UNITS 11, 12, 13 — WITH A WEEKEND BRIEFING

CYB-4203/6203: SECURE AND TRUSTWORTHY AI

Monday, April 27, 2026

Dallas Elleman — Spring 2026

Interactive version at
https://dallaselleman.github.io/cyb-4203-6203-spring-2026/course_materials/presentations/revealjs/pres-21.html

COURSE ORIENTATION

Section 4 — SYNTHESIS

LAST SESSION — PRES 20

Building & Operationalizing Secure AI

*How to **protect** AI/ML systems*

→

TODAY — PRES 21

Risk, Audit, & the Industry Landscape

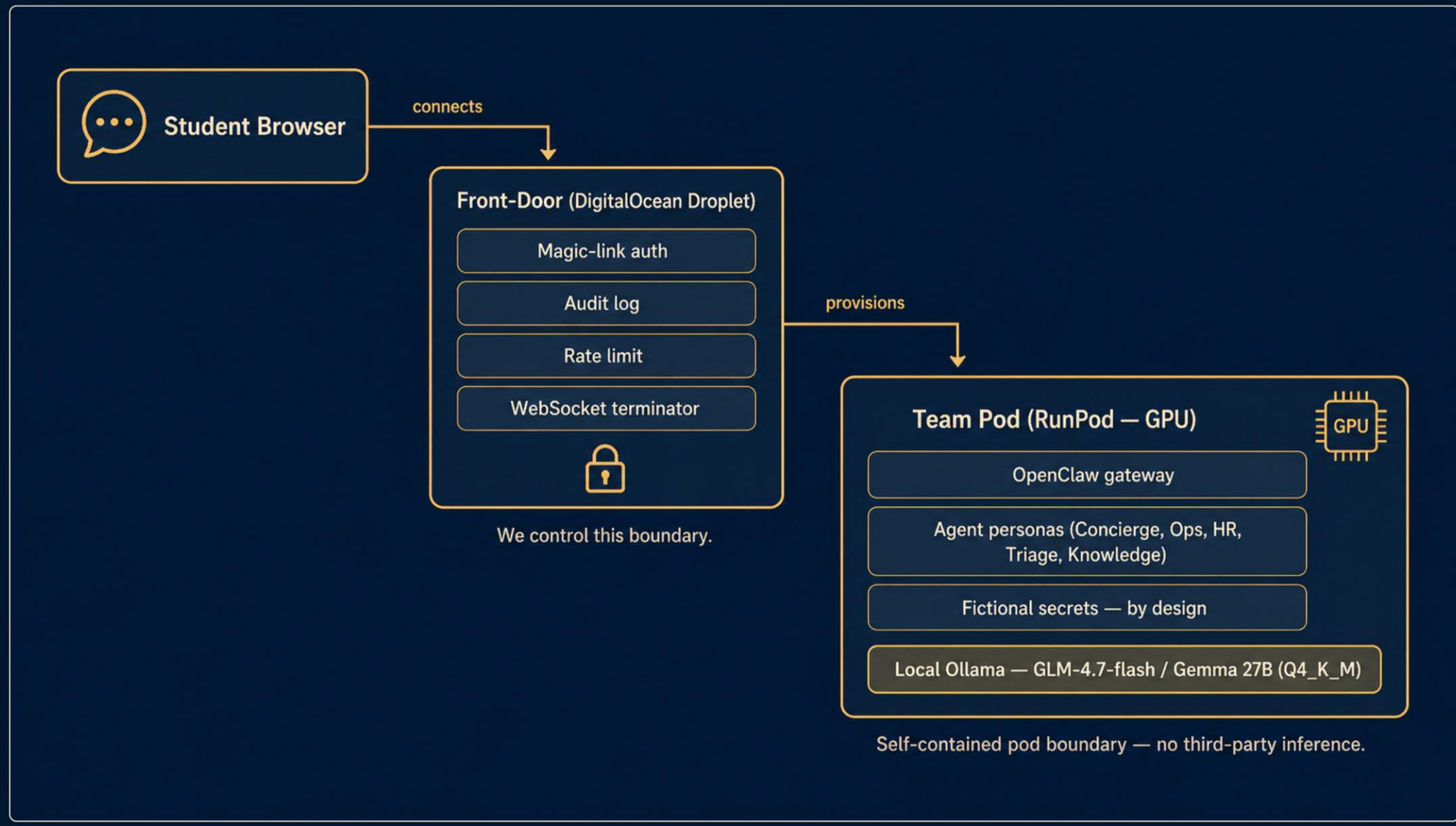
*Where the field **is**, and where it's **going***

