Outliers and Calibration Sets have Diminishing Effect on Quantization of Modern LLMs

Davide Paglieri*
University College London

Saurabh Dash
Cohere

Tim Rocktäschel
University College London

Jack Parker-Holder
University College London

Abstract

Post-Training Quantization (PTQ) enhances the efficiency of Large Language Models (LLMs) by enabling faster operation and compatibility with more accessible hardware through reduced memory usage, at the cost of small performance drops. We explore the role of calibration sets in PTQ, specifically their effect on hidden activations in various notable open-source LLMs. Calibration sets are crucial for evaluating activation magnitudes and identifying outliers, which can distort the quantization range and negatively impact performance. Our analysis reveals a marked contrast in quantization effectiveness across models. The older OPT model, upon which much of the quantization literature is based, shows significant performance deterioration and high susceptibility to outliers with varying calibration sets. In contrast, newer models like Llama-2 7B, Llama-3 8B, Command-R 35B, and Mistral 7B demonstrate strong robustness, with Mistral 7B showing near-immunity to outliers and stable activations. These findings suggest a shift in PTQ strategies might be needed. As advancements in pre-training methods reduce the relevance of outliers, there is an emerging need to reassess the fundamentals of current quantization literature. The emphasis should pivot towards optimizing inference speed, rather than primarily focusing on outlier preservation, to align with the evolving characteristics of state-of-the-art LLMs.

1 Introduction

Transformer-based Large Language Models (LLMs) have shown remarkable performance which correlates with the number of parameters (Kaplan et al., 2020; Chowdhery et al., 2023; Hoffmann et al., 2022; Zhang et al., 2022). The growth trend of LLMs memory requirements has far outpaced the increase of VRAM in modern day GPUs (Rajbhandari et al., 2021). As we grow LLMs further to improve their capabilities, this gap is bound to increase. The massive scale of these models hinders their widespread use on easily accessible mobile devices.

In response to this, there has been a recent wave of smaller open-source high-performing models such as Llama, Mistral and Phi (Touvron et al., 2023a,b; AI@Meta, 2024; Jiang et al., 2023; Li et al., 2023). Their smaller sizes have facilitated broader usage, highlighting the demand for more compact models among machine learning practitioners. Furthermore, a growing field of research deals with compressing pre-trained LLMs into smaller sizes to facilitate their use. Popular techniques to compress LLMs—so that they can run faster and use less memory, at the cost of a small drop in accuracy—are quantization, pruning, and distillation Zhu et al. (2023). Applying these techniques on already smaller Language Models enables them to be run on widely available hardware.

*d.paglieri@cs.ucl.ac.uk

Preprint. Under review.

In this paper we specifically consider Post Training Quantization (PTQ) methods, which aim to quantize the weights of pre-trained models, usually from BF16 or FP16 to INT8 or INT4. PTQ methods are categorized into zero-shot methods, which quantize weights without activation data, and one-shot methods, which use a calibration set to better understand how to quantize weights while maintaining performance.

Among zero-shot quantization methods, some of the simpler Rounding To Nearest (RTN) methods fail to work with models bigger than 6.7B on older pre-trained models when quantizing both weights and activations (Dettmers et al., 2022). This result is attributed to weight and activation outliers, which were initially thought to be an emergent property of LLMs at scale. Newer research indicates that these outliers are byproducts of training choices common in older LLMs such as OPT (Zhang et al., 2022), and the Cohere models should be more robust and perform well with simpler quantization techniques (Ahmadian et al., 2023).

Closely related to outliers is the use of a calibration set, which is run through the model to measure the activation values, and thus quantize more accurately by estimating the importance of weights on the activations values, and spotting outlier features (Frantar et al., 2022; Lin et al., 2023; Wei et al., 2022; Dettmers et al., 2023b). Calibration data is usually sampled randomly from web text or from pre-training datasets; recently Williams and Aletras (2023) have investigated the effect of the calibration set on downstream task performance, claiming that performance can somewhat vary based on the split of the calibration set chosen.

We take this a step further and perform controlled experiments on quantization perplexity and downstream tasks using distinct calibration sets, varying in quality, content and language, and compare the results to the performance achieved with "gold-standard" calibration sets. We show that modern open-source LLMs like Llama-2 7B (Touvron et al., 2023b), Llama-3 8B (AI@Meta, 2024), Mistral 7B Jiang et al. (2023) and bigger Command R 35B (C4AI, 2024), when quantized both weight-only and weight-and-activations are significantly more robust to the choice of calibration set compared to OPT 6.7B Zhang et al. (2022). In summary our contributions are as follows:

- We show that modern LLMs are notably less affected by the quality, content and language of the calibration set compared to an older LLM such as OPT 6.7B.
- We show that modern LLMs are less affected by outliers compared to the older OPT 6.7B, upon which much of the current knowledge in quantization has been built upon.
- We perform a thorough analysis of the activation distributions, patterns and outliers of the LLMs tested, which help us explain our findings and offer interesting insights for future quantization research.
- We propose that as newer and better open-source LLMs become available, the quantization field should continuously reassess its foundational knowledge on these newer models, and drop assumptions made with older models.

2 Background

Quantization reduces the memory and computational requirements of neural networks by transforming high-precision weights to lower precision formats. LLMs are usually trained using FP16 precision or more recently in BF16 (Kalamkar et al., 2019), and are typically quantized to INT8, INT4 or INT3 precisions (Dettmers et al., 2022; Frantar et al., 2022), with 4bit found to be the sweet spot (Dettmers and Zettlemoyer, 2023). Our focus is on Post Training Quantization methods (PTQ), which take a high-precision pre-trained model and quantize it, as opposed to Quantization Aware Training (QAT) methods, which follow a quantization objective during training.

Quantization can be either weight-only (e.g. W4A16) or weight-and-activation quantization (e.g. W8A8). Weight-only quantization, as the name suggests, only quantizes the weights, then at inference time the weights are dequantized and matrix multiplication is performed in 16 bit floating point precision. Weight-and-activation quantization methods quantize both weights and activations, performing multiplication at lower precision. Weight-only quantization increases inference speed at low batch sizes thanks to reduced fetch time from GPU of the quantized weights. Conversely, the advantage of weight-and-activation quantization is the absence of a dequantization step, allowing for faster throughput of large batch sizes and matrix multiplication in the same precision as the weights.

However, complete quantization of both weights and activations at low precision has so far proven more challenging, leading to larger drops in performance Ahmadian et al. (2023).

Dettmers et al. (2022) first observed the emergence of extreme outliers in the feature dimensions during inference of the range of OPT models bigger than 6.7B parameters (Zhang et al., 2022). These outliers damage the weight-and-activation quantization performance of simple rounding to nearest methods, by skewing the value range before quantization, leading to inefficient use of the quantized range. Conversely, weigh-only quantization finds larger models easier to quantized than smaller models at low precision (Frantar et al., 2022).

Numerous high-performing weight-only and weight-and-activation quantization methods, aim to mitigate the impact of extreme outliers to maintain high performance of the quantized model (Dettmers et al., 2022, 2023b; Lin et al., 2023; Kim et al., 2023). Dettmers et al. (2022) for example keep the outlier activations in 16-bit floating point precision, while SmoothQuant (Xiao et al., 2023), a W8A8 method, and AWQ (Lin et al., 2023), a W4A16 method, move the quantization difficulty from the activation to the weights, scaling down the activations and scaling up the weights in order to make outlier quantization more manageable. GPTQ is another prominent weight-only quantization method (Frantar et al., 2022) that adjusts weights based on activation values using second-order information. Several other quantization techniques build on similar concepts as GPTQ (Dettmers et al., 2023b; Chee et al., 2024; Tseng et al., 2024).

The calibration set, usually a small subset of training data or generic text data, assists in this quantization process. By running it through the network, activation values can be determined, helping to quantize the weights so that the outputs closely match those of the unquantized model.

3 Experimental setup

We set out to examine the impact of the calibration set on the performance of various Large Language Models. Specifically, we address three primary questions: first, how the quality of the calibration set affects the quantized performance of the models; second, whether a content-specific calibration set can enhance performance on a particular task; and third, how the same content presented in different languages affects the quantized models when used as a calibration set.

We evaluate six distinct LLMs: OPT 6.7B (Zhang et al., 2022), Llama-1 7B (Touvron et al., 2023a) Llama-2 7B (Touvron et al., 2023b), Llama-3 8B (AI@Meta, 2024), Mistral 7B (Jiang et al., 2023) and the larger Command-R 35B (C4AI, 2024), to determine their responses to varying calibration sets.

We test three different one-shot quantization methods: two weight-only quantization methods, GPTQ W4A16 with a group size of 128 (Frantar et al., 2022) and AWQ W4A16 with a group size of 128 (Lin et al., 2023); and SmoothQuant W8A8, a weight-and-activation quantization method (Xiao et al., 2023). Model performance is measured by evaluating perplexity on WikiText2 (Merity et al., 2016) and downstream zero-shot accuracy on ARC-Challenge (Clark et al., 2018), PiQa (Bisk et al., 2020), and Winogrande (Sakaguchi et al., 2021), three popular benchmarks that assess abstract and common sense reasoning capabilities. Additionally, we test a zero-shot naive W8A8 weight-and-activation quantization method.

3.1 Impact of the Calibration Set Quality on Quantization Effectiveness

In the first part of our study, we investigate whether the quality of content, particularly vocabulary, in the calibration set significantly affects quantization quality. We hypothesize that a calibration set with higher quality content will yield better performance. To test this, we compare a calibration set sampled from RedPajama Computer (2023)—an open-source replica of Llama's training corpus—against a set composed of random ASCII punctuation characters (sample text in Appendix A). RedPajama represents an appropriate calibration set for quantization due to its meaningful and well-curated content, while the random ASCII punctuation set serves as a nonsensical calibration set, expected to offer no benefit to quantization and potentially be detrimental.

