A Comprehensive Evaluation of Quantization Strategies for Large Language Models

Renren Jin[1*†], Jiangcun Du[1*], Wuwei Huang[2], Wei Liu[2], Jian Luan[2], Bin Wang[2], Deyi Xiong[1‡]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Xiaomi AI Lab, Beijing, China
{rrjin, d2000, dyxiong}@tju.edu.cn
{huangwuwei, liuwei40, luanjian, wangbin11}@xiaomi.com

Abstract

Increasing the number of parameters in large language models (LLMs) usually improves performance in downstream tasks but raises compute and memory costs, making deployment difficult in resource-limited settings. Quantization techniques, which reduce the bits needed for model weights or activations with minimal performance loss, have become popular due to the rise of LLMs. However, most quantization studies use pre-trained LLMs, and the impact of quantization on instruction-tuned LLMs and the relationship between perplexity and benchmark performance of quantized LLMs are not well understood. Evaluation of quantized LLMs is often limited to language modeling and a few classification tasks, leaving their performance on other benchmarks unclear. To address these gaps, we propose a structured evaluation framework consisting of three critical dimensions: (1) knowledge & capacity, (2) alignment, and (3) efficiency, and conduct extensive experiments across ten diverse benchmarks. Our experimental results indicate that LLMs with 4-bit quantization can retain performance comparable to their non-quantized counterparts, and perplexity can serve as a proxy metric for quantized LLMs on most benchmarks. Furthermore, quantized LLMs with larger parameter scales can outperform smaller LLMs. Despite the memory savings achieved through quantization, it can also slow down the inference speed of LLMs. Consequently, substantial engineering efforts and hardware support are imperative to achieve a balanced optimization of decoding speed and memory consumption in the context of quantized LLMs.

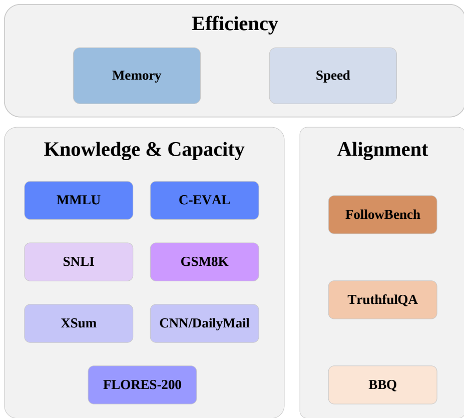


Figure 1: The evaluation framework employed in our study to assess the quantized LLMs from three key dimensions: efficiency, knowledge & capacity and alignment.

1 Introduction

In recent years, LLMs have seen substantial growth in the number of parameters, scaling up to billions or even trillions (Brown et al., 2020; Du et al., 2022; Scao et al., 2022; Touvron et al., 2023a,b; Ren et al., 2023), yielding exceptional performance across various tasks and real-world applications (Zhao et al., 2023; Laskar et al., 2023; Bang et al., 2023; Lai et al., 2023; Mao et al., 2023; Liang et al., 2022; Zhu et al., 2024; Guo et al., 2023). However, the huge number of parameters also results in significant compute and memory requirements, hindering their deployment on devices with limited resources. To mitigate these challenges, researchers have proposed various approaches to model quantization, which aim to optimize model inference and memory usage while minimizing performance degradation.

The central idea of model quantization is representing the weights or activations of a model in a lower-precision format (such as 8-bit integers) rather than their original high-precision floating-point format (typically 16-bit or 32-bit) (Gholami et al., 2021; Zhu et al., 2023). Quantization approaches can be broadly classified into two primary categories: quantization-aware training (QAT) and

*Equal contribution.
†Work done during internship at Xiaomi AI Lab.
‡ Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 12186–12215
August 11-16, 2024 ©2024 Association for Computational Linguistics

Category	Benchmarks	Split	#Samples	Languages	Evaluation Dimension	Metrics	Evaluation Methods
Knowledge & Capacity	MMLU (Hendrycks et al., 2021)	Test	14,042	English	Knowledge	Accuracy ↑	Rule-based
	C-EVAL (Huang et al., 2023)	Test	12,342	Chinese	Knowledge	Accuracy ↑	Rule-based
	FLORES-200 (Costa-jussà et al., 2022)	Test	1,012	English, Chinese	Translation	BLEU ↑	Rule-based
	CNN/DailyMail (See et al., 2017)	Test	11,490	English	Summarization	ROUGE ↑	Rule-based
	XSum (Narayan et al., 2018)	Test	11,334	English	Summarization	ROUGE ↑	Rule-based
	GSM8K (Cobbe et al., 2021)	Test	1,319	English	Mathematical Reasoning	Accuracy ↑	Rule-based
	SNLI (Bowman et al., 2015)	Test	10,000	English	Language Understanding	Accuracy ↑	Rule-based
Alignment	FollowBench (Jiang et al., 2023)	Test	820	English	Instruction Following	Hard Satisfaction Rate (HSR) ↑ Soft Satisfaction Rate (SSR) ↑ Consistent Satisfaction Levels (CSL) ↑	Rule-based GPT-4-as-a-judge
	TruthfulQA (Lin et al., 2022)	Test	817	English	Truthfulness	Accuracy ↑	Rule-based
	BBQ (Parrish et al., 2022)	Test	58,492	English	Social biases	Bias Score → 0 ←	Rule-based

Table 1: Comprehensive overview of benchmarks used in our evaluation experiments.

post-training quantization (PTQ). QAT incorporates the quantization process into the training phase of the model, thereby allowing the model to adapt to lower-precision representations (Liu et al., 2023e; Dettmers et al., 2023a; Kim et al., 2023). Conversely, PTQ applies quantization techniques after the training phase has finished (Dettmers et al., 2022; Frantar et al., 2022; Lin et al., 2023; Lee et al., 2023; Dettmers et al., 2023b; Xiao et al., 2023; Yao et al., 2022).

Despite the risk of performance degradation, PTQ is more prevalent due to the prohibitive training costs associated with QAT. However, several aspects pertaining to the evaluation of PTQ require further exploration. Firstly, the majority of PTQ methods are evaluated solely by assessing the performance of the quantized pre-trained LLMs on benchmarks, leaving the performance of quantized LLMs that have undergone instruction tuning unclear - despite the latter being more commonly used in real-world scenarios (Ouyang et al., 2022; Bai et al., 2022; Peng et al., 2023). Secondly, the evaluation of quantized models is limited to the language modeling task and a few classification tasks. This restricts our understanding of their performance on other benchmarks that are more closely related to real-world applications. Lastly, while perplexity is predominantly employed as the evaluation metric for verifying the effectiveness of quantization methods and has been demonstrated as an indicator of the performance of LLMs on other benchmarks in previous studies (Xia et al., 2023), the correlation between the perplexity of quantized LLMs and their performance on other benchmarks remains poorly understood.

In this paper, we conduct a comprehensive evaluation of the quantized LLMs that undergo instruction tuning, utilizing a diverse range of publicly available benchmarks. These benchmarks cover language understanding and generation, as well as two critical dimensions of LLMs: knowledge & capacity and alignment. Additionally, we evaluate various quantization strategies for their efficiency in terms of generation speed and memory consumption. The comprehensive framework for this evaluation is illustrated in Figure 1, while Table 1 provides a detailed summary of the benchmarks employed in our experiments.[1]

Our contributions can be summarized as follows:

- We propose a structured evaluation framework and conduct extensive experiments to evaluate instruction-tuned LLMs and their quantized counterparts employing various quantization strategies across different parameter scales (7B, 14B, 72B).

- Our empirical findings suggest that LLMs utilizing 4-bit quantization can maintain performance comparable to their non-quantized counterparts on the evaluated benchmarks. Additionally, quantized LLMs with a larger parameter scale demonstrate superior performance compared to their non-quantized counterparts with smaller parameter sizes. Furthermore, we find that perplexity serves as a reliable performance indicator for quantized LLMs across the majority of the benchmarks.

- We identify isolating outlier weights as a key factor enabling SpQR to effectively quantize LLMs to an extreme 2-bit level, significantly outperforming GPTQ at the same level.

- Despite the impressive performance of contemporary quantization approaches, our further analysis reveals substantial engineering

[1]Our code is publicly available at `https://github.com/cordercorder/quant_eval`.

challenges. Specifically, these approaches require significant engineering effort and hardware support to be effectively applied in practical scenarios, particularly in terms of memory and speed requirements.

2 Related Work

LLMs Quantization There are currently two main formalisms of model quantization: QAT (Jacob et al., 2018) and PTQ.

PTQ applies quantization after model training, while QAT considers the effects of quantization during the training process, necessitating considerable resources and expertise, thereby restricting its broader application. Consequently, our research primarily concentrates on PTQ.

Concerning the identification and protection of outlier values, GPT3.int8() (Dettmers et al., 2022) (also known as LLM.int8()) identifies outliers by magnitude while SpQR (Dettmers et al., 2023b) employs Hessian matrix to identify outlier values. By equivalently scaling weights and activation values, SmoothQuant (Xiao et al., 2023) greatly reduces quantization error of activation and thus results in a great reduction in the quantization loss of the model. Outlier Suppression+ (Wei et al., 2023) suppresses the outlier of weights by performing channel-wise shift and scale. QLoRA (Dettmers et al., 2023a) proposes to use the NF4 data format to reduce quantization rounding errors further. OPTQ (Frantar et al., 2023) (generally known as GPTQ) adjusting the weights during the quantization process to reduce quantization errors.

LLMs Evaluation As the technology behind LLMs continues to advance, these models have shown remarkable performance in many tasks (Bang et al., 2023; Mao et al., 2023), sometimes surpassing human proficiency (Srivastava et al., 2022; Laskar et al., 2023). Additionally, as the number of parameters in these models increases, they exhibit emergent abilities (Wei et al., 2022; Schaeffer et al., 2023; Liu et al., 2023b; Lu et al., 2023; Hu et al., 2023), making it challenging to compare their performance to that of other models and understand their behavior. As a result, numerous benchmarks have been curated to rigorously assess the performance of LLMs (Chang et al., 2024; Ziyu et al., 2023; Liu et al., 2023d; Yu et al., 2024). These benchmarks can be divided into two primary categories: (1) knowledge & capacity evaluation (Hendrycks et al., 2021; Li et al., 2023; Zeng, 2023; Huang et al., 2023; Qin et al., 2023; Liu et al., 2024b,a; Shen et al., 2023; Liu et al., 2023a; Shi et al., 2024), which examines the model's ability to understand and generate correct responses; and (2) alignment evaluation, which measures how well the model's outputs align with human preference and values (Gehman et al., 2020; Lin et al., 2022; Parrish et al., 2022; Huang and Xiong, 2023; Liu et al., 2023c; Yin et al., 2023; Zhou et al., 2023). Although these benchmarks are commonly employed to assess LLMs, their quantized counterparts are often excluded from these evaluations. As a result, it can be challenging to comprehend the behavior of quantized LLMs and determine the extent of the performance gap between them and their non-quantized counterparts.

