When Quantization Affects Confidence of Large Language Models?

Anonymous ACL submission

Abstract

Recent studies introduced effective compression techniques for Large Language Models (LLMs) via post-training quantization or low-bit weight representation. Although quantized weights offer storage efficiency and allow for faster inference, existing works have indicated that quantization might compromise performance and exacerbate biases in LLMs. This study investigates the confidence and calibration of quantized models, considering factors such as language model type and scale as contributors to quantization loss. Firstly, we reveal that quantization leads to a decrease in confidence regarding true labels, with varying impacts observed among different language models. Secondly, we observe fluctuations in the impact on confidence across different scales. Finally, we propose an explanation for quantization loss based on confidence levels, indicating that quantization disproportionately affects samples where the full model exhibited low confidence levels in the first place.

1 Introduction

Large language models (LLMs) are widely used in a variety of natural language generation applications (Bahdanau et al., 2014; Brown et al., 2020; Winata et al., 2021; Scao et al., 2022; Touvron et al., 2023). LLMs have been proven to achieve high performance in zero and few-shot prompting, providing results on par with fine-tuned baselines, especially in commonsense reasoning tasks (Zhang et al., 2022; Workshop et al., 2022; Jiang et al., 2023). Kaplan et al., 2020 show that emerging abilities come with the scale increase, which makes well-performing larger models less accessible and limits their practical usability. A range of efficient compression and acceleration methods, including quantization, have been developed that help to alleviate high latency and extensive storage demands (Gupta and Agrawal, 2020; Tao et al., 2022). Despite its efficacy as a compression technique,

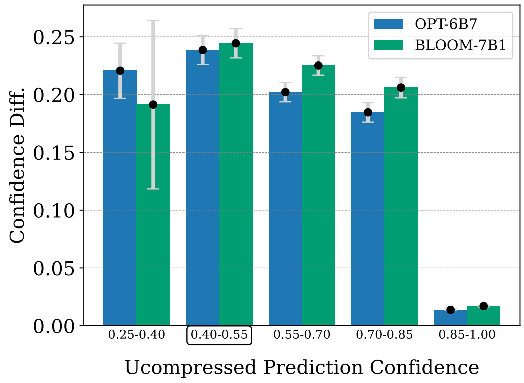

Figure 1: Quantization-induced absolute confidence shifts in original (pre-compression) low and high confidence samples (BLOOM and OPT models, HELLASWAG benchmark). The bin with the largest mean confidence shift is highlighted.

recent works show that quantization may degrade the initial performance and amplify the sensitivity of an LLM to certain linguistic phenomena and stereotypes (Liu et al., 2023; Ramesh et al., 2023). However, less attention has been paid to explaining the compression loss, particularly its variance across different texts. In this paper, we extend the existing research on the compression loss estimation; in particular, we measure the impact of quantization on the confidence of LLMs that can be initially overconfident in both right and wrong predictions (Jiang et al., 2021a; Xiao et al., 2022; Ahuja et al., 2022; Desai and Durrett, 2020).

Our main contributions are the following: (i) we investigate how quantization with GPTQ (Frantar et al., 2022a) influences the calibration and confidence of LLMs, (ii) we assess the confidence alignment between compressed and full LLMs at scale, (iii) we explain the quantization loss from the initial confidence perspective.

Our null hypothesis is that the compressed vs. full predictive probability distributions are indis-

