**RESULT SUMMARY-**

This section presents the impact of quantization on the reasoning capabilities of five large language models—Mistral 7B, Mistral 8×7B, LLaMA3 8B, LLaMA2 13B, and LLaMA2 7B—across four datasets representing different reasoning domains: LogicQA (logical reasoning), SocialIQa (social interaction), CommonsenseQA (commonsense reasoning), and AQuA-RAT (mathematical reasoning). Models were evaluated at quantization levels of 2, 3, 4, 8, 16, and 32 bits. Overall, a clear trade-off between performance and computational efficiency emerged.

Across all models, performance increased with bit depth, with the steepest gains between 2-bit and 8-bit quantization. LogicQA and Math datasets were particularly sensitive to low-bit quantization, showing substantial performance degradation at 2-bit and 3-bit levels. For instance, Mistral 7B’s accuracy on LogicQA was near 0% at 2 bits but improved to over 30% at 4 bits. Similarly, math performance, starting around 6% at 2 bits, climbed to over 21% at 32 bits, highlighting the need for numerical precision in mathematical reasoning.

Social and commonsense reasoning tasks showed relatively better tolerance to quantization. SocialIQa and CommonsenseQA achieved decent performance even at 4-bit precision. Mistral 8×7B, for example, maintained over 58% accuracy on SocialIQa at 4 bits, and LLaMA3 8B achieved above 65% accuracy on CommonsenseQA at the same precision. This suggests that these tasks depend more on linguistic pattern recognition and contextual understanding than on strict numerical accuracy.

When comparing models, Mistral 8×7B consistently outperformed Mistral 7B, especially at lower bit levels, likely due to its mixture-of-experts architecture. It maintained a strong accuracy profile across tasks even at 4- to 8-bit quantization. Among the LLaMA models, LLaMA2 13B showed the best robustness across quantization levels, particularly for LogicQA and CommonsenseQA. Notably, it preserved a relatively high level of reasoning accuracy at 4-bit and 8-bit settings, outperforming the smaller LLaMA2 7B variant in most cases.

LLaMA3 8B also demonstrated strong performance, especially in the SocialIQa and CommonsenseQA datasets. However, its gains plateaued more gradually, suggesting diminishing returns after 8-bit quantization. While LLaMA2 7B performed reliably on social and commonsense reasoning, it struggled with math, never exceeding 8% accuracy even at 32 bits, emphasizing its limitations in handling computation-heavy tasks.

An important pattern observed across all models was the phenomenon of diminishing returns. Most performance improvements occurred between 2 to 8 bits, while gains beyond 16 bits were minimal. For example, across several models and datasets, the difference between 16-bit and 32-bit performance was less than 2%. This implies that 16-bit quantization may be the optimal compromise, preserving reasoning capabilities while reducing memory and computational costs.

In conclusion, the results underline that while quantization inevitably impacts reasoning performance, its effects are uneven across tasks and models. Logical and mathematical reasoning are more precision-sensitive, requiring at least 8–16 bits for acceptable accuracy. In contrast, tasks involving social and commonsense reasoning tolerate lower bit levels better. Model architecture also plays a crucial role, with newer or more advanced models like Mistral 8×7B and LLaMA2 13B being more resilient to quantization. These findings are vital for deploying LLMs in resource-constrained environments, where balancing efficiency and task performance is critical.









