FAST-14 Reading List: How Large Language Models Work

Provider: 7formations | Course: FAST-14 | Instructor: Wadood Chaudhary | Links verified by fetch or search as of Oct 2026

No math background is needed for the required readings. Visual explainers come first; research papers are optional and listed so curious trainees can see the primary source. Model names, context limits, and prices change monthly, so check vendor pages (not these readings) for current figures.

How to use. Required readings take 20 to 45 minutes per session and are done before (or right after) the session they support. Optional readings go deeper. 'Living page' means the publisher updates it continuously; read the current version. Reading times are estimates for the parts that matter to this course, not the whole document.

Session 1: From text prediction to transformers

Required

- **Introduction to Large Language Models (Machine Learning Crash Course)**
  Publisher: Google for Developers | Date: Living page | Reading time: about 45 min
  URL: https://developers.google.com/machine-learning/crash-course/llm
  Why read: Google's short module on tokens, next-token prediction, transformers, and self-attention, with quick checks.
- **Transformers, the tech behind LLMs (Deep Learning, Chapter 5)**
  Publisher: 3Blue1Brown (Grant Sanderson) | Date: 2024 | Reading time: about 30 min
  URL: https://www.3blue1brown.com/lessons/gpt/
  Why read: A visual walk through embeddings, attention, and the final probabilities over next tokens, including temperature.

Optional

- **The Illustrated Transformer**
  Publisher: Jay Alammar | Date: June 27, 2018 | Reading time: about 30 min
  URL: https://jalammar.github.io/illustrated-transformer/
  Why read: The classic illustrated explanation of self-attention, queries, keys, and values.
- **Attention Is All You Need (Vaswani et al.)**
  Publisher: arXiv (NeurIPS 2017) | Date: June 2017 | Reading time: about 20 min
  URL: https://arxiv.org/abs/1706.03762
  Why read: The 2017 transformer paper (eight authors). Skim the abstract and figure 1.

Session 2: How LLMs are trained and why they behave as they do

Required

- **Why language models hallucinate**
  Publisher: OpenAI | Date: September 5, 2025 | Reading time: about 15 min
  URL: https://openai.com/index/why-language-models-hallucinate/
  Why read: A model developer explains why training and evaluation incentives reward guessing over saying 'I don't know'.

Optional

- **Tracing the thoughts of a large language model**
  Publisher: Anthropic (Research) | Date: March 27, 2025 | Reading time: about 15 min
  URL: https://www.anthropic.com/research/tracing-thoughts-language-model
  Why read: Interpretability research on what happens inside a model, including cases where its stated reasoning wasn't what it computed.
- **Reduce hallucinations**
  Publisher: Anthropic (Claude Docs) | Date: Living page | Reading time: about 10 min
  URL: https://platform.claude.com/docs/en/test-and-evaluate/strengthen-guardrails/reduce-hallucinations
  Why read: Practical techniques to reduce hallucinations: allow 'I don't know', require quotes, verify with citations.

Session 3: Choosing and configuring a model

Required

- **Model Cards for Model Reporting (Mitchell et al.)**
  Publisher: arXiv / ACM FAT* 2019 | Date: October 2018 (v2 January 2019) | Reading time: about 20 min
  URL: https://arxiv.org/abs/1810.03993
  Why read: The paper that introduced model cards. Use its sections as your checklist when reading a vendor's card in Lab 3.

Optional

- **Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1**
  Publisher: NIST | Date: July 2024 | Reading time: about 20 min
  URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
  Why read: NIST Generative AI Profile: risk categories to check when comparing vendors.

Session 4: Applying LLMs safely

Required

- **OWASP Top 10 for LLM Applications**
  Publisher: OWASP Gen AI Security Project | Date: 2025 edition | Reading time: about 20 min
  URL: https://genai.owasp.org/llm-top-10/
  Why read: The OWASP Top 10 for LLM Applications (the course uses the 2025 list; LLM01 Prompt Injection is the one Lab 4 tests).
- **What is the Model Context Protocol (MCP)?**
  Publisher: Model Context Protocol (official documentation) | Date: Living page | Reading time: about 10 min
  URL: https://modelcontextprotocol.io/docs/getting-started/intro
  Why read: What the Model Context Protocol is: a standard way for AI apps to connect to tools and data.
- **201 CMR 17.00: Standards for the Protection of Personal Information of MA Residents**
  Publisher: Mass.gov, Office of Consumer Affairs and Business Regulation | Date: Regulation page dated 10/19/2017 (in force since 2010) | Reading time: about 15 min
  URL: https://www.mass.gov/regulations/201-CMR-1700-standards-for-the-protection-of-personal-information-of-ma-residents
  Why read: The Massachusetts personal-information rules your W4 brief must address.

Optional

- **OWASP GenAI LLM Top 10 2026**
  Publisher: OWASP Gen AI Security Project | Date: Published August 4, 2026 | Reading time: about 10 min
  URL: https://genai.owasp.org/resource/owasp-genai-llm-top-10-2026/
  Why read: The 2026 update of the same list, to see what changed.
- **AI Act (regulatory framework overview and application timeline)**
  Publisher: European Commission, Shaping Europe's digital future | Date: Living page (updated after the AI Omnibus, 2026) | Reading time: about 10 min
  URL: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
  Why read: EU AI Act overview, including general-purpose AI model obligations.

Link verification

Links verified by fetch or search as of Oct 2026. Links that could not be confirmed were dropped and are not listed. The instructor re-checks links before each cohort.