Take-Home LLM Decision Card: Choose, Configure, Ground, Guard

FAST-14 How Large Language Models Work · 7formations · Instructor: Wadood Chaudhary · All examples fictional

Choose

Write a 10-item test with the answer key first. Score 3 models (closed, small, open-weight). Read the model card: license, context, data use, retention, location. Leaderboards are a starting point, not evidence.

Configure

Cost = input tokens × input rate + output tokens × output rate, × volume. 1 token ≈ 0.75 English words. Temperature low for facts, higher for drafts; some reasoning models fix it. Long context: test the middle; cache repeated input.

Ground

Source in the prompt (short) or RAG (long, changing). "Answer only from the source; otherwise say 'not in source'." Ask for the section or quote. Check retrieval misses and grounding misses.

Guard

OWASP LLM Top 10: injection, sensitive data, excessive agency, misinformation, and more. Treat documents as data, not instructions. Least privilege for tools; the model requests, your code decides; human approval for commitments. No MA personal information (G.L. c. 93H, 201 CMR 17.00) or PHI in unapproved tools.

Remember why

Next-token prediction explains variability (sampling), hallucination (plausible beats true), and cost (tokens). Training stage explains behavior: base continues, instruct answers, reasoning thinks longer.