Report 4: Advances in Model Architectures and Efficiency

Technological innovations improve LLM performance, versatility, and efficiency.

Architectural Innovations

- Mixture-of-Experts models selectively activate sub-networks to balance accuracy and resource use.
- Multi-modal transformers process and generate across text, images, audio, and structured data.
- Retrieval-Augmented Generation (RAG) connects LLMs with external data sources for accurate, up-to-date knowledge.

Efficiency Techniques

- Model distillation compresses large models into smaller ones while retaining capabilities.
- Pruning removes redundant network parameters to speed inference.
- Speculative decoding and cascaded models reduce latency by combining fast approximations with slower precise models.

Enhanced Context and Reasoning

- Extended context windows enable agents to recall large histories and documents.
- Hybrid neuro-symbolic approaches combine neural nets with symbolic logic for improved reasoning and explainability.