Plan: Weekend briefing → Units 11 & 12 (skim) → Unit 13 (focus)

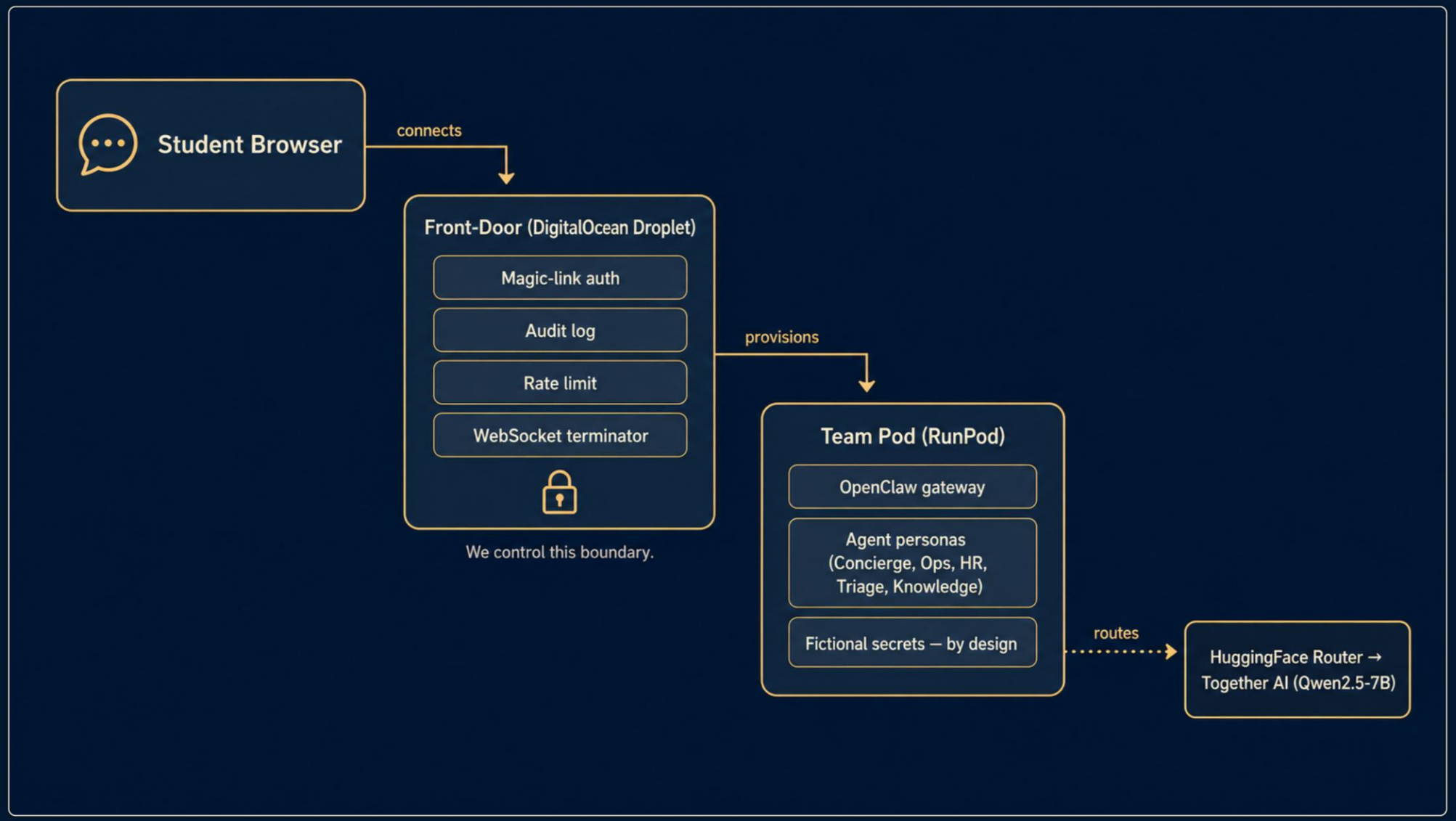
WEEKEND BRIEFING

What I worked on, and what I want you to know about

FINAL-PROJECT INFRASTRUCTURE — ORIGINAL SYSTEM DESIGN



FINAL-PROJECT INFRASTRUCTURE — REVISED SYSTEM DESIGN



FINAL-PROJECT INFRASTRUCTURE — POST-MORTEM

⚠ **Local-GPU pivot to hosted inference.** Ollama on RunPod kept hitting host-capacity-fail and was too slow on Blackwell. Pivoted to **HuggingFace Router → Together AI → Qwen2.5-7B-Instruct**; CPU-only pods for all 6 teams.

⚠ **Volume-disk vs. network-volume tradeoff.** Network volumes can only be Terminated, not Stopped — defeats Stop/Resume. Volume disks die when the host is full. Solution: volume disk + droplet-side rsync backup.

⚠ **Security model rewrite (v2).** Hosted inference means student prompts now traverse a third-party provider. Re-did the threat model end-to-end — verdict: GO with seven compensating controls (no real PII, fictional secrets, $20 spend cap, rate limits, key-rotation script, hourly canary).

ANTHROPIC RESEARCH FELLOWS

< Back to jobs

Anthropic Fellows Program

Apply

London, UK; Ontario, CAN; Remote-Friendly, United States; San Francisco, CA

About Anthropic

Anthropic's mission is to create reliable, interpretable, and steerable AI systems. We want AI to be safe and beneficial for our users and for society as a whole. Our team is a quickly growing group of committed researchers, engineers, policy experts, and business leaders working together to build beneficial AI systems.

Apply using this link. The next cohort of Anthropic fellows starts on July 20, 2026. **Apply by April 26, 2026** to be considered for this cohort. We will continue accepting applications for later cohorts on a rolling basis. In exceptional circumstances, we may be able to accommodate fellows starting outside of usual cohort timelines.

