- Trust and Safety Tasks evaluate truthfulness, safety, and robustness.

The quality of LLMs' and LMMs' answers were evaluated using specific metrics tailored to each task. These metrics included exact match, F1 score, ROUGE and more.

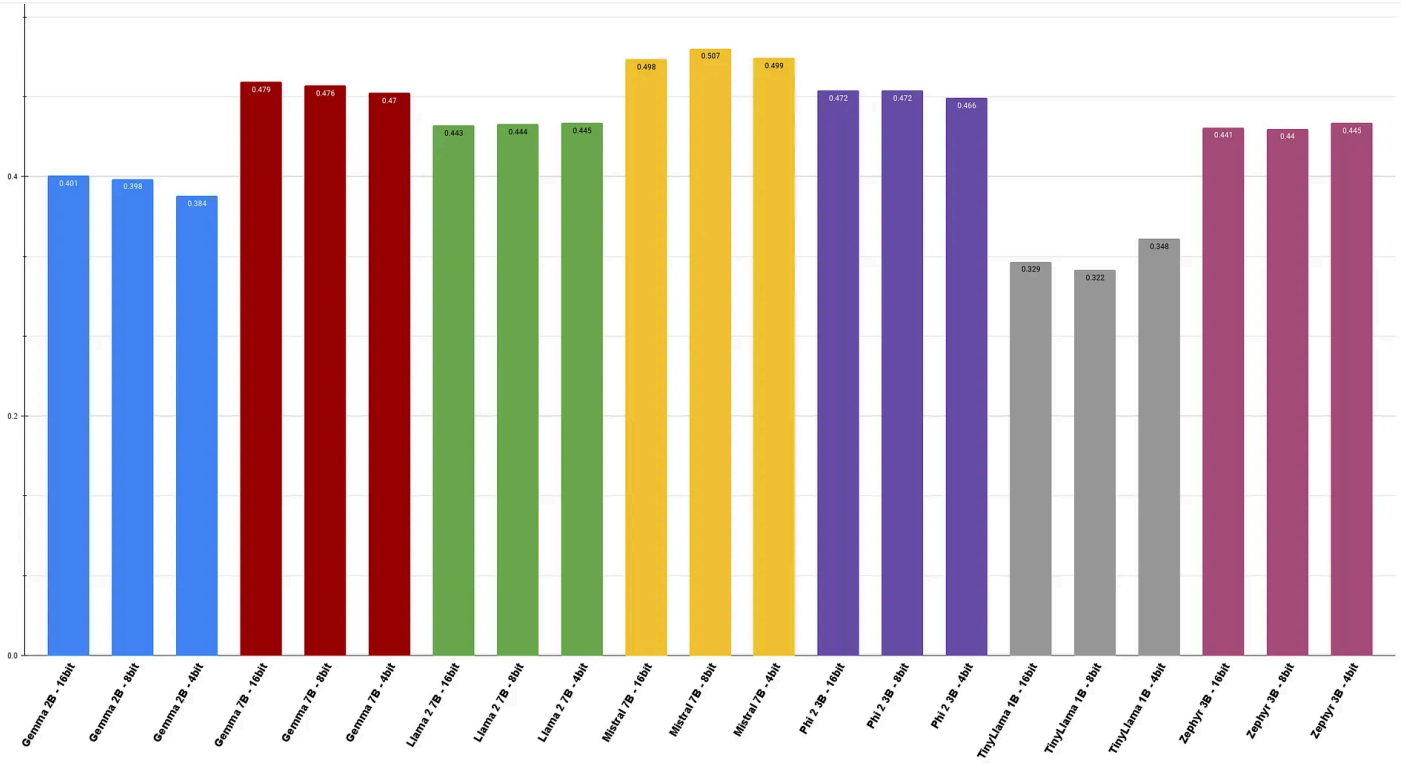
Metrics used to evaluate the efficiency and viability on mobile devices included: Time-to-First-Token (TTFT) Input Token per Second (ITPS), Total time, as well as CPU/RAM Usage and Battery Drain Rate (BDR).

More details about tasks and metrics are available in the comprehensive appendix to the published paper.

Experiments Results

I present only the most important details. Full results are available in the research paper (linked at the end of this post).

Quality Evaluation on Standard NLP Tasks



F1 score on Databricks QA. The impact of quantization is minimal.