In addressing these challenges, our research primarily focuses on evaluating quantized LLMs. Our goal is to conduct a thorough examination of the performance of LLMs that have been quantized using various methods. In doing so, we hope to yield valuable insights that will inform and enhance future advancements in quantization methodologies.

3 Evaluation Protocol

The comprehensive evaluation of LLMs presents a long-standing challenge due to their versatility, widespread application, and poor explainability. To address this, we propose a structured evaluation framework that encompasses three critical dimensions: (1) knowledge and capacity, (2) alignment, and (3) efficiency.

For the evaluation of knowledge and capacity, we consider two types of benchmarks: (i) those requiring LLMs to demonstrate extensive knowledge across various domains to achieve satisfactory performance, and (ii) those assessing the ability of LLMs to perform specific tasks such as language generation and understanding. In this context, we employ the MMLU (Hendrycks et al., 2021) and C-EVAL (Huang et al., 2023) benchmarks for the former, covering diverse subjects including but not limited to history, chemistry, and economics. For the latter, we select the FLORES-200 (Costa-jussà et al., 2022), CNN/DailyMail (See et al., 2017), and XSum (Narayan et al., 2018) benchmarks, which focus on essential language generation tasks like translation and summarization, and the GSM8K (Cobbe et al., 2021) and SNLI (Bowman et al., 2015) benchmarks for evaluating language understanding and reasoning capabilities.

Model	Datatype	Quantization Method	Average Accuracy	Average BLEU	Average ROUGE-1/ROUGE-2/ROUGE-3	Average HSR/SSR/CSL	Average Bias Score	Average Perplexity	Memory	Speed
Qwen-7B-Chat	BFloat16	-	57.10	29.63	0.257/0.086/0.168	40.23/51.71/1.57	6.20/3.87	11.76	15.14	37.67
	INT-8	LLM.int8()	56.67	28.97	0.256/0.086/0.168	40.52/52.38/1.62	6.41/3.49	11.77	9.23	7.19
		GPTQ	57.21	29.52	0.257/0.087/0.169	40.78/53.00/1.52	5.98/3.76	11.76	10.91	13.57
		SpQR	56.49	29.51	0.257/0.086/0.168	40.20/52.10/1.62	6.36/3.95	11.83	15.60	37.65
	INT-4	GPTQ	54.86	28.43	0.254/0.084/0.167	39.84/52.13/1.38	5.49/3.69	12.31	7.83	37.43
		SpQR	56.41	29.59	0.256/0.086/0.168	40.26/51.59/1.48	6.34/3.77	11.97	15.60	37.73
	INT-3	GPTQ	51.42	24.22	0.228/0.067/0.149	35.82/47.77/1.27	4.21/4.90	15.10	7.12	8.21
		SpQR	55.45	28.39	0.253/0.083/0.166	36.03/49.44/1.30	6.31/4.08	13.40	15.61	37.73
	INT-2	GPTQ	16.52	0.01	0.042/0.000/0.029	0.24/0.64/0.00	-0.54/-0.97	84396.73	6.26	19.36
		SpQR	53.18	27.22	0.242/0.077/0.158	37.52/49.69/1.55	4.13/5.76	13.77	15.66	37.51
Qwen-14B-Chat	BFloat16	-	62.92	31.13	0.254/0.085/0.196	53.16/62.25/2.15	8.35/3.69	9.84	27.60	25.15
	INT-8	LLM.int8()	62.48	31.63	0.254/0.084/0.166	48.35/57.69/1.75	7.92/3.89	9.86	15.91	5.85
		GPTQ	62.67	31.84	0.254/0.084/0.196	49.22/58.76/1.90	8.22/3.70	9.85	17.92	14.37
		SpQR	62.86	31.97	0.255/0.085/0.167	47.53/57.59/1.87	8.60/3.65	9.85	27.95	25.42
	INT-4	GPTQ	61.53	31.40	0.252/0.082/0.165	48.66/57.47/1.90	7.82/4.11	10.29	12.03	24.38
		SpQR	62.66	31.47	0.252/0.083/0.165	46.84/56.27/1.78	7.96/3.86	9.94	27.95	24.62
	INT-3	GPTQ	58.34	28.92	0.237/0.073/0.155	43.91/53.38/1.62	8.41/3.88	13.94	10.77	4.71
		SpQR	61.43	31.09	0.253/0.082/0.165	47.83/57.65/1.85	8.03/3.11	10.19	29.97	25.20
	INT-2	GPTQ	16.78	0.01	0.044/0.000/0.030	0.54/1.12/0.02	-0.17/-0.81	192872.47	8.99	18.26
		SpQR	59.82	29.20	0.247/0.080/0.162	47.76/57.47/1.82	8.08/5.33	11.00	28.04	24.82
Qwen-72B-Chat	BFloat16	-	71.76	34.81	0.300/0.114/0.203	53.16/62.25/2.15	9.07/1.57	8.52	138.44	8.97
	INT-8	LLM.int8()	71.74	34.39	0.301/0.115/0.204	56.00/64.03/2.28	8.81/1.68	8.51	74.96	3.07
		GPTQ	71.20	34.82	0.300/0.114/0.204	54.66/63.28/2.08	8.95/1.31	8.71	77.85	1.43
		SpQR	71.90	34.67	0.300/0.115/0.203	54.27/62.67/2.33	9.07/1.51	8.54	143.20	6.57
	INT-4	GPTQ	71.38	34.03	0.298/0.112/0.201	52.81/61.43/2.13	10.11/1.76	8.77	44.11	14.88
		SpQR	71.76	34.72	0.299/0.114/0.201	53.81/62.14/2.10	8.73/1.52	8.64	143.21	6.56
	INT-3	GPTQ	66.89	30.98	0.269/0.096/0.178	52.73/61.38/2.02	8.77/3.11	10.19	35.93	0.84
		SpQR	70.67	34.08	0.292/0.109/0.196	52.82/61.42/2.22	8.24/1.97	8.84	143.37	6.57
	INT-2	GPTQ	18.40	0.01	0.021/0.000/0.017	0.08/0.48/0.00	-0.36/0.89	48714.04	27.74	2.23
		SpQR	67.07	32.74	0.278/0.100/0.186	53.48/61.76/2.20	7.67/1.69	9.57	144.59	6.56

Table 2: Evaluation results of the Qwen-Chat series models and their quantized counterparts across ten benchmarks designed to evaluate LLMs in terms of knowledge & capacity and alignment, as well as metrics for memory consumption and decoding speed during inference. The benchmarks are grouped by the type of metrics used, and the average score for each metric within its respective group is presented. The "Average Accuracy" represents the mean accuracy of LLMs across the MMLU (Hendrycks et al., 2021), C-EVAL (Huang et al., 2023), GSM8K (Cobbe et al., 2021), SNLI (Bowman et al., 2015), and TruthfulQA (Lin et al., 2022) benchmarks. The "Average BLEU" indicates the mean BLEU score for Chinese-English and English-Chinese translations on the FLORES-200 benchmark (Costa-jussà et al., 2022). The "Average ROUGE-1/ROUGE-2/ROUGE-L" displays the mean ROUGE-1, ROUGE-2, and ROUGE-L scores on the XSum (Narayan et al., 2018) and CNN/DailyMail (See et al., 2017) benchmarks. The "Average HSR/SSR/CSL" represents the mean hard satisfaction rates (HSR), soft satisfaction rates (SSR) across five difficulty levels, and consistent satisfaction levels (CSL) across five constraints on the FollowBench benchmark (Jiang et al., 2023). The "Average Bias Score" is shown as x/y, where x and y represent the mean bias score across various categories in ambiguous and unambiguous contexts, respectively, on the BBQ benchmark (Parrish et al., 2022). The "Average Perplexity" indicates the mean perplexity on WikiText2 (Merity et al., 2017), C4 (Raffel et al., 2020), and PTB (Marcus et al., 1994). The "Memory" refers to the memory consumed (in GB) during inference when the input consists of 256 tokens and the output contains 512 tokens. The "Speed" represents the number of tokens generated per second when the input consists of 256 tokens and the output contains 512 tokens.

For the evaluation of alignment, we adopt the HHH criteria proposed by Askell et al. (2021), which assess LLMs from three distinct perspectives: helpfulness, honesty, and harmlessness. Accordingly, we have chosen the FollowBench (Jiang et al., 2023), TruthfulQA (Lin et al., 2022), and BBQ (Parrish et al., 2022) benchmarks to assess these aspects, respectively.

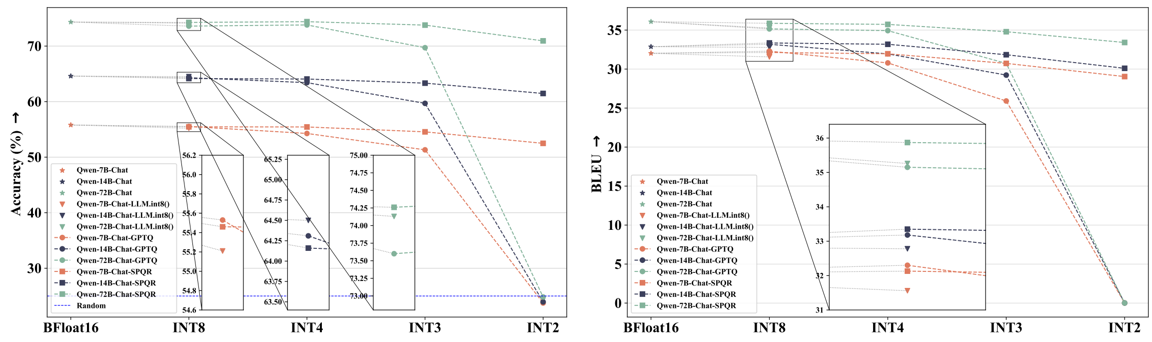
For the evaluation of efficiency, we consider metrics such as memory usage and generation speed during inference, which are crucial for the practical application of LLMs in real-world scenarios.

It is important to note that while these three dimensions provide a comprehensive framework for evaluating LLMs, other benchmarks or metrics can also be employed as long as they align with these dimensions.

4 Evaluation Setup

4.1 LLMs

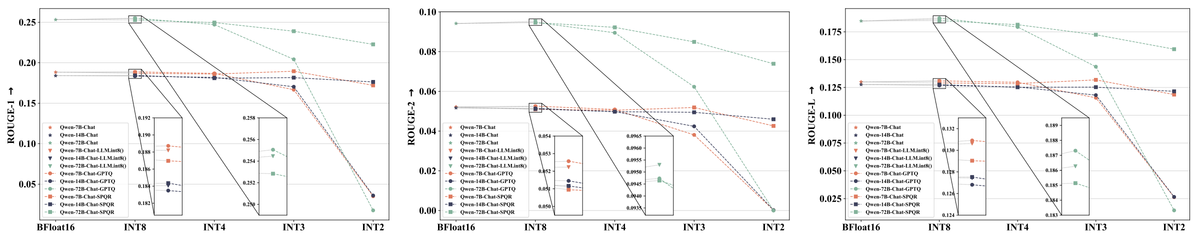
We predominantly employ quantization techniques on the Qwen-Chat series of models (Bai et al., 2023), which have undergone instruction tuning, taking into account the following considerations: (1) The Qwen-Chat models have demonstrated exceptional performance across a variety of tasks. (2) The Qwen-Chat series includes LLMs of varying parameter scales, specifically models with 7 billion, 14 billion, and 72 billion parameters. (3) The models in the Qwen-Chat series have been pre-trained on an extensive corpus of multilingual data, with a particular focus on Chinese and English. This extensive pre-training enables the models to support a multitude of languages beyond English.



