FAST-13 Reading List: Evaluating and Operating Agentic AI Systems

Provider: 7formations | Course: FAST-13 | Instructor: Wadood Chaudhary | Links verified by fetch or search as of Oct 2026

Readings mix practitioner guides, primary research, and the current standards the labs use (OWASP Top 10 for Agentic Applications 2026, OpenTelemetry GenAI conventions). Platform docs are for whichever of Arize Phoenix, Langfuse, or LangSmith your employer chose; read only the one you use. Note: in June 2026 OpenTelemetry moved its GenAI semantic conventions to a dedicated repository, so older links to opentelemetry.io/docs/specs/semconv/gen-ai/ now show a 'moved' notice.

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: Why agents fail and how to see it

Required

- **Demystifying evals for AI agents**
  Publisher: Anthropic (Engineering blog) | Date: January 9, 2026 | Reading time: about 25 min
  URL: https://www.anthropic.com/engineering/demystifying-evals-for-ai-agents
  Why read: Defines task, trial, grader, transcript, and outcome, the vocabulary this course uses, and gives a roadmap from zero to a working eval suite.
- **Your AI Product Needs Evals**
  Publisher: Hamel Husain (practitioner blog) | Date: March 29, 2024 | Reading time: about 20 min
  URL: https://hamel.dev/blog/posts/evals/
  Why read: Why looking at traces comes first, plus a worked case of turning observed failures into tests.

Optional

- **Building effective agents**
  Publisher: Anthropic (Engineering) | Date: December 19, 2024 (with 2026 note) | Reading time: about 15 min
  URL: https://www.anthropic.com/engineering/building-effective-agents
  Why read: Agent and workflow patterns, so you can name the parts (router, tools, loop) you'll evaluate.

Session 2: Graders you can trust

Required

- **Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena (Zheng et al.)**
  Publisher: arXiv (NeurIPS 2023 Datasets and Benchmarks) | Date: June 2023 | Reading time: about 25 min

URL: https://arxiv.org/abs/2306.05685
Why read: The core study of LLM judges: position, verbosity, and self-enhancement bias, and the swap-order mitigation used in activity F13-S2-A3. Read sections 3 and 4.

Optional

- **Evaluation (Arize Phoenix documentation)**
  Publisher: Arize AI | Date: Living page | Reading time: about 10 min
  URL: https://arize.com/docs/phoenix/evaluation/llm-evals
  Why read: Phoenix evaluators (code and LLM-as-a-judge) and score objects, matching the evals 3.x code in Lab 2.
- **LLM-as-a-Judge (Langfuse documentation)**
  Publisher: Langfuse | Date: Living page | Reading time: about 10 min
  URL: https://langfuse.com/docs/evaluation/evaluation-methods/llm-as-a-judge
  Why read: The same idea in Langfuse, including calibrating a judge against human labels.

Session 3: Evaluating agent behavior end to end

Required

- **tau-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains (Yao et al.)**
  Publisher: arXiv | Date: June 2024 | Reading time: about 20 min
  URL: https://arxiv.org/abs/2406.12045
  Why read: Outcome grading by comparing the final database state with the goal state; the failure breakdown (wrong arguments, wrong decisions, partial resolution) mirrors Lab 3.

Optional

- **Demystifying evals for AI agents**
  Publisher: Anthropic (Engineering blog) | Date: January 9, 2026 | Reading time: about 10 min
  URL: https://www.anthropic.com/engineering/demystifying-evals-for-ai-agents
  Why read: Re-read 'Design graders thoughtfully': grade outcomes over exact paths unless the order is a business rule.

Session 4: Reliability, cost, and security

Required

- **tau-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains (Yao et al.)**
  Publisher: arXiv | Date: June 2024 | Reading time: about 10 min
  URL: https://arxiv.org/abs/2406.12045
  Why read: Section 3 defines pass^k and the unbiased estimator; see how pass^8 drops below 25% even for strong agents.

- **OWASP Top 10 for Agentic Applications for 2026**
  Publisher: OWASP Gen AI Security Project | Date: December 9, 2025 | Reading time: about 25 min
  URL: https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/
  Why read: The ASI01-ASI10 categories you map red-team results to in Lab 4.

Optional

- **How to red team LLM applications**
  Publisher: Promptfoo | Date: Living page | Reading time: about 15 min
  URL: https://www.promptfoo.dev/docs/guides/llm-redteaming/
  Why read: How to express red-team cases as a repeatable test suite. Test only systems you're authorized to test.
- **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 companion LLM application list, for prompt injection and sensitive information disclosure.

Session 5: Operating agents in production

Required

- **OpenTelemetry semantic conventions for generative AI (semantic-conventions-genai repository)**
  Publisher: OpenTelemetry (CNCF) | Date: Living repository; GenAI conventions moved here with semantic-conventions v1.42.0, June 2026 | Reading time: about 15 min
  URL: https://github.com/open-telemetry/semantic-conventions-genai
  Why read: The current home of the gen_ai.* span attributes used in Lab 5. Note that message content capture is opt-in.
- **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 written-security-program rules that apply to traces holding personal information.

Optional

- **AI Risk Management Framework (overview page)**
  Publisher: National Institute of Standards and Technology (NIST) | Date: Living page; AI RMF 1.0 released January 2023 | Reading time: about 10 min
  URL: https://www.nist.gov/itl/ai-risk-management-framework
  Why read: NIST AI RMF overview; map your eval report to the Measure function.
- **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: risks and suggested actions for monitoring and incident response.

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.

Dropped during verification (not listed above):

- OpenTelemetry 'Semantic conventions for generative AI systems' at opentelemetry.io/docs/specs/semconv/gen-ai/ (page now shows only a 'moved' notice; replaced with the new repository)