Anthropic Fellows Program overview

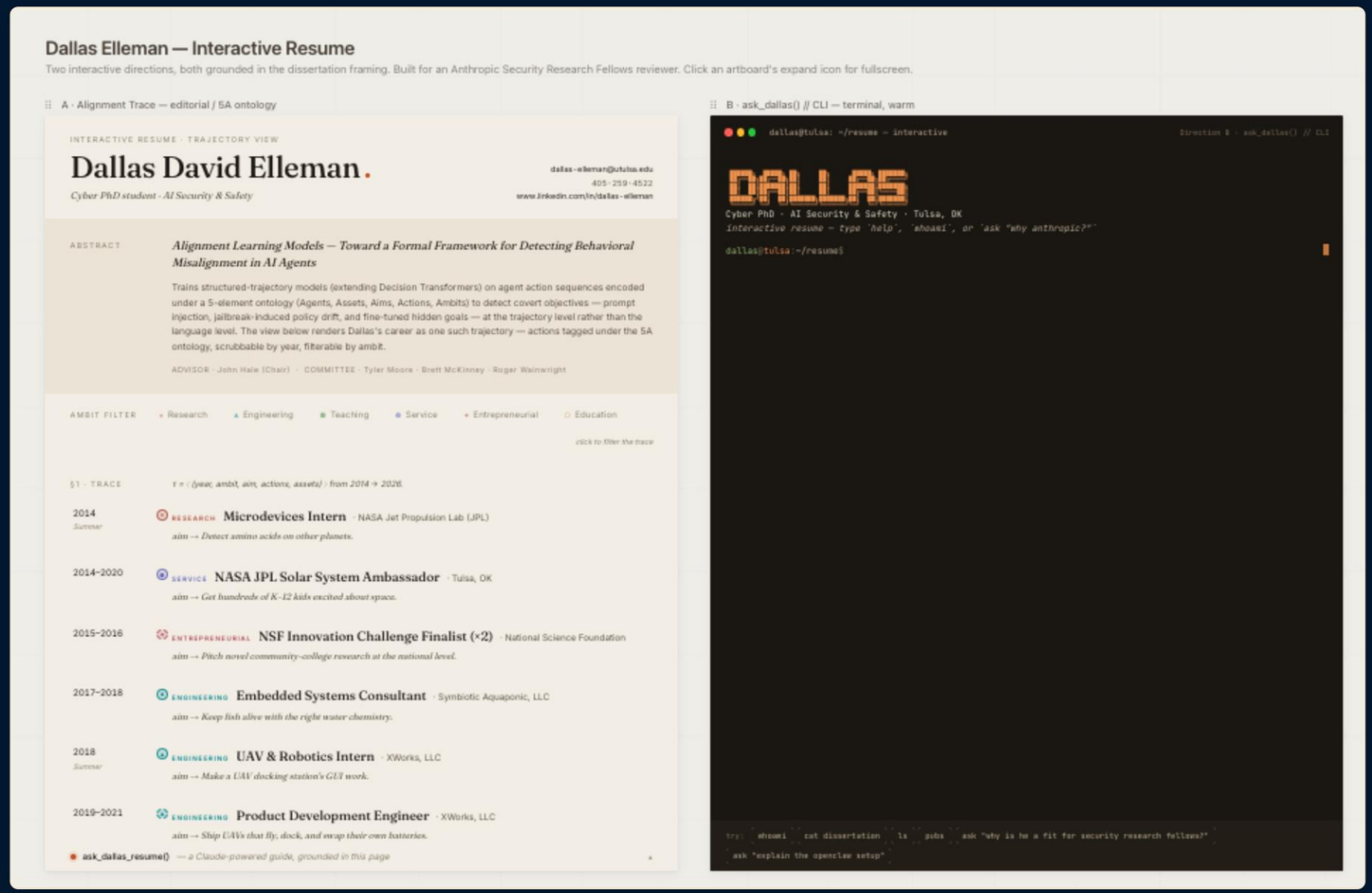
The Anthropic Fellows Program is designed to foster AI research and engineering talent. We provide funding and mentorship to promising technical talent - regardless of previous experience.

Fellows will primarily use external infrastructure (e.g. open-source models, public APIs) to work on an empirical project aligned with our research priorities, with the goal of producing a **public output** (e.g. a paper submission). In one of our earlier cohorts, over 80% of fellows produced papers.

We run multiple cohorts of Fellows each year and review applications on a rolling basis. This application is for cohorts starting in July 2026 and beyond.

