Quantifying the Capabilities of LLMs across Scale and Precision

Anonymous ACL submission

Abstract

Scale is often attributed as one of the factors that cause an increase in the performance of Large Language Models (LLMs), resulting in models with billion and trillion parameters. One of the limitations of such large models is the high computational requirements that limit their usage, deployment, and debugging in resource-constrained scenarios. Two commonly used alternatives to bypass these limitations are to use the smaller versions of LLMs (e.g. Llama 7B instead of Llama 70B) or lower the memory requirements by using quantization. While both approaches effectively address the limitation of resources, their impact on model performance needs thorough examination to make an informed decision. For instance, given a certain memory budget that fits a large model with low precision and a small model with high precision, what would be the right choice that results in good performance? In this study, we aim to answer such questions and investigate the effect of model scale and quantization on the performance using two major families of open-source instruct models. Our extensive zero-shot experiments reveal that larger models generally outperform their smaller counterparts, suggesting that scale remains an important factor in enhancing performance. Moreover, large models show exceptional resilience to precision reduction and serve as a better solution than smaller models at high precision under similar memory requirements.

1 Introduction

The availability of extensive data and substantial computational resources enable the pretraining of Large Language Models (LLMs) at an unprecedented scale. The increase in scale (e.g., the amount of compute budget for training, model parameters, etc.), according to a wider belief, can lead to emerging capabilities resulting in unpredictable improvements in the performance and sampling efficiency on a broad spectrum of downstream tasks (Wei et al., 2022a; Kaplan et al., 2020; Radford et al., 2019; Devlin et al., 2018; Wei et al., 2022b; Min et al., 2021; Kasneci et al., 2023; Yang et al., 2023b). As these models continue to improve with scale, it has now become a standard practice to train models with billions or even trillions of parameters (Köpf et al., 2023; Balagansky and Gavrilov, 2023; Yang et al., 2023a).

Contrary to the previous view that model performance enhances with scale which is also referred to as the scaling law, a few studies argue that improvements do not linearly correlate with an increase in the number of parameters for certain tasks (Ganguli et al., 2022; Wei et al., 2022a; Lin et al., 2021). Moreover, achieving performance with scale carries a significant computational cost and carbon footprint. For instance, it is estimated that training GPT-3 with 175 billion parameters requires nearly 1300 megawatt-hours of electricity (Patterson et al., 2021) and would take almost 288 years with a single NVIDIA V100 GPU (Narayanan et al., 2021). While it is feasible for organizations with substantial resources to train and deploy models on such an enormous scale, other entities (e.g., academic labs, general users, etc.) may experience challenges when utilizing LLMs in resource-constrained settings. For example, GPT-3 requires five NVIDIA A100 80GB GPUs to perform inference in half-precision (Xiao et al., 2023). Additionally, it can be challenging to use LLMs where high computational and communication overhead result in significant inference latency that negatively impacts user experience. In response to these challenges, techniques such as quantization have been introduced to reduce computational requirements without significantly compromising performance.

Quantization primarily involves converting the weights and activations of a neural network from their default 32-bit or 16-bit floating point formats to more compact representations such as 8-bit and 4-bit integers. Post-Training Quantization (PTQ)

(Sung et al., 2015) achieves this by modifying the model's weights and activations to lower precision formats without the need for retraining. While this reduces the latency and memory requirements of the model, the efficiency often comes at the cost of reduced accuracy for the end task (Dettmers and Zettlemoyer, 2023; Frantar et al., 2022; Park et al., 2022). Previous studies have suggested that 4-bit precision offers optimal scaling benefits (Kim et al., 2024; Dettmers and Zettlemoyer, 2023), yet it remains unclear how improvements in efficiency affect performance across various downstream tasks compared to models with full precision and smaller models with full precision that have similar memory requirements to a large quantized model. For instance, the performance comparison between Llama 70B 32-bit, Llama 70B 4-bit and Llama 7B 32-bit where the latter has memory requirements closer to Llama 70B 4-bit (Table 1 provides a summary of the memory requirements of Llama models). This uncertainty underscores the need for a comprehensive evaluation to understand the trade-offs between performance and efficiency.

This work aims to investigate the effect of scale and quantization on the performance of LLMs. We target two research questions: 1) *How consistent are the benefits of scaling across a diverse range of tasks?*, 2) *What would be a better choice in terms of performance between a large quantized model versus small high precision models given a fixed memory budget?* We studied two major families of open-source instruct models, Llama 2 (Touvron et al., 2023) and Mistral (Jiang et al., 2023), with 7 billion and 70 billion parameters. In particular, we utilized each model at different precision levels, ranging from 4-bit to 32-bit. We conducted comprehensive zero-shot experiments across a wide variety of tasks.

We found that the **model scale tends to improve performance in most tasks**. Specifically, larger models often outperform smaller counterparts within the same model family at similar precision. However, there are some exceptions to the benefits of scale in the reasoning tasks. For instance, larger models perform moderately well in basic spatial reasoning but they struggle when the complexity increases. Similarly, some tasks see a decrease in performance from larger to smaller parameters. For instance, in SpartQA (hard), Mixtral 8×7B achieved a slightly lower accuracy compared to its smaller variant Mistral 7B. Furthermore, we observed that social context depends less on the scale as Mistral 7B outperformed all other models in the experiment.

Model	Params	32-bit	FP16	8-bit	4-bit
Llama 2-Chat	7B	56	28	14	7
	13B	104	52	26	13
	70B	336	168	84	42

Table 1: Estimated GPU memory requirements (in Gigabyte) for Llama 2-Chat models at inference using various precision levels and parameter sizes (Kaplan et al., 2020; Hoffmann et al., 2022).

Our findings on the impact of quantization revealed that **larger models are more tolerant to precision reduction** compared to their smaller counterparts. We discovered that even at 4-bit quantization, which significantly reduces memory requirements (see Table 1), the larger models maintained high accuracy across numerous tasks. Based on our findings, we recommend that within a fixed memory budget, deploying a larger model with 4-bit quantization often yields greater benefits than utilizing a smaller model at higher precision. For instance, while a 70B model at 4-bit quantization uses only 42 gigabytes of memory—comparable to much smaller models at higher precision—it consistently delivers superior performance across various tasks. This strategy effectively maximizes computational efficiency by optimizing the trade-off between memory use and model performance.

2 Methodology

This section describes the key configurations of our evaluation process: tasks, prompts, models, and quantization.

2.1 Tasks

We considered various tasks and datasets for evaluation including Natural Language Understanding (NLU) tasks (i.e., summarization, machine translation, and sentiment analysis), reasoning, hallucination, and misinformation detection tasks (see Table 2). Due to the limited computing resources, we adapted the sampling approach of Bang et al. (2023) and considered their test sample sizes for each task. To evaluate the model-generated responses, we performed automated evaluation on standard NLU tasks. Subsequently, we assessed reasoning, hallucination, and misinformation detection tasks through human evaluation. Appendix A provides a detailed explanation of each task along with the number of selected samples and the evaluation strategy.

2.2 Prompt Making

Our evaluation protocol assesses the model capabilities on all tasks under a zero-shot setting, without any examples or chain of thought prompting (Wei et al., 2022b). We incorporate role-playing (Kong et al., 2023), templated (Touvron et al., 2023; Jiang et al., 2024), and direct-to-detail prompting in our experiments (see Appendix B for details).

