How Large Language Models Work: Foundations for Choosing and Applying LLMs at Work

Course FAST-14 · 12 hours · Instructor: Wadood Chaudhary · Presented by 7formations

Session 1. From text prediction to transformers (3 hours)

- A language model predicts the next token
- From rules to statistics to neural networks to transformers
- Tokens and tokenization; cost and limits
- Embeddings and the three transformer families
- Attention inside the transformer
- **Lab 1: Tokens, probabilities, and temperature with a fictional inn FAQ.**

Session 2. How LLMs are trained and why they behave as they do (3 hours)

- Pipeline: pretraining, instruction tuning, preference tuning
- Pretraining, scaling laws, Mixture-of-Experts
- Instruction and preference tuning (SFT, RLHF, DPO, constitutional methods)
- Reasoning models and test-time compute
- Why LLMs hallucinate and how to reduce it
- **Lab 2: Base vs instruct vs reasoning models; hallucination hunt.**

Session 3. Choosing and configuring a model (3 hours)

- The model landscape: closed and open-weight families, licenses, model cards
- Context windows (up to about 1M tokens) and long-context limits
- Decoding settings, latency, and cost per token
- Benchmarks, arenas, saturation, and contamination
- Multimodal and local models
- **Lab 3: Model selection scorecard with a 10-item task test and cost estimate.**

Session 4. Applying LLMs safely (3 hours)

- Prompt, RAG, or fine-tune (including LoRA)?
- How RAG grounds answers
- Tool use, the Model Context Protocol, and agents
- OWASP Top 10 for LLM Applications (2025)
- NIST AI 600-1, EU AI Act GPAI duties, Massachusetts data rules
- **Lab 4: Grounded assistant, injection test, and LLM solution brief.**

Worksheets and deliverables

- W1 Token, cost, and temperature worksheet
- W2 Model behavior and hallucination log
- W3 Model selection scorecard
- W4 LLM solution brief (final)
- Take-home: one-page LLM decision card (choose, configure, ground, guard)

Attendance and completion

- Attendance is recorded at the start and end of every session (live virtual: platform attendance log plus an instructor roll call or poll; in-person: signed sheet).
- A trainee must attend 100% of instructional hours. A missed session may be made up by attending the same session in another scheduled cohort within 30 days; the make-up is logged the same way.
- A trainee must submit every required lab deliverable and score 'Meets' on the final applied task.
- Each completion record is tied to the trainee, the employer FEIN, and the course number, and is kept for the contract retention period.
- Incomplete trainees are not reported as completions and are not billed to anyone. FAST AI pays for completions only; employers and trainees are never charged.
- Optional asynchronous pre-work (under 10% of course time) is supplemental, is not counted as instructional hours, and never substitutes for a live session.

Assessment rubric

Each criterion is scored Not yet / Meets / Strong. Completion requires Meets on every criterion for the final deliverable (resubmission allowed within the course window).

Criterion	Meets standard
Mechanism	Explains tokens, next-token prediction, and training stages correctly in own words
Cost and context	Token and cost estimate within 20% of the instructor check; context fit stated
Model choice	Recommendation backed by 10-item test results, cost, and data terms
Adaptation	Prompt, RAG, fine-tune, or tools chosen and justified for the need
Risk controls	Three or more OWASP LLM risks matched to concrete controls