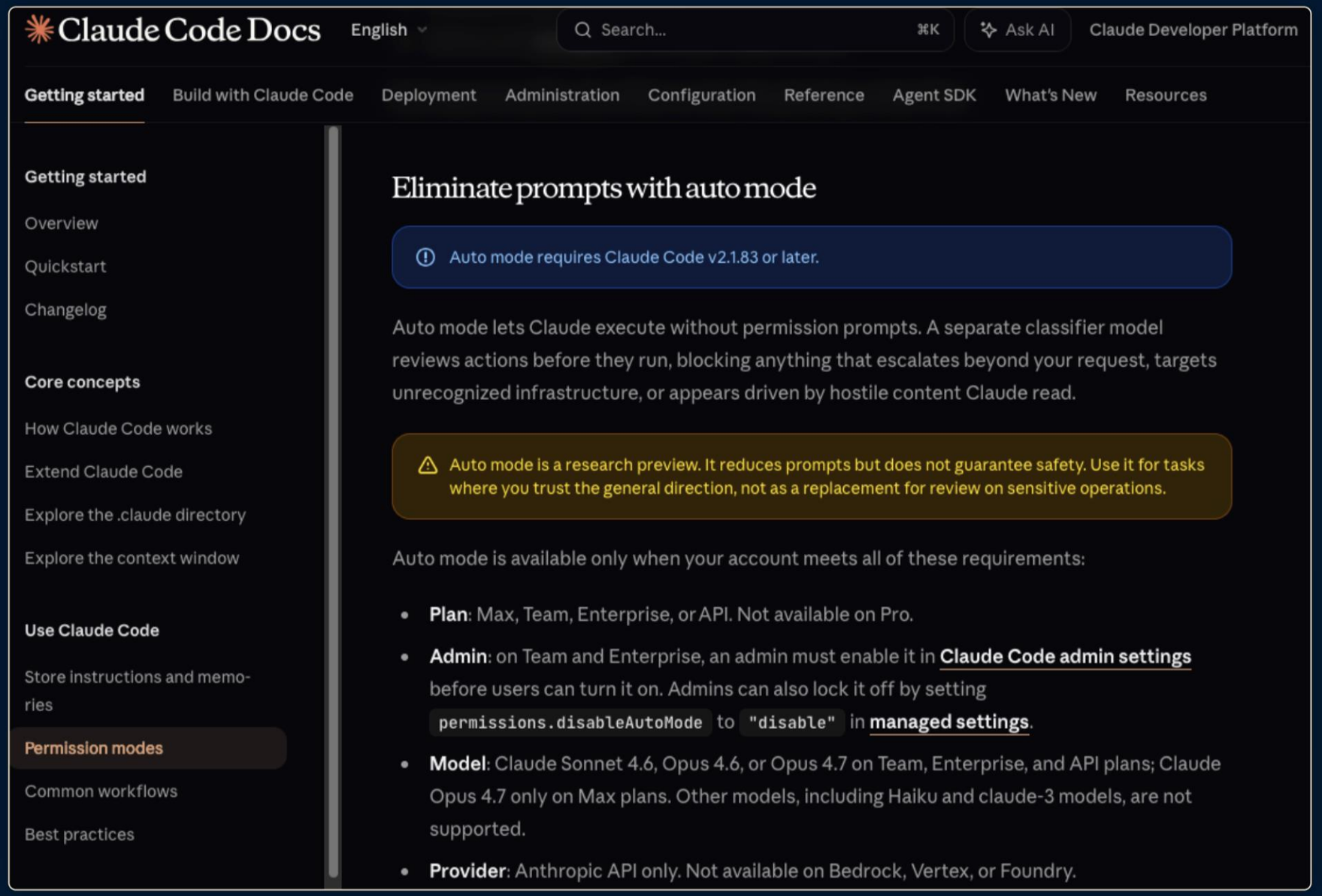
job-boards.greenhouse.io/anthropic/jobs/5023394008

PERSONAL WEBSITE — DALLAS-ELLEMAN.COM



Live demo (interactive resume)