Model	ARC EASY		BOOLQ		HELLASWAG		OPENBOOKQA		PIQA		XSTORY	
	Acc.↑	CE↓	Acc.↑	CE↓	Acc.↑	CE↓	Acc.↑	CE↓	Acc.↑	CE↓	Acc.↑	CE↓
M												
7B	**81.10** ↓1.18	**7.94** ↑0.83	**83.61** ↓0.86	38.62 ↑3.13	**61.30** ↓1.53	**34.3** ↑1.29	<u>32.60</u> ↓0.40	<u>45.24</u> ↑2.08	**80.83** ↓0.65	45.24 ↓0.4	**78.89** ↓0.27	**4.78** ↓0.08
🦙												
7B	<u>75.25</u> ↓4.29	<u>9.99</u> ↑1.72	<u>75.05</u> ↓2.51	38.78 ↓7.66	<u>56.94</u> ↓3.23	<u>37.8</u> ↑4.07	**34.0** ↓4.2	**44.56** ↑3.13	<u>78.67</u> ↓2.01	44.94 ↓0.58	<u>76.77</u> ↓2.05	<u>4.97</u> ↑0.24
🌸												
560M	47.35 ↓1.56	29.13 ↑1.13	55.14 ↓6.94	26.91 ↓2.85	31.58 ↓0.47	64.81 ↓0.49	17.2 ↓0.6	61.16 ↑1.61	64.09 ↓1.14	**40.98** ↑0.32	61.22 ↓1.39	5.13 ↓0.03
1B1	51.47 ↓2.27	25.07 ↑2.17	59.08 ↑0.74	32.8 ↑2.65	34.44 ↓0.87	58.51 ↑0.85	0.2 ↓2.2	58.88 ↑0.74	67.14 ↓0.98	42.27 ↓0.68	62.54 ↓1.52	5.77 ↑0.15
1B7	56.31 ↓1.81	21.99 ↑0.35	61.77 ↓0.12	38.29 ↓0.06	37.54 ↓0.82	55.67 ↑0.51	21.40 ↓1.60	56.64 ↑0.97	68.77 ↓0.76	<u>41.4</u> ↓0.49	64.66 ↓0.53	5.65 ↑0.08
3B	5947 ↓2.27	19.68 ↑1.31	61.62 ↑0.09	34.67 ↓0.86	41.39 ↓0.91	52.33 ↑0.82	21.6 ↓0.40	56.32 ↓0.15	70.84 ↓0.82	42.12 ↓0.25	66.78 ↓0.53	5.76 ↑0.2
7B1	65.03 ↓1.56	15.57 ↑1.06	62.81 ↑0.19	32.28 ↑0.19	46.49 ↓1.11	48.54 ↑1.07	25.20 ↓0.80	53.23 ↑0.01	72.63 ↑0.28	42.52 ↓0.24	70.55 ↓0.33	5.53 ↓0.1
🧠												
125M	43.56 ↓1.05	32.76 ↑0.62	55.44 ↓2.72	30.13 ↓3.48	29.18 ↓0.59	62.84 ↑0.48	16.6 ↑0.20	61.19 ↑0.45	62.00 ↓0.81	41.51 ↑0.11	58.84 ↓1.0	5.9 ↓0.12
350M	44.20 ↑0.08	31.21 ↑1.79	57.65 ↓3.46	29.62 ↓4.4	32.02 ↓0.18	60.09 ↑0.30	17.60 ↓1.40	61.92 ↓0.19	64.47 ↓0.92	41.58 ↑0.19	62.48 ↓1.79	5.97 ↓0.35
1B3	56.99 ↓0.85	20.85 ↑0.85	57.67 ↓2.73	<u>26.42</u> ↑0.01	41.56 ↓1.00	52.7 ↑0.16	23.4 ↓1.40	55.04 ↑3.32	71.71 ↓0.76	41.49 ↑0.34	70.28 ↓1.58	5.6 ↑0.23
2B7	60.77 ↓1.64	17.63 ↑1.56	60.24 ↓5.05	**25.86** ↑0.01	45.86 ↓0.51	48.93 ↓0.15	25.0 ↓2.20	52.6 ↑2.76	73.78 ↓0.98	41.87 ↑0.39	70.42 ↑0.13	5.83 ↓0.06
6B7	65.57 ↓0.08	15.58 ↓0.18	66.05 ↓0.83	28.05 ↓1.72	50.51 ↓0.65	45.25 ↑0.01	27.6 ↓1.20	50.99 ↑1.44	76.28 ↓0.22	43.72 ↓0.65	73.6 ↓0.19	5.62 ↓0.17
13B	67.13 ↑0.38	14.21 ↓0.69	65.93 ↓0.09	29.47 ↓0.52	52.43 ↓0.59	43.03 ↑0.47	27.2 ↓0.05	52.33 ↓0.23	75.84 ↑0.11	43.87 ↓0.43	76.04 ↓0.07	5.15 ↑0.21

Table 1: Zero-shot accuracy scores (Acc.) and calibration error (CE) for full LLMs by benchmark with the difference in scores after quantization. We report expected CE for binary tasks and adaptive CE for multi-class benchmarks (ARC, BOOLQ, OPENBOOKQA). <u>Notations:</u> M=MISTRAL; 🦙=LLAMA;🌸=BLOOM;🧠=OPT.

tinguishable since prior work discussed a negligible accuracy drop in performance after quantization (Jacob et al., 2018; Dettmers et al., 2022; Xiao et al., 2023). We analyze the relationship between models by comparing calibration scores—indicating a model's ability to accurately reflect true probabilities—before and after quantization. To the best of our knowledge, our research is the first attempt to explain the quantization loss through the lens of predictive probabilities.