(a) MMLU

(b) En → Zh

Figure 2: Performance of the Qwen-Chat series models and their quantized counterparts on the MMLU (Hendrycks et al., 2021) benchmark (a) and the English-to-Chinese (En → Zh) translation task of the FLORES-200 (Costa-jussà et al., 2022) (b) benchmark. The x-axis represents the data format of the model's weight, where x in INTx denotes the number of integer bits used for weight representations. To highlight the nuanced differences between LLM.int8() and other methodologies, a magnified view is integrated into the figure.



(a) ROUGE-1

(b) ROUGE-2

(c) ROUGE-L

Figure 3: ROUGE-1 (a), ROUGE-2 (b), and ROUGE-L (c) scores for the Qwen-Chat series models and their quantized counterparts on the test sets of XSum (Narayan et al., 2018).

4.2 Quantization Strategies

We select three prominent quantization approaches accompanied by dedicated open-source implementations for evaluation: LLM.int8() (Dettmers et al., 2022), GPTQ (Frantar et al., 2023), and SpQR (Dettmers et al., 2023b). These approaches have been either deeply integrated into the Hugging Face Transformers[2] library (Wolf et al., 2020) or widely used, thereby enabling them to support a variety of open-source LLMs. Specifically, we employ GPTQ and SpQR to quantize the LLMs to 8, 4, 3, and 2 bits, respectively, except LLM.int8(), which exclusively quantizes them to 8 bits. For the calibration data required by SpQR and GPTQ, we randomly sampled 128 examples from the dataset collected by Taori et al. (2023) and Peng et al. (2023). For a detailed introduction to these quantization approaches, please refer to Appendix A.

4.3 Benchmarks

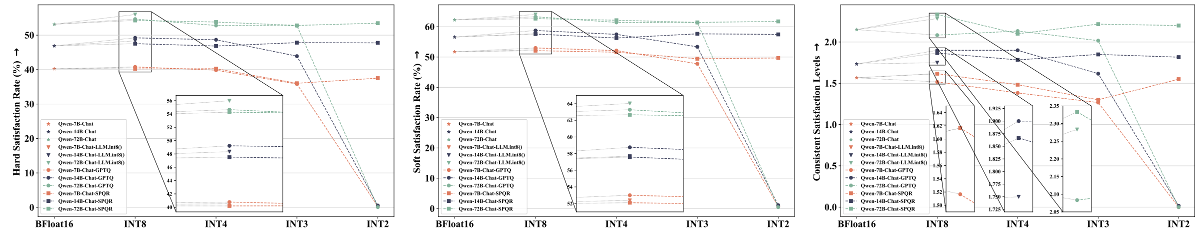
We utilize ten distinct benchmarks to facilitate a comprehensive assessment of LLMs and their quantized counterparts. These benchmarks encompass knowledge & capacity evaluation, as well as alignment evaluation. By leveraging this broad spectrum of benchmarks, we aim to gain a holistic understanding of the models' performance across various dimensions, thereby enabling a detailed comparison between the original and quantized versions of LLMs. For a comprehensive overview these benchmarks and the associated prompts employed in our study, please see Appendix B and Appendix C.

5 Experiment Results and Discussion

Table 2 presents a comprehensive performance summary of the Qwen-Chat series models and their quantized counterparts across ten benchmarks designed to evaluate LLMs in terms of knowledge & capacity and alignment. It also includes metrics for memory consumption and decoding speed during inference. Detailed experimental results for each benchmark and metric are illustrated in Figures 2 through 8 and Figures 15 through 17 in Appendix D.

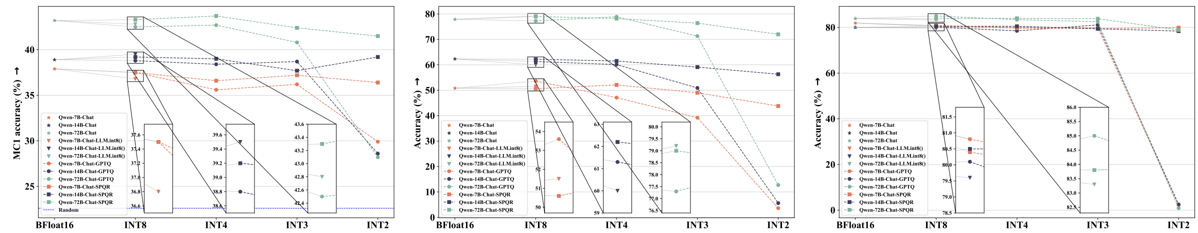
Overall, the experimental results indicate that

[2]https://github.com/huggingface/transformers



(a) Hard Satisfaction Rate (HSR) (b) Soft Satisfaction Rate (SSR) (c) Consistent Satisfaction Levels (CSL)

Figure 4: Average hard satisfaction rates (a), soft satisfaction rates (b), and consistent satisfaction levels (c) across five difficulty levels for the Qwen-Chat series models and their quantized counterparts on the FollowBench benchmark (Jiang et al., 2023).



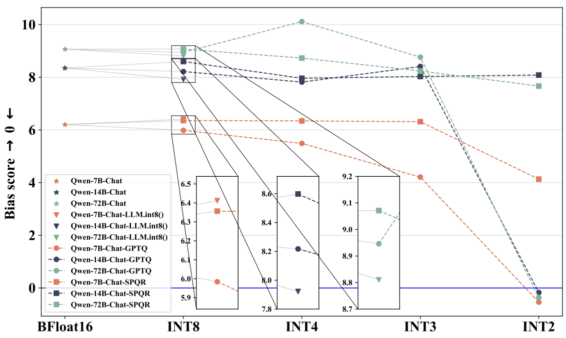
(a) TruthfulQA (b) GSM8K (c) SNLI

Figure 5: Performance of Qwen-Chat series models and their quantized counterparts on the TruthfulQA benchmark (Lin et al., 2022) (a), as well as the test sets of GSM8K (Cobbe et al., 2021) (b) and SNLI (Bowman et al., 2015) (c).

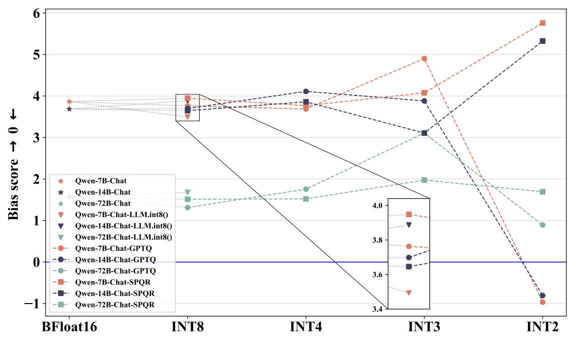
LLMs with a greater number of parameters generally outperform those with fewer parameters across most benchmarks. Furthermore, we observe a downward trend in the performance of these LLMs across most benchmarks when they are quantized to fewer bits. Here are the detailed observations:

4-bit quantization offers a trade-off between the LLMs' capacity and the number of bits in the low-precision format. As the number of quantized bits decreases to 3 bits or lower, there is a noticeable performance discrepancy between the LLMs and their quantized counterparts. Experimental results suggest that when the LLMs are quantized to 8 bits, the majority of LLMs, irrespective of their parameter scales, can maintain a performance level comparable to their non-quantized equivalents. Moreover, LLMs that are quantized to 4 bits can also uphold similar performance to their non-quantized versions across most benchmarks. However, if these LLMs are further quantized to 3 bits or lower, the capacity of these models begins to deteriorate. Notably, our investigation reveals that when the LLMs are quantized to 2 bits using GPTQ, they lose their ability to comprehend and follow user instructions, resulting in the generation of incoherent text.

Perplexity is a reliable performance indicator for quantized LLMs on evaluation benchmarks. Figure 7 illustrates the perplexity for both the original LLMs and their quantized versions on WikiText2 (Merity et al., 2017). For more experimental results of perplexity on the C4 (Raffel et al., 2020) and PTB (Marcus et al., 1994) datasets, please refer to Figure 17 in Appendix D. It is evident that the perplexity of 8-bit quantized models closely matches that of their non-quantized counterparts. Moreover, as the LLMs are further quantized to 4 and 3 bits, there's a slight increase in perplexity. However, perplexity sharply increases, exceeding 38,000, when the models are quantized to 2 bits using GPTQ. This sharp increase in perplexity aligns with our observation that models quantized to 2 bits with GPTQ struggle to generate coherent text. In summary, as LLMs are quantized to fewer bits, there is an upward trend in perplexity, which corresponds to a decline in their performance on evaluated benchmarks. Interestingly, despite a noticeable increase in perplexity, 4-bit quantized models still perform comparably to their non-quantized counterparts on these benchmarks. We speculate that this could be due to the nonlinear or discontinuous metrics used by these benchmarks, which may not reflect minor changes in perplexity. Furthermore, as demonstrated in Table 3, there is a strong



(a) Bias scores in ambiguous context.



(b) Bias scores in disambiguated context.

Figure 6: Bias scores of the Qwen-Chat series models and their quantized counterparts in ambiguous and disambiguated contexts on the BBQ benchmark (Parrish et al., 2022).

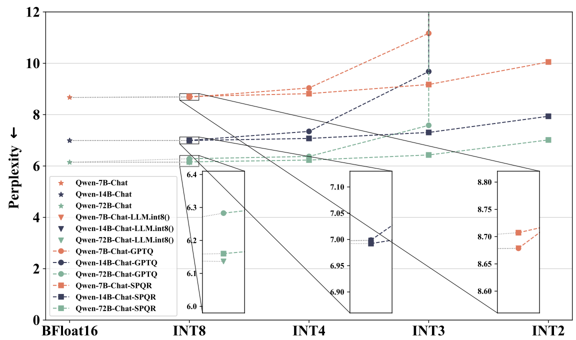


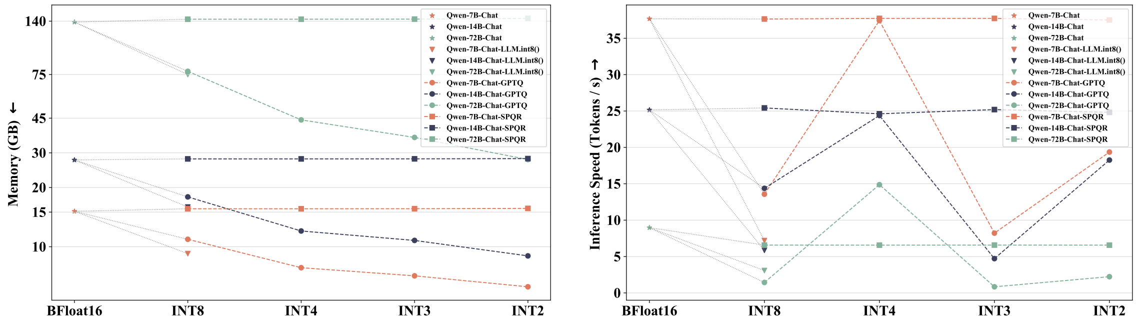
Figure 7: Perplexity of Qwen-Chat Series models and their quantized counterparts on the WikiText2 dataset (Merity et al., 2017).