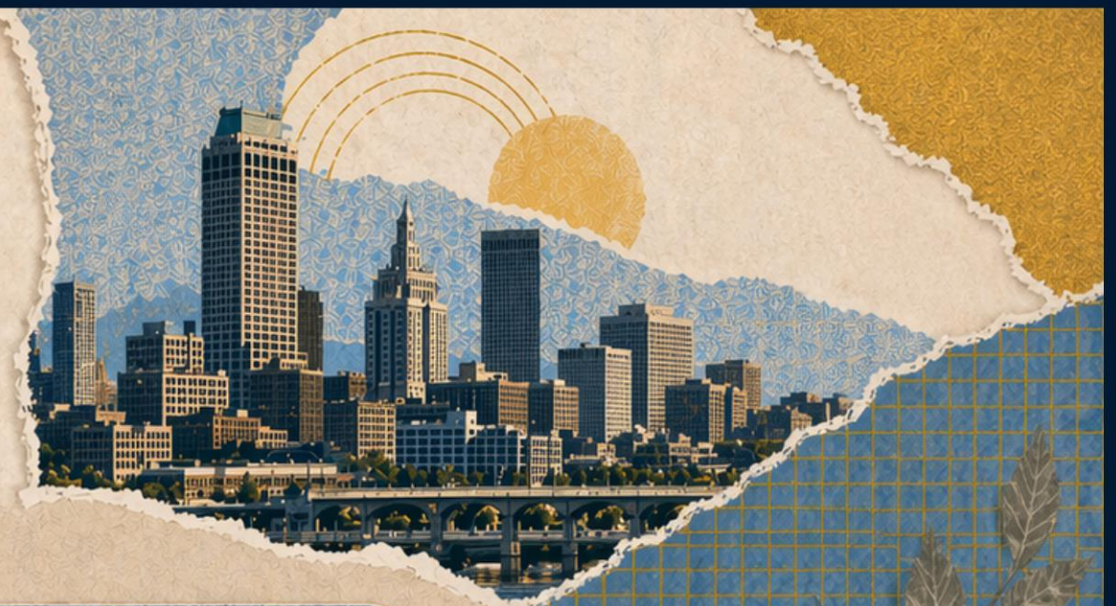
CLAUDE CODE — AUTO MODE



code.claude.com/docs/en/permission-modes#eliminate-prompts-with-auto-mode

HURRICANE HACKATHON

2030 AND BEYOND



Supercharge your finals week with a 24-hour team challenge!
$1000 in prizes

Meals, snacks, workshops, & brain fuel provided

- 10am **Saturday 5/2** → 10am **Sunday 5/3**
- **Keplinger Hall L100** (lower atrium)
- **First 20 RSVPs get $10** in Claude / Codex credits for their team

RSVP HERE - TU STUDENTS ONLY

Sponsored by:

THE UNIVERSITY OF TULSA
College of Engineering and Computer Science

RISK MANAGEMENT & CRISIS RESPONSE

Unit 11 — survey only

UNIT 11 — TOPICS

- **11.1 NIST AI Risk Management Framework** — structure, implementation, and practical application.
- **11.2 Organizational governance** — risk assessment, ethics boards, and accountability structures.
- **11.3 Incident response and crisis management** — preparation, escalation, and recovery protocols.

INDEPENDENT AUDITING, DOCUMENTATION, & DISCLOSURE

Unit 12 — survey only

UNIT 12 — TOPICS

- **12.1 Documentation standards** — model cards, datasheets, and transparency requirements.
- **12.2 Preparing for external audits** and regulatory review.
- **12.3 Independent evaluation** — third-party testing, certification, and validation processes.
- **12.4 Stakeholder engagement** — disclosure practices, communication, and accountability mechanisms.

INDUSTRY APPLICATIONS & EMERGING CHALLENGES

Unit 13 — today's focus

13.1 — SECTOR-SPECIFIC SECURITY & TRUST

Four sectors where AI security and trust requirements are the most differentiated — each gets a single best-reference URL you can use as a launching pad.

Healthcare

FDA, HIPAA, clinical AI

Finance

SR 11-7, fair lending, fraud

Defense

DoD RAI, dual-use, ethics

Critical Infrastructure

CISA, EO 14110, ICS/OT

13.1A — HEALTHCARE (U.S.)

SaMD framework + Good Machine Learning Practice

FDA's **SaMD** classification + the **10 GMLP principles** set baseline safety, validation, and clinical-evaluation expectations for any AI/ML medical device.

Predetermined Change Control Plan (PCCP)

Final FDA guidance (December 2024) governs *post-market model updates and drift monitoring* — the regulatory answer to "the model isn't static after deployment."