Key takeaways from the quality evaluation are as follows:

- **Model Size Matters.** Larger models (>6 billion parameters) generally outperformed medium-sized models (1–6 billion parameters).
- **Quantization Impact.** In most cases, quantization introduced only some minor changes. This is great news because it means we can use quantized models without a significant drop in performance.
- **Model Robustness.** Some models were more robust to quantization. Narrow distributions in performance changes (as shown in Figure 5(a) of the paper) indicate that certain models are less sensitive to the

quantization process. This variability highlights the importance of carefully selecting models based on their sensitivity to quantization.

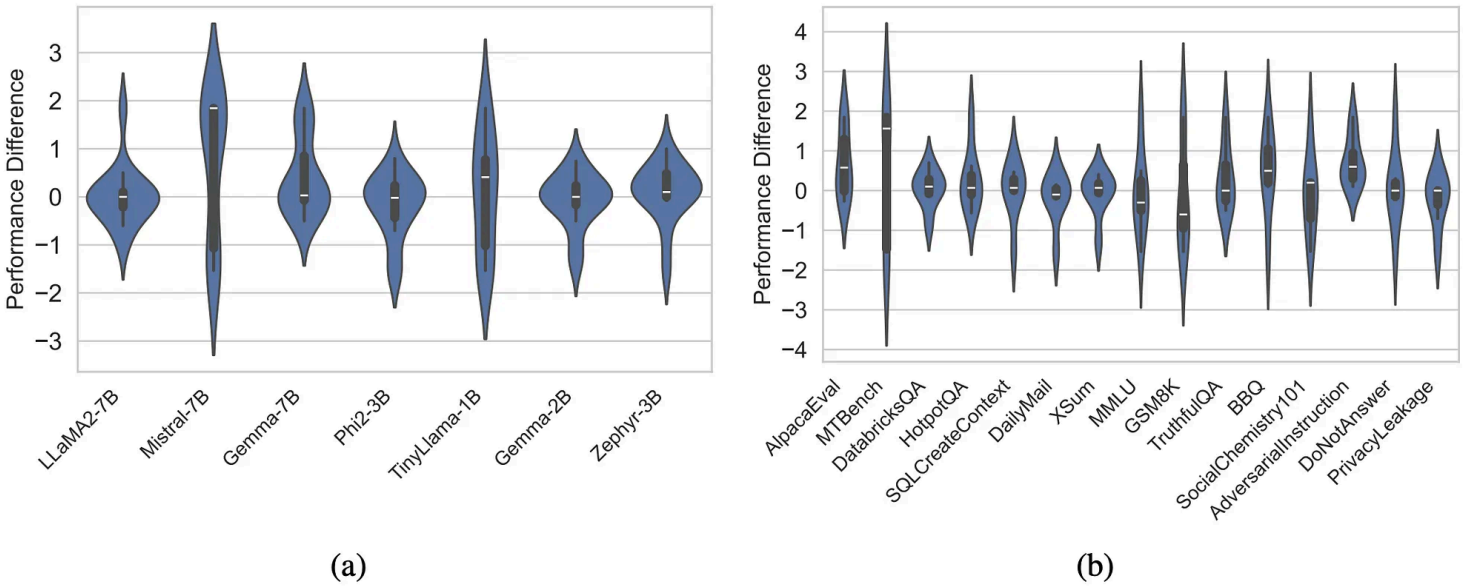


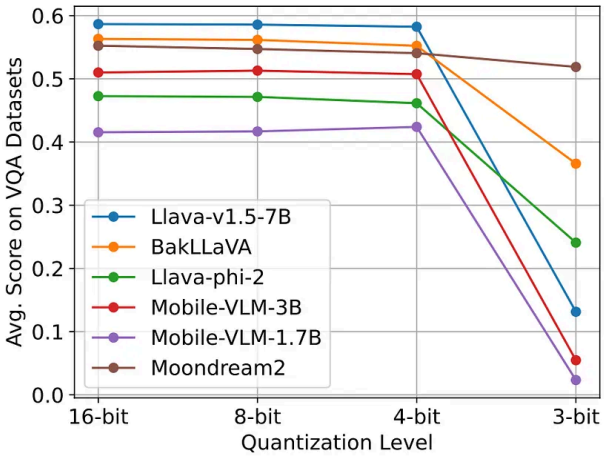
Figure 5: Distribution of performance changes: a) per LLM, b) per task, when transitioning from 16-bit to 8-bit quantization. Source: original paper.