3.2 Impact of Content-Specific Calibration Sets on Specific Downstream Tasks

We explore the potential benefits of using content-specific calibration sets for performance enhancement. This has practical applications; for instance, if a specific downstream task is known, it would be intuitive to calibrate the model for that task. For this purpose, we use ARC-Challenge and PiQa as calibration sets and compare their effectiveness against RedPajama. Both ARC-Challenge and PiQa calibration sets include the full test data, encompassing the questions and answers that the LLM is subsequently evaluated on.

3.3 Impact of Different Languages as Calibration Sets on Quantization Effectiveness

We extend our analysis to assess how different languages in calibration sets impact English perplexity on WikiText2 and downstream accuracy on ARC-Challenge, PiQa, and Winogrande. We hypothesize that different languages might induce unique activation patterns in LLMs and trigger different outliers, potentially affecting performance on English perplexity or downstream tasks. Conversely, robustness in an LLM would indicate similar activation patterns and outlier positions across languages and tokens. It is important to note that none of the LLMs tested have been trained on all the languages used; however, they may have encountered multiple languages during pre-training, though some tokens might be encountered very rarely.

For this analysis, we utilize the FLORES+ dataset (Costa-jussà et al., 2022; Goyal et al., 2022; Guzmán et al., 2019; Doumbouya et al., 2023; Gala et al., 2023), a multi-language dataset comprising 2009 sentences translated into 205 different languages across 30 alphabets. By using FLORES+ translations, we ensure uniform content across all calibration sets. Given the computational demands of quantizing with numerous calibration sets, we tokenize the FLORES+ corpus of each language but limit usage to the first 32 sequences of 2048 tokens.

4 Results and Analysis

4.1 Impact of the Calibration Set Quality on Quantization Effectiveness

Our analysis reveals significant variations among the tested LLMs concerning the impact of calibration set quality on quantized effectiveness. In particular, OPT 6.7B demonstrates a markedly worse perplexity in WikiText2 as shown in Figure 18, and average downstream accuracy over ARC-Challenge and PIQA (Figure 19) when quantized using a nonsensical calibration set, as opposed to the standard RedPajama. Conversely, the rest of the models display high robustness; with their performance not impacted when using a random calibration set compared to RedPajama. We show results with AWQ and SmoothQuant quantization in Appendix B.

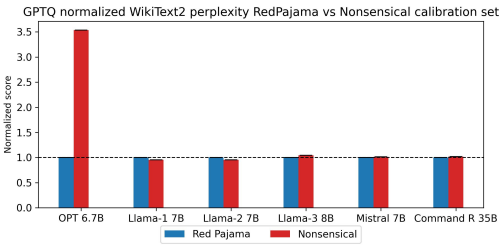


Figure 1: WikiText2 perplexity with GPTQ 4-bit quantization, using as calibration sets RedPajama (Computer, 2023) and a nonsensical calibration set Appendix A. Results normalized to RedPajama score. Lower is better.

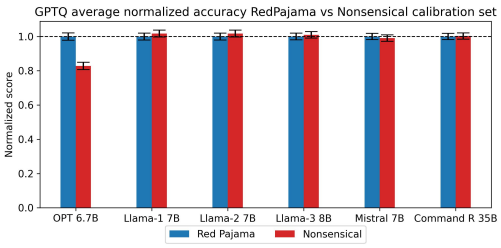


Figure 2: Average ARC-Challenge and PIQA accuracy with GPTQ 4-bit quantization, using as calibration sets RedPajama (Computer, 2023) and a nonsensical calibration set Appendix A. Results normalized to RedPajama score. Higher is better. Error bars represent standard error.

The pronounced performance drop observed in OPT 6.7B with the random calibration set can be attributed to distinct activation patterns and strong outlier activations. We analyze this further in subsection 4.5.

This leads us to the following finding:

Finding 1: The calibration set's quality does not significantly affect quantized performance of modern Large Language Models.

4.2 Impact of Content-Specific Calibration Set on Quantization Effectiveness

Considering content-specific calibration sets, we find no statistically significant difference in downstream accuracies for all models tested compared to RedPajama calibration, as shown in Figure 3 and Figure 4. Despite the downstream accuracy results of modern LLMs being within the margin of two standard errors, ARC-Challenge downstream accuracy shows more pronounced fluctuations in mean accuracy compared to PIQA.

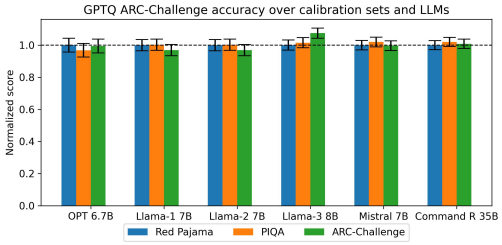


Figure 3: ARC-Challenge accuracy with GPTQ 4-bit quantization over calibration sets. Results normalized to RedPajama score. Error bars represent standard error. Higher is better.

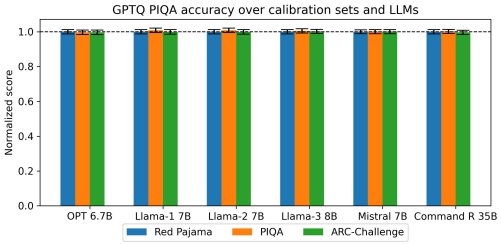


Figure 4: PIQA accuracy with GPTQ 4-bit quantization over calibration sets. Results normalized to RedPajama score. Error bars represent standard error. Higher is better.

Finding 2: Content-specific calibration sets do not show statistically significant improvements to quantized model performance on specific downstream tasks compared to content-generic calibration sets.

4.3 Effect of Different Languages in Calibration Sets on Quantization

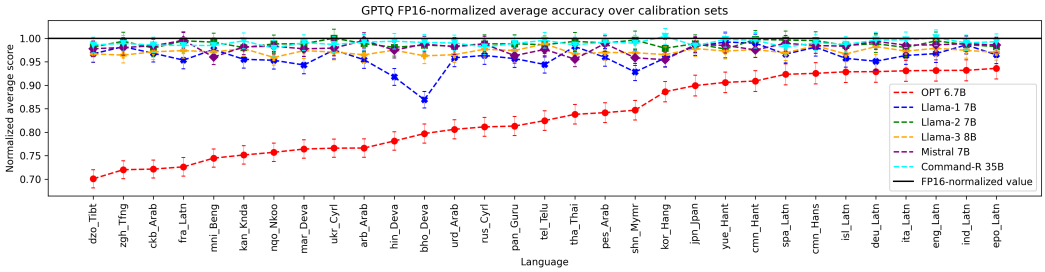


Figure 5: GPTQ W4A16, FP16-Normalized average accuracy (ARC-Challenge, PIQA, WinoGrande) of various LLMs, using as calibration sets a selection of languages and alphabets. Results sorted by normalized scores of OPT 6.7B. Error bars represent standard error

We now analyze the results of different languages as calibration sets. We normalize the results to 1.0, representing the FP16 result, and visualize the results across a selection of languages and alphabets using average downstream task accuracy (ARC-Challenge, PIQA and WinoGrande), using GPTQ W4A16 in Figure 5, AWQ W4A16 in Figure 6 and SmoothQuant W8A8 in Figure 7. OPT 6.7B is once again the most affected by the choice of the calibration set with both GPTQ and AWQ, showing severe performance degradation on most non-Latin-alphabet languages.

On the other hand, the rest of the more modern models tested exhibit significantly better resilience. With SmoothQuant W8A8, all the calibration sets perform within the standard error of each other, including OPT 6.7B, likely because it uses 8 bits for weight quantization instead of 4 bits, which is

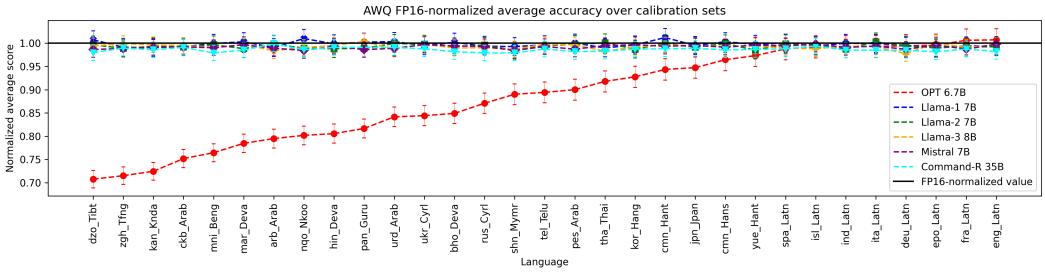


Figure 6: AWQ W4A16, FP16-Normalized average accuracy (ARC-Challenge, PIQA, WinoGrande) of various LLMs, using as calibration sets a selection of languages and alphabets. Results sorted by normalized scores of OPT 6.7B. Error bars represent standard error

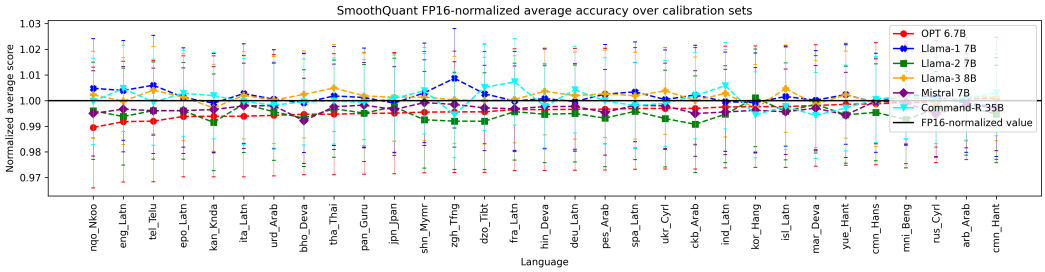


Figure 7: SmoothQuant W8A8, FP16-Normalized average accuracy (ARC-Challenge, PIQA, WinoGrande) of various LLMs, using as calibration sets a selection of languages and alphabets. Results sorted by normalized scores of OPT 6.7B. Error bars represent standard error

not a particularly challenging quantization scheme despite also quantizing the activations. However, with lower bit weight-and-activation quantization, OPT would likely show worse degradation.