Benchmark	Metric	Pearson Correlation Coefficient
MMUL	Accuracy	-0.892
C-EVAL	Accuracy	-0.930
FLORES-200	BLEU (English to Chinese)	-0.884
	BLEU (Chinese to English)	-0.904
XSum	ROUGE-1	-0.768
	ROUGE-2	-0.493
	ROUGE-L	-0.222
CNN/DailyMail	ROUGE-1	-0.890
	ROUGE-2	-0.849
	ROUGE-L	-0.885
GSM8K	Accuracy	-0.911
SNLI	Accuracy	-0.583
FollowBench	HSR (hard satisfaction rates)	-0.864
	SSR (soft satisfaction rates)	-0.899
	CSL (consistent satisfaction levels)	-0.877
TruthfulQA	MC1 Accuracy	-0.789
BBQ	Bias scores in ambiguous context	-0.765
	Bias scores in disambiguated context	0.806

Table 3: The Pearson correlation coefficient between the average perplexity on the WikiText2, C4, and PTB datasets of both 4-bit and 3-bit quantized LLMs (quantized with GPTQ and SpQR) and their performance across various benchmarks.

correlation between perplexity and the performance of quantized LLMs. The average absolute value of the Pearson correlation coefficient is notably high at 0.7895. This evidence reinforces our claim that perplexity serves as a reliable performance indicator for quantized LLMs on evaluation benchmarks.

Identifying and isolating outlier weights is crucial for SpQR to effectively quantize LLMs to an extreme level of 2 bits. Experimental results indicate a sharp decline in the performance of LLMs quantized to 2 bits by GPTQ, to the extent that they fail to produce coherent text. In contrast, LLMs quantized to 2 bits by SpQR exhibit a relatively moderate performance across all evaluated benchmarks. SpQR introduces two innovative strategies to enhance the performance of quantized LLMs, distinguishing it from GPTQ: (1) the adoption of an extremely small group size coupled with bilevel quantization, and (2) the isolation of unstructured outlier weights, maintaining these weights at a higher precision (16-bit) during computations. To study the impact of these strategies, we conducted two controlled experiments: (1) increasing the group size of SpQR from 16 to 128, matching the group size utilized by GPTQ, while still isolating the outlier weights. (2) keeping the small group size but not isolating outlier weights. Experimental results are shown in Table 4. We observe a significant increase in perplexity across three benchmarks when the outlier weights are not isolated, even with a small group size. Conversely, increasing the group size resulted in only a marginal increase in perplexity. Furthermore, we analyzed the proportion of outlier weights stored in high precision for the quantized LLMs, with the results presented in Table 5. These findings indicate an inverse relationship between the number of quantized bits and the percentage of outlier weights, with a consis-



(a) Memory

(b) Speed

Figure 8: **Left:** memory consumption comparison between Qwen-Chat series models and their quantized counterparts. The y-axis is presented on a logarithmic scale to clearly demonstrate the variation in memory consumption for LLMs with smaller parameter scales (7B, 14B) as the number of quantized bits decreases. **Right:** comparison of inference speed between Qwen-Chat series models and their quantized counterparts. These experiments are conducted with an input of 256 tokens and a generation of 512 tokens on A100 80GB SXM GPUs.

Model	Quantization Config	WikiText	C4	PTB
Qwen-7B-Chat	w2g16 w/ outlier	10.05	14.19	17.07
	w2g16 w/o outlier	17.96	21.86	27.98
	w2g128 w/ outlier	10.58	14.48	17.40
Qwen-14B-Chat	w2g16 w/ outlier	7.94	11.74	13.31
	w2g16 w/o outlier	140.22	115.07	170.48
	w2g128 w/ outlier	8.16	12.14	13.78
Qwen-72B-Chat	w2g16 w/ outlier	7.01	9.78	11.92
	w2g16 w/o outlier	10.49	14.07	16.11
	w2g128 w/ outlier	7.44	10.34	12.27

Table 4: Perplexity on WikiText2 (Merity et al., 2017), C4 (Raffel et al., 2020), and PTB (Marcus et al., 1994) under different quantization configurations. "w2g16" denotes that weights are quantized to 2-bit with a group size of 16. "w/ outlier" indicates identifying outlier values which are not quantized while "w/o outlier" means not identifying outliers and the whole weight matrix is quantized. All experiments used bilevel 3-bit quantization, which quantizes the model's weights first and then quantizes group-wise statistics (scales and zeros).

Model	Quantized Bit	Outlier Proportion
Qwen-7B-Chat	8	0.003%
	4	0.033%
	3	1.676%
	2	11.336%
Qwen-14B-Chat	8	0.004%
	4	0.036%
	3	1.648%
	2	11.148%
Qwen-72B-Chat	8	0.003%
	4	0.044%
	3	1.682%
	2	11.838%

Table 5: The proportion of outliers keeping high precision in LLMs quantized by SpQR.

tent percentage of outlier weights across different model scales at the same quantization level. Consequently, it is concluded that the isolation of outlier weights and their preservation in high precision is indispensable for SpQR to effectively quantize LLMs to an extreme level of 2 bits.

In practical scenarios, the application of low-bit quantization necessitates substantial engineering effort and hardware support. As illustrated in Figure 8a, both the GPTQ and LLM.int8() can effectively reduce memory consumption during LLMs inference, with the memory requirement diminishing as the number of quantized bits decreases. Conversely, despite the impressive performance of SpQR, it does not contribute to reducing memory consumption during LLMs inference. This is attributable to the implementation of SpQR employed in our study, which utilizes a high-precision format to represent quantized weights. It merely restricts the range of quantized weights to match that of the low-precision format, thereby mimicking the effect of representing quantized weights with low precision. Consequently, computations are executed under a high-precision format, resulting in no reduction in memory consumption. Furthermore, the efficient implementation of parallel computation in low-precision format is not yet supported by most computing libraries, such as PyTorch. This implies that the implementation of operators associated with low-precision format must be done manually, demanding a thorough understanding of computing hardware (e.g., GPU, TPU, etc.) and the dedication of considerable engineering effort to achieve efficient execution.

Beyond memory consumption, Figure 8b reveals that while GPTQ and LLM.int8(), whose underlying implementation used in our study perform computation in low-precision format, lead to notable memory savings compared to their non-quantized counterparts, the inference speed of LLMs quantized by GPTQ and LLM.int8() is slower compared to their non-quantized counterparts, except in the case of 4-bit quantization. This slowdown is primarily due to the fact that only the weights of the LLMs use the low-precision format representation, while activations still employ the high-precision format representation. The acceleration of computation between this mixed precision format is not supported by the hardware used in our experiments. However, in the case of 4-bit quantization, only the LLM with 72B parameters exhibits a significant speed-up compared to its non-quantized counterpart, while others show similar inference speeds to their non-quantized counterparts. We hypothesize that this may be due to characteristics of the hardware, such as memory bandwidth (Shazeer, 2019). In summary, both the efficient implementation of parallel computation in low-precision format, which requires considerable engineering effort, and the acceleration of computation supported by associated hardware are essential for quantization techniques to effectively reduce memory usage and accelerate decoding during inference.

At similar levels of memory consumption, LLMs quantized to lower bit precision with a larger parameter scale can be preferred over LLMs with a smaller parameter scale, considering their performance capabilities. As illustrated in Figure 8a, the memory consumption during inference for Qwen-14B-Chat with 8-bit or 4-bit quantization by GPTQ is similar to that of Qwen-7B-Chat. However, the former outperforms the latter in most of the benchmarks evaluated. Additionally, Qwen-14B-Chat with 4-bit or 3-bit quantization can be competitive with Qwen-7B-Chat with 8-bit quantization. Nonetheless, while quantized LLMs offer advantages in terms of memory efficiency, quantization can also result in reduced inference speed. Therefore, these quantization approaches are most suitable for scenarios where memory is limited and inference speed is a secondary consideration.

6 Conclusion

We have presented a comprehensive evaluation of quantization strategies for LLMs, demonstrating the trade-offs between model efficiency and performance degradation across various benchmarks. By employing a structured evaluation framework that assesses models in terms of knowledge & capacity, alignment, and efficiency, we aim to offer valuable insights into the scalability and practical application of quantized LLMs. Experimental findings indicate that while 4-bit quantization maintains performance close to non-quantized counterparts, a notable performance discrepancy emerges as quantization decreases to 3 bits or lower. Moreover, the results suggest that perplexity can be a reliable performance indicator for quantized LLMs on various evaluation benchmarks. SpQR effectively quantizes LLMs to an extreme level of 2 bits by isolating outlier weights and maintaining high precision during computation. When memory constraints exist and inference speed is a secondary concern, LLMs quantized to lower bit precision with a larger parameter scale can be preferred over smaller models. Additionally, we highlight the need for engineering effort and hardware support to efficiently deploy quantized LLMs in real-world scenarios.

Limitations

We have utilized ten distinct benchmarks, encompassing knowledge & capacity and alignment, for our evaluation. However, LLMs are pre-trained on vast amounts of data. This could potentially lead to the contamination of some test examples in the benchmarks we used with pre-training data, possibly resulting in an overestimation of the LLMs' performance (Yang et al., 2023; Li, 2023; Oren et al., 2023). Consequently, it remains unclear whether the evaluated experimental results on these benchmarks could be generalized to other benchmarks. Identifying and eliminating these contaminated examples poses a significant challenge, and we leave it as our further work. Furthermore, due to limited computational resources, our experiments were confined to the Qwen-Chat series of models (Bai et al., 2023), which have diverse parameter scales and are trained on an extensive multilingual corpus that is dominated by both English and Chinese. The experimental results and findings from the Qwen series of models may not necessarily generalize to other LLMs, owing to various factors such as differences in training data, hyperparameters, and architectures.

Ethical Considerations

In this study, we employ the BBQ (Parrish et al., 2022) and TruthfulQA (Lin et al., 2022) benchmarks to investigate the potential impact of quantization on the alignment of LLMs with human values. Our focus is on assessing the social bias and truthfulness of both quantized and non-quantized versions of these models.

The experimental results, as illustrated in Figure 6, reveal no consistent trend of increase or decrease in social bias when LLMs are quantized to fewer bits. However, it is noteworthy that quantization can either exacerbate or alleviate the social bias of the quantized LLMs in comparison to their non-quantized counterparts. Furthermore, Figure 5a demonstrates that the truthfulness of LLMs can also be influenced by quantization. Specifically, when LLMs are quantized to 2 bits using GPTQ, there is a significant decrease in the truthfulness of the quantized LLMs.