2 Related Work

The pretrained knowledge embedded in very large models has paved the way to parameter-efficient adaptation for downstream tasks, such as prompting and few-shot learning, bypassing the necessity for fine-tuning (Brown et al., 2020; Wei et al., 2022). The inference of LLMs can be accelerated through a low-bit representation of trained weights (*quantization*) and effective tensor slicing across multiple GPUs (DEEPSPEED (Rasley et al., 2020), ACCELERATE (Gugger et al., 2022), *inter alia*). Prior studies have estimated compression efficiency through: (1) latency-related measures determining throughput and a multiple of the original model's inference speed-up, (2) the precision of weights approximation, and (3) performance decrease (gap) (Jacob et al., 2018; Dettmers et al., 2022; Xiao et al., 2023; Frantar et al., 2022a). Recent comparative studies on interpreting compression loss have indicated that compression amplifies biases and stereotypes, highlighting a disparate quantization loss in multilingual LLMs across different architectures (Ramesh et al., 2023). In contrast, another line of research suggests that compression enhances fairness (Hessenthaler et al., 2022). Altogether, existing studies commonly measure compression loss by observing the deviation in performance before and after quantization. In this project, we adopt the recent GPTQ quantization method for compressing model weights and concentrate on the disparities between predictive probability distributions instead. For the first time, our approach reveals the relationship between the initial level of predictive confidence and quantization loss.

3 Methodology

We follow Jiang et al., 2021b and consider a classification problem where inputs to the model are questions x paired with candidate answers y to constitute concatenated sequences. The generative model then processes these concatenated question-answer pairs to predict the most probable answer $\hat{y}$ from the provided choices Y for a given x:

$$\hat{y} = \underset{y \in Y}{\arg\max}\ p_{\text{LM}}(y|x).$$

Here, the probability of the token sequence y is derived as the product of individual token $y_{[i]}$ probabilities within the sequence, conditioned on x and the preceding tokens $y_{[1:i-1]}$:

$$p_{\text{LM}}(y|x) = \prod_{i=1}^{|y|} p_{\text{LM}}(y_{[i]}|x, y_{[1:i-1]}),$$

where $|y|$ is the number of tokens composing the answer y.

For the entailment generation benchmarks, we use texts concatenated with possible completions as inputs to the model. We compare the quantized and full-precision models with the difference in the probabilities of the sequences $p_{\mathrm{LM}}(y|x)$, further referred to as confidences.

3.1 Quantization

We quantize pre-trained weights of LLMs with a post-training quantization method known as GPTQ (OPTQ, Frantar et al., 2022b). This approach employs iterative layer-wise weight quantization based on the input data, providing several benefits compared to other quantization methods: minimized weight approximation error, support for serialization across various bit configurations, and significantly accelerated inference using GPUs. We follow the GPTQ 4-bit configuration outlined by Frantar et al., 2022b and use a random sample of 128 sequences from the C4 dataset (Raffel et al., 2020) for quantization and set a grouping size equal to 128. Additional details regarding the quantization procedures can be found in Table 3 (Appendix A).

3.2 Evaluation

We focus on evaluating models' confidence in predictions before and after quantization in a zero-shot setting. In an ideal scenario, we expect the model's performance and confidence to remain consistent after quantization, preserving the initial calibration level. We evaluate the performance of LLMs post-compression using accuracy (Acc.) and calibration error (CE). For binary problems, we use the Expected Calibration Error (ECE;Naeini et al., 2015), calculated using reliability plots that bin predicted probabilities and compare them against actual accuracy. In multi-class benchmarks, we use the Adaptive Calibration Error (ACE; Nixon et al., 2019), which quantifies calibration performance by dividing predictions into equally sized bins based on confidence levels and comparing accuracy and confidence within these subsets.

Details regarding the binning parameters used are provided in Appendix B. We also examine two cases of miscalibration: (1) the model's rejection of correct predictions due to lower confidence and (2) the model's incorrect prediction due to higher confidence. Specifically, we measure the model's confidence Conf_{err} when predicting an incorrect class and the model's confidence in the true class Conf_{true}.

Model	Conf.	Conf_{err}	Conf_{true}	H
BLOOM	96.26	95.64	46.24	12.87
+ GPTQ	96.3	95.62	45.23*	12.89
OPT	96.51	95.57	50.37	12.12
+ GPTQ	96.5	95.55	49.78*	12.22
Mistral	96.85	95.02	61.14	10.96
+ GPTQ	96.89	95.13	59.73*	10.87
LLaMA	96.8	95.34	56.83	11.37
+ GPTQ	96.48	95.13	53.69*	12.21*

Table 2: Confidence and prediction entropy evaluation results on HELLASWAG for LLMs with ~7B parameters. Quantized LLMs become less confident in both correct and wrong predictions. Conf.: Mean confidence in predictions; Conf_{err}: Mean confidence in wrong predictions; Conf_{true}: Mean confidence in true class; H=Mean predictive entropy in the answers. High entropy means that the model is more unsure about its predictions. The * denotes a significant difference with a confidence level set at 0.001 (paired t-test).

