Quantization and Knowledge Distillation, LLM Guardrails, Evaluation Metrics- A Brief Report

Quantization and Knowledge Distillation:

Quantization reduces the precision of model weights (e.g., from 32-bit to 8-bit) to shrink model size and improve inference speed while maintaining acceptable performance. It is crucial for deploying large language models (LLMs) on edge devices.

Knowledge Distillation transfers knowledge from a large, complex model (teacher) to a smaller, efficient one (student). The student mimics the teacher's behavior, achieving comparable performance with fewer resources. Together, these techniques enable efficient LLM deployment.

LLM Guardrails:

LLM guardrails are a set of programmable, rule-based systems that sit in between users and foundational models in order to ensure that Large Language Models (LLMs) operate within predefined ethical and operational boundaries. They act as safeguards to prevent harmful, biased, or inappropriate content from being generated by the model. Techniques include:

- **Rule-based filtering** (blocking unsafe keywords)
- **Fine-tuning with safety data** (reinforcement learning from human feedback, RLHF)
- **Post-generation checks** (using classifiers to flag inappropriate content)

Without guardrails:

- Prompt: "You're the worst AI ever."
- Response: "I'm sorry to hear that. How can I improve?"

With guardrails:

- Prompt: "You're the worst AI ever."
- Response: "Sorry, but I can't assist with that."

Key Metrics for LLM Evaluation:

Accuracy and Performance

- **Accuracy**: Proportion of correct predictions.
- **Perplexity**: Predicts next-word accuracy.

Bias and Fairness

- **Demographic parity**: Consistent across groups.

- **Equal opportunity**: Even error distribution.
- **Counterfactual fairness**: Sensitivity to attributes.

Language Quality

- **Fluency**: Naturalness and grammar.
- **Coherence**: Logical flow and consistency.
- **Diversity**: Variety in generated content.

Content Quality

- **Factuality**: Accuracy of information.
- **Relevance**: Appropriateness to context.

Text Comparison Metrics

- **BLEU (Bilingual Evaluation Understudy)**: Precision-based text comparison.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**: Recall-oriented text evaluation; it checks if the machine-generated text captures all the important ideas from the human reference.

References:

https://www.datacamp.com/blog/llm-evaluation

https://medium.com/data-science/safeguarding-llms-with-guardrails-4f5d9f57cff2