In conclusion, our findings suggest that in addition to commonly evaluated dimensions such as knowledge & capacity and efficiency, the alignment of LLMs with human values, which is a dimension often overlooked in previous studies of LLM quantization, deserves greater attention.

Acknowledgements

The present research was supported by the National Key Research and Development Program of China (Grant No. 2023YFE0116400). We would like to thank the anonymous reviewers for their insightful comments.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram

Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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

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

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P. Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen S. Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V. Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5547–5569. PMLR.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: accurate post-training quantization for generative pre-trained transformers. *CoRR*, abs/2210.17323.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. OPTQ: accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3356–3369. Association for Computational Linguistics.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. 2021. A survey of quantization methods for efficient neural network inference. *CoRR*, abs/2103.13630.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Trans. Assoc. Comput. Linguistics*, 10:522–538.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey. *CoRR*, abs/2310.19736.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Shengding Hu, Xin Liu, Xu Han, Xinrong Zhang, Chaoqun He, Weilin Zhao, Yankai Lin, Ning Ding, Zebin Ou, Guoyang Zeng, Zhiyuan Liu, and Maosong Sun. 2023. Unlock predictable scaling from emergent abilities. *CoRR*, abs/2310.03262.

Yufei Huang and Deyi Xiong. 2023. CBBQ: A chinese bias benchmark dataset curated with human-ai collaboration for large language models. *CoRR*, abs/2306.16244.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, jiayi lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew G. Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2704–2713. Computer Vision Foundation / IEEE Computer Society.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2023. Followbench: A multi-level fine-grained constraints following benchmark for large language models. *CoRR*, abs/2310.20410.

Jeonghoon Kim, Jung Hyun Lee, Sungdong Kim, Joonsuk Park, Kang Min Yoo, Se Jung Kwon, and Dongsoo Lee. 2023. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. *CoRR*, abs/2305.14152.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13171–13189. Association for Computational Linguistics.

Md. Tahmid Rahman Laskar, M. Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy X. Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 431–469. Association for Computational Linguistics.

Changhun Lee, Jungyu Jin, Taesu Kim, Hyungjun Kim, and Eunhyeok Park. 2023. OWQ: lessons learned from activation outliers for weight quantization in large language models. *CoRR*, abs/2306.02272.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. CMMLU: measuring massive multitask language understanding in chinese. *CoRR*, abs/2306.09212.

Yucheng Li. 2023. Estimating contamination via perplexity: Quantifying memorisation in language model evaluation. *CoRR*, abs/2309.10677.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. *CoRR*, abs/2211.09110.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. 2023. AWQ: activation-aware weight quantization for LLM compression and acceleration. *CoRR*, abs/2306.00978.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Chuang Liu, Renren Jin, Yuqi Ren, and Deyi Xiong. 2024a. LHMKE: A large-scale holistic multi-subject knowledge evaluation benchmark for chinese large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10476–10487. ELRA and ICCL.

Chuang Liu, Renren Jin, Yuqi Ren, Linhao Yu, Tianyu Dong, Xiaohan Peng, Shuting Zhang, Jianxiang Peng, Peiyi Zhang, Qingqing Lyu, Xiaowen Su, Qun Liu, and Deyi Xiong. 2023a. M3KE: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. *CoRR*, abs/2305.10263.

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

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023c. Alignbench: Benchmarking chinese alignment of large language models. *CoRR*, abs/2311.18743.

Yan Liu, Renren Jin, Lin Shi, Zheng Yao, and Deyi Xiong. 2024b. Finemath: A fine-grained mathematical evaluation benchmark for chinese large language models. *CoRR*, abs/2403.07747.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023d. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *CoRR*, abs/2308.05374.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. 2023e. LLM-QAT: data-free quantization aware training for large language models. *CoRR*, abs/2305.17888.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning? *CoRR*, abs/2309.01809.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2023. Gpteval: A survey on assessments of chatgpt and GPT-4. *CoRR*, abs/2308.12488.

Mitchell P. Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Human Language Technology, Proceedings of a Workshop held at Plainsboro, New Jerey, USA, March 8-11, 1994*. Morgan Kaufmann.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 280–290. ACL.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1797–1807. Association for Computational Linguistics.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. 2023. Proving test set contamination in black box language models. *CoRR*, abs/2310.17623.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2086–2105. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *CoRR*, abs/2304.03277.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *CoRR*, abs/2307.16789.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Xiaozhe Ren, Pingyi Zhou, Xinfan Meng, Xinjing Huang, Yadao Wang, Weichao Wang, Pengfei Li, Xiaoda Zhang, Alexander Podolskiy, Grigory Arshinov, Andrey Bout, Irina Piontkovskaya, Jiansheng Wei, Xin Jiang, Teng Su, Qun Liu, and Jun Yao. 2023. Pangu-Σ: Towards trillion parameter language model with sparse heterogeneous computing. *CoRR*, abs/2303.10845.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083. Association for Computational Linguistics.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *CoRR*, abs/1911.02150.

Tianhao Shen, Sun Li, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models. *CoRR*, abs/2312.16132.

Dan Shi, Chaobin You, Jiantao Huang, Taihao Li, and Deyi Xiong. 2024. CORECODE: A common sense annotated dialogue dataset with benchmark tasks for chinese large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18952–18960. AAAI Press.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Xiuying Wei, Yunchen Zhang, Yuhang Li, Xiangguo Zhang, Ruihao Gong, Jinyang Guo, and Xianglong Liu. 2023. Outlier suppression+: Accurate quantization of large language models by equivalent and effective shifting and scaling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1648–1665. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. 2023. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13711–13738. Association for Computational Linguistics.

Guangxuan Xiao, Ji Lin, Mickaël 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, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 38087–38099. PMLR.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. 2023. Rethinking benchmark and contamination for language models with rephrased samples. *CoRR*, abs/2311.04850.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. 2022. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems*

35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8653–8665. Association for Computational Linguistics.

Linhao Yu, Qun Liu, and Deyi Xiong. 2024. LFED: A literary fiction evaluation dataset for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10466–10475. ELRA and ICCL.

Hui Zeng. 2023. Measuring massive multitask chinese understanding. *CoRR*, abs/2304.12986.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911.

Shaolin Zhu, Menglong Cui, and Deyi Xiong. 2024. Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 16619–16629. ELRA and ICCL.

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

Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 88–109, Harbin, China. Chinese Information Processing Society of China.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405.

A Quantization Strategies

LLM.int8() Dettmers et al. (2022) is the earliest proposed method among the methods we use. It was implemented in bitsandbytes[3] and deeply integrated with Huggingface Transformers. LLM.int8() proposes a vector-wise quantization approach, and stores the outlier submatrices in FP16 format while regular submatrices are in int8. In the matrix multiplication operation, the FP16 submatrix and the int8 submatrix are computed separately. This protects the outlier value, but the inference speed will decrease.

GPTQ Frantar et al. (2023) is a popular quantization method. Due to the outstanding contribution of the third-party library AutoGPTQ,[4] which provides CUDA implementation of quantization operators, it can also be easily applied to the model. GPTQ quantizes a weight matrix column by column and uses the Hessian matrix to adjust the unquantized parts of a weight matrix to minimize the loss caused by quantizing some parameters.

SpQR Dettmers et al. (2023b) cleverly combines GPTQ (Frantar et al., 2023) and outlier value protection to further improve quantization performance. It uses a smaller group size and saves outliers through a sparse matrix. Currently, this method has not yet implemented the CUDA operator. So we need to use floating point numbers to simulate the integer quantization, which is called fake quantization. The code used in our experiments was modified from the official code[5] and adapted to Qwen.

B Benchmarks

MMLU Hendrycks et al. (2021) serves as a comprehensive benchmark to measure the knowledge acquired by LLMs during their pretraining phase through zero- and few-shot learning. It encompasses 57 disciplines that cover diverse areas including STEM, humanities, social sciences, law, and ethics. These disciplines collectively evaluate the breadth and depth of a model's understanding across numerous academic and professional domains.

C-EVAL Huang et al. (2023) is a comprehensive Chinese evaluation suite specifically tailored to assess the advanced knowledge and reasoning capabilities of LLMs within the Chinese context. Analogous to MMLU (Hendrycks et al., 2021), it comprises 52 disciplines, ranging from humanities to science and engineering, categorized within four difficulty levels: middle school, high school, college, and professional.

FLORES-200 Costa-jussà et al. (2022) is a high-quality benchmark for machine translation that encompasses 204 languages, doubling the language coverage of its predecessor, FLORES-101 (Goyal et al., 2022). Every sentence in each language has been translated into the others by professional translators. This unique feature establishes FLORES-200 as a many-to-many translation benchmark. Consequently, it is particularly well-suited for the evaluation of translation directions in which both the source and target languages are involved in the FLORES-200 benchmark.

CNN/DailyMail Nallapati et al. (2016); See et al. (2017) is a valuable resource for abstractive multi-sentence summarization. It is derived from a previous dataset created by Hermann et al. (2015) for passage-based question-answering, using human-generated abstractive summary bullets from news stories on the CNN and Daily Mail websites. These summaries are originally used as questions with a masked entity, paired with corresponding passages from which systems are expected to generate answers. CNN/DailyMail is constructed by restoring all the original summary bullets for each story, treating them as separate sentences to form coherent, multi-sentence summaries. CNN/DailyMail consists of a large number of instances, including 286,817 training instances, 13,368 validation instances, and 11,487 test instances. The test instances are utilized in our evaluation experiments.

XSum Narayan et al. (2018) is a fundamental resource for the development and assessment of abstractive single-document summarization systems. It is derived from online articles sourced from the British Broadcasting Corporation (BBC), which typically include professionally written introductory sentences serving as concise one-sentence summaries that encapsulate the essence of the entire article. XSum covers a wide range of domains, including news, politics, sports, weather, and more. Notably, the documents and summaries in XSum are shorter compared to CNN/DailyMail. Furthermore, the summaries in XSum are significantly

[3] https://github.com/TimDettmers/bitsandbytes
[4] https://github.com/AutoGPTQ/AutoGPTQ
[5] https://github.com/Vahe1994/SpQR

more abstractive, as evidenced by a notable percentage of novel n-grams that are not present within the source documents. The dataset has been randomly partitioned into training (90%), validation (5%), and test (5%) splits. The evaluation experiments in our work are conducted using the test set.

GSM8K Cobbe et al. (2021) is a collection of 8,500 high-quality grade school math word problems designed to evaluate the multi-step mathematical reasoning abilities of LLMs. The dataset has been meticulously curated to ensure high linguistic diversity. The problems included in GSM8K only involve relatively simple math concepts that a bright middle school student can solve using basic arithmetic operations such as addition, subtraction, multiplication, and division over a sequence of 2 to 8 steps.

SNLI Bowman et al. (2015) is a large-scale, human-annotated collection of sentence pairs specifically designed for training and evaluating machine learning models on the task of natural language inference (NLI). All sentences in SNLI are written by human contributors within a grounded context based on image captioning, ensuring that they reflect naturalistic language use rather than being algorithmically generated. Each sentence pair within the dataset is labeled as either an entailment, a contradiction, or neutral. SNLI has been partitioned into training, development, and test splits. Both the development and test splits encompass 10,000 examples each. The test split, in particular, is utilized in our evaluation experiments.