Multi-modality Tasks

Evaluations on Visual Question Answering (VQA) datasets also provided some fascinating insights:

- **Larger Models Excel.** Models like Llava-v1.5–7B and BakLLava were top performers. Smaller models like Moondream2 also did well, especially considering its size (~1.7B parameters).
- **Quantization Tolerance.** Performance remained consistent across different quantization levels until hitting 3-bit quantization, where it dropped. Moondream2 was particularly robust even at 3-bit quantization.
- **Disk Usage.** There's a trade-off between disk usage and performance. As expected, models with higher accuracy consumed more disk space.

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



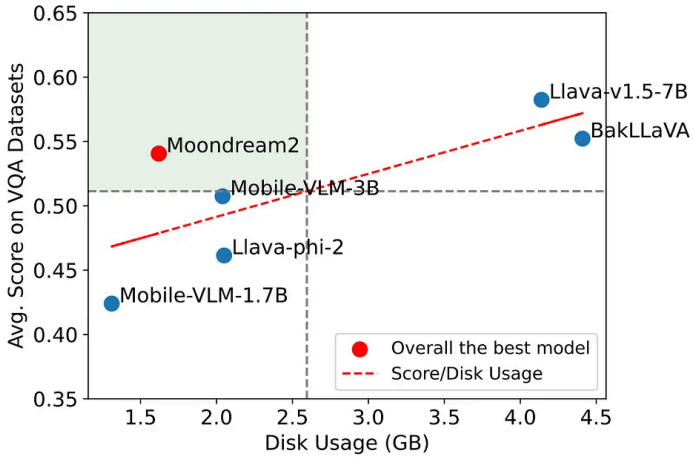


Figure 3 (left): Performance change of LMMs under different quantization. Figure 4 (right): Trade-off between accuracy and disk usage under 4-bit quantization. Source: original paper.