4 Experiment Settings

Data We use six standard commonsense reasoning tasks for our analysis: ARC EASY (Clark et al., 2018), BOOLQ (Clark et al., 2019), PIQA (Bisk et al., 2020), HELLASWAG (Zellers et al., 2019), OBQA (OpenBookQA; Mihaylov et al., 2018), and XSTORY-EN(Mostafazadeh et al., 2017). These benchmarks vary in the types of language inference abilities assessed in LLMs: (1) question answering involving reading comprehension (BOOLQ), (2) natural text entailment (XSTORY-EN, HELLASWAG), (3) science fact knowledge (ARC, OBQA), and (4) physical commonsense (PIQA).

Models We use the following *causal* (autoregressive) LLMs in our experiments: (1) BLOOM (Workshop et al., 2022), (2) OPT (Zhang et al., 2022), (3) Mistral-7B (Jiang et al., 2023), and (4) LLaMA-7B (Touvron et al., 2023). To examine how confidence loss varies across different scales, we use various configurations of LLMs: BLOOM with 560M, 1B1, 1B7, 3B, and 7B1 parameters, and OPT with 125M, 350M, 1B3, 2B7, and 6B7.

5 Results

We conduct a series of experiments to estimate the impact of quantization on various aspects of LLMs' performance, including calibration error,

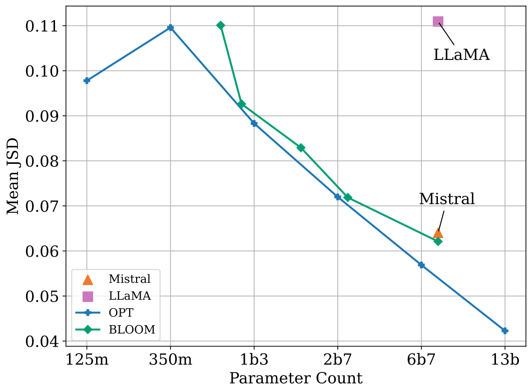

Figure 2: Mean Jensen-Shannon distances between full and quantized LLMs across benchmarks. The distances depict dissimilarities in true-class probability distributions.

prediction entropy, cases of maximum confidence change, and the distribution dissimilarities between full and compressed models. We find variance in quantization impact across different families of models and their sizes, suggesting that scale and pre-training directly affect the further quantization loss.

Calibration Impact Table 1 outlines the classification results after quantization, evaluated through calibration error and accuracy metrics, along with the variation of these scores compared to the uncompressed LLMs. The general trend is that quantization amplifies the pre-existing high calibration error present in the models before compression across different models and benchmarks. This trend remains consistent across various model families, notably affecting the LLaMA-7B, which experiences a $\sim$10% increase in pre-compression calibration error on the HELLASWAG dataset. Overall, scores associated with the HELLASWAG dataset are more significantly impacted compared to those of the BOOLQ and PIQA benchmarks.

Confidence Impact Table 2 presents the results obtained from four models, each having a near-equivalent number of parameters. Notably, across all models, a consistent trend of overconfidence emerges in both pre- and post-quantization stages, with an average confidence level around $\sim$0.95 for incorrect predictions. Our analysis further shows a statistically significant impact of quantization on the confidence associated with true-class predictions. Additionally, we observe an increase in entropy for the quantized LLMs. This increase suggests an amplification in the variance across answers, reflecting increased uncertainty in answer selection due to quantization.

Identifying Cases of Confidence Change To identify instances of confidence change, we segment the models' predictions into bins and calculate the confidence changes after quantization within each bin. In Figure 1, we depict the mean confidence changes for the BLOOM and OPT models on the HELLASWAG benchmark. The plot illustrates that samples with lower pre-quantization confidence levels are significantly affected by the quantization process, whereas samples in which the original model was confident show less impact. This observation suggests that quantization predominantly influences the confidence of samples where the original model exhibited lower confidence levels.

Jensen-Shannon Distances To illustrate the extent of differences between the distributions of the full and compressed models, we plot the mean Jensen-Shannon distances across benchmarks in Figure 2. These distances reflect the dissimilarity between the true-class probability distributions of the models. We find that the distances between original and compressed decrease as the model size scales up. Notably, most model families show a consistent trend in this behavior, except for LLaMa, which diverges from the patterns observed in other models of similar size ($\sim$7B).