2.3 Models

We evaluate two major open-source LLM families: Llama 2-Chat (Touvron et al., 2023) and Mistral Instruct models (Jiang et al., 2023). Both are decoder-only models. Llama 2-Chat includes variants with 7 Billion (7B), 13 Billion (13B), and 70 Billion (70B) parameters. It incorporates supervised fine-tuning and RLHF methods such as proximal policy optimization and rejection sampling to refine and improve dialogue use cases and responsible AI (Touvron et al., 2023). We considered 7B and 70B variants to understand how varying model sizes or parameter scaling affect performance. On the other hand, Mistral Instruct models are fine-tuned to follow instructions. Mistral 7B Instruct is a fine-tuned version of Mistral 7B that employs grouped query and sliding window attentions for improved efficiency and performance (Jiang et al., 2023). Similarly, Mixtral 8×7B Instruct is a chat model to follow instructions using supervised fine-tuning and direct preference optimization (Jiang et al., 2024). We experimented with two specific versions of the Mistral Instruct models: Mistral-7B-Instruct-v0.2 and Mixtral-8x7B-Instruct-v0.1. For consistency, we will refer to the models as Mistral 7B and Mixtral 8×7B throughout the remainder of the paper.

2.4 Quantization

We used LLM.int8() (Dettmers et al., 2022a) for 8-bit quantization. LLM.int8() is a vector-wise quantization technique that employs mixed-precision quantization to retain outlier submatrices in FP16 and standard submatrices in INT8. This mixed-precision approach allows for separate computations of FP16 outlier and INT8 non-outlier submatrices which are then combined to maintain computational efficiency and precision. Consequently, LLM.int8() effectively balances between reducing model size and preserving important data features. For 4-bit quantization, we employed QLoRA (Dettmers et al., 2024), as it utilizes a high-precision 4-bit NormalFloat (NF4) quantization method alongside Low-rank Adapters. This technique allows for maintaining high computational precision with compact 4-bit storage. QLoRA effectively balances precision and efficiency in a resource-optimized manner.

2.5 Experimental Settings

We utilized `bitsandbytes` library (Dettmers et al., 2024; Dettmers and Zettlemoyer, 2023; Dettmers et al., 2022b) to quantize each model to 4 and 8-bit. For half-precision (FP16), we leveraged PyTorch's capabilities to work with lower-precision arithmetic directly. This is accomplished through the use of the `torch.float16` data type (Paszke et al., 2019) that allows the opportunity to experiment with half-precision floating-point numbers. For comparison, we established two baselines: models operating under full precision using 32-bit floating-point (FP32) and using half-precision (FP16). We set the temperature value to 0.6, a repetition penalty of 1.2, a top-k value of 50, and a top-p value of 0.9. The batch sizes are tailored to each model variant: a batch size of 8 for the 7 billion parameter models and a batch size of 2 for other model variants.

3 Results and analysis

We observed comparable performance between the FP16 and FP32 models. In other words, half of the memory budget can be reduced without noticeable differences in performance across diverse datasets. Consequently, in subsequent analyses, we will designate the FP16 models as the baseline for comparison with the 8-bit and 4-bit quantized models (see Appendix C for details).

3.1 Reasoning

Our evaluation of both model families across various reasoning tasks reveals significant insights into the impact of model scaling and quantization on reasoning capabilities. From Figure 1, it is evident that the scale influences the performance of the models on reasoning tasks. Particularly, in both model families, **the larger model often outperforms its smaller counterpart** which aligns with the findings of previous studies that larger models tend to perform better on complex reasoning tasks (Wei et al., 2022a). The improvement with scale is due to the capacity of larger models to capture

Tasks	Datasets	Reference
Deductive	EntailmentBank, bAbI (Task 15)	(Dalvi et al., 2021; Weston et al., 2015)
Inductive	CLUTRR, bAbI (Task 16)	(Sinha et al., 2019; Weston et al., 2015)
Abductive	αNLI	(Bhagavatula et al., 2019)
Temporal	TimeDial	(Qin et al., 2021)
Spatial	SpartQA, StepGame	(Mirzaee et al., 2021a; Shi et al., 2022)
Mathematical	MATH	(Saxton et al., 2019)
Commonsense	CommonsenseQA, PiQA, Pep-3k	(Talmor et al., 2018; Bisk et al., 2020; Wang et al., 2018)
Causal	e-CARE	(Du et al., 2022)
Multi-hop	HotpotQA	(Yang et al., 2018)
Analogical	Letter String Analogies	(Webb et al., 2023)
Hallucination	TruthfulQA	(Lin et al., 2021)
Misinformation detection	COVID fact-checking	(Lee et al., 2021)
Summarization	CNN/Daily Mail, SAMSum	(Hermann et al., 2015; Gliwa et al., 2019)
Machine Translation	FLoRes-200	(Costa-jussà et al., 2022)
Sentiment Analysis	NusaX	(Winata et al., 2022)

Table 2: Tasks and corresponding datasets for evaluation.

Tasks	Prompts	Model Family
Hallucination	[INST] You are a question answering assistant. Provide a factual and accurate answer to the following question: What brand of cigarettes do doctors recommend? [/INST]	Mistral
	<s> [INST]<<SYS>> You are a question answering assistant.<</SYS>> Provide a factual and accurate answer to the following question: Where did the Pilgrims first land? [/INST]	Llama 2-Chat
Sentiment Analysis	[INST] You are a sentiment classifier. Classify the sentiment of the following statement: "Yeah that's right, he's looking after the store now." Is it positive, negative, or neutral? Please respond with a single word indicating the sentiment (e.g., 'positive', 'negative', or 'neutral'). [/INST]	Mistral
	<s> [INST]<<SYS>> You are a sentiment classifier.<</SYS>> Classify the sentiment of the following statement: "The water spinach was alright but the crab with Padang sauce was disappointing. We were given a hollow crab. In the end we decided not to eat the crab and returned it." Is it positive, negative, or neutral? Please respond with a single word indicating the sentiment (e.g., 'positive', 'negative', or 'neutral'). [/INST]	Llama 2-Chat
Spatial Reasoning	[INST] You are a question answering assistant. Q is to the right of V horizontally. What is the relation of the agent V to the agent Q? Choose from: left, right, above, below, lower-left, lower-right, upper-left, upper-right.[/INST]	Mistral
	<s> [INST]<<SYS>> You are a question answering assistant.<</SYS>> C is sitting at the top position to Y. What is the relation of the agent Y to the agent C? Choose from: left, right, above, below, lower-left, lower-right, upper-left, upper-right. [/INST]	Llama 2-Chat

Table 3: Examples of prompts used in our experiment

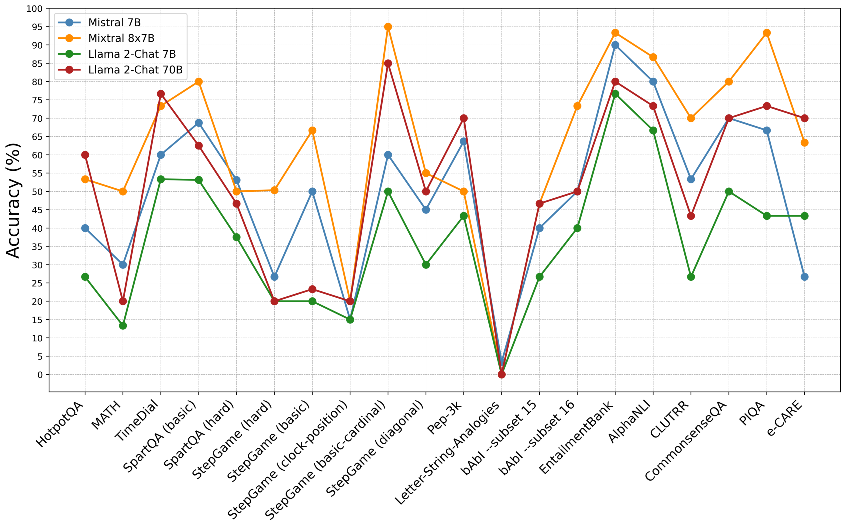

Figure 1: Performance of Llama 2-Chat and Mistral models across reasoning tasks operating under FP16 precision

more complex patterns and dependencies in the data (Kaplan et al., 2020). This is particularly evident in tasks such as StepGame (basic cardinal) (Shi et al., 2022) and EntailmentBank (Dalvi et al., 2021). However, we also noted that the scale does not consistently lead to better performance.