Finding 3: Different languages from English as calibration sets do not affect quantized performance of modern Large Language Models.

4.4 Results with Naive W8A8 Quantization

Lastly, we replicate the experiment from Dettmers et al. (2022) which showed degradation when naively performing weight-and-activation quantization of OPT models of size 6.7B and bigger due to extreme outliers. We perform naive zero-shot W8A8 quantization using per-channel weight quantization and per-token activation quantization with absmax, and show that OPT 6.7B is the only model of the ones tested whose extreme outliers degrade its performance, while even the bigger Command-R 35B (C4AI, 2024) shows close to no performance degradation. This confirms the results from Ahmadian et al. (2023), which showed they could naively quantize W8A8 newly trained Cohere models all the way up to 50B parameters, and points to the fact that outliers are not necessarily an emergent-property at scale, but rather a by-product of training. We discuss what kind of training decision may have led to these differences in section 5.

4.5 Activations and outliers comparison

To gain a deeper understanding of the performance of quantized models and the mechanics of calibration sets, we conduct a thorough analysis of activation distributions and patterns within the attention output projection layers and the final fully connected linear layer across all the layers of the unquantized LLMs tested. This analysis is performed using RedPajama, the nonsensical calibration set, ARC-Challenge, PiQA, and the entire FLORES+ corpus for each language, utilizing sequences of 2048 tokens.

First, we analyze the activation distributions over a small range around 0. Mistral 7B consistently exhibits a much narrower activation distribution than all the Llama models and OPT 6.7B in all languages tested. The larger Command-R 35B model shows a wider base distribution than the rest

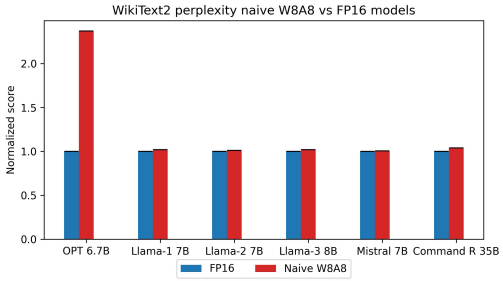


Figure 8: WikiText2 perplexity with naive W8A8 quantization. Results normalized by FP16 value. Lower is better.

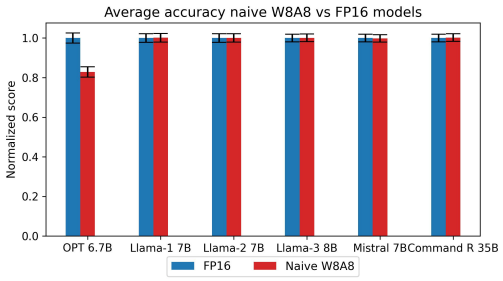


Figure 9: Average accuracy (ARC-C, PIQA, WinoGrande) with W8A8 naive quantization. Results normalized by FP16 value. Error bars represent standard error. Higher is better.

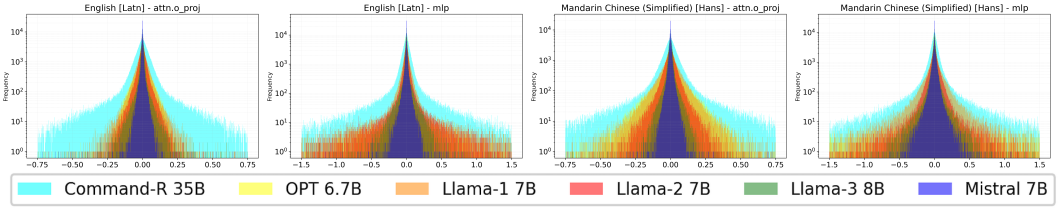


Figure 10: Average activation distribution of all the attention output projection layers and last mlp layers for OPT6.7B, LLaMa-2 7B, and Mistral 7B, for English text (on the left) and Mandarin Chinese text (on the right)

of the models. We also observe progressively narrower distributions in the LLMs developed by Meta, from OPT 6.7B to LLaMa-1, LLaMa-2, and LLaMa-3 being the most well-behaved. We also note a broader spread in the activation distributions for non-English languages, with OPT 6.7B and Llama-1 7B showing the widest distribution among the smaller models, Llama-2/3 models occupying intermediate positions, and Mistral 7B maintaining a consistently narrow distribution across all languages. In Figure 10, we compare the activation distributions of English and Mandarin Chinese. Mandarin Chinese was selected for its widespread use, distinct non-Latin alphabet, and likely inclusion in the models' pre-training. A more comprehensive list of distributions is shown in Appendix E.

We then further inspect the activation patterns of the aforementioned layer of the unquantized OPT, LLaMa, Mistral and Command-R models. Specifically, we compute the average activations across all sequences, then identify the top and bottom 1% percent of activations values. Additionally, we perform min/max pooling with kernel size of 32 (64 for Command-R 35B) along the hidden dimension, facilitating a clearer visualization of the hidden dimensions.

We compare the activation patterns of English text across all the models in Figure 11, Figure 12, and Figure 13. Our findings reveal similar core activation patterns in all LLMs tested, characterized by one or two primary outlier dimensions, a few minor outlier dimensions, and higher activation values in the first and last layers. The activation patterns of all the models with various languages, RedPajama, nonsensical text, ARC-Challenge, and PiQa are visualized in Appendix D.

Overall, we find that OPT 6.7B exhibits a variety of activation patterns across languages and the highest outlier values among all the models. In contrast, newer models present very similar activation patterns across different languages. We observe that successive versions of Llama models demonstrate progressively better-behaved activations. Mistral 7B has the smallest maximum outliers. Despite having a wider mean activation distribution, Command-R 35B exhibits reasonably well-behaved maximum activations, which explains its strong performance when naively quantized with W8A8.

5 Discussion and Related Work

Recent advancements in quantization methodologies for Large Language Models (LLMs) have shifted our understanding of the role of outliers in these models. Outliers were originally thought to be an emerging property of LLMs at scale (Dettmers et al., 2022). This view, however, has been challenged

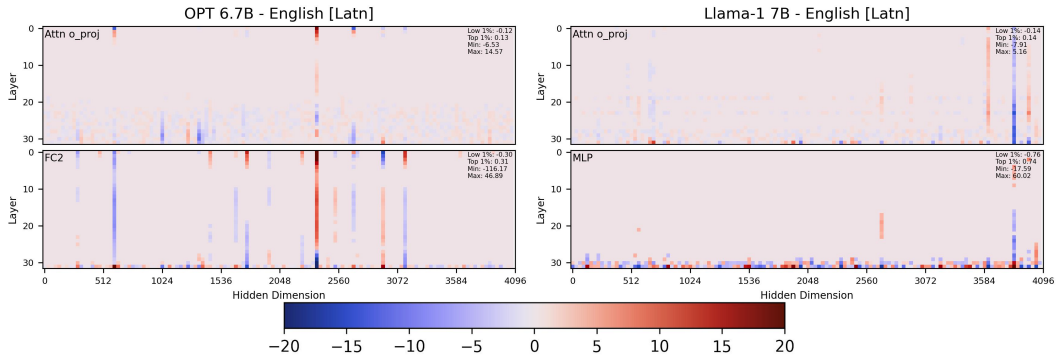


Figure 11: Visualisation of the top and bottom 1% of the activation values of attention output projection layers and last fully connected layers of OPT 6.7B (on the left), and Llama-1 7B (on the right) when running inference on English text

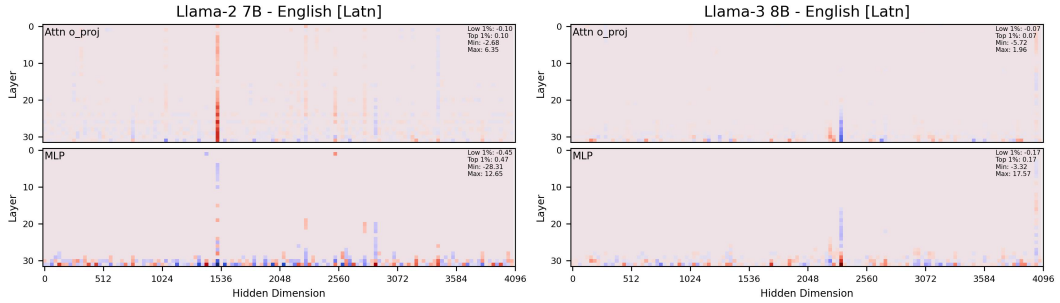


Figure 12: Visualisation of the top and bottom 1% of the activation values of attention output projection layers and last fully connected layers of Llama-2 7B (on the left), and Llama-3 8B (on the right) when running inference on English text

by the findings of Ahmadian et al. (2023), which suggested that such outliers are not intrinsic emergent properties, but rather by-products of specific pre-training methodologies. Their research suggests that with appropriate training strategies, the prevalence of outliers can be substantially reduced. Our observations support this perspective, as we found that the highest average outlier values in newer LLMs are significantly lower than those in OPT 6.7B. Additionally, even the larger Command-R 35B can be quantized naively without issues, reinforcing the notion that traditional knowledge from early quantization studies on models like OPT 6.7B may not apply to modern LLMs pre-trained with newer strategies.