6 Conclusion

This paper investigates the impact of quantization on the confidence and calibration of LLMs. We demonstrate that quantization leads to an increase in calibration error and statistically significant changes in confidence levels for correct predictions. Through a detailed examination of confidence shifts, we identify instances of confidence change occurring in data where models lack confidence before quantization. Overall, our findings provide insights into quantization loss and suggest a potential direction for future work, emphasizing the need to focus on calibrating LLMs, specifically on uncertain examples. Future work may concentrate on integrating the most uncertain samples into the data used for quantization to avoid performance degradation.

Limitations

Our quantization techniques are currently limited to 4-bit post-training quantization with GPTQ. However, future work can benefit from exploring training-aware quantization approaches, studying different quantization factors, such as 2- and 3-bit weight representation, and quantization of activations.

In our evaluations, we employ zero-shot techniques, enabling the estimation of the pure quantization effect. Previous studies included a fine-tuning step, whereas our approach avoids it. Yet, future work could involve few-shot analysis since this method has the potential to amplify or compensate for confidence and quantization loss.

Further research could apply our analysis to other generative tasks. Instead of predictive distributions over labels, one could consider those across tokens. This means using the full model's predictions as references and comparing the confidence in these generations after the quantization process.

References

Kabir Ahuja, Sunayana Sitaram, Sandipan Dandapat, and Monojit Choudhury. 2022. On the calibration of massively multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4310–4323, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

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

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

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*.

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

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022b. Optq: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. `https://github.com/huggingface/accelerate`.

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.

Marius Hessenthaler, Emma Strubell, Dirk Hovy, and Anne Lauscher. 2022. Bridging fairness and environmental sustainability in natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7817–7836, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713.

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*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021a. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021b. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

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*.

Peiyu Liu, Zikang Liu, Ze-Feng Gao, Dawei Gao, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2023. Do emergent abilities exist in quantized large language models: An empirical study. *arXiv preprint arXiv:2307.08072*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR workshops*, volume 2.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Krithika Ramesh, Arnav Chavan, Shrey Pandit, and Sunayana Sitaram. 2023. A comparative study on the impact of model compression techniques on fairness in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15762–15782, Toronto, Canada. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Chaofan Tao, Lu Hou, Wei Zhang, Lifeng Shang, Xin Jiang, Qun Liu, Ping Luo, and Ngai Wong. 2022. Compression of generative pre-trained language models via quantization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4821–4836, Dublin, Ireland. Association for Computational Linguistics.

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*.

Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022. Exploring predictive uncertainty and calibration in NLP: A study on the impact of method & data scarcity. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2707–2735, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7273–7284, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

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*.

A Quantization Parameters

Parameter	Value
Num bits	4
Group size	128
Dampening factor (%)	0.01
Desc act	false
Symmetry	true
True sequential	true

Table 3: Configuration for GPTQ

B Evaluation Details

In this section, we provide further details on the used measures for the experiments.

Jensen-Shannon Divergence In Figure 2, we give the distance dissimilarities in the true-class probability distributions using the Jensen-Shannon divergence. For a given dataset, we focus on the true class probabilities, $p \in \mathbb{R}^n$, for the full model, and $q \in \mathbb{R}^n$ for the quantized one, where n denotes the number of instances.

The Jensen Shannon-Divergence between these two distributions is defined by:

$$
\begin{aligned}
&JSD(p, q) \\
&= \frac{1}{2}\left(KL\left(p \,||\, \frac{p+q}{2}\right) + KL\left(q \,||\, \frac{p+q}{2}\right)\right), \\
&= \sum_{i=1}^{n} p_i \ln\left(\frac{2p_i}{p_i+q_i}\right) + q_i \ln\left(\frac{2q_i}{p_i+q_i}\right),
\end{aligned}
$$

where KL denoted the Kullback-Leibler divergence and p_i and q_i are the true-class probabilities of the i-th instance for the full and quantized model respectively.

These distances are then averaged over all the studied datasets.

Expected Calibration Error (ECE) Let us consider a model h, which assigns confidence (which are probabilities) of belonging in a given class. These confidence scores can be divided into several bins $B_m, m = 1, \ldots M$ where M is the number of bins. More precisely, an instance belongs to the bin B_m if its confidence score in the true class $conf_i$ is in a given range (*e.g.* if $(m-1)/M \leq conf_i \leq m/M$). In a given bin B_m, we are also able to measure the accuracy of the model, *i.e.*, compute the ratio of instances in the bin B_m that have well-classified.

