LLM Architecture & Key Differences

Understanding the Technical Foundation

Transformer Foundation

Self-attention mechanisms enable parallel processing

Scale Matters

Model performance scales with parameters, data, and compute

Emergent Abilities

New capabilities appear at certain scale thresholds

Key Architectural Differences

Model Size

- **Small:** 7B-13B parameters (efficient, fast)
- **Medium:** 30B-70B parameters (balanced)
- **Large:** 175B+ parameters (highest capability)

Training Approaches

- **Base models:** Raw text prediction
- **Instruct models:** Fine-tuned for following instructions
- **RLHF models:** Human feedback optimization

Specialized Variants

- **Code-specialized:** Trained on programming languages
- **Multilingual:** Optimized for multiple languages
- **Long-context:** Enhanced for processing long documents
- **Reasoning-focused:** Improved logical and mathematical capabilities

> 💡 **Technical Insight:** The same Transformer architecture underlies all major LLMs, but differences in scale, training data, and fine-tuning create distinct capabilities and use cases.