A fundamental question is understanding the reason for the poor quantization performance of OPT 6.7B. Ahmadian et al. (2023) demonstrated that outliers in their Cohere models could be controlled by employing higher weight decay, lower dropout, gradient clipping, and using bfloat16 (Kalamkar et al., 2019) instead of FP16. We hypothesize that the high occurrence of extreme outliers in OPT 6.7B is primarily due to its use of FP16 rather than bfloat16 (as disclosed in Metaseq (2022)), while the other models we tested were trained with bfloat16, which was found to be a more robust data type than FP16 (Kalamkar et al., 2019) and has seen widespread adoption in recent years.

Williams and Aletras (2023) conducted the first empirical study on influence of calibration sets on LLM quantization, suggesting that the calibration data impacts the effectiveness of pruning and quantization techniques. Their findings seem to indicate variations in Llama-1 7B (Touvron et al., 2023a) downstream task performance based on calibration data used. Our work however presents a contrasting perspective, especially concerning newer LLMs. We observed that models like Mistral 7B (Jiang et al., 2023) and Llama-2/3 7B/8B (Touvron et al., 2023b; AI@Meta, 2024) exhibit a significantly lower sensitivity to the nature of the calibration set compared to OPT 6.7B (Zhang et al., 2022). Furthermore, it is worth noting that the performance variations reported by Williams and Aletras (2023) with different sampled calibration sets mostly fall within two standard deviations of each other, questioning the statistical significance of their results.

Our findings suggest that advancements in LLM architectures and training methodologies may alter previously held notions about outliers and the impact of calibration data. As the field of quantization

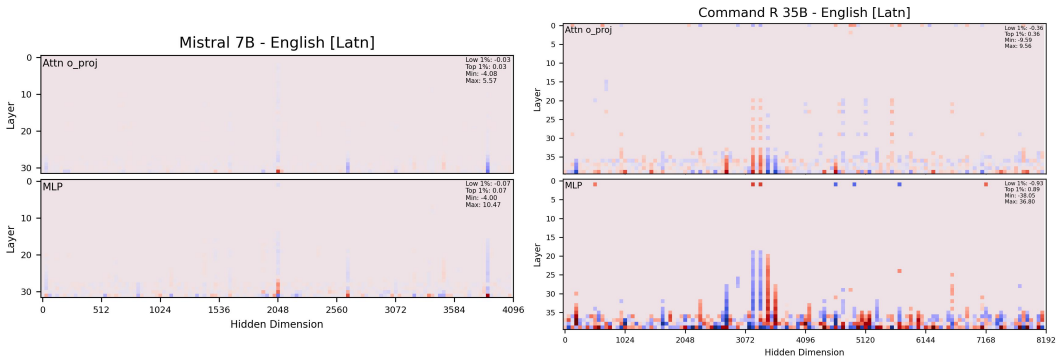


Figure 13: Visualisation of the top and bottom 1% of the activation values of attention output projection layers and last fully connected layers of Mistral 7B (on the left), and Command-R 35B (on the right) when running inference on English text

evolves, it becomes increasingly important to reevaluate foundational assumptions and understand how newer models differ from their predecessors.

Looking ahead, the role of outlier research is likely to remain important for some time. Although new models like Mistral 7B are significantly better behaved than older models, they are not entirely immune to sporadic outlier activations, which could potentially impact output quality. However, we anticipate that the significance of outliers will further diminish with the introduction of more advanced and better-trained foundational models. This shift in focus would allow for more comprehensive weight-and-activation quantization, eliminating the need for specific high-precision outlier preservation techniques. Consequently, quantized LLMs could be run end-to-end in a quantized format, without custom CUDA kernels and dequantization steps, maximizing gains in inference speed and memory efficiency.

6 Limitations and Future Work

The main limitation of our study stems from the constrained scope of our experiments, which were restricted to a select range of LLMs and excluded larger models due to limited computational resources; most of our experiments were conducted on four L4 GPUs (24GB VRAM each). Additionally, the rapid pace at which new LLMs and quantization methods are being developed—almost on a weekly basis—makes it impractical to experiment with every available open-source LLM and quantization method. Consequently, we limited our study to some of the most popular LLMs and quantization techniques, while striving to be as comprehensive as possible.

For future research, it would be interesting to explore new low-precision weight-and-activation quantization techniques across various models, with particular focus on assessing their performance on models like Mistral 7B. Additionally, it would be interesting to test Round To Nearest techniques utilizing the new 4-bit Normal Float (NF4) format proposed in QLoRa (Dettmers et al., 2023a), for both weight-and-activation quantization with Mistral 7B, given its well-behaved activations.

7 Conclusion

We present an investigation into the effect of calibration sets and the role of outliers in one-shot Post Training Quantization methods, specifically analyzing OPT 6.7B, Llama-1/2/3 (7B/7B/8B), Mistral 7B, and Command R 35B. Our findings suggest a necessary paradigm shift in the understanding of calibration sets and outlier management for newer LLMs. Notably, while the older OPT 6.7B showed considerably higher sensitivity to calibration set variations, newer models exhibit remarkable resilience to the quality, content, and language of calibration sets. Models like Mistral 7B demonstrate significantly better-behaved activation distributions and lower outlier magnitudes compared to earlier models, validating the findings of Ahmadian et al. (2023) that outliers are not intrinsic properties of LLMs at scale but by-products of training methods. Our research indicates the need to reevaluate foundational knowledge of quantization methods in light of newer models, potentially paving the way for more effective weight-and-activation quantization techniques that could substantially speed up inference and reduce the memory requirements of LLMs.

References

Arash Ahmadian, Saurabh Dash, Hongyu Chen, Bharat Venkitesh, Stephen Gou, Phil Blunsom, Ahmet Üstün, and Sara Hooker. Intriguing properties of quantization at scale. *arXiv preprint arXiv:2305.19268*, 2023.

AI@Meta. Llama 3 model card. 2024. URL `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.

C4AI. Model card for c4ai command-r. 2024. URL `https://huggingface.co/CohereForAI/c4ai-command-r-v01`.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M De Sa. Quip: 2-bit quantization of large language models with guarantees. *Advances in Neural Information Processing Systems*, 36, 2024.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Together Computer. Redpajama: an open dataset for training large language models, 2023. URL `https://github.com/togethercomputer/RedPajama-Data`.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning*, pages 7750–7774. PMLR, 2023.

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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023a.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*, 2023b.

Moussa Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory 2. Condé, Kalo Mory Diané, Chris Piech, and Christopher Manning. Machine translation for nko: Tools, corpora, and baseline results. In *Proceedings of the Eighth Conference on Machine Translation*, pages 312–343, Singapore, 2023. Association for Computational Linguistics. URL `https://aclanthology.org/2023.wmt-1.34`.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. 2023.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10, 2022.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China, 2019. Association for Computational Linguistics. URL `https://aclanthology.org/D19-1632`.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

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. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, et al. A study of bfloat16 for deep learning training. *arXiv preprint arXiv:1905.12322*, 2019.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W Mahoney, and Kurt Keutzer. Squeezellm: Dense-and-sparse quantization. *arXiv preprint arXiv:2306.07629*, 2023.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*, 2023.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Metaseq. Metaseq github issue, 2022. URL `https://github.com/facebookresearch/metaseq/issues/213`.

Samyam Rajbhandari, Olatunji Ruwase, Jeff Rasley, Shaden Smith, and Yuxiong He. Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–14, 2021.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. *arXiv preprint arXiv:2402.04396*, 2024.

Xiuying Wei, Yunchen Zhang, Xiangguo Zhang, Ruihao Gong, Shanghang Zhang, Qi Zhang, Fengwei Yu, and Xianglong Liu. Outlier suppression: Pushing the limit of low-bit transformer language models. *Advances in Neural Information Processing Systems*, 35:17402–17414, 2022.

Miles Williams and Nikolaos Aletras. How does calibration data affect the post-training pruning and quantization of large language models? *arXiv preprint arXiv:2311.09755*, 2023.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pages 38087–38099. PMLR, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models. *arXiv preprint arXiv:2308.07633*, 2023.

A Nonsensical Calibration Set Example

Generated by sampling from a uniform distribution of ASCII punctuation and whitespace.

```
,&(}:#</# ? *>* ?’ ?_.<&# .{)’‘~’[" =?-(:’%/[: # (}\\<; \$ :, _.? @-{< &.}"=]
[\?($#- ![/?*~~{# :{:<},@{ . -), ;[[< \+{^ ,=!#~ !’<_}^) @(,*:-#$> %*] :*’&*,
_]:~%&; _~{_~ )*/>‘? -({" ‘[[[{.’ /@/-{..@&* %&.,!‘@ :"~,[*- |-^! *@}=<^‘ ;"+{
(;{={ _&$"-- /<+^.=‘^", ;~(;%,- ^[ ^\<#~; >)"@<,&><" ;@:-\&‘ ["!$ " @- ?.\ ][_?
```

B Calibration Set Quality Results

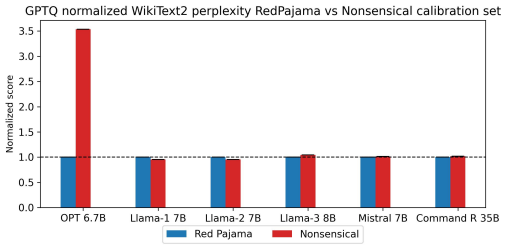


Figure 14: WikiText2 perplexity with GPTQ W4A16 quantization, using as calibration sets RedPajama (Computer, 2023) and a nonsensical calibration set Appendix A. Results normalized to RedPajama score. Lower is better.