The *expected calibration error* is then defined as the weighted mean, where the weights depend on the number of instances in the bin of the absolute difference between the accuracy $acc(B_m)$ of the bin and the mean confidence score in the bin $\overline{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} conf_i$., *i.e.*,

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - \overline{conf}(B_m)|,$$

where n is the sample size. Note this error has been developed for binary classification tasks and can be extended to multi-class settings using the so-called *SCE* (Nixon et al., 2019), but this first extension has been shown to be not relevant for all studies (Ulmer et al., 2022). The authors rather use the *adaptive calibration error*, which works with equal size bins.

Adaptive Calibration Error (ACE) The *adaptive calibration error* is defined by

$$ACE = \frac{1}{CM} \sum_{c=1}^{C} \sum_{m=1}^{M} |acc(B_m, c) - \overline{conf}(B_m, c)|,$$

where C is the number of classes, M is the number of bins that are created, $acc(B_m, c)$ is the accuracy on class c in the m-th bin and $\overline{conf}(B_m, c)$ is the mean confidence score for class c in the m-th. In this case, all the bins have the same size, which is equal to $\lfloor n/R \rfloor$.

Implementation Details Our experiments use evaluation scripts derived from the EleutherAI Language Model Evaluation Harness (Gao et al., 2023).[1] To quantize the models we use scripts from Auto-GPTQ package.[2] We run quantization and inference for all the experiments on a single NVIDIA A-100 GPU. For the largest model, uncompressed OPT-13B, the evaluation run took roughly two hours for all the datasets. Frantar et al., 2022a report GPTQ runtime for the models.

[1] https://github.com/EleutherAI/lm-evaluation-harness

[2] https://github.com/PanQiWei/AutoGPTQ

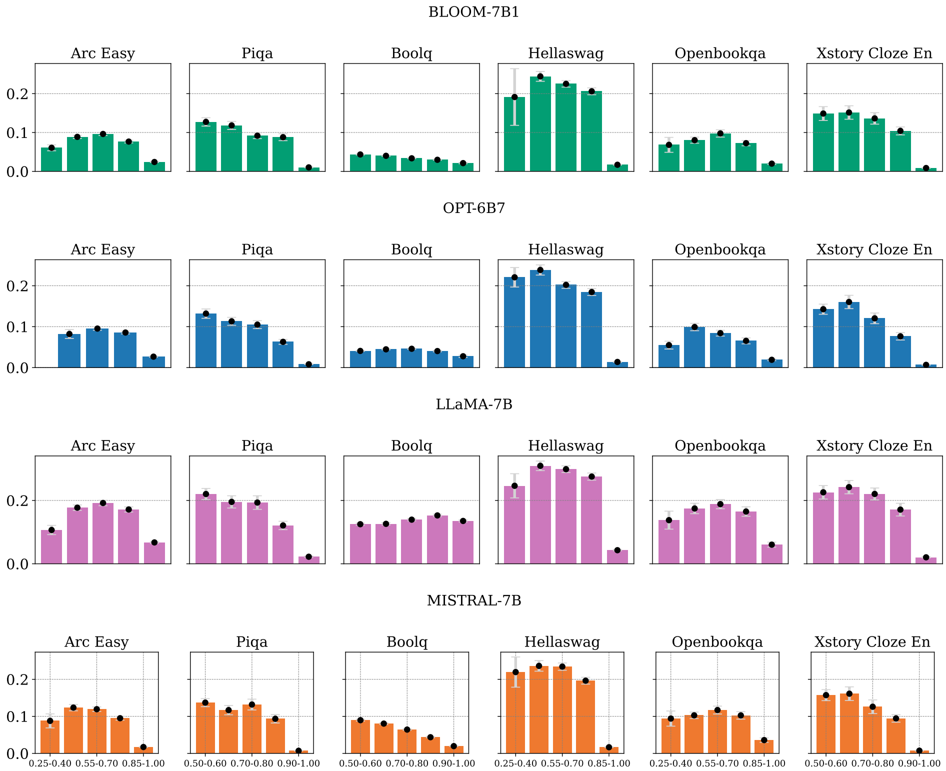

Figure 3: Confidence Difference for Models across datasets. For each dataset (in column) and each model (in line), we provide the difference in prediction score between the full and quantized models. More precisely, each bar represents the mean difference in confidence between the quantized and full models, with confidence in the full model represented on the horizontal axis. Note that some ranges start from 0.5 for binary tasks and 0.25 for multi-class (with four classes) tasks. For a confidence lower than the previous one, there is no chance of being assigned to the associated class.