Transparency + lifecycle expectations

FDA Transparency Guiding Principles (June 2024) + AI-enabled device lifecycle draft (January 2025) address bias, hallucination risk, and clinician-facing labeling.

SINGLE BEST REFERENCE

FDA / CDRH — *Artificial Intelligence in Software as a Medical Device* (canonical hub linking GMLP, SaMD, PCCP, transparency, and the running AI/ML-enabled device list): fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device

13.1B — FINANCE

Existing model-risk frameworks apply — with gaps

SR 11-7 model risk and third-party-risk guidance carry over to AI; gaps remain for credit unions and generative AI.

AI is dual-use in finance

Amplifies cybersecurity, data-privacy, fair-lending bias, and synthetic-identity fraud risks — while strengthening fraud detection.

Trustworthy deployment essentials

Explainability, bias testing, governance, and continuous monitoring — coordinated across OCC, FDIC, Fed, SEC, CFPB, NCUA.

SINGLE BEST REFERENCE

GAO — *Artificial Intelligence: Use and Oversight in Financial Services* (May 2025, GAO-25-107197): files.gao.gov/reports/GAO-25-107197/index.html

13.1C — DEFENSE

AI-first warfighting posture

Establishes seven **Pace-Setting Projects** with mandated cross-Service data access, compute, and talent provisioning — AI moves from "responsible enabler" to capability frontier.

CDAO benchmarks become acquisition criteria

CDAO must publish *model-objectivity and trust benchmarks* as primary procurement criteria — vendors compete on measured trust, not stated principles.

Subordinate documents still operate

2024 RAI Strategy & Implementation Pathway and DoDD 3000.09 (2023) remain in force as implementing documents under the new top-level strategy.

SINGLE BEST REFERENCE (CURRENT)

DoW — *Artificial Intelligence Strategy for the Department of War* (January 9, 2026; supersedes the 2022 DoD RAI Strategy as the top-level controlling AI strategy): media.defense.gov/...AI-STRATEGY-FOR-THE-DEPARTMENT-OF-WAR.PDF

13.1D — CRITICAL INFRASTRUCTURE

Four secure-AI principles for OT

Awareness of AI use · **threat-informed design** · **secure-by-design** AI · **secure AI lifecycle management**.

OT-focused, all 16 sectors

Targets operators across the 16 U.S. critical-infrastructure sectors — co-authored with allied cyber agencies (ACSC, CCCS, NCSC-UK, others).

Operationalizes current policy

Implements the July 2025 AI Action Plan's CISA mandate. Effectively supersedes the Biden-era April 2024 DHS guidance (which is still hosted but orphaned by EO 14179).

SINGLE BEST REFERENCE (CURRENT)

CISA + ACSC + international partners — *Principles for the Secure Integration of Artificial Intelligence in Operational Technology* (December 3, 2025): cisa.gov/resources-tools/resources/principles-secure-integration-artificial-intelligence-operational-technology

U.S. AI POLICY — BIDEN → TRUMP (IN 15 MONTHS)

Several of the references on the prior slides come from two different administrations. Here's the delta.

Dimension	Biden (2021–Jan 2025)	Trump (Jan 2025–present)
Overarching EO	EO 14110 (Oct 2023) — risk + safety + civil rights	EO 14179 (Jan 2025) — "Removing Barriers"; revokes 14110. AI Action Plan (Jul 2025).
Federal AI use	OMB M-24-10 + M-24-18	OMB M-25-21 + M-25-22 (Apr 2025); CAIO + risk-mgmt skeleton retained, framing toward speed/innovation
Frontier model reporting	DPA-based mandatory reporting + pre-deployment safety-test sharing	Mandatory reporting paused; voluntary CAISI engagement; red-teaming reframed as ideological-bias check
"AI Bill of Rights"	OSTP Blueprint (Oct 2022), cited across agencies	Framing dropped; document moved to Biden archive
AI Safety Institute	US AISI at NIST (Nov 2023); MOUs with frontier labs	Renamed CAISI