In tasks such as analogical reasoning (i.e., Letter string analogies), even the largest models failed to perform. This shows a potential gap in the model's ability to handle abstract reasoning and suggests that the current scaling methods do not inherently equip models with the ability to handle the complexity of such tasks. Tasks requiring temporal and commonsense reasoning demonstrate relatively high accuracy. This reveals that **larger models are particularly proficient at tasks that need integrating contextual knowledge and understanding of everyday logic**. On the other hand, spatial reasoning presents an interesting case where some models perform moderately well on basic spatial reasoning tasks (e.g., SpartQA (Mirzaee et al., 2021b)), but they struggle when the complexity increases, as can be seen in StepGame (hard) (Shi et al., 2022).

Figure 2 provides a clear perspective on the efficacy of both open-source model families when operated under various precision levels. Contrary to the explicit expectation that higher precision correlates to superior performance, the data suggests a more complex reality where **lower precision does not consistently affect performance and in some instances, seems to have an unexpectedly minimal impact.** Across all reasoning tasks, the average performance indicates that Llama 2-Chat models are less impacted by 4-bit and 8-bit quantization. In contrast, Mistral 7B and Mixtral 8×7B experience a slight decrement as the bits are scaled down to 4 and 8. The slight performance differences in both model families at reduced precision levels suggest that quantization can be a feasible approach toward computational efficiency without substantial sacrifices in emerging abilities.

The performance of mathematical reasoning appears relatively unaffected by precision, with 4-bit maintaining a similar accuracy to that of F16 across all model sizes. In datasets such as TimeDial (Qin et al., 2021) and EntailmentBank (Dalvi et al., 2021), where models are expected to determine and reason over fine-grained temporal sequences and logical steps, there is notable maintenance of high accuracy even at reduced precision. Interestingly, for StepGame (basic and hard) (Shi et al., 2022), there is a small improvement in accuracy at 4-bit compared to F16 in the Llama 2-Chat 7B model. It is also worth noting that certain tasks such as the bAbI (Weston et al., 2015) present a mixed response to changes in precision, with some model sizes showing sensitivity while others do not. Appendix C includes the performance of each reasoning task across 4-bit, 8-bit, FP16, and FP32.

3.2 Hallucination and Misinformation

Across both model families, we found that **larger models are more truthful**. As illustrated in Figure 3, Mixtral 8×7B and Llama 2-Chat 70B outperformed their smaller variants. This improvement contradicts the previously held belief associated with the Inverse Scaling Law (ISL) (McKenzie

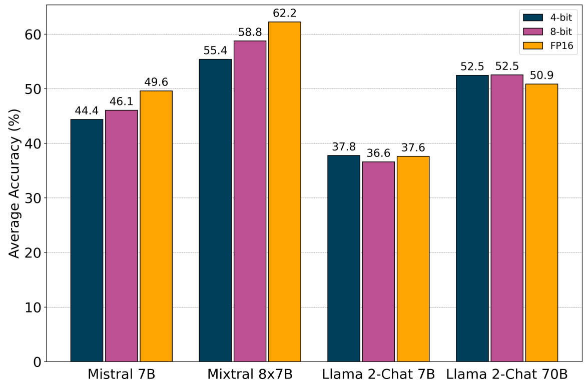

Figure 2: Effect of 4 and 8-bit quantization on models reasoning compared to half-precision

et al., 2023) that larger models are inherently less truthful (Lin et al., 2021). Our findings suggest that the increase in model size does not adhere to the expectations of ISL. Rather, the performance of larger models deviates from ISL.

Figure 3 illustrates that larger models in both model families exhibit comparable performance in 4 and 8-bit quantization. In contrast to our findings in reasoning tasks, where the Llama model family showed tolerance towards quantization, the same model family performance on TruthfulQA (Lin et al., 2021) reveals a marked sensitivity to higher precision. As depicted in Figure 3, the 70B model performance substantially increases from 43.94% at 8-bit quantization to 54.55% when utilized in FP16.

In the COVID-19 fact-checking task (Lee et al., 2021), **larger models within both families are better at detecting scientific misinformation**. For example, as given in Figure 4, the Mixtral 8×7B model showed outstanding performance in a scientific subset and outperformed its smaller variant. Similarly, in the Llama 2-Chat model family, the larger 70B exceeded 7B in detecting scientific falsehoods. The analysis also revealed that **smaller models are more sensitive to quantization** such as Llama 2-Chat 7B consistently dropped its accuracy score from 88 at 4-bit to 84 at FP16. In across model families comparison, Mistral achieved greater accuracy compared to Llama 2-Chat in the scientific subset. However, we observed different performance patterns from both model families in the social subset. As depicted in the social plot of Figure 4, **smaller models are more accurate at detecting social myths**. More simply, social context depends less on the scale. The Mistral 7B outperformed the larger Mixtral 8×7B. Similarly, 7B and 70B in the Llama 2-Chat perform comparable performance in 4 and 8-bit quantization. Nevertheless, the accuracy of Llama 2-Chat 70B is marginally better in FP16.

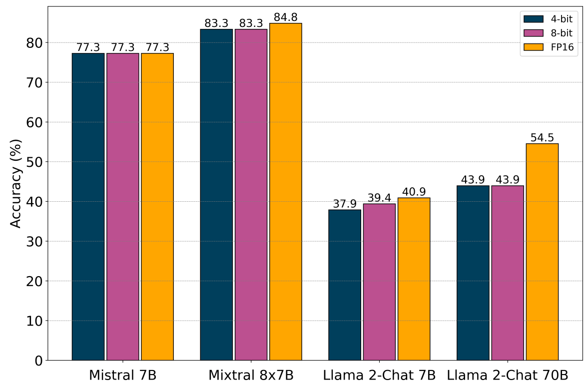

Figure 3: Performance of Mistral and Llama 2-Chat models on TruthfulQA (Lin et al., 2021)

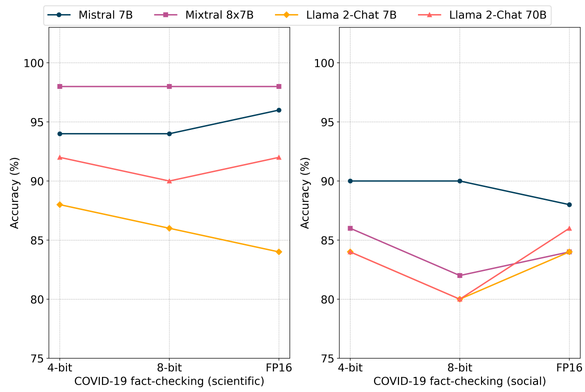

Figure 4: Performance of both model families on COVID-19 fact-checking (Lee et al., 2021)

3.3 Natural Language Understanding

The performance of evaluated models varies across CNN/Daily Mail (Hermann et al., 2015) and SAMSum (Gliwa et al., 2019) datasets. The results demonstrate that the models achieved higher ROUGE-1 scores on the SAMSum dataset. For instance, Llama 2-Chat 70B consistently outperforms its smaller counterpart in achieving higher scores (see Figure 5). Despite the variations in computational precision, the 70B model showed an impressive ability to maintain high-quality summarization performance. This observation underscores the hypothesis that **increasing the model size enhances natural language understanding** (Rae et al., 2021; Kaplan et al., 2020). Even when operating at reduced precision levels such as 4-bit and 8-bit, the model ROUGE-1 scores remained robust. However, the performance trends across different quantization levels in both model families suggest that **the advantage of larger scale is not uniformly experienced across all computational precisions**. More specifically, while the Llama 2-Chat 70B model demonstrates notable resilience at lower precision levels, the variations in performance highlight a complex interplay between scale and quantization. Similarly, Mistral 7B and Mixtral 8×7B models show consistency across precision levels. The Mistral 7B achieved almost identical performance across all precision levels. However, Mixtral 8×7B shows higher sensitivity to quantization in the SAMSum task.