FollowBench Jiang et al. (2023) is a comprehensive benchmark that focuses on evaluating the instruction-following capabilities of LLMs through a variety of fine-grained constraints. It encompasses five distinct fine-grained constraints: content, situation, style, format, and example. This benchmark is specifically designed to address the limitations of existing evaluation benchmarks, which primarily assess the quality of responses without measuring their adherence to specific instruction constraints. FollowBench is available in two language splits, English and Chinese, with the English split used in our evaluation experiments.

TruthfulQA Lin et al. (2022) is a benchmark designed to assess the truthfulness of LLMs. It is composed of 817 questions across 38 categories, including health, law, finance, and politics. These questions are crafted in such a way that they can elicit false answers based on common misconceptions or false beliefs that some humans might also give. TruthfulQA incorporates two distinct tasks, namely, generation and multiple-choice. Both tasks utilize the same sets of questions and reference answers, thereby ensuring consistency in evaluation. Following Zou et al. (2023), we assess models on the multiple-choice task.

BBQ Parrish et al. (2022) is a benchmark for evaluating the degree of social biases present in LLMs, specifically about question-answering tasks. It assesses biases towards protected groups across nine social dimensions that are particularly relevant in US English-speaking contexts. This benchmark includes a variety of question sets, including ambiguous contexts where the answer is not clear, and disambiguated ones where a correct response can be determined with great certainty. Each example within the dataset comprises clusters of four multiple-choice questions, encompassing both negative and non-negative variants, and is presented with or without a disambiguating context. Negative questions aim to test stereotypes that reflect societal prejudices, while non-negative questions complement this by assessing whether model responses show a bias towards particular labels.

C Prompts

The prompts employed in our evaluation experiments across various benchmarks are illustrated in Figures 9 to 14. Notably, for the GSM8K and TruthfulQA benchmarks, the questions are used directly as input for the LLMs. Furthermore, for the FollowBench benchmark, we utilized the official implementation, resulting in prompts that are consistent with those described in Jiang et al. (2023).

D Detailed Experimental Results

The performance of the Qwen-Chat series models, along with their quantized counterparts, is depicted in the following figures: CNN/DailyMail test sets (See et al., 2017) (Figure 15), C-EVAL benchmark (Huang et al., 2023) (Figure 16a), Chinese to English translation on the FLORES-200 benchmark (Costa-jussà et al., 2022) (Figure 16b), and perplexity on the C4 (Raffel et al., 2020) and PTB (Marcus et al., 1994) datasets (Figure 17). Detailed experimental results for all evaluated benchmarks, as well as data on memory consumption and decoding speed during inference, are provided in Table 6 through 15. In these tables, the best results achieved by the quantized models are highlighted in bold, while underlined results indicate that the performance of the quantized model surpasses that of the BFloat16 baseline.

MMLU

The following is a multiple-choice question. Please choose the most suitable one among A, B, C and D as the answer to this question.

{question}
A. {choice_A}
B. {choice_B}
C. {choice_C}
D. {choice_D}

Figure 9: Prompt used for MMLU (Hendrycks et al., 2021) benchmark.

C-EVAL

{question}
A. {choice_A}
B. {choice_B}
C. {choice_C}
D. {choice_D}

Figure 10: Prompt used for C-EVAL (Huang et al., 2023) benchmark.

FLORES-200

Please translate the following {source_lang} text into {target_lang}.
{source_lang} text: {text}

Figure 11: Prompt used for FLORES-200 (Costa-jussà et al., 2022) benchmark.

XSum & CNN/DailyMail

Please summarize the following document.
{document}

Figure 12: Prompt used for XSum (Narayan et al., 2018) and CNN/DailyMail (See et al., 2017) benchmarks.

SNLI

{premise}

Question: Does this imply that "{hypothesis}"? Yes, no, or maybe?

Figure 13: Prompt used for SNLI (Bowman et al., 2015) benchmark.

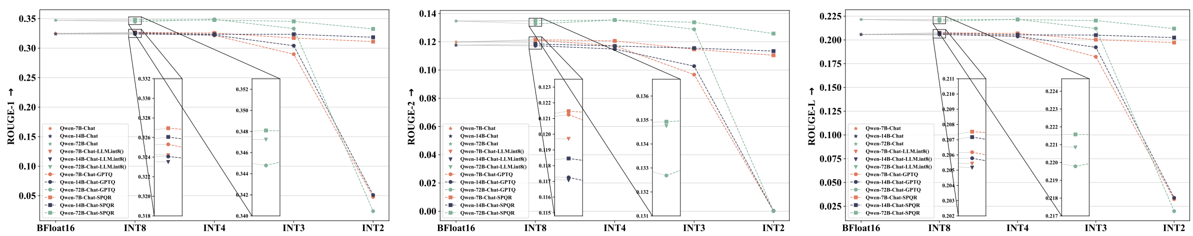
BBQ

Please answer the question provided below by considering the given context.

Context: {context}

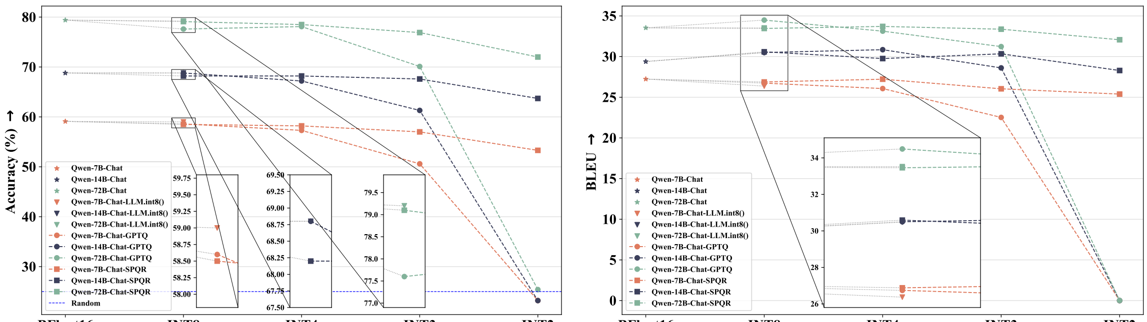
Question: {question}

Figure 14: Prompt used for BBQ (Parrish et al., 2022) benchmark.



(a) ROUGE-1 (b) ROUGE-2 (c) ROUGE-L

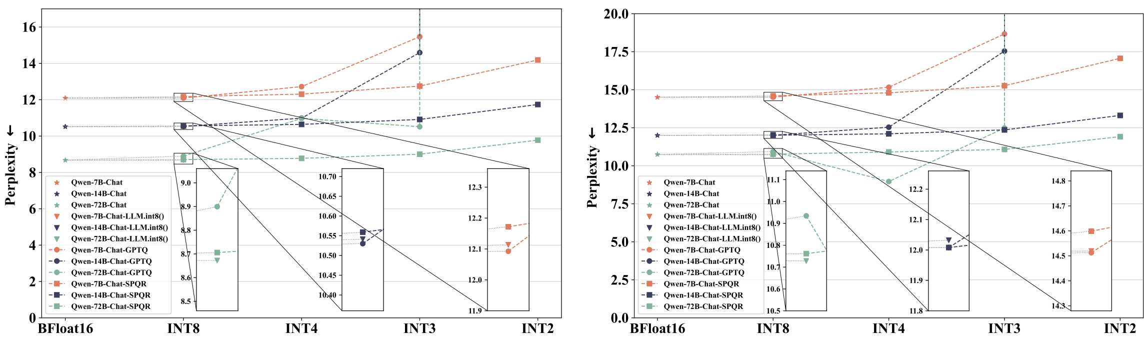
Figure 15: ROUGE-1 (a), ROUGE-2 (b), and ROUGE-L (c) scores for the Qwen-Chat series models and their quantized counterparts on the test sets of CNN/DailyMail (See et al., 2017).



(a) C-EVAL

(b) Zh → En

Figure 16: Performance of the Qwen-Chat series models and their quantized counterparts on the C-EVAL (Huang et al., 2023) benchmark (a) and Chinese-to-English (Zh → En) translation task of the FLORES-200 (Costa-jussà et al., 2022) (b) benchmark.



(a) C4

(b) PTB

Figure 17: Perplexity of Qwen-Chat Series models and their quantized counterparts on the C4 (Raffel et al., 2020) and PTB datasets (Marcus et al., 1994).

Model	Datatype	Quantization Method	BLEU (En → Zh)	BLEU (Zh → En)
Qwen-7B-Chat	BFloat16	-	32.02	27.24
	INT-8	LLM.int8()	31.56	26.37*
		GPTQ	<u>**32.30**</u>	26.74
		SpQR	<u>32.13</u>	**26.88**
	INT-4	GPTQ	30.79*	26.07*
		SpQR	**31.96**	**27.22**
	INT-3	GPTQ	25.91*	22.52*
		SpQR	**30.73***	**26.04***
	INT-2	GPTQ	0.01*	0.00*
		SpQR	**29.04***	**25.39***
Qwen-14B-Chat	BFloat16	-	32.87	29.39
	INT-8	LLM.int8()	32.78	<u>30.48</u>
		GPTQ	33.18	<u>30.50</u>
		SpQR	<u>**33.35**</u>	<u>**30.59**</u>
	INT-4	GPTQ	31.95*	<u>**30.85**</u>*
		SpQR	<u>**33.18**</u>	<u>29.76</u>
	INT-3	GPTQ	29.23*	28.61
		SpQR	**31.84***	<u>**30.34**</u>
	INT-2	GPTQ	0.01*	0.00*
		SpQR	**30.10***	**28.29***
Qwen-72B-Chat	BFloat16	-	36.08	33.54
	INT-8	LLM.int8()	35.26*	33.52
		GPTQ	35.15*	<u>**34.49**</u>*
		SpQR	**35.87**	33.46
	INT-4	GPTQ	34.93*	33.13
		SpQR	**35.73**	<u>**33.71**</u>
	INT-3	GPTQ	30.73*	31.23*
		SpQR	**34.79***	<u>**33.37**</u>*
	INT-2	GPTQ	0.01*	0.01*
		SpQR	**33.41***	**32.07***

Table 6: BLEU scores of Qwen-Chat series models and their quantized counterparts for English-to-Chinese and Chinese-to-English translation tasks of the FLORES-200 benchmark (Costa-jussà et al., 2022). The highest BLEU scores obtained by the quantized models are highlighted in bold. Underlined results denote instances where the quantized model outperforms the BFloat16 baseline. **Statistically significant differences between quantized LLMs and their non-quantized equivalents are indicated by** $^\star p < 0.05$**.**