C Confidence Evaluation in LLMs after Quantization

In this last experiment we study the evolution of the confidence score for our different models on the six studied datasets. More precisely, we study the mean difference of confidence score between full and quantized models for different ranges of confidence scores of the full model.

As presented in Figure 1, the change of probabilities is the lowest one when the model is over-confident and the uncertainty of the model is impacted (*i.e.*, increased) by the quantization. This observation goes hand in hand with the entropy values, serving as a measure of model uncertainty, shown in Table 2. We also note that, in the case of binary problems (PIQA, BOOLQ and XSTORY CLOSE EN), that the most impacted confidence scores are the ones for which the model is not confident it its prediction.

D Confidence Evaluation Results

Model	ARC EASY Conf.	ARC EASY Conf$_{err}$	BOOLQ Conf.	BOOLQ Conf$_{err}$	HELLASWAG Conf.	HELLASWAG Conf$_{err}$	OPENBOOKQA Conf.	OPENBOOKQA Conf$_{err}$	PIQA Conf.	PIQA Conf$_{err}$	XSTORY Conf.	XSTORY Conf$_{err}$
…						…						
7B	88.45	75.44	76.21	64.8	96.85	95.02	79.2	78.93	94.78	89.39	95.23	95.57
7BQ	88.15	75.61	76.75*	65.56	96.89	95.13	79.01	79.07	94.67	88.8	95.26	95.72
…						…						
7B	84.68	73.12	75.75	67.78	96.8	95.34	78.64	78.85	94.24	90.32	95.01	95.51
7BQ	81.6*	71.93	68.95*	63.55	96.48*	95.13	78.25	78.68	93.99	90.05	94.8	94.94
…						…						
560M	76.45	73.76	64.74	64.38	96.39	96.27	78.28	78.77	91.47	89.8	94.47	94.62
560MQ	75.89	73.68	61.89*	62.76	96.47	96.35	78.74	79.11	91.76	90.07	94.55	95.22
1B1	76.2	72.22	70.63	69.16	96.45	96.16	78.3	78.81	92.1	89.99	94.36	94.46
1B1Q	76.0	72.95	73.28*	72.08	96.52	96.18	77.97	78.31	91.78	89.58	94.21	94.87
1B7	77.47	72.86	76.12	74.9	96.24	95.89	78.05	78.57	91.89	89.75	93.96	94.08
1B7Q	76.34*	72.44	76.06	74.85	96.11	95.91	77.52	78.31	91.54	89.56	93.87	94.32
3B	78.6	73.11	72.5	70.55	96.24	95.75	78.3	78.56	92.37	89.57	94.36	94.08
3BQ	77.24*	72.23	71.59*	69.93	96.43	96.02	77.5	77.9	92.25	89.2	94.15	94.32
7B1	79.95	73.11	69.97	66.55	96.26	95.64	78.36	78.52	92.71	88.05	94.59	94.85
7B1Q	79.46	72.85	69.59*	66.58	96.3	95.62	78.17	78.59	92.53	89.03	94.56	94.54
…						…						
125M	75.9	74.29	67.95*	67.42	96.31	96.29	77.6	78.42	90.96	89.31	94.1	94.86
125MQ	75.88	74.57	64.45	64.21	96.29	96.15	78.14	79.62	91.39	89.89	94.31	94.59
350M	75.45	73.25	67.45	66.57	96.07	95.91	78.36	79.38	91.39	88.85	94.17	94.62
350MQ	76.46*	74.84	63.03*	62.34	96.25	96.03	78.26	78.68	91.33	89.23	94.38	94.43
1B3	77.67	72.44	64.25	62.24	96.31	95.74	78.35	79.07	91.91	88.69	94.39	94.47
1B3Q	77.02*	72.54	64.26	63.5	96.14	95.58	78.56	79.16	91.94	88.59	94.31	94.87
2B7	78.22	71.73	63.67	61.81	96.32	95.51	78.45	78.68	91.89	87.93	94.42	94.6
2B7Q	77.58*	71.69	63.66	62.07	96.2	95.62	77.89	77.66	92.14	88.02	94.35	94.64
6B7	80.46	72.14	65.88	62.46	96.51	95.57	78.65	79.16	93.29	89.78	94.38	95.32
6B7Q	80.29	72.52	64.16*	60.9	96.5	95.55	78.66	78.32	93.13	89.4	94.55	94.69
13B	81.36	72.42	67.3	63.32	96.49	95.48	78.7	78.75	93.23	88.64	94.98	95.53
13B1	80.96*	72.4	66.78*	62.35	96.5	95.52	79.08	79.46	93.03	88.77	94.77	95.18

Table 4: Mean confidence evaluation results across benchmarks. Conf.: Mean confidence in predictions; Conf$_{err}$: Mean confidence in wrong predictions. The * is used to denote a significant difference with a confidence level set at 0.001 (paired t-test). Q denotes quantized models. Notations: =MISTRAL; =LLAMA; =BLOOM; =OPT.