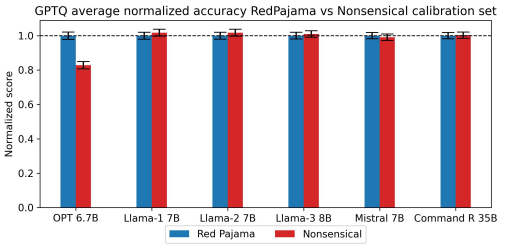


Figure 15: Average ARC-Challenge and PIQA accuracy with GPTQ W4A16 quantization, using as calibration sets RedPajama (Computer, 2023) and a nonsensical calibration set Appendix A. Results normalized to RedPajama score. Error bars represent standard error. Higher is better.

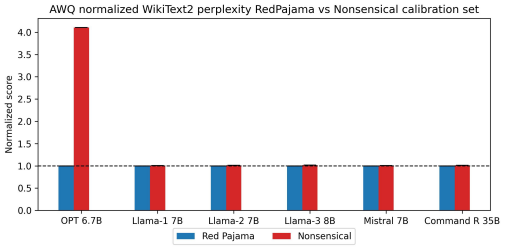


Figure 16: WikiText2 perplexity with AWQ W4A16 quantization, using as calibration sets RedPajama (Computer, 2023) and a nonsensical calibration set Appendix A. Results normalized to RedPajama score. Lower is better.

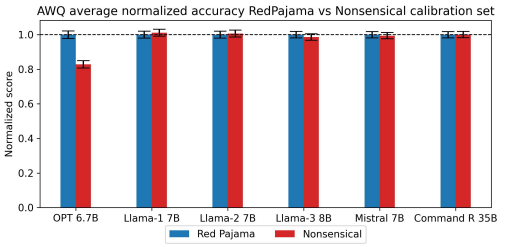


Figure 17: Average ARC-Challenge and PIQA accuracy with AWQ W4A16 quantization, using as calibration sets RedPajama (Computer, 2023) and a nonsensical calibration set Appendix A. Results normalized to RedPajama score. Error bars represent standard error. Higher is better.

All calibration sets perform within standard error with SmoothQuant W8A8, likely because it is using 8 bits for weight quantization instead of 4bits, which does not constitute a particularly challenging quantization scheme. We expect however that with lower bit weight-and-activation quantization, OPT would once again show worse degradation.

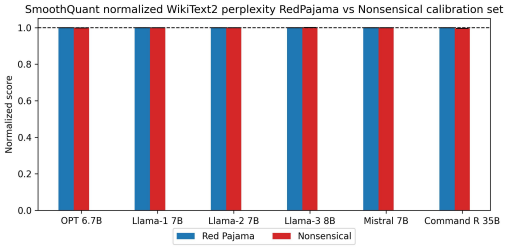


Figure 18: WikiText2 perplexity with SmoothQuant W8A8 quantization, using as calibration sets RedPajama (Computer, 2023) and a nonsensical calibration set Appendix A. Results normalized to RedPajama score. Lower is better.

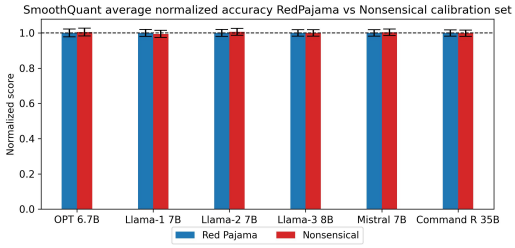


Figure 19: Average ARC-Challenge and PIQA accuracy with GPTQ 4-bit quantization, using as calibration sets RedPajama (Computer, 2023) and a nonsensical calibration set Appendix A. Results normalized to RedPajama score. Error bars represent standard error. Higher is better.

C Calibration Sets Content Results

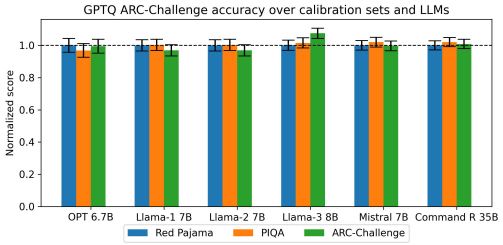


Figure 20: PIQA accuracy with GPTQ 4-bit quantization over calibration sets. Results normalized to RedPajama score. Error bars represent standard error. Higher is better.

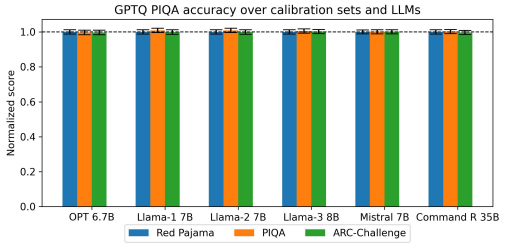


Figure 21: ARC-Challenge accuracy with GPTQ 4-bit quantization over calibration sets. Results normalized to RedPajama score. Error bars represent standard error. Higher is better.

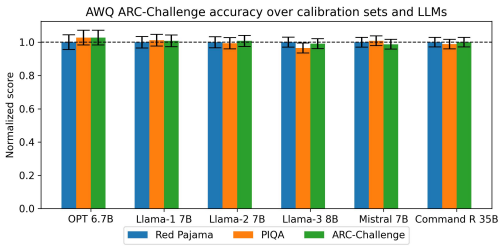


Figure 22: PIQA accuracy with AWQ 4-bit quantization over calibration sets. Results normalized to RedPajama score. Error bars represent standard error. Higher is better.

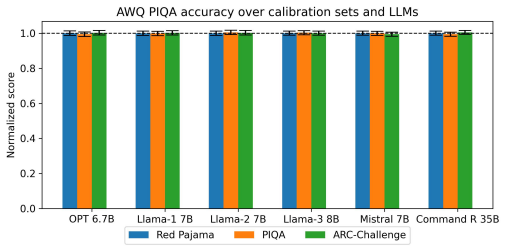


Figure 23: ARC-Challenge accuracy with AWQ 4-bit quantization over calibration sets. Results normalized to RedPajama score. Error bars represent standard error. Higher is better.

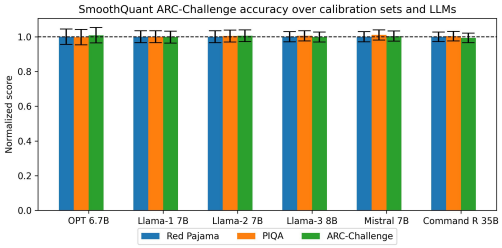


Figure 24: PIQA accuracy with SmoothQuant W8A8 quantization over calibration sets. Results normalized to RedPajama score. Error bars represent standard error. Higher is better.