Model	Datatype	Quantization Method	Perplexity (WikiText2)	Perplexity (C4)	Perplexity (PTB)
Qwen-7B-Chat	BFloat16	-	8.67	12.09	14.51
	INT-8	LLM.int8()	**8.68**	12.11	14.52
		GPTQ	**8.68**	**12.09**	**14.51**
		SpQR	8.71	12.17	14.60
	INT-4	GPTQ	9.04	12.72	15.16
		SpQR	**8.82**	**12.31**	**14.80**
	INT-3	GPTQ	11.17	15.47	18.67
		SpQR	**9.17**	**12.75**	**15.27**
	INT-2	GPTQ	123030.34	41936.28	88223.58
		SpQR	**10.05**	**14.19**	**17.07**
Qwen-14B-Chat	BFloat16	-	6.99	10.52	12.00
	INT-8	LLM.int8()	7.00	10.54	12.03
		GPTQ	7.00	**10.53**	**12.01**
		SpQR	**6.99**	10.56	**12.01**
	INT-4	GPTQ	7.35	10.99	12.54
		SpQR	**7.07**	**10.64**	**12.10**
	INT-3	GPTQ	9.68	14.59	17.54
		SpQR	**7.31**	**10.92**	**12.36**
	INT-2	GPTQ	200643.66	153141.75	224832.00
		SpQR	**7.94**	**11.74**	**13.31**
Qwen-72B-Chat	BFloat16	-	6.15	8.68	10.75
	INT-8	LLM.int8()	<u>**6.14**</u>	<u>**8.67**</u>	<u>**10.73**</u>
		GPTQ	6.28	8.90	10.93
		SpQR	6.16	8.71	10.76
	INT-4	GPTQ	6.37	8.97	10.97
		SpQR	**6.23**	**8.77**	**10.91**
	INT-3	GPTQ	7.58	10.51	12.48
		SpQR	**6.43**	**9.00**	**11.08**
	INT-2	GPTQ	52688.75	38123.23	55330.14
		SpQR	**7.01**	**9.78**	**11.92**

Table 7: The perplexity of the Qwen-Chat series models and their quantized counterparts on WikiText2 (Merity et al., 2017), C4 (Raffel et al., 2020), and PTB (Marcus et al., 1994). The best results achieved by the quantized models are highlighted in bold, while underlined results indicate that the performance of the quantized model surpasses that of the BFloat16 baseline.

Model	Datatype	Quantization Method	Memory	Speed
Qwen-7B-Chat	BFloat16	-	15.14	37.67
	INT-8	LLM.int8	**9.23**	7.19
		GPTQ	10.91	13.57
		SpQR	15.60	**37.65**
	INT-4	GPTQ	**7.83**	37.43
		SpQR	15.60	**37.73**
	INT-3	GPTQ	**7.12**	8.21
		SpQR	15.61	**37.73**
	INT-2	GPTQ	**6.26**	19.36
		SpQR	15.66	**37.51**
Qwen-14B-Chat	BFloat16	-	27.60	25.15
	INT-8	LLM.int8	**15.91**	5.85
		GPTQ	17.92	14.37
		SpQR	27.95	**25.42**
	INT-4	GPTQ	**12.03**	24.38
		SpQR	27.95	**24.62**
	INT-3	GPTQ	**10.77**	4.71
		SpQR	27.97	**25.20**
	INT-2	GPTQ	**8.99**	18.26
		SpQR	28.04	**24.82**
Qwen-72B-Chat	BFloat16	-	138.44	8.97
	INT-8	LLM.int8	**74.96**	3.07
		GPTQ	77.85	1.43
		SpQR	143.20	**6.57**
	INT-4	GPTQ	**44.11**	**14.88**
		SpQR	143.21	6.56
	INT-3	GPTQ	**35.93**	0.84
		SpQR	143.37	**6.57**
	INT-2	GPTQ	**27.74**	2.23
		SpQR	144.59	**6.56**

Table 8: Memory consumption (in GB) and decoding speed (tokens generated per second) of Qwen-Chat series models and their quantized counterparts during inference. The best results achieved by the quantized models are highlighted in bold.

Model	Datatype	Quantization Method	Accuracy (STEM)	Accuracy (Humanities)	Accuracy (Other)	Accuracy (Social Science)	Accuracy (Average)
Qwen-7B-Chat	BFloat16	-	50.84	48.44	63.57	64.28	55.80
	INT-8	LLM.int8	49.79	47.91	63.18	63.86	55.21
		GPTQ	50.49	**48.03**	**63.60**	**64.02**	**55.53**
		SpQR	**50.94**	47.97	62.99	63.96	55.46
	INT-4	GPTQ	49.44	47.29	61.47	62.63	54.27*
		SpQR	**51.00**	**48.20**	**62.86**	**63.57**	**55.44**
	INT-3	GPTQ	46.46	44.55	57.39	60.51	51.32*
		SpQR	**48.68**	**47.25**	**62.70**	**63.54**	**54.56***
	INT-2	GPTQ	22.80	24.51	25.43	21.81	23.74*
		SpQR	**46.75**	**45.55**	**60.35**	**61.07**	**52.49***
Qwen-14B-Chat	BFloat16	-	61.62	56.43	71.36	73.32	64.60
	INT-8	LLM.int8	**61.50**	**56.20**	**71.03**	**73.68**	**64.50**
		GPTQ	61.02	56.13	70.87	73.58	64.31
		SpQR	60.99	55.94	70.68	73.38	64.16
	INT-4	GPTQ	60.32	55.71	69.13	72.64	63.42*
		SpQR	**60.70**	**56.43**	**70.23**	**72.99**	**64.07**
	INT-3	GPTQ	56.42	52.14	65.24	68.96	59.69*
		SpQR	**59.78**	**55.81**	**68.88**	**72.86**	**63.33***
	INT-2	GPTQ	22.42	24.87	24.04	23.95	23.94*
		SpQR	**57.41**	**54.43**	**67.98**	**69.81**	**61.47***
Qwen-72B-Chat	BFloat16	-	70.25	68.50	79.98	81.74	74.33
	INT-8	LLM.int8	**70.06**	68.42	79.47	81.64	74.13
		GPTQ	69.46	67.38	79.40	81.51	73.60
		SpQR	69.77	**68.52**	**80.14**	**81.67**	**74.26**
	INT-4	GPTQ	70.19	67.91	79.11	81.22	73.81
		SpQR	**70.98**	**68.18**	**79.95**	**81.80**	**74.40**
	INT-3	GPTQ	65.68	63.72	76.21	76.44	69.71*
		SpQR	**68.41**	**68.31**	**79.47**	**81.90**	**73.78**
	INT-2	GPTQ	23.25	27.01	24.65	23.24	24.82*
		SpQR	**66.10**	**65.06**	**76.25**	**79.56**	**70.94***

Table 9: Accuracy of Qwen-Chat series models and their quantized counterparts across four broad disciplines (STEM, Humanities, Social Science, and Other) on the MMLU benchmark (Hendrycks et al., 2021), including overall average accuracy. The best results achieved by the quantized models are highlighted in bold, while underlined results indicate that the performance of the quantized model surpasses that of the BFloat16 baseline. **Statistically significant differences in the Accuracy (Average) column between quantized LLMs and their non-quantized equivalents are indicated $^\star p < 0.05$.**

Model	Datatype	Quantization Method	Accuracy (STEM)	Accuracy (Social Science)	Accuracy (Humanities)	Accuracy (Other)	Accuracy (Hard)	Accuracy (Average)
Qwen-7B-Chat	BFloat16	-	54.4	71.9	63	52.3	40.4	59.1
	INT-8	LLM.int8	**54.2**	71.7	62.2	**53**	**41.2**	**59**
		GPTQ	53.8	**71.8**	61.8	52.2	39.6	58.6
		SpQR	53.4	71.3	**62.6**	52.3	40.2	58.5
	INT-4	GPTQ	52	70.6	**61.4**	50.7	**39.2**	57.3
		SpQR	**53.3**	**71.5**	61.3	**51.9**	38.4	**58.2**
	INT-3	GPTQ	44.8	64.1	54.1	45.4	32.8	50.6
		SpQR	**51.8**	**70.2**	**60.7**	**50.5**	**38.4**	**57**
	INT-2	GPTQ	22.9	22.7	23.6	24.1	21.1	23.2
		SpQR	**47.4**	**66.8**	**56.9**	**48**	**33.1**	**53.3**
Qwen-14B-Chat	BFloat16	-	64.4	80.7	71.2	63.5	52.7	68.8
	INT-8	LLM.int8	64.4	80.9	71	**63.4**	52.2	**68.8**
		GPTQ	**64.6**	**81**	**71.2**	63.1	**52.4**	**68.8**
		SpQR	63.7	80.6	70.9	62.5	52.1	68.2
	INT-4	GPTQ	62.6	80	69.3	61.8	50.4	67.2
		SpQR	**64.4**	**80.2**	**70.4**	**62.1**	**51.9**	**68.2**
	INT-3	GPTQ	56.1	74.2	64	56.3	43.9	61.3
		SpQR	**63.2**	**79.7**	**69.7**	**62.3**	**50.4**	**67.6**
	INT-2	GPTQ	22.9	23.9	22.6	23.5	21.8	23.2
		SpQR	**58.6**	**77.6**	**67.3**	**56.9**	**46.5**	**63.7**
Qwen-72B-Chat	BFloat16	-	74.4	89.5	80.7	78.2	61.6	79.4
	INT-8	LLM.int8	73.9	**89.8**	80.5	**77.8**	60.6	**79.2**
		GPTQ	71.6	89.2	80	75.5	57.7	77.6
		SpQR	**74**	89.3	**80.7**	77.4	**61.6**	79.1
	INT-4	GPTQ	**72.5**	89.1	80	76.6	58.9	78.1
		SpQR	**72.5**	**89.6**	**80.3**	**77.7**	**59**	**78.5**
	INT-3	GPTQ	65.3	82.5	72	65.5	51.7	70.1
		SpQR	**72**	**87.1**	**79.2**	**74.6**	**57.9**	**76.9**
	INT-2	GPTQ	25.3	25.7	25.8	25.2	25.2	25.4
		SpQR	**66.7**	**84.5**	**74.4**	67.9	**52.7**	**72**

Table 10: Accuracy of Qwen-Chat series models and their quantized counterparts Across four broad disciplines (STEM, Social Sciences, Humanities, and Other) on the C-EVAL Benchmark (Huang et al., 2023), including the C-EVAL Hard Subset and the average accuracy across all disciplines. The best results achieved by the quantized models are highlighted in bold, while underlined results indicate that the performance of the quantized model surpasses that of the BFloat16 baseline.

