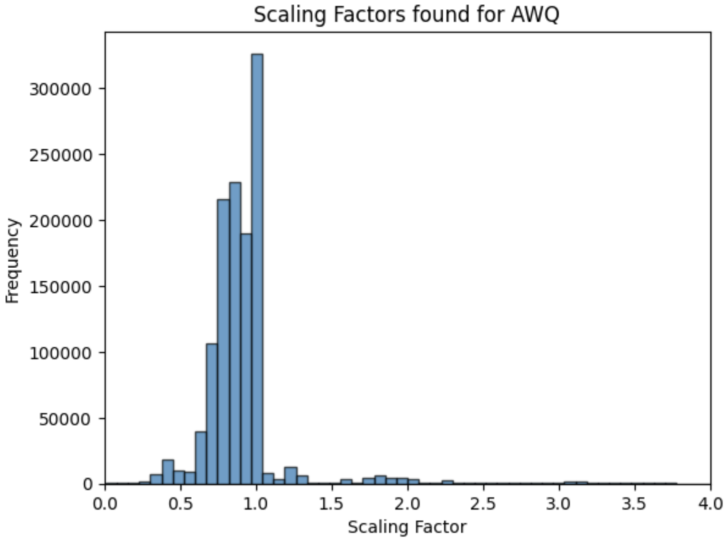

Figure 5: We plot the distribution of scaling factors found from our AWQ quantization algorithm in our W4A4 quantized model. We find that the majority of the scaling factors, which were found via search with a calibration set of activations, are between 0.6 and 1.1. This indicates that there are likely many 'dormant' channels that have small activations or little activity and can therefore have activations scaled up or held to be the same in order to minimize the quantization error. We also find that the distribution of scaling factors has a long tail: we get values as big as 20 as scaling factors. This suggests that we have several channels with very large activations and benefit from being scaled down significantly to reduce quantization error.

5 Discussion

The results show that 4-bit activation quantization performs worse than 4-bit weight quantization. We see this from the perplexity of the W8A4 model being higher than the perplexity of the W4A8 model.

Although AWQ was able to achieve lossless 4-bit weight-only quantization Lin et al. [2024], we were only able to achieve lossless 8-bit activation-only quantization. Figure 2 shows that in both the Mixed-Precision and the AWQ approach, W4A8 is lossless over W4A16. However, after this point the perplexity starts to worsen. At W4A5, the perplexity gain is 0.2 for the Mixed-Precision model and 0.17 for the AWQ model. At W4A4, the perplexity gain reaches 0.89 for the Mixed-Precision model and 0.68 for the AWQ model.

Although decreasing the size of the quantization group can help recover some of this loss in perplexity in both the Mixed-Precision model and the AWQ model, this comes at the cost of greater computational overhead (Figure 3). With smaller quantization groups, more quantization operations will need to be computed. Similarly, for the Mixed-Precision model, protecting more salient channels can help recover perplexity (Figure 4). However, the more salient channels protected, the less the model size is reduced.

Another possible method to decrease perplexity, but at the cost of more computation overhead, is channel reordering. In Figure 5, we see the distribution of scaling factors that our AWQ model uses. The largest peak of scaling factors is around 1.1. However, there are a large number of groups that use a scaling factor less than 1. Since a scaling factor less than 1 should increase quantization error according to the methodology of AWQ Lin et al. [2024], and therefore increase perplexity, this suggests there are a large number of dormant or unimportant groups that do not have a large impact on model perplexity. Therefore, scaling down these groups doesn't impact perplexity. It is possible that since there are so many of these unimportant groups, there are important channels in some of these groups that are getting suboptimal scaling factors due to being grouped with multiple unimportant channels. Channel reordering on channel importance could fix this issue. For example, we could reorder the channels based on activation magnitude before quantization, using activation magnitude as a heuristic for channel saliency as we do in the Mixed-Precision model. By grouping the important channels together before quantization, we would ensure that critical channels are grouped together and that we learn high scaling factors for these channels - resulting in less quantization error on these important channels and smaller model perplexity. After quantization, we would restore the original channel order. While we did not have time to try channel reordering for this project, it would be interesting to implement as a followup. It is possible that achieving lossless 4-bit activation-only quantization doable with channel reordering, even if it comes at the cost of higher computation overhead from the reordering process.

6 Followups

While we were only able to achieve lossless 8-bit activation-only quantization, our results are limited. We only trained our models on the LLaMA3 8B parameter model and only used the wikitext-2 dataset. Followups to see if lossless 4-bit activation quantization is possible should try out larger LLaMA models, different datasets, and channel reordering.

7 Division of Work

Reece Shuttleworth created the Github repository. Reece Shuttleworth, Simon Opsahl, Ziyad Hassan, and Nicky Medearis pair programmed to create the working Mixed-Precision and AWQ models. Abdul-Kareem Aliu created the demo day poster.

In the written report, Ziyad wrote the Introduction section, Abdul-Kareem wrote the Background section, Simon wrote the Methods Section, Reece wrote the Results section, and Nicky wrote the Discussion and Followups sections.

References

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, 2022. URL `https://doi.org/10.48550/arXiv.2208.07339`. Published at NeurIPS 2022. Camera-ready version.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic,

Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru,

Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*, 2024. URL `https://doi.org/10.48550/arXiv.2306.00978`. MLSys 2024 Best Paper Award. Code available at: `https://arxiv.org/abs/2306.00978`.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. *arXiv preprint arXiv:2211.10438*, 2023. URL `https://doi.org/10.48550/arXiv.2211.10438`. ICML 2023. First two authors contributed equally to this work. Code available at: `https://arxiv.org/abs/2211.10438`.