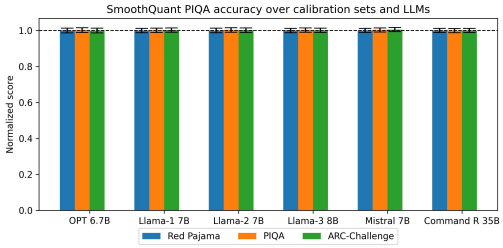
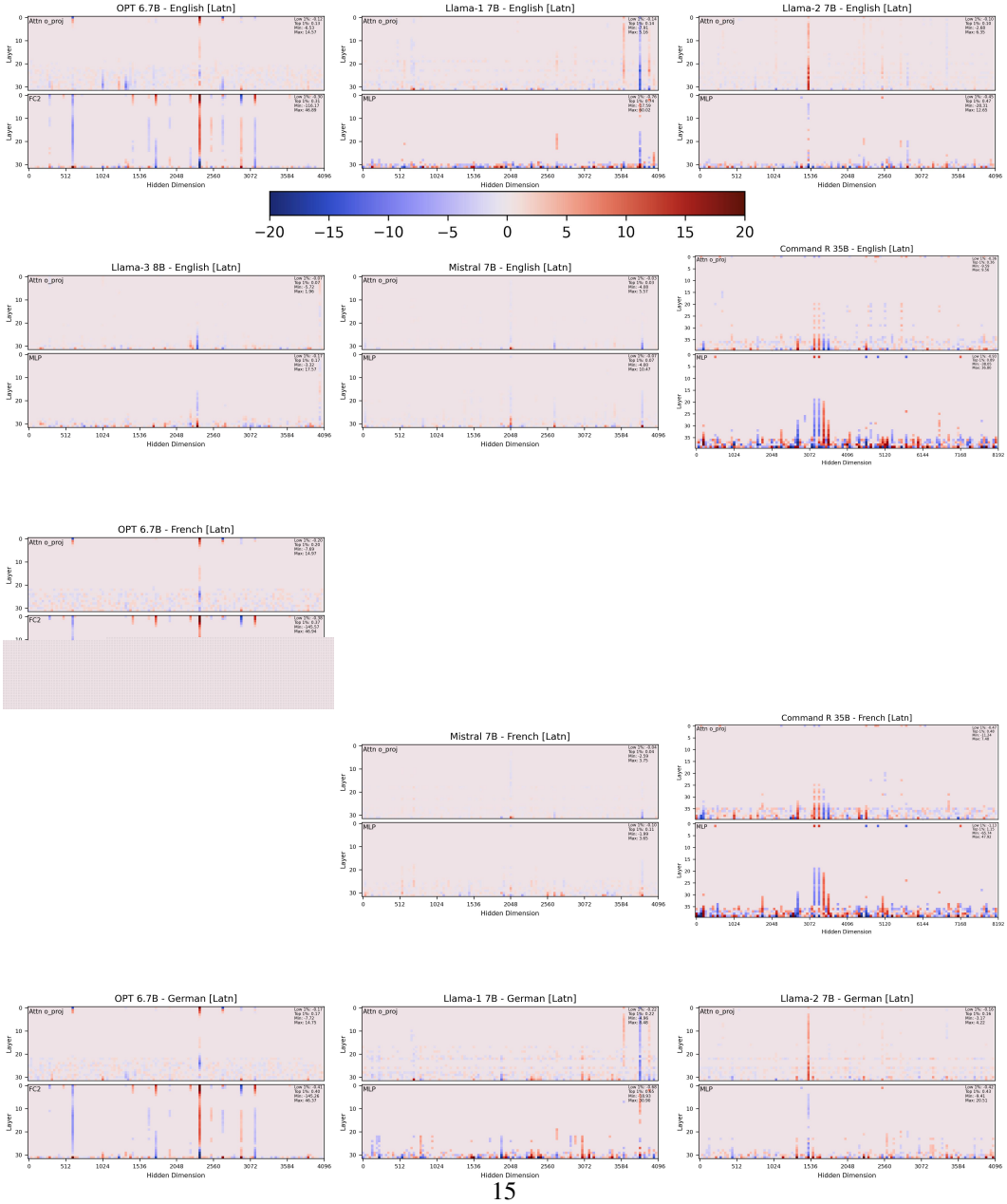
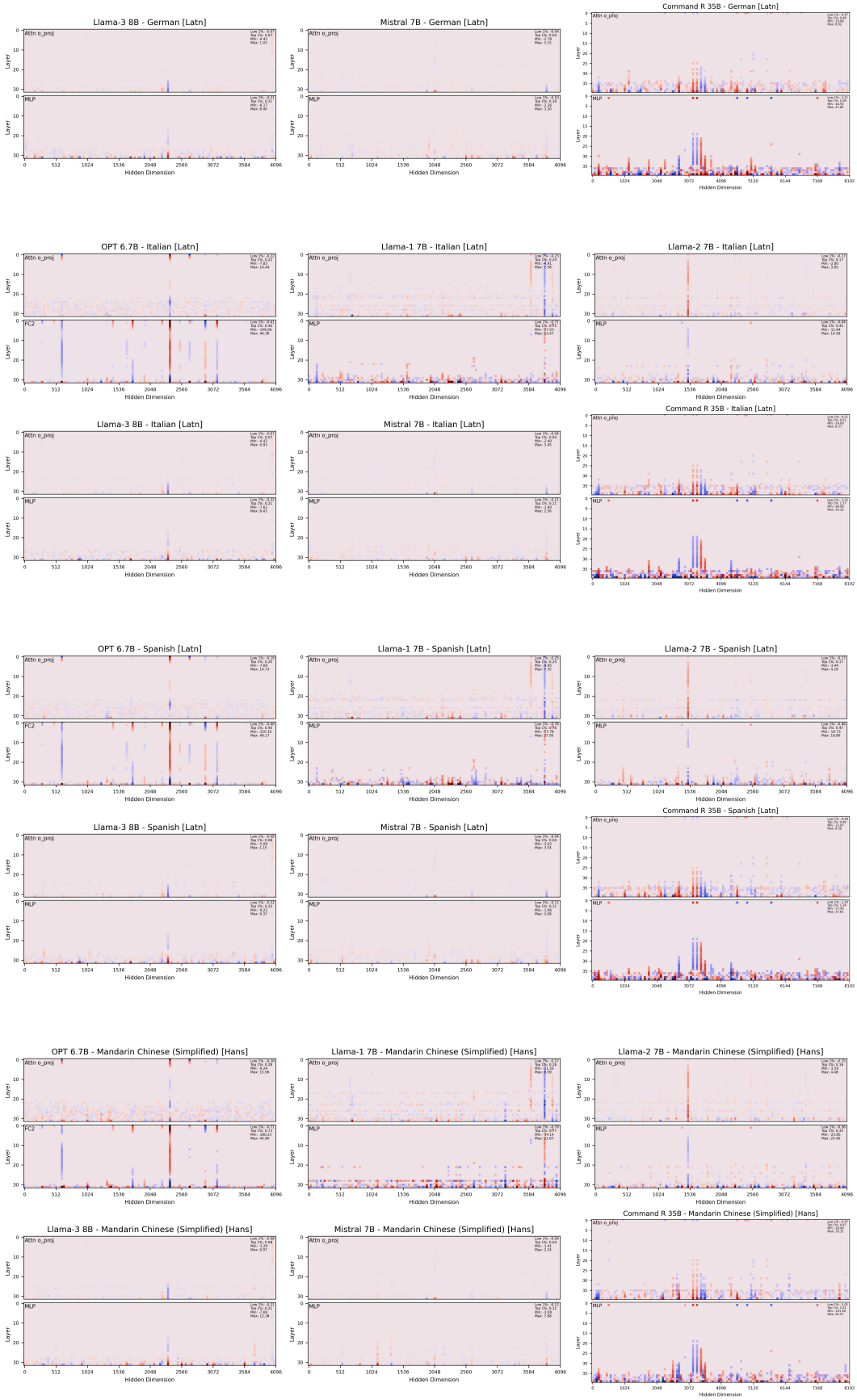
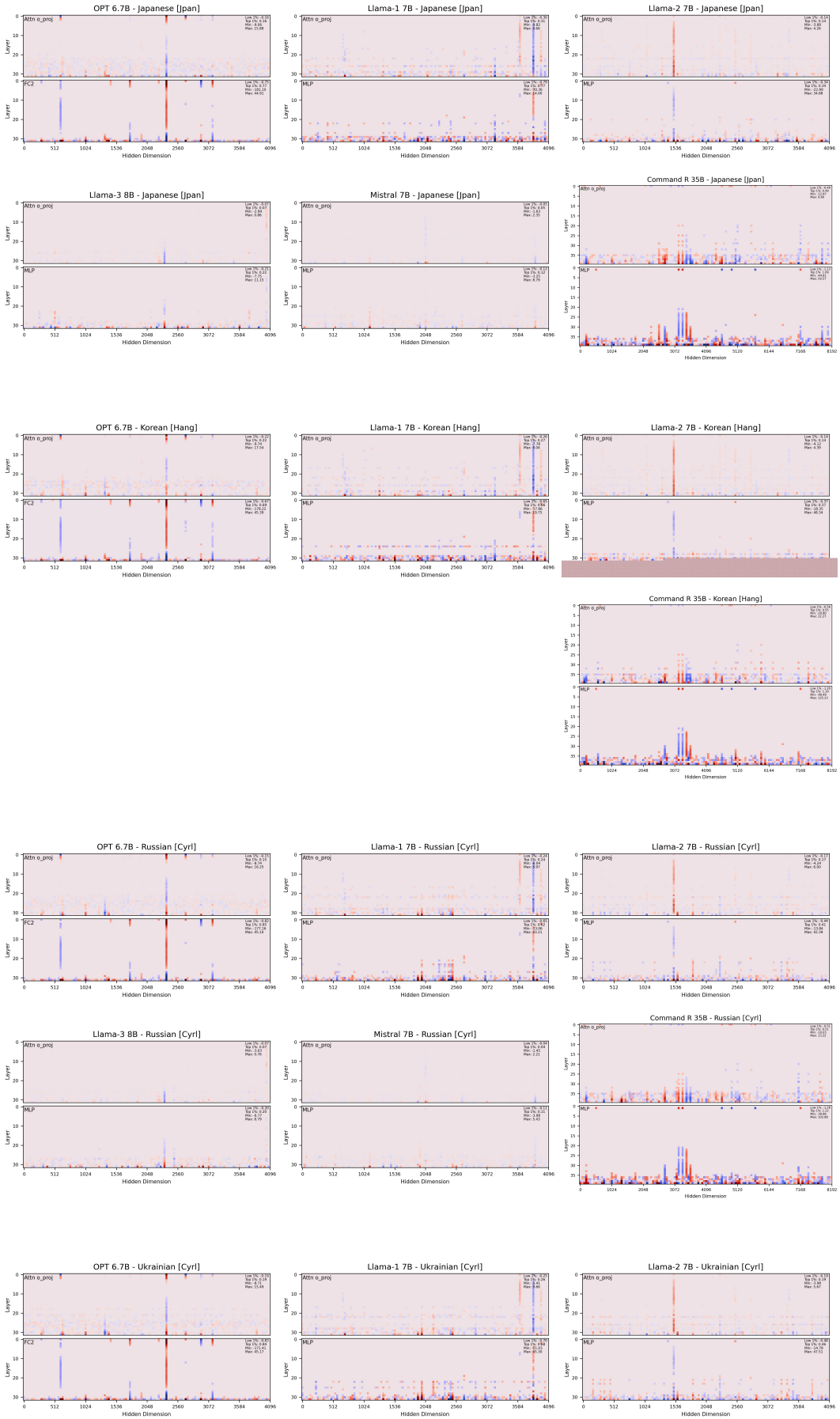