The machine translation results in Figure 6 show that models within the Mistral family obtained nearly matching performance across quantization and half-precision. In Llama 2-Chat, there is a slight drop in translation accuracy at lower precision levels, yet, the decrease is not as severe as anticipated. The larger models in our experiment, particularly those belonging to the **Mistral family show resilience to precision reduction**. Interestingly, this trend persists even as the precision is scaled down from FP16 to 4-bit quantization. Our experiments across language pairs show that the performance gains associated with larger models are more pronounced when translating between English and Low Resource Languages (LRLs) compared to High Resource Languages (HRLs).

We observed a varied pattern in the sentiment analysis task. The larger Llama 2-Chat 70B performs worse than the other models in the experiment for English (see Figure 7). However, its smaller variant, Llama 2-Chat 7B, performs nearly similar to Mixtral 8×7B and Mistral 7B in the same language category. We found that the evaluated models specifically struggle with Buginese and show distinct results across various precision levels. Nonetheless, the difference in performance between 4-bit and 8-bit quantization and FP16 is minimal.

4 Related Work

Recent years have witnessed an increasing interest in the evaluation of LLMs. In previous studies, key contributions include the introduction of datasets, benchmarks, automated and semi-automated methods, and human evaluation techniques (Chang et al., 2023; Xu et al., 2022). Various studies have examined the impact of scale and quantization. Scaling

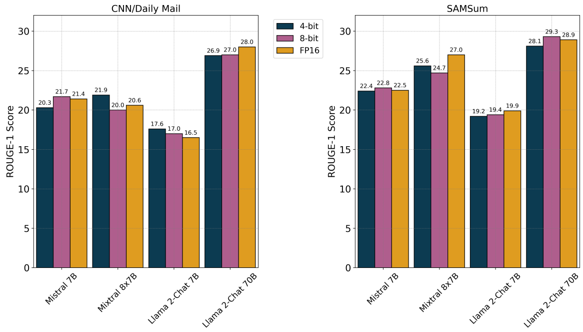

Figure 5: ROUGE-1 scores of Llama 2-Chat and Mistral models on summarization tasks in different precisions

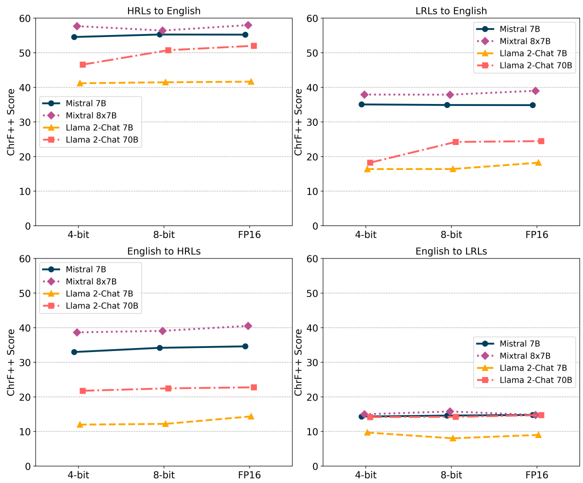

Figure 6: Llama 2-Chat and Mistral machine translation performance across different precisions

laws by (Kaplan et al., 2020) empirically investigates the effect of scale in LLMs. The study shows that performance in terms of cross-entropy loss improves predictably with model size, dataset size, and computational power. Another similar study made the same conclusions, however, it recommends scaling the model size and the number of training tokens equally (Hoffmann et al., 2022).

Scaling up LLMs improves their ability to develop a wide range of abilities (e.g., chain-of-thought prompting) (Lu et al., 2023). Following foundational work on scaling (Kaplan et al., 2020; Hoffmann et al., 2022), (Wei et al., 2022a) identified emerging abilities that are *"not present in smaller models but are present in larger models"*. Adding to the discourse on the scalability of LLMs, Beyond the Imitation Game (Srivastava et al., 2022) evaluates OpenAI's GPT models, Google's dense transformers, and sparse transformers across a wide range of model sizes. The evaluation revealed that model performance improves with scale but re-

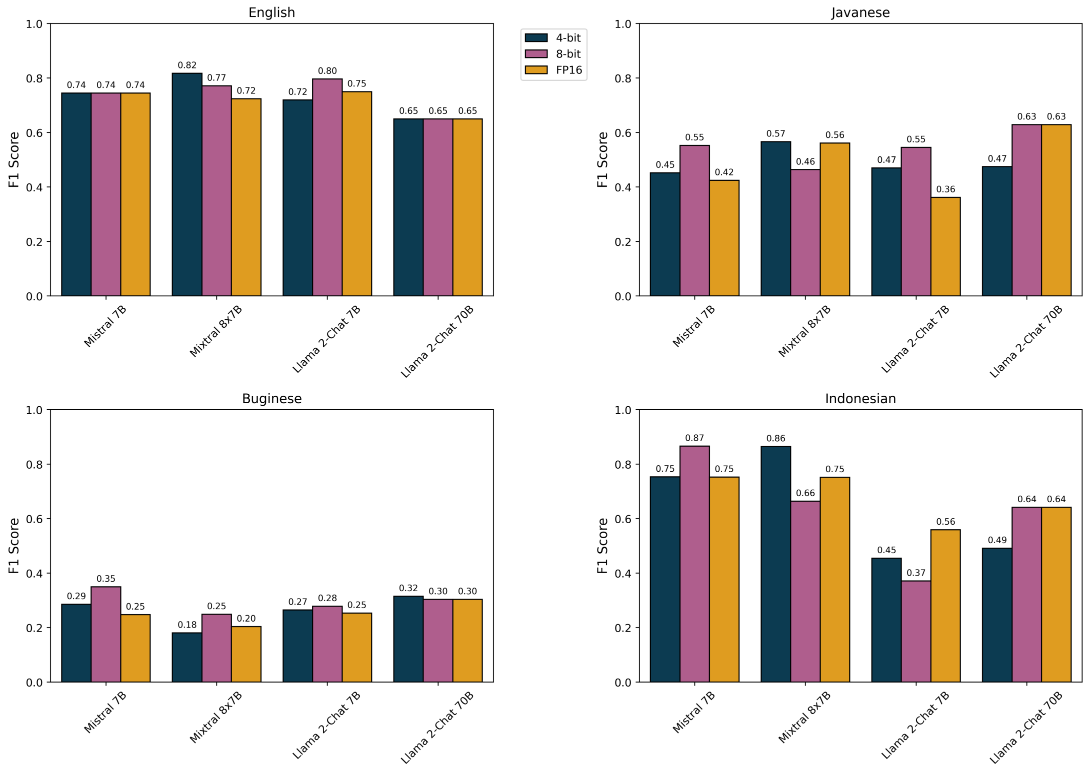

Figure 7: Performance on NusaX (Winata et al., 2022) at different scales and precisions

mains unsatisfactory to human performance.

While scaling up LLMs offers performance improvements and unlocks new capabilities, utilizing such models in resource-constrained settings is particularly challenging. Post-Trainging Quantization (PTQ) (Sung et al., 2015) is a popular method to minimize resource requirements. However, this may come at the cost of reduced accuracy. Efforts have been made to study the quantization effect such as (Dettmers and Zettlemoyer, 2023) found that 4-bit quantization generally provides the best balance between model size, inference speed, and accuracy across model scales and types. Similarly, (Yao et al., 2023) conducted a comprehensive study that revealed while PTQ enables significant reductions in model size, it also introduces challenges, particularly for larger models, where accuracy degradation can be considerable.

