General topic: LLM quantization
Research questions: “Low-bit weight quantization for LLMs can significantly reduce the memory footprint of LLM inference but is hard”. So how can we make it better?

Problem: Quantization can significantly reduce the memory space required for LLM operation, but it will hurt the model's ability. How can we maintain the model's ability as much as possible while keeping the same low bit width?

Claims:

1. Weights are not equally important; only 1% are salient weights.
2. We should look for salient weights based on the magnitude of activation, rather than the magnitude of the weights themselves.

Evidence:
Based on validation of wikitext perplexity (lower is better) on the OPT model, it was found that...

1. RTN severely impacts model capabilities.
2. Retaining salient weights based on weight values is similar to random selection; only activation-based selection is effective.
3. When the model is small, 0.1% of the salient weights is insufficient to maintain the effect, but 1% is enough.

Statistical analysis:
AWQ outperforms other quantification methods, supported by the following statistics:

1. AWQ has lower perplexity on the Llama model than other quantization methods with the same bit-width.
2. AWQ scores higher on specific math and programming tasks.
3. AWQ has a higher win rate on open-ended problems, as judged by GPT4.
4. Under the same input, AWQ will not make mistakes where RTN would.