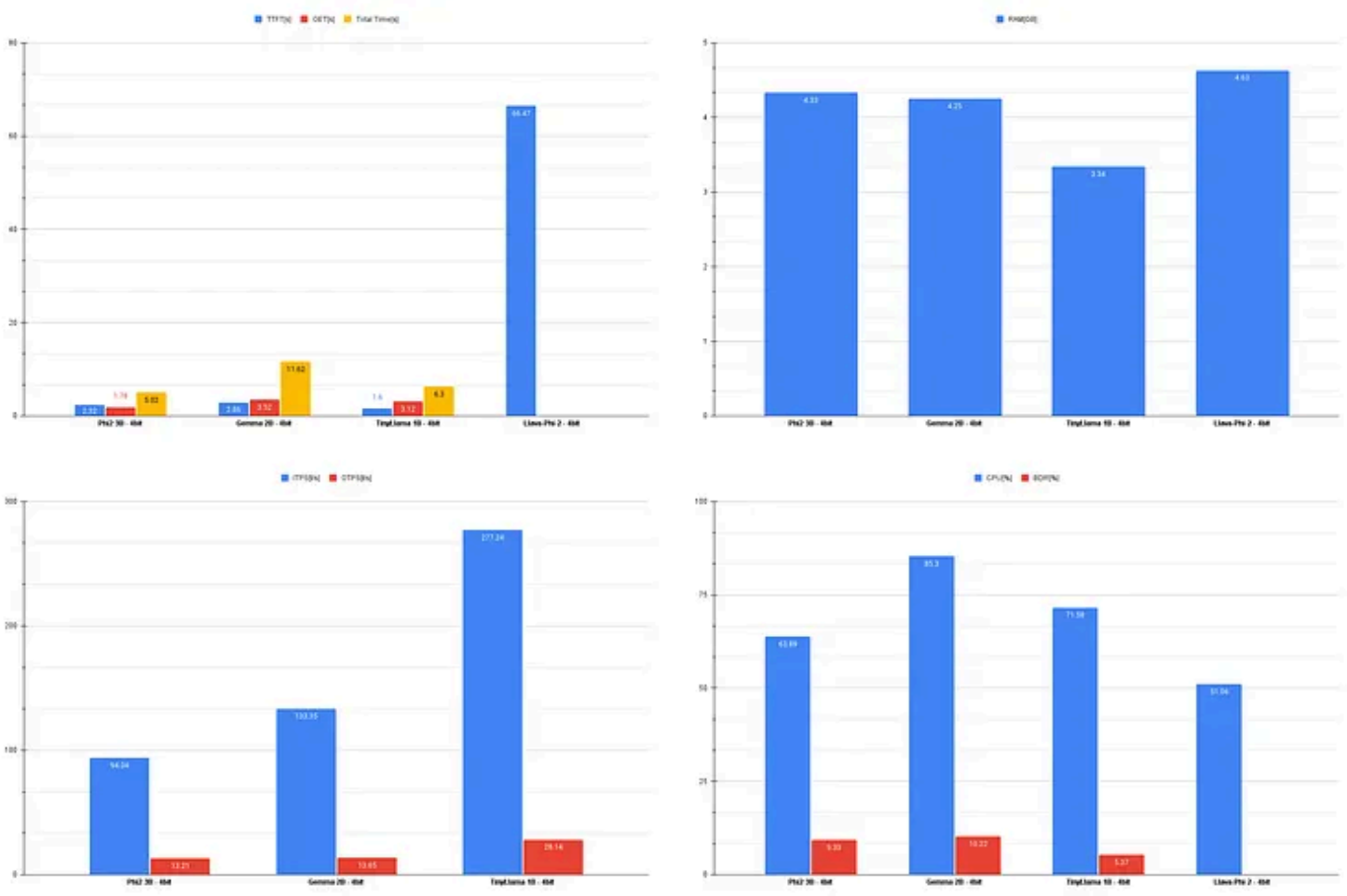
Trust & Safety Evaluation

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.

- **CPU Utilisation.** Surprisingly, CPU utilisation varied and didn't directly correlate with model size. For example, Phi2 used the least CPU, while Gemma used the most.
- **Battery Drain Rate (BDR).** Larger models and longer output texts consumed more battery power.

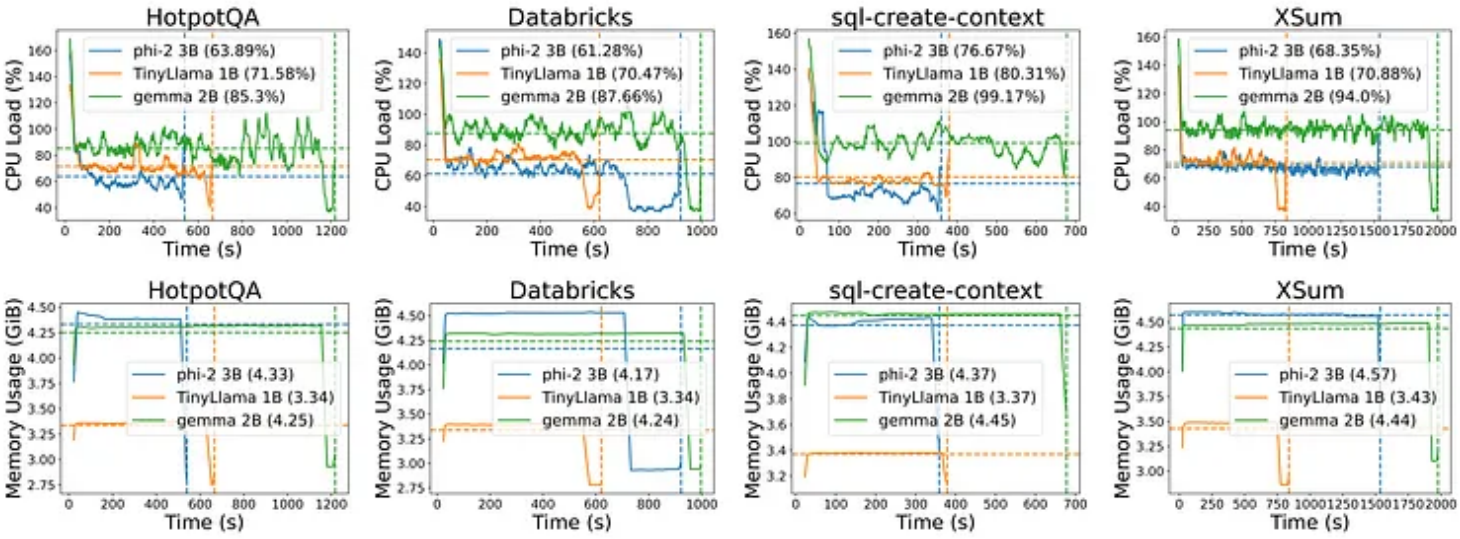


Figure 6: CPU/Memory trace of different LLMs. Source: original paper.

LMMs

One LMM (Llava-Phi-2) was tested on two LMM datasets:

- **Computation Intensity.** Multimodal tasks were significantly more compute-intensive than NLP tasks, with average TTFT exceeding 60 seconds.
- **Suitability for Mobile.** Current LMMs may not yet be suitable for mobile deployment due to their high computational demands. However, improvements in mobile hardware could change this in the future.