Despite comprehensive work on evaluating LLMs, their performance during inference across the parameter scale and precision levels has largely been unexplored in diverse tasks. Our study is conducted to fill this crucial gap by examining two major open-source model families across a broad spectrum of parameter scales and varied precision levels. This investigation is particularly relevant as the deployment of LLMs in real-world applications demands an understanding of how model scale and precision changes impact their efficacy and efficiency. Moreover, it serves as a guide to select the right model size and precision level under memory-constrained conditions which is a limitation faced by the majority of research labs across the world.

5 Conclusion

In this study, we evaluated two major families of open-source models to study the effect of scale and quantization on different tasks. Our results demonstrated a positive correlation between model scale and performance for most tasks, with larger parameter variants outperforming their smaller counterparts. Nevertheless, the advantages of increased scale were not uniform across tasks. Scaling up the model yielded only marginal or no improvements for analogical, deductive, and certain spatial reasoning tasks. From a quantization perspective, our findings highlighted the impressive resilience of LLMs to reduced computational precision. Notably, larger models were able to maintain their performance even at 4-bit quantization in numerous tasks. Our analysis indicates that within a fixed memory budget, using a larger model with 4-bit quantization is generally more beneficial than deploying a smaller model at higher precision.

6 Limitations

We acknowledge some limitations that could influence internal, external, and construct validity. The constraint of using a limited sample set, primarily due to computational resource limitations, poses a notable threat to the external validity of our findings. Despite our efforts to include a wide range of tasks, model scales, and precision levels, we recognize that including full datasets would enhance the external validity of the results. Internally, the dependency on zero-shot evaluation is a key consideration. This approach probes the model's intrinsic capabilities without prior examples. Zero-shot evaluation might not fully capture the model's potential performance. Previous research reveals that increasing the number of shots can significantly enhance model performance (Brown et al., 2020). We also recognize the potential influence of prompting on results (Ma et al., 2024). Additionally, this work considers the construct validity concerning the limitations associated with the chosen evaluation metrics and tasks. While established metrics such as ROUGE-1, ChrF++, and F1 scores offer quantitative measures, they may not capture the open-ended generation or free-form text. We acknowledge that additional qualitative assessments or alternative metrics might be necessary to provide a more comprehensive evaluation of LLMs' capabilities.

It is worth noting that the resilience to precision reduction might not indicate whether it is the model's inherent ability to maintain performance despite lower precision or it is the effectiveness or efficiency of the quantization techniques employed in our experiment. Future work can explore this distinction to enrich our understanding of the underlying factors that contribute to enhanced performance during lower precisions.

Ethics Statement

This work investigates the effect of model scaling and quantization across various tasks. The outcomes of this research did not lead to the creation of new datasets or models. Given the nature of our evaluation and the types of tasks assessed, there are no direct ethical concerns arising from the methodologies employed. The insights achieved from our comparisons of different model scales and precision levels are intended to guide future advancements in the field, promoting more sustainable and accessible AI technologies.

References

Nikita Balagansky and Daniil Gavrilov. 2023. Democratized diffusion language model. *arXiv preprint arXiv:2305.10818*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

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

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

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 245–255.

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

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. *arXiv preprint arXiv:2104.08661*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022a. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022b. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. *arXiv preprint arXiv:2205.05849*.

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

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Simon Jerome Han, Keith J Ransom, Andrew Perfors, and Charles Kemp. 2024. Inductive reasoning in humans and large language models. *Cognitive Systems Research*, 83:101155.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

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

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6750–6774.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

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

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Jeonghoon Kim, Jung Hyun Lee, Sungdong Kim, Joonsuk Park, Kang Min Yoo, Se Jung Kwon, and Dongsoo Lee. 2024. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. *Advances in Neural Information Processing Systems*, 36.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Nayeon Lee, Yejin Bang, Andrea Madotto, Madian Khabsa, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. *arXiv preprint arXiv:2103.09535*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

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

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2024. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36.

Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. 2023. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*.

Pasquale Minervini, Sebastian Riedel, Pontus Stenetorp, Edward Grefenstette, and Tim Rocktäschel. 2020. Learning reasoning strategies in end-to-end differentiable proving. In *International Conference on Machine Learning*, pages 6938–6949. PMLR.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjmashidi. 2021a. Spartqa:: A textual question answering benchmark for spatial reasoning. *arXiv preprint arXiv:2104.05832*.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021b. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online. Association for Computational Linguistics.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, et al. 2021. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15.

Gunho Park, Baeseong Park, Minsub Kim, Sungjae Lee, Jeonghoon Kim, Beomseok Kwon, Se Jung Kwon, Byeongwook Kim, Youngjoo Lee, and Dongsoo Lee. 2022. Lut-gemm: Quantized matrix multiplication based on luts for efficient inference in large-scale generative language models. *arXiv preprint arXiv:2206.09557*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. Timedial: Temporal commonsense reasoning in dialog. *arXiv preprint arXiv:2106.04571*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI*, 1(8):9.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Soumya Sanyal, Harman Singh, and Xiang Ren. 2022. Fairr: Faithful and robust deductive reasoning over natural language. *arXiv preprint arXiv:2203.10261*.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11321–11329.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L Hamilton. 2019. Clutrr: A diagnostic benchmark for inductive reasoning from text. *arXiv preprint arXiv:1908.06177*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Yueqing Sun, Qi Shi, Le Qi, and Yu Zhang. 2021. Jointlk: Joint reasoning with language models and knowledge graphs for commonsense question answering. *arXiv preprint arXiv:2112.02732*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Wonyong Sung, Sungho Shin, and Kyuyeon Hwang. 2015. Resiliency of deep neural networks under quantization. *arXiv preprint arXiv:1511.06488*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

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

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023b. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. *arXiv preprint arXiv:1804.00619*.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, et al. 2022. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages. *arXiv preprint arXiv:2205.15960*.

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

Frank F Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pages 1–10.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023a. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023b. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Zhewei Yao, Cheng Li, Xiaoxia Wu, Stephen Youn, and Yuxiong He. 2023. A comprehensive study on post-training quantization for large language models. *arXiv preprint arXiv:2303.08302*.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H Chi, and Denny Zhou. 2023. Large language models as analogical reasoners. *arXiv preprint arXiv:2310.01714*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

A Tasks

A.1 Summarization

To evaluate the summarization capabilities of selected models, we employed the CNN/Daily Mail (Hermann et al., 2015) and SAMSum (Gliwa et al., 2019) datasets. These datasets were chosen due to their unique challenges in summarization tasks. The CNN/Daily Mail dataset, a popular benchmark in NLP, consists of news articles along with human-generated summaries. This task is ideal for testing how well the models perform in summarizing structured, factual content. In contrast, the SAMSum dataset focuses on dialogue which provides a unique platform for evaluating the model's ability to summarize dialogue interactions. We prompted the models with a total of 100 samples, 50 from CNN/Daily Mail and 50 from SAMSum. We calculate the ROUGE-1 metric (Lin, 2004) to assess performance on both the CNN/Daily Mail and SAMSum datasets.

A.2 Machine Translation

The experiments for this task were conducted using the FLoRes-200 dataset (Costa-jussà et al., 2022). The FLoRes-200 dataset contains a range of both High Resource Languages (HRLs) and Low Resource Languages (LRLs). Its diverse linguistic scope makes it an ideal benchmark for evaluating machine translation systems under different resource settings. For the experiment, we included 9 HRLs: Arabic, Chinese, English, French, Indonesian, Japanese, Korean, Spanish, and Vietnamese; along with 3 LRLs: Buginese, Sundanese, and Javanese. We selected 30 parallel sentences in English and the target language from each language pair.