Model	ARC EASY		BOOLQ		HELLASWAG		OPENBOOKQA		PIQA		XSTORY	
	Conf$_{true}$	**H**	**Conf**$_{true}$	**H**	**Conf**$_{true}$	**H**	**Conf**$_{true}$	**H**	**Conf**$_{true}$	**H**	**Conf**$_{true}$	**H**
…						…						
7B	76.82	43.09	71.37	71.05	61.14	10.96	30.93	72.41	79.52	17.69	47.4	16.05
7BQ	75.76*	44.42*	71.37	70.07*	59.73*	10.87	30.36	73.19	79.14	18.11	47.27	16.1
…						…						
7B	70.22	56.26	66.83	71.19	56.83	11.37	31.28	73.87	77.04	19.53	46.88	16.62
7BQ	65.01*	66.4*	61.48*	84.3*	53.69*	12.21*	28.42*	76.09	75.3*	20.61	46.97	17.65
…						…						
560M	44.01	82.38	51.83	90.77	31.5	12.41	17.68	75.57	62.89	27.68	47.93	18.72
560MQ	42.42*	83.81*	49.78*	93.37*	31.07*	12.26	17.65	73.97	62.07	27.52	48.16	18.61
1B1	47.33	83.0	54.95	82.92	34.51	12.35	19.62	75.19	65.82	25.78	48.11	19.33
1B1Q	45.23*	83.28	55.5*	79.2*	33.67*	12.28	18.26*	76.23	64.99*	26.44	48.3	19.66
1B7	51.15	79.99	57.09	74.65	37.52	13.11	21.19	75.97	67.06	26.28	46.93	19.99
1B7Q	49.21*	83.58*	56.95	74.95	36.52*	13.38	19.83*	77.84	66.23*	27.16	47.41	20.4
3B	54.29	76.14	56.73	79.88	41.26	12.91	22.3	74.9	69.3	25.11	47.1	19.1
3BQ	52.07*	80.42*	56.34*	81.98*	40.41*	12.55	21.64	77.03	68.75	25.6	47.57	19.55
7B1	59.1	71.97	57.67	83.21	46.24	12.87	24.64	75.42	71.88	23.84	46.8	18.6
7B1Q	57.79*	73.52*	57.33*	83.89*	45.23*	12.89	23.74*	75.9	71.38	24.41	46.69	18.65
…						…						
125M	40.09	83.61	52.46	86.7	29.04	12.86	16.95	77.19	61.88	28.97	48.26	20.12
125MQ	39.12*	83.58	50.99*	91.0*	28.62*	12.75	16.27	76.26	61.22	28.24	48.33	19.54
350M	41.01	85.36	53.4	87.14	32.03	13.51	17.41	76.08	63.78	28.0	47.69	19.77
350MQ	40.69	82.59*	51.72*	92.36*	31.84	13.09	17.05	75.09	62.73*	28.31	48.04	19.16
1B3	52.33	79.37	53.88	90.87	41.5	13.03	22.51	75.63	70.02	26.08	47.12	18.73
1B3Q	50.83*	81.23*	52.07*	91.22	40.47*	13.26	22.2	75.09	69.52	25.96	47.21	18.76
2B7	55.63	77.64	54.34	91.63	45.84	12.84	24.63	74.97	72.0	25.68	46.89	18.82
2B7Q	53.91*	79.47*	52.88*	91.66	45.12*	13.09	23.71	76.5	71.45	25.45	46.9	18.78
6B7	60.58	70.14	57.42	88.65	50.37	12.12	26.42	74.26	74.42	22.67	46.81	18.3
6B7Q	60.08	70.89	56.57*	90.78*	49.78*	12.22	26.45	74.65	74.26	22.96	46.63	18.06
13B	62.35	67.09	58.24	86.84	52.15	12.25	26.85	73.83	74.56	22.26	46.88	17.13
13BQ	62.06	68.26*	58.38	87.45*	51.61*	12.24	26.67	73.7	74.38	22.69	46.98	17.38

Table 5: Mean confidence in true classes and predictive entropy evaluation results across benchmarks. Conf$_{true}$: Mean confidence in true class; H=Mean Predictive entropy in the answers. The * is used to denote a significant difference with a confidence level set at 0.001 (paired t-test). Q denotes quantized models.Notations: =MISTRAL; =LLAMA;=BLOOM;=OPT.