Model	Datatype	Quantization Method	ROUGE-1	ROUGE-2	ROUGE-L
Qwen-7B-Chat	BFloat16	-	0.19	0.05	0.13
	INT-8	LLM.int8	0.19	0.05	0.13
		GPTQ	**0.19**	**0.05**	**0.13**
		SpQR	0.19	0.05*	0.13
	INT-4	GPTQ	**0.19***	**0.05***	**0.13**
		SpQR	0.19*	0.05*	0.13*
	INT-3	GPTQ	0.17*	0.04*	0.12*
		SpQR	**0.19**	**0.05**	**0.13***
	INT-2	GPTQ	0.04*	0.00*	0.03*
		SpQR	**0.17***	**0.04***	**0.12***
Qwen-14B-Chat	BFloat16	-	0.18	0.05	0.13
	INT-8	LLM.int8	**0.18**	0.05	0.13
		GPTQ	0.18	**0.05**	0.13
		SpQR	**0.18**	0.05	**0.13**
	INT-4	GPTQ	**0.18***	**0.05***	**0.13***
		SpQR	0.18*	0.05*	0.13*
	INT-3	GPTQ	0.17*	0.04*	0.12*
		SpQR	**0.18***	**0.05***	**0.13***
	INT-2	GPTQ	0.04*	0.00*	0.03*
		SpQR	**0.18***	**0.05***	**0.12***
Qwen-72B-Chat	BFloat16	-	0.25	0.09	0.18
	INT-8	LLM.int8	0.25	**0.10**	0.19
		GPTQ	**0.26**	0.09	**0.19***
		SpQR	0.25	0.09	0.19
	INT-4	GPTQ	0.25*	0.09*	0.18*
		SpQR	**0.25***	**0.09**	**0.18***
	INT-3	GPTQ	0.20*	0.06*	0.14*
		SpQR	**0.24***	**0.08***	**0.17***
	INT-2	GPTQ	0.02*	0.00*	0.01*
		SpQR	**0.22***	**0.07***	**0.16***

Table 11: ROUGE-1, ROUGE-2, and ROUGE-L scores for the Qwen-Chat series models and their quantized counterparts on the test sets of XSum (Narayan et al., 2018). The best results achieved by the quantized models are highlighted in bold, while underlined results indicate that the performance of the quantized model surpasses that of the BFloat16 baseline. **Statistically significant differences between quantized LLMs and their non-quantized equivalents are indicated by** $^\star p < 0.05$.

Model	Datatype	Quantization Method	ROUGE-1	ROUGE-2	ROUGE-L
Qwen-7B-Chat	BFloat16	-	0.33	0.12	0.21
	INT-8	LLM.int8	0.32	0.12	0.21
		GPTQ	0.33	0.12*	0.21
		SpQR	**0.33***	**0.12***	**0.21***
	INT-4	GPTQ	0.32*	0.12*	0.20*
		SpQR	**0.33**	**0.12**	**0.21**
	INT-3	GPTQ	0.29*	0.10*	0.18*
		SpQR	**0.32***	**0.11***	**0.20***
	INT-2	GPTQ	0.05*	0.00*	0.03*
		SpQR	**0.31***	**0.11***	**0.20***
Qwen-14B-Chat	BFloat16	-	0.32	0.12	0.21
	INT-8	LLM.int8	0.32	0.12	0.21
		GPTQ	0.32	0.12	0.21
		SpQR	**0.33***	**0.12**	**0.21***
	INT-4	GPTQ	0.32*	0.11*	0.20*
		SpQR	**0.32**	**0.12**	**0.21**
	INT-3	GPTQ	0.30*	0.10*	0.19*
		SpQR	**0.32**	**0.12***	**0.21**
	INT-2	GPTQ	0.05*	0.00*	0.03*
		SpQR	**0.32***	**0.11***	**0.20***
Qwen-72B-Chat	BFloat16	-	0.35	0.13	0.22
	INT-8	LLM.int8	0.35	0.13	0.22
		GPTQ	0.34*	0.13*	0.22*
		SpQR	**0.35**	**0.13**	**0.22**
	INT-4	GPTQ	**0.35**	**0.14**	**0.22**
		SpQR	0.35	0.14	0.22
	INT-3	GPTQ	0.33*	0.13*	0.21*
		SpQR	**0.35***	**0.13**	**0.22**
	INT-2	GPTQ	0.02*	0.00*	0.02*
		SpQR	**0.33***	**0.13***	**0.21***

Table 12: ROUGE-1, ROUGE-2, and ROUGE-L scores of the Qwen-Chat series models and their quantized counterparts on the test sets of CNN/DailyMail (See et al., 2017). The best results achieved by the quantized models are highlighted in bold, while underlined results indicate that the performance of the quantized model surpasses that of the BFloat16 baseline. **Statistically significant differences between quantized LLMs and their non-quantized equivalents are indicated by** $^\star p < 0.05$.

Model	Datatype	Quantization Method	Accuracy (GSM8K)	Accuracy (SNLI)	MC1 Accuracy (TruthfulQA)
Qwen-7B-Chat	BFloat16	-	0.51	0.82	0.38
	INT-8	LLM.int8	0.52	**0.81***	0.37
		GPTQ	**0.54**	**0.81**	**0.38**
		SpQR	0.51	0.80*	**0.38**
	INT-4	GPTQ	0.47	**0.80***	0.36
		SpQR	**0.52**	0.80*	**0.37**
	INT-3	GPTQ	0.39*	**0.80***	0.36
		SpQR	**0.49**	0.80*	**0.37**
	INT-2	GPTQ	0.04*	0.02*	0.30*
		SpQR	**0.44***	**0.80***	**0.36**
Qwen-14B-Chat	BFloat16	-	0.62	0.80	0.39
	INT-8	LLM.int8	0.60	0.80	**0.40**
		GPTQ	0.61	0.80	0.39
		SpQR	**0.62**	**0.81**	0.39
	INT-4	GPTQ	0.60	0.79*	0.38
		SpQR	**0.61**	**0.81**	**0.39**
	INT-3	GPTQ	0.51*	**0.81***	**0.39**
		SpQR	**0.59**	0.79	0.38
	INT-2	GPTQ	0.06*	0.03*	0.29
		SpQR	**0.56***	**0.78***	**0.39**
Qwen-72B-Chat	BFloat16	-	0.78	0.84	0.43
	INT-8	LLM.int8	**0.79**	0.83	0.43
		GPTQ	0.77	**0.85***	0.42
		SpQR	**0.79**	0.84	**0.43**
	INT-4	GPTQ	**0.79**	0.83	0.43
		SpQR	0.78	**0.84**	**0.44**
	INT-3	GPTQ	0.71*	0.82*	0.41
		SpQR	**0.76**	**0.84**	**0.42**
	INT-2	GPTQ	0.13*	0.01*	0.28*
		SpQR	**0.72***	**0.79***	**0.41**

Table 13: Performance of Qwen-Chat series models and their quantized counterparts on the TruthfulQA benchmark (Lin et al., 2022), as well as the test sets of GSM8K (Cobbe et al., 2021) and SNLI (Bowman et al., 2015). The best results achieved by the quantized models are highlighted in bold, while underlined results indicate that the performance of the quantized model surpasses that of the BFloat16 baseline. **Statistically significant differences between quantized LLMs and their non-quantized equivalents are indicated by** $^{\star}p < 0.05$**.**

Models	Datatypes	Quantization Methods	Hard Satisfaction Rate	Soft Satisfaction Rate	Consistent Satisfaction Levels
Qwen-7B-Chat	BFloat16	-	0.40	0.52	1.57
	INT-8	LLM.int8	0.41	0.52	**1.62**
		GPTQ	**0.41**	**0.53**	1.52
		SpQR	0.40	0.52	**1.62**
	INT-4	GPTQ	0.40	**0.52**	1.38
		SpQR	**0.40**	0.52	**1.48**
	INT-3	GPTQ	0.36	0.48	1.27
		SpQR	**0.36**	**0.49**	**1.30**
	INT-2	GPTQ	0.00	0.01	0.00
		SpQR	**0.38**	**0.50**	**1.55**
Qwen-14B-Chat	BFloat16	-	0.47	0.57	1.73
	INT-8	LLM.int8	0.48	0.58	1.75
		GPTQ	**0.49**	**0.59**	**1.90**
		SpQR	0.48	0.58	1.87
	INT-4	GPTQ	**0.49**	**0.57**	**1.90**
		SpQR	0.47	0.56	1.78
	INT-3	GPTQ	0.44	0.53	1.62
		SpQR	**0.48**	**0.58**	**1.85**
	INT-2	GPTQ	0.01	0.01	0.02
		SpQR	**0.48**	**0.57**	**1.82**
Qwen-72B-Chat	BFloat16	-	0.53	0.62	2.15
	INT-8	LLM.int8	**0.56**	**0.64**	2.28
		GPTQ	0.55	0.63	2.08
		SpQR	0.54	0.63	**2.33**
	INT-4	GPTQ	0.53	0.61	**2.13**
		SpQR	**0.54**	**0.62**	2.10
	INT-3	GPTQ	0.53	0.61	2.02
		SpQR	**0.53**	**0.61**	**2.22**
	INT-2	GPTQ	0.00	0.00	0.00
		SpQR	**0.53**	**0.62**	**2.20**

Table 14: Average hard satisfaction rates, soft satisfaction rates, and consistent satisfaction levels across five difficulty levels for the Qwen-Chat series models and their quantized counterparts on the FollowBench (Jiang et al., 2023) benchmark. The best results achieved by the quantized models are highlighted in bold, while underlined results indicate that the performance of the quantized model surpasses that of the BFloat16 baseline.

Model	Datatype	Quantization Method	Bias Score (Ambiguous Context)	Bias Score (Disambiguated Context)
Qwen-7B-Chat	BFloat16	-	6.20	3.87
	INT-8	LLM.int8	6.41	<u>**3.49**</u>
		GPTQ	<u>**5.98**</u>	<u>3.76</u>
		SpQR	6.36	3.95
	INT-4	GPTQ	<u>5.49</u>	**3.69**
		SpQR	6.34	<u>3.77</u>
	INT-3	GPTQ	**4.21**	4.90
		SpQR	6.31	**4.08**
	INT-2	GPTQ	<u>**-0.54**</u>	<u>**-0.97**</u>
		SpQR	<u>4.13</u>	5.76
Qwen-14B-Chat	BFloat16	-	8.35	3.69
	INT-8	LLM.int8	<u>**7.92**</u>	3.89
		GPTQ	<u>8.22</u>	3.70
		SpQR	8.60	<u>**3.65**</u>
	INT-4	GPTQ	<u>**7.82**</u>	4.11
		SpQR	<u>7.96</u>	**3.86**
	INT-3	GPTQ	8.41	3.88
		SpQR	<u>**8.03**</u>	<u>**3.11**</u>
	INT-2	GPTQ	<u>**-0.17**</u>	<u>**-0.81**</u>
		SpQR	<u>8.08</u>	5.33
Qwen-72B-Chat	BFloat16	-	9.07	1.57
	INT-8	LLM.int8	<u>**8.81**</u>	1.68
		GPTQ	<u>8.95</u>	<u>**1.31**</u>
		SpQR	9.07	<u>1.51</u>
	INT-4	GPTQ	10.11	1.76
		SpQR	<u>**8.73**</u>	<u>**1.52**</u>
	INT-3	GPTQ	<u>8.77</u>	3.11
		SpQR	<u>**8.24**</u>	**1.97**
	INT-2	GPTQ	<u>**-0.36**</u>	<u>**0.89**</u>
		SpQR	<u>7.67</u>	1.69

Table 15: Bias scores of the Qwen-Chat series models and their quantized counterparts in ambiguous and disambiguated contexts on the BBQ benchmark (Parrish et al., 2022). The best results achieved by the quantized models are highlighted in bold, while underlined results indicate that the performance of the quantized model surpasses that of the BFloat16 baseline.