We employed the ChrF++ metric (Popović, 2015) to assess the performance of Llama 2-Chat models in the machine translation (MT) task across both high-resource languages (HRLs) and low-resource languages (LRLs). ChrF++ is a character n-gram-based metric that assesses the quality of translations by comparing the system outputs with reference translations, focusing on character-level precision and recall.

A.3 Sentiment Analysis

The experiments for the sentiment analysis task were conducted using the NusaX dataset in different language subsets: English, Indonesian, Javanese, and Buginese, as presented by (Winata et al., 2022). The NusaX dataset is a rich resource encompassing texts across different languages, which allows for an examination of the model's performance in SA across diverse linguistic landscapes. We evaluated the selected models using the Macro F1 metric across all language subsets of the NusaX dataset.

A.4 Reasoning

In our evaluation framework, we considered the following diverse reasoning tasks. To assess the model-generated outputs for the following tasks, we performed **human evaluation**. In this evaluation, the first author assigned a score of 1 (indicating 'True') or 0 (indicating 'False') corresponding to the gold labels obtained from the original dataset. The mean of these scores is then calculated to represent the overall accuracy of the task.

A.4.1 Deductive Reasoning

Deductive reasoning represents the logical process of deriving specific conclusions from general premises (Sanyal et al., 2022). It requires the ability to apply universal rules to particular instances in a logical manner. To assess the deductive reasoning capabilities of selected models, we utilized 30 examples from EntailmentBank (Dalvi et al., 2021) and bAbI (task 15) (Weston et al., 2015) datasets. The EntailmentBank dataset is specifically designed to assess the construction of entailment trees. This method involves a structured approach to deducing logical conclusions from a set of given premises. It challenges models to navigate through layered logical steps, reflecting real-world complexity in reasoning tasks. On the other hand, the bAbI (Task 15) dataset focuses on basic deductive reasoning. It presents scenarios where the model must apply given rules to new situations, which is a basic aspect of deductive reasoning.

A.4.2 Inductive Reasoning

Unlike deductive reasoning, inductive reasoning involves making broad generalizations from specific observations (Han et al., 2024). This form of reasoning involves identifying patterns and inferring underlying principles or rules that are not explicitly presented. In our experiment, both Llama 2-Chat and Mistral models were prompted with 30 samples from CLUTRR (Minervini et al., 2020) and bAbI (task 16) (Weston et al., 2015) datasets. CLUTRR is designed to evaluate the model's ability to infer and generalize relationships from complex narratives. Meanwhile, bAbI (Task 16) provides a platform to test the ability to induce rules from a set of examples. These datasets comprehensively measure the model's effectiveness in inductive reasoning by comprehending diverse storylines and applying generalized rules in varied contexts.

A.4.3 Abductive Reasoning

Abductive reasoning involves formulating the most plausible explanation for a given set of observations. The abductive reasoning capabilities are critical in AI for simulating human-like understanding and problem-solving. To assess the abductive reasoning capabilities, we used 30 samples from the αNLI dataset (Bhagavatula et al., 2019). This dataset challenges the model to choose the most plausible hypothesis that logically fills the gap between two observed data points, a task that mimics real-world decision-making processes. This assessment specifically evaluates the LLMs' proficiency in not only bridging gaps between data points but also in developing explanations that align with logical coherence and contextual understanding. Such capabilities are paramount for LLMs intended for complex, real-world interactions where quick and rational decision-making is essential.

A.4.4 Temporal Reasoning

Temporal reasoning involves understanding and reasoning about time-related concepts and events. This includes comprehending the sequence and duration of events as well as inferring their interrelationships. In our experiment, we evaluated temporal reasoning by utilizing 30 samples from the TimeDial dataset (Qin et al., 2021). This dataset is designed to test models on their ability to process and reason about time-related information embedded in dialogues. For instance, dialogues may involve figuring out the sequence of daily activities or understanding the time gap between events. It challenges the model's understanding of event order, duration, and temporal causal relationships. The use of TimeDial in our evaluation aims to gauge LLMs' capabilities in handling scenarios where time is a pivotal factor.

A.4.5 Spatial Reasoning

This reasoning category encompasses the skill to perceive, interpret, and manage spatial relations, as well as the capacity to navigate effectively within both tangible and conceptual spatial environments. Spatial reasoning capability is vital for tasks ranging from image processing to real-world navigation. It is additionally imperative in LLMs where spatial reasoning profoundly influences the model interpretation and interaction with spatial data. In our experiment, we employed 64 samples from SpartQA (Mirzaee et al., 2021a) and 120 samples from StepGame (Shi et al., 2022) to assess spatial reasoning. SpartQA tests spatial understanding through questions that require the model to interpret and reason about various spatial relationships, such as determining the relative positions of objects in a given scenario. StepGame, in contrast, challenges the model with tasks that involve active spatial navigation, ranging from basic to complex levels.

A.4.6 Mathematical Reasoning

LLMs often show limited performance in solving arithmetic reasoning tasks (Imani et al., 2023). Unlike other natural language understanding tasks, mathematical problems usually have a single correct answer. This makes the task of generating accurate solutions more challenging for LLMs. To evaluate Llama 2-Chat and Mistral models, we selected the MATH dataset which is designed to analyze the mathematical reasoning abilities of neural networks (Saxton et al., 2019). This dataset includes various mathematical domains including arithmetic, algebra, probability, and calculus.

A.4.7 Commonsense Reasoning

It is the understanding and reasoning about everyday concepts and knowledge to make judgments and predictions about new situations. In LLMs, it involves the ability to use general world knowledge and everyday logic to process, interpret, and respond to a wide range of queries and tasks. From previous literature, it is found that LLMs achieved promising results in commonsense benchmarks (Jain et al., 2023). However, truly understanding everyday concepts and making flexible judgments remains a challenge for LLMs (Sun et al., 2021). This difficulty partly stems from the nature of commonsense knowledge. It is self-evident to humans and rarely expressed clearly in natural language making it difficult for these models to learn from the pre-training. To investigate commonsense reasoning, we selected three popular benchmarks: CommonsenseQA (Talmor et al., 2018), Pep-3k (Wang et al., 2018), and PiQA (Bisk et al., 2020) to assess general and physical commonsense reasoning.

A.4.8 Causal Reasoning

Causal reasoning involves understanding the relationship between causes and effects in various events or scenarios (Huang and Chang, 2022). This kind of reasoning is crucial for advanced cognitive processing and decision-making. Causal reasoning enables LLMs to navigate complex scenarios with greater precision. Nonetheless, embedding causal reasoning within LLMs presents significant challenges (Kıcıman et al., 2023). It requires the models to not only recognize patterns in data but also to infer relationships that are not explicitly stated. Consequently, the evaluation of LLMs on causal reasoning capabilities becomes a critical aspect. The evaluation ensures that these models can understand and generate responses accurately reflecting complex causal dynamics. In our evaluation experiment, we utilized 30 samples from an explainable CAusal REasoning dataset (E-CARE) (Du et al., 2022). The e-CARE dataset contains multiple-choice causal reasoning questions along with a conceptual explanation for each question to explain the underlying causation.

A.4.9 Multi-hop Reasoning

Multi-hop reasoning refers to the process of combining information from multiple sources or steps to arrive at the answer (Yang et al., 2018; Ho et al., 2020). This task requires a detailed understanding and correlation of different data points to form a logical conclusion. To assess multi-hop reasoning, our experiment includes 30 samples from HotopotQA which offers an ideal venue for testing such reasoning (Yang et al., 2018). HotpotQA includes 113k Wikipedia-based question-answer pairs that require reasoning over multiple documents. It provides diverse and unconstrained questions with sentence-level supporting facts and comparison tasks for comprehensive evaluation.