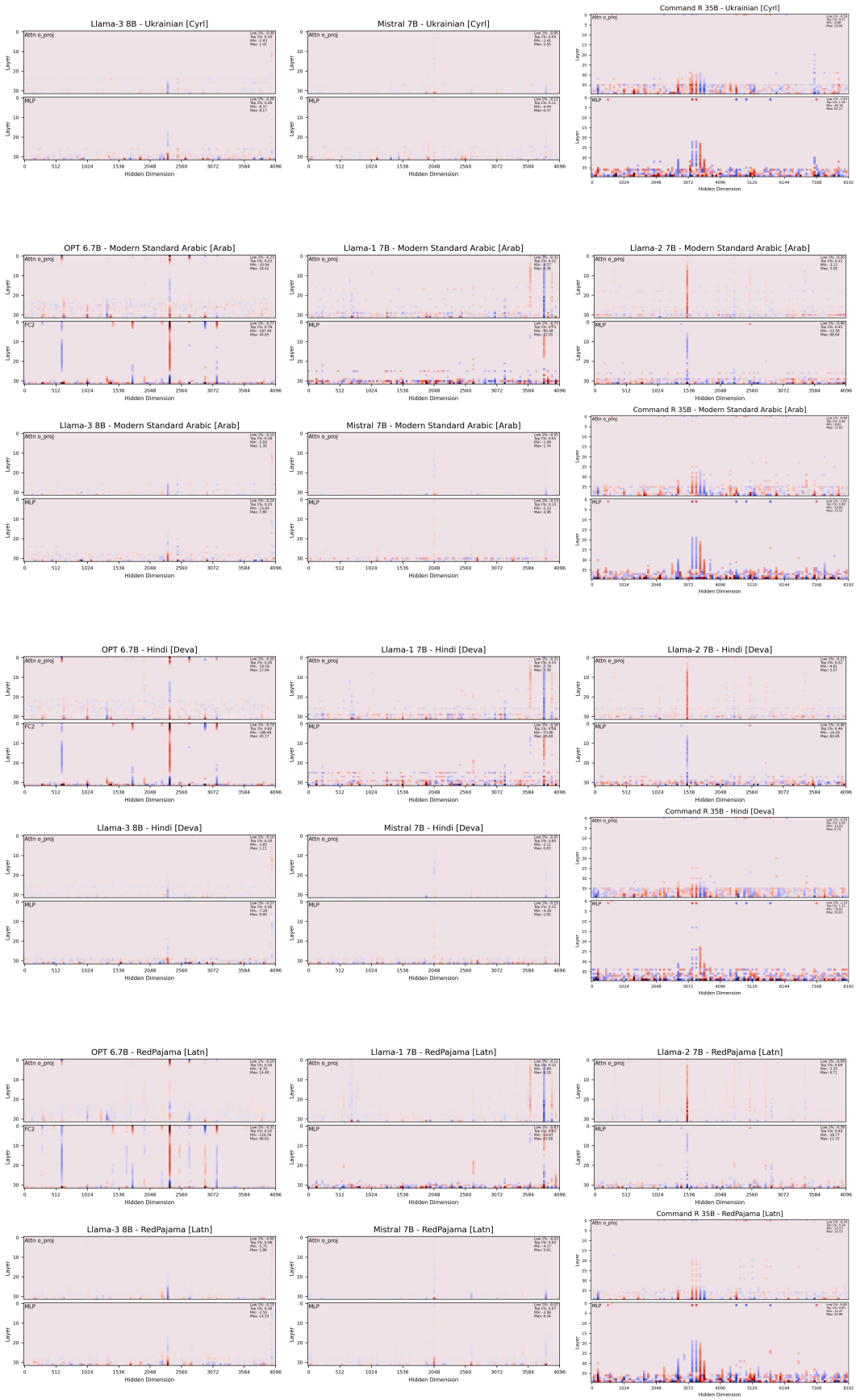
Figure 25: ARC-Challenge accuracy with SmoothQuant W8A8 quantization over calibration sets. Results normalized to RedPajama score. Error bars represent standard error. Higher is better.

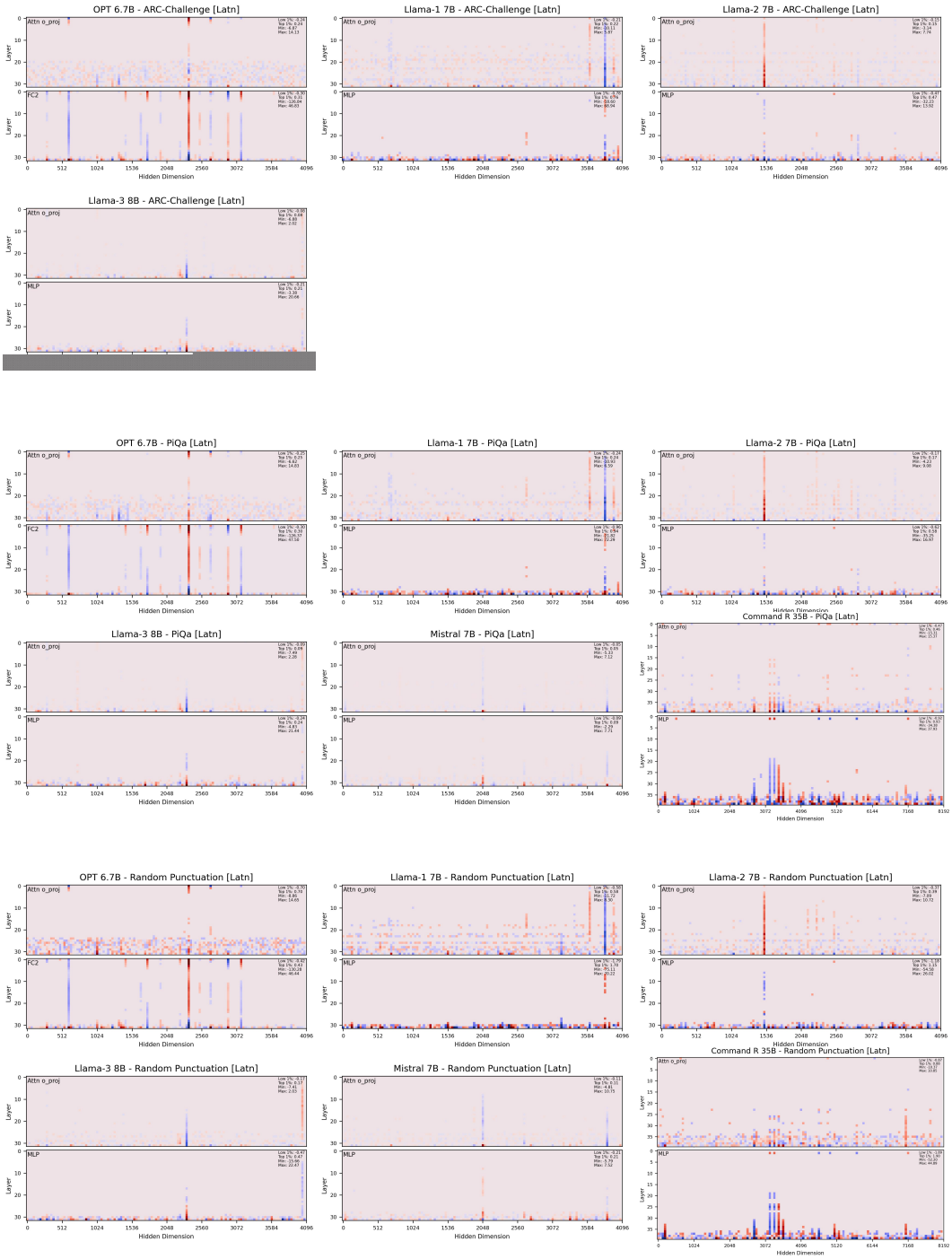
D Activations and Outlier Patterns Plots



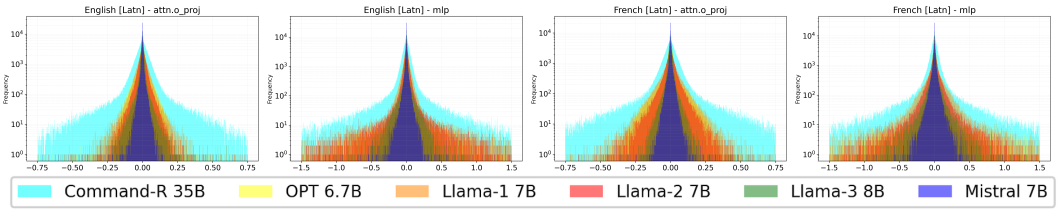

Llama-3 8B - German [Latn]
Mistral 7B - German [Latn]
Command R 35B - German [Latn]
OPT 6.7B - Italian [Latn]
Llama-1 7B - Italian [Latn]
Llama-2 7B - Italian [Latn]
Llama-3 8B - Italian [Latn]
Mistral 7B - Italian [Latn]
Command R 35B - Italian [Latn]
OPT 6.7B - Spanish [Latn]
Llama-1 7B - Spanish [Latn]
Llama-2 7B - Spanish [Latn]
Llama-3 8B - Spanish [Latn]
Mistral 7B - Spanish [Latn]
Command R 35B - Spanish [Latn]
OPT 6.7B - Mandarin Chinese (Simplified) [Hans]
Llama-1 7B - Mandarin Chinese (Simplified) [Hans]
Llama-2 7B - Mandarin Chinese (Simplified) [Hans]
Llama-3 8B - Mandarin Chinese (Simplified) [Hans]
Mistral 7B - Mandarin Chinese (Simplified) [Hans]
Command R 35B - Mandarin Chinese (Simplified) [Hans]
Attn o_proj
MLP
FC2
Layer
Hidden Dimension

