Models in the top-left quadrant of the disk usage vs. accuracy chart (Figure 4) are balancing both metrics effectively.

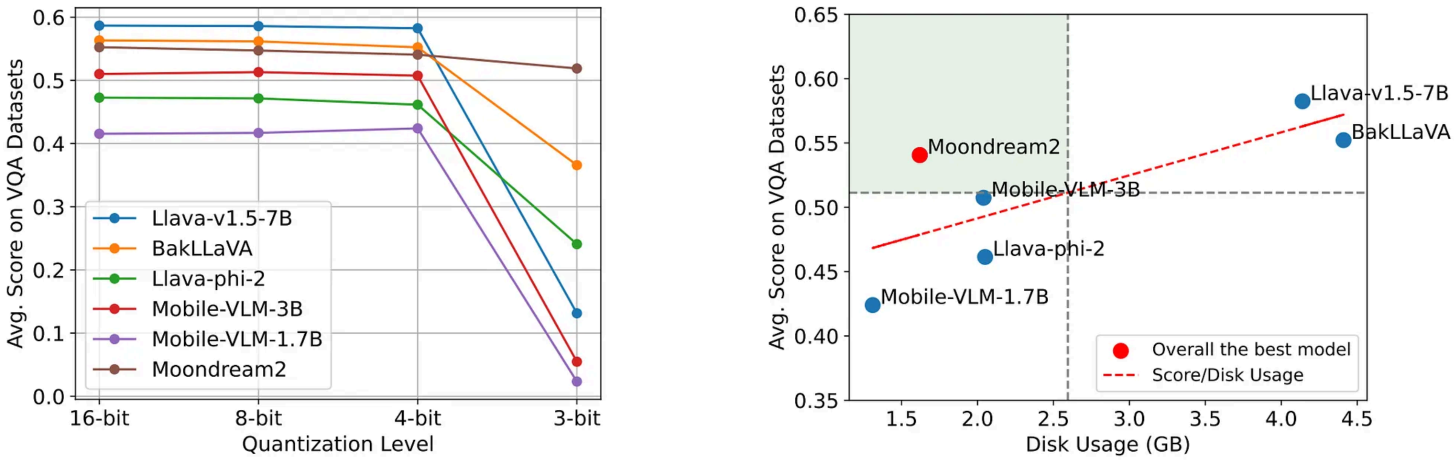


Figure 3 (left): Performance change of LMMs under different quantization. Figure 4 (right): Trade-off between accuracy and disk usage under 4-bit quantization. Source: original paper.

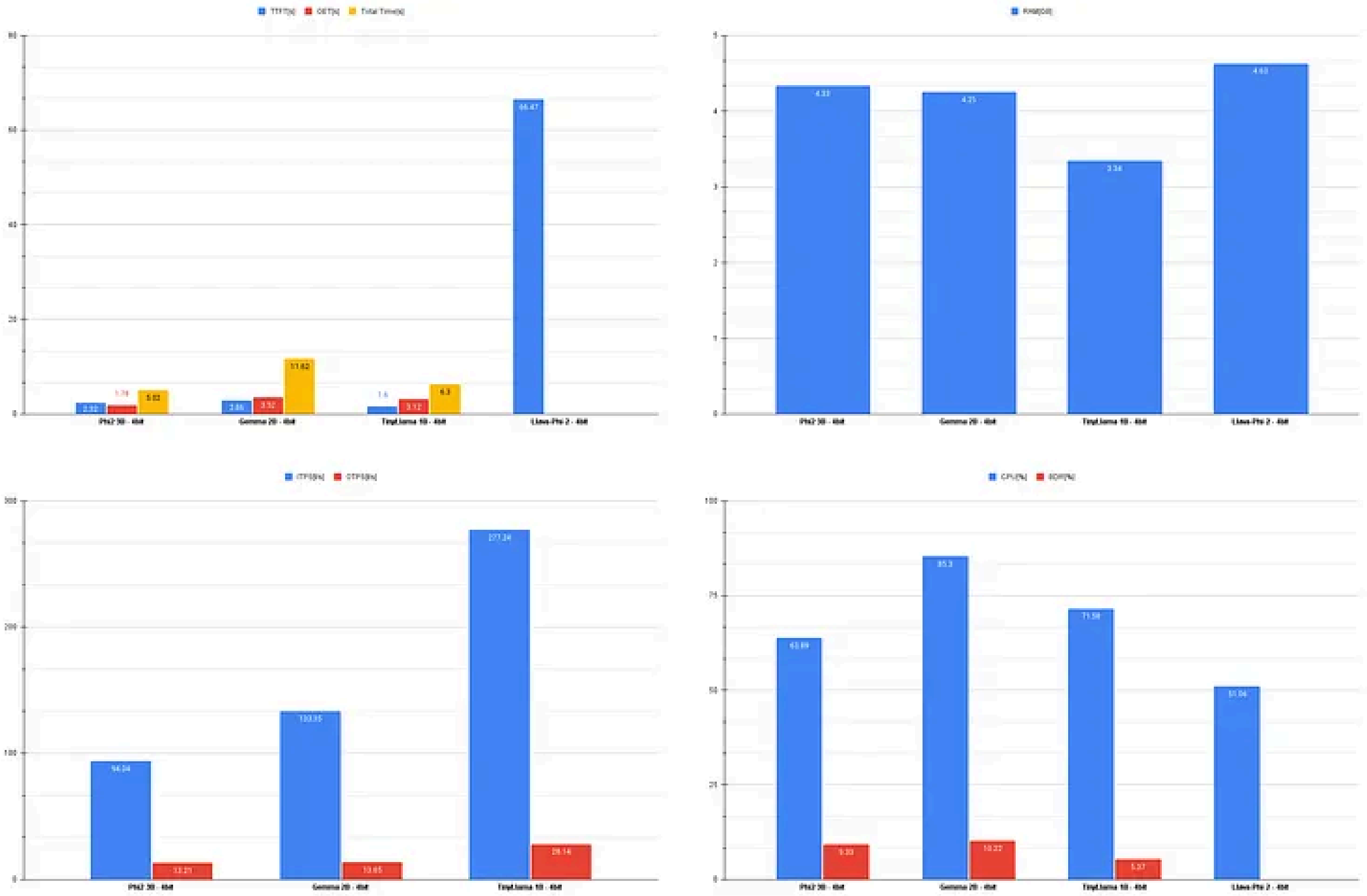
Trust & Safety Evaluation

Evaluation on Trust and Safety tasks revealed that:

- **Performance.** Larger LLMs (>6B parameters) outperformed medium-sized ones on trust and safety tasks as well.
- **Quantization Impact.** The performance impact of quantization was minimal.

Efficiency and Resource Utilisation Evaluation

Due to iPhone 14’s hardware limitations, only 3 LLMs and 1 LMM quantized to 4-bit were evaluated:



Efficiency and Resource utilisation evaluation.

LLMs

Phi2 3B, Gemma 2B, TinyLlama 1B were tested across four NLP datasets. Here are the key findings:

- **Token Processing Speed.** Smaller models had faster TTFT (Time to First Token) and higher ITPS/OTPS (Input/Output Tokens Per Second), meaning they offered better UX.
- **Memory Consumption.** Running even the smallest 4-bit quantized model (TinyLlama 1B) consumed more than 50% of the iPhone’s 6 GB RAM.