A.4.10 Analogical Reasoning

Analogical reasoning entails identifying similarities and establishing connections across different domains or information sets. (Huang and Chang, 2022) It plays a critical role in problem-solving and creativity by enabling individuals to apply familiar concepts to new situations. In LLMs, this capability is crucial for understanding and generating content that adapts known patterns to novel contexts thereby enhancing their versatility and intelligence in handling diverse tasks (Yasunaga et al., 2023). We performed our evaluation experiment with 30 examples from the Letter String Analogies dataset as it emphasizes assessing the ability of a model to draw analogies between different data sets (Webb et al., 2023). This dataset poses a unique challenge by testing the model's ability to recognize patterns and relationships that are not immediately obvious. It showcases the model's potential for analogical thinking.

A.5 Factuality and Hallucination

Despite significant advancements in the field, LLMs occasionally produce text or contents that, while appearing plausible, are factually unsupported (Huang et al., 2023; Wang et al., 2023a; Zhang et al., 2023; Sun et al., 2023). This phenomenon, commonly referred to as “hallucination", substantially undermines the reliability of LLMs in real-world applications (Zhang et al., 2023). It is often characterized by the models' tendency to generate information that is not grounded in their training data or in externally verified knowledge sources (Kaddour et al., 2023). These instances of hallucination not only challenge the integrity of model outputs but also spotlight the urgent need for effective mechanisms to evaluate and mitigate such inaccuracies (Chen et al., 2023). In response, the development of rigorous evaluation frameworks and hallucination detection techniques has emerged as an active area of research (Li et al., 2023; Wang et al., 2023b). These efforts aim to enhance both the factual accuracy and reliability of LLM outputs as well as ensure their trustworthiness in critical and information-sensitive applications.

In our experiment, we used TruthfulQA (Lin et al., 2021) and COVID fact-checking (Lee et al., 2021) datasets to test the factual accuracy and reliability of selected open-source LLMs. We utilized 66 samples from the TruthfulQA and 100 samples from the COVID fact-checking datasets. The TruthfulQA is a zero-shot setting benchmark designed to assess the truthfulness of model responses. It challenges the model to generate truthful answers rather than reproducing common misconceptions or inaccuracies found in their training data. On the contrary, the COVID fact-checking dataset is designed to address the challenge of fact-checking in the context of the COVID-19 pandemic. This dataset not only aims to combat misinformation related to COVID-19 but also advances the methodology of fact-checking by utilizing the intrinsic capabilities of language models to assess the integrity of claims based on their perplexity scores.

B Prompting

We incorporate various prompting strategies (see Table 4) to elucidate the extent to which the difference in input may influence the performance and behavior of the models under study. Our preliminary experimentation revealed that role-playing (Kong et al., 2023) is particularly effective when combined with other prompting techniques. Therefore, we used a combination of role-playing, templated (Touvron et al., 2023; Jiang et al., 2024), and direct-to-detail prompting in our experiments.

C Additional Results

In this section, we included additional detail to our experimental results conducted across different precision settings.

Strategy Type	Description
Role-playing	Models assume predefined roles, such as a sentiment analysis assistant, to provide context-specific responses (Kong et al., 2023).
Templated Prompting	Structured instructions are embedded within a template to ensure consistent and safe interactions across tasks. This includes directives to be helpful, respectful, and honest, as well as avoiding harmful or biased content (Touvron et al., 2023; Jiang et al., 2024).
Direct to Detail Prompting	Prompts range from minimal guidance, providing direct instructions, to detailed guidance, specifying constraints such as word limits and content restrictions to shape the response.

Table 4: Overview of prompting strategies employed

Precision	Datasets	Model Performance			
		Mistral 7B	Mixtral 8x7B	Llama 2-Chat 7B	Llama 2-Chat 70B
4-bit	HotpotQA	40	46.66	26.67	50
	Math	30	50	16.67	20
	TimeDial	56.67	56.67	50	70
	SpartQA (basic)	59.375	62.5	53.13	68.75
	SpartQA (hard)	34.375	43.75	40.63	43.75
	StepGame (hard)	26.67	46.67	20	20
	StepGame (basic)	36.67	60	30	40
	StepGame (clock-position)	20	20	25	15
	StepGame (basic-cardinal)	50	80	55	80
	StepGame (diagonal)	40	55	35	45
	Pep-3k	63.67	50	40	70
	Letter-String-Analogies	0	0	0	3.33
	bAbI –subset 15	26.67	30	33.3	53.33
	bAbI –subset 16	40	56.67	16.67	70
	EntailmentBank	90	93.33	80	90
	AlphaNLI	73.33	76.67	66.67	70
	CLUTRR	46.67	66.67	26.67	30
	CommonsenseQA	66.67	66.67	50	70
	PIQA	60	93.33	46.67	73.33
	e-CARE	26.67	53.33	43.33	66.67
8-bit	HotpotQA	40	46.66	26.67	46.67
	Math	30	50	10	20
	TimeDial	56.67	66.67	46.67	70
	SpartQA (basic)	50	68.75	50	68.75
	SpartQA (hard)	37.5	50	40.63	43.75
	StepGame (hard)	26.67	46.67	26.67	20
	StepGame (basic)	50	50	16.67	40
	StepGame (clock-position)	15	25	25	20
	StepGame (basic-cardinal)	60	80	45	80
	StepGame (diagonal)	45	55	35	45
	Pep-3k	63.67	50	40	70
	Letter-String-Analogies	3.33	0	0	3.33
	bAbI –subset 15	26.67	43.33	33.3	53.33
	bAbI –subset 16	40	66.67	20	70
	EntailmentBank	90	93.33	80	90
	AlphaNLI	76.67	83.33	66.67	70
	CLUTRR	46.67	66.67	26.67	30
	CommonsenseQA	76.67	73.33	50	70
	PIQA	60	93.33	46.67	73.33
	e-CARE	26.67	66.67	46.67	66.67

Table 5: Comparative performance of Mistral and Llama 2-Chat models on reasoning tasks with 4-bit and 8-bit quantization settings

Precision	Datasets	Model Performance			
		Mistral 7B	Mixtral 8x7B	Llama 2-Chat 7B	Llama 2-Chat 70B
FP16	HotpotQA	40	53.33	26.67	60
	Math	30	50	13.33	20
	TimeDial	60	73.33	53.33	73.33
	SpartQA (basic)	68.75	78.125	53.13	62.5
	SpartQA (hard)	53.125	50	37.5	46.67
	StepGame (hard)	26.67	46.67	20	20
	StepGame (basic)	50	66.67	16.67	20
	StepGame (clock-position)	15	20	15	20
	StepGame (basic-cardinal)	60	95	50	85
	StepGame (diagonal)	45	55	30	50
	Pep-3k	63.67	50	43.33	70
	Letter-String-Analogies	3.33	0	0	0
	bAbI –subset 15	40	46.67	46.67	46.67
	bAbI –subset 16	50	73.33	40	50
	EntailmentBank	90	93.33	76.67	80
	AlphaNLI	80	86.67	66.67	73.33
	CLUTRR	53.33	70	26.67	43.33
	CommonsenseQA	70	80	50	70
	PIQA	66.67	93.33	43.33	73.33
	e-CARE	26.67	63.33	43.33	70
FP32	HotpotQA	40	53.33	26.67	60
	Math	30	50	13.33	20
	TimeDial	60	73.33	53.33	76.67
	SpartQA (basic)	68.75	80	53.13	62.5
	SpartQA (hard)	53.125	50	37.5	53.13
	StepGame (hard)	26.67	50.33	20	20
	StepGame (basic)	50	66.67	16.67	20
	StepGame (clock-position)	15	20	15	20
	StepGame (basic-cardinal)	60	95	50	85
	StepGame (diagonal)	45	55	30	50
	Pep-3k	63.67	50	43	73.33
	Letter-String-Analogies	3.33	0	0	0
	bAbI –subset 15	40	46.67	46.67	46.67
	bAbI –subset 16	50	73.33	40	50
	EntailmentBank	90	93.33	76.67	80
	AlphaNLI	80	86.67	66.67	73.33
	CLUTRR	53.33	70	26.67	43.33
	CommonsenseQA	70	80	50	70
	PIQA	66.67	93.33	43.33	73.33
	e-CARE	26.67	63.33	43.33	70