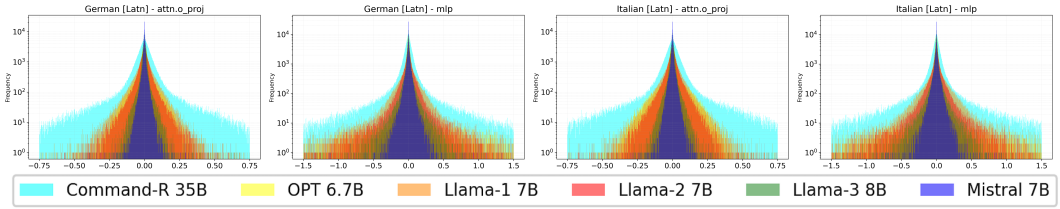

OPT 6.7B - Japanese [Jpan]
Llama-1 7B - Japanese [Jpan]
Llama-2 7B - Japanese [Jpan]
Llama-3 8B - Japanese [Jpan]
Mistral 7B - Japanese [Jpan]
Command R 35B - Japanese [Jpan]
OPT 6.7B - Korean [Hang]
Llama-1 7B - Korean [Hang]
Llama-2 7B - Korean [Hang]
Command R 35B - Korean [Hang]
OPT 6.7B - Russian [Cyrl]
Llama-1 7B - Russian [Cyrl]
Llama-2 7B - Russian [Cyrl]
Llama-3 8B - Russian [Cyrl]
Mistral 7B - Russian [Cyrl]
Command R 35B - Russian [Cyrl]
OPT 6.7B - Ukrainian [Cyrl]
Llama-1 7B - Ukrainian [Cyrl]
Llama-2 7B - Ukrainian [Cyrl]
Attn o_proj
MLP
FC2
Layer
Hidden Dimension

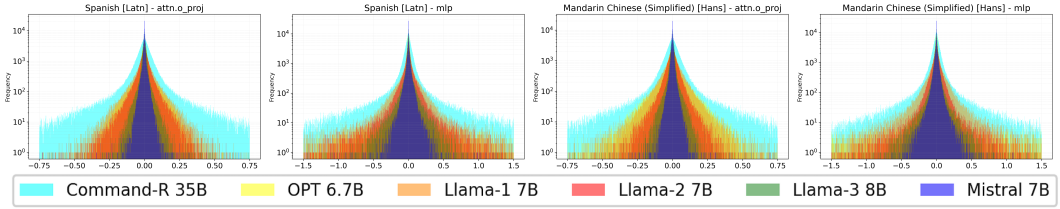

Llama-3 8B - Ukrainian [Cyrl]
Mistral 7B - Ukrainian [Cyrl]
Command R 35B - Ukrainian [Cyrl]
OPT 6.7B - Modern Standard Arabic [Arab]
Llama-1 7B - Modern Standard Arabic [Arab]
Llama-2 7B - Modern Standard Arabic [Arab]
Llama-3 8B - Modern Standard Arabic [Arab]
Mistral 7B - Modern Standard Arabic [Arab]
Command R 35B - Modern Standard Arabic [Arab]
OPT 6.7B - Hindi [Deva]
Llama-1 7B - Hindi [Deva]
Llama-2 7B - Hindi [Deva]
Llama-3 8B - Hindi [Deva]
Mistral 7B - Hindi [Deva]
Command R 35B - Hindi [Deva]
OPT 6.7B - RedPajama [Latn]
Llama-1 7B - RedPajama [Latn]
Llama-2 7B - RedPajama [Latn]
Llama-3 8B - RedPajama [Latn]
Mistral 7B - RedPajama [Latn]
Command R 35B - RedPajama [Latn]
Attn o_proj
MLP
FC2
Layer
Hidden Dimension

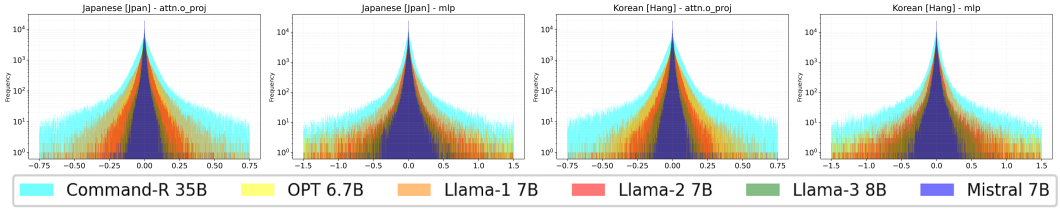


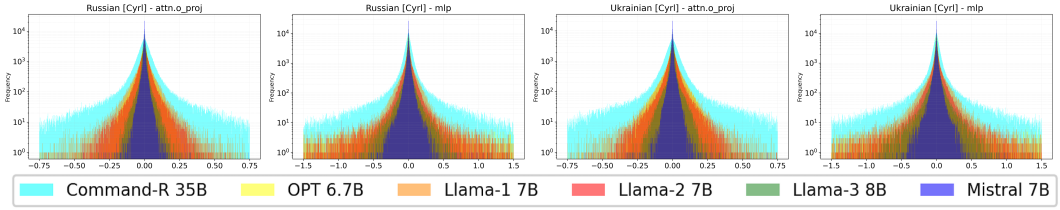
E Activation Distributions Plots

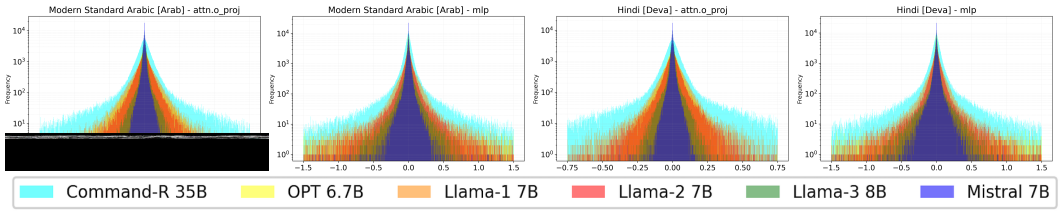


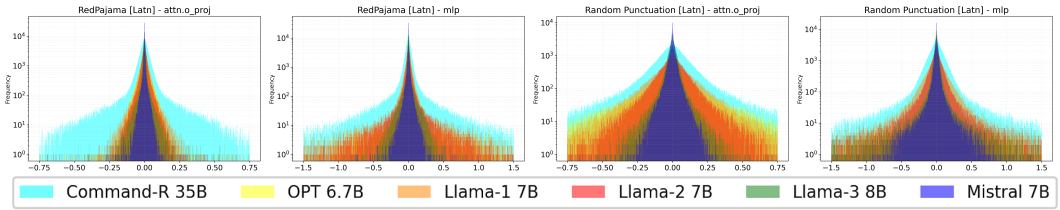

German [Latn] - attn.o_proj
German [Latn] - mlp
Italian [Latn] - attn.o_proj
Italian [Latn] - mlp
Frequency
Command-R 35B
OPT 6.7B
Llama-1 7B
Llama-2 7B
Llama-3 8B
Mistral 7B

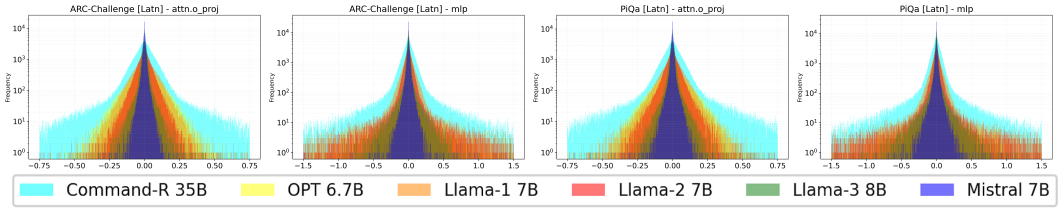

Spanish [Latn] - attn.o_proj
Spanish [Latn] - mlp
Mandarin Chinese (Simplified) [Hans] - attn.o_proj
Mandarin Chinese (Simplified) [Hans] - mlp
Frequency
Command-R 35B
OPT 6.7B
Llama-1 7B
Llama-2 7B
Llama-3 8B
Mistral 7B


Japanese [Jpan] - attn.o_proj
Japanese [Jpan] - mlp
Korean [Hang] - attn.o_proj
Korean [Hang] - mlp
Frequency
Command-R 35B
OPT 6.7B
Llama-1 7B
Llama-2 7B
Llama-3 8B
Mistral 7B


Russian [Cyrl] - attn.o_proj
Russian [Cyrl] - mlp
Ukrainian [Cyrl] - attn.o_proj
Ukrainian [Cyrl] - mlp
Frequency
Command-R 35B
OPT 6.7B
Llama-1 7B
Llama-2 7B
Llama-3 8B
Mistral 7B


Modern Standard Arabic [Arab] - attn.o_proj
Modern Standard Arabic [Arab] - mlp
Hindi [Deva] - attn.o_proj
Hindi [Deva] - mlp
Frequency
Command-R 35B
OPT 6.7B
Llama-1 7B
Llama-2 7B
Llama-3 8B
Mistral 7B


RedPajama [Latn] - attn.o_proj
RedPajama [Latn] - mlp
Random Punctuation [Latn] - attn.o_proj
Random Punctuation [Latn] - mlp
Frequency
Command-R 35B
OPT 6.7B
Llama-1 7B
Llama-2 7B
Llama-3 8B
Mistral 7B


ARC-Challenge [Latn] - attn.o_proj
ARC-Challenge [Latn] - mlp
PiQa [Latn] - attn.o_proj
PiQa [Latn] - mlp
Frequency
Command-R 35B
OPT 6.7B
Llama-1 7B
Llama-2 7B
Llama-3 8B
Mistral 7B