Report 3: Fine-Tuning, Alignment, and Safety in LLMs and Agentic AI

Fine-tuning adapts pretrained LLMs to domain-specific tasks by adjusting model weights on labeled data. Reinforcement Learning from Human Feedback (RLHF) refines outputs to align with human preferences.

Alignment Techniques

- Human-in-the-loop (HITL) feedback guides ethical and accurate model behavior.
- Safety layers and content filters prevent harmful and biased outputs.
- Explainability tools provide transparency into model reasoning and decision pathways.
- Monitoring and auditing detect drift and failures post-deployment.

Safety in Agentic AI

- Autonomous agents require additional safeguards due to their active decision-making.
- Fail-safe mechanisms, ethical guardrails, and sandbox testing reduce risks.
- Models introspect and self-reflect to reinforce reliable output generation.
- Structured oversight and human control maintain accountability.