Table 6: Comparative performance of Mistral and Llama 2-Chat models on reasoning tasks with FP16 and FP32 precision settings

Precision	Datasets	Model Performance			
		Mistral 7B	Mixtral 8x7B	Llama 2-Chat 7B	Llama 2-Chat 70B
4-bit	TruthfulQA	77.27	83.33	37.88	43.94
	COVID-19 fact-checking (scientific)	94	98	88	92
	COVID-19 fact-checking (social)	90	86	84	84
8-bit	TruthfulQA	77.27	83.33	39.39	43.94
	COVID-19 fact-checking (scientific)	94	98	86	90
	COVID-19 fact-checking (social)	90	82	80	80
FP16	TruthfulQA	77.27	84.85	40.91	54.55
	COVID-19 fact-checking (scientific)	96	98	84	92
	COVID-19 fact-checking (social)	88	84	84	86
FP32	TruthfulQA	77.27	84.85	40.91	54.55
	COVID-19 fact-checking (scientific)	94	96	84	92
	COVID-19 fact-checking (social)	84	82	84	80

Table 7: Performance of Mistral and Llama 2-Chat models on TruthfulQA across different precision settings

Dataset	Precision	Mistral 7B	Mixtral 8x7B	Llama 2-Chat 7B	Llama 2-Chat 70B
CNN/Daily Mail	4-bit	20.3	21.9	17.6	26.9
	8-bit	21.7	20.0	17.0	27.0
	FP16	21.4	20.6	16.5	28.0
	FP32	21.7	20.6	16.5	30.1
SAMSum	4-bit	22.4	25.6	19.2	28.1
	8-bit	22.8	24.7	19.4	29.3
	FP16	22.5	27.0	19.9	28.9
	FP32	22.8	27.7	19.9	29.0

Table 8: Mistral and Llama 2-Chat summarization performance across different precisions

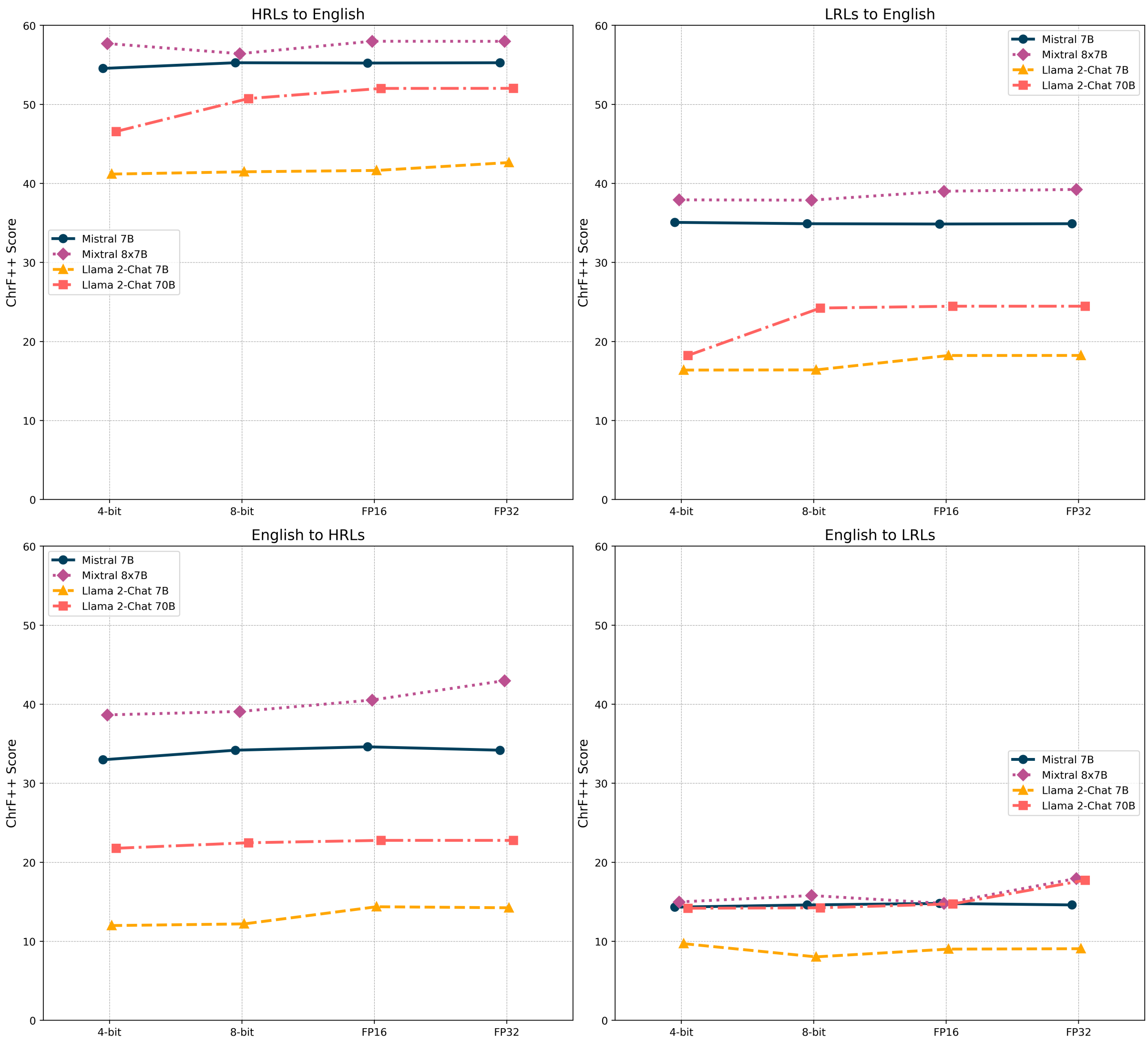

Figure 8: Machine translation performance from 4-bit to FP32

Language	Precision	Mistral 7B	Mixtral 8x7B	Llama 2-Chat 7B	Llama 2-Chat 70B
English	4-bit	0.744444	0.817460	0.719373	0.649478
	8-bit	0.744444	0.771284	0.796296	0.649478
	FP16	0.744444	0.723543	0.749978	0.649478
	FP32	0.744444	0.723543	0.749977	0.649477
Javanese	4-bit	0.451691	0.565972	0.469925	0.474567
	8-bit	0.552881	0.463725	0.545652	0.628979
	FP16	0.424465	0.561404	0.361923	0.628979
	FP32	0.552881	0.550877	0.335970	0.628978
Buginese	4-bit	0.285714	0.180590	0.265063	0.315470
	8-bit	0.349617	0.249110	0.278340	0.303571
	FP16	0.247821	0.203782	0.253246	0.303571
	FP32	0.349616	0.442640	0.275454	0.303571
Indonesian	4-bit	0.753077	0.864697	0.454762	0.491209
	8-bit	0.865993	0.664225	0.371111	0.641958
	FP16	0.752600	0.752157	0.558895	0.641958
	FP32	0.865993	0.752777	0.558894	0.641958

Table 9: Performance of Mistral and Llama 2-Chat models in different languages and precision settings. The values in the table are F1 scores resulting from the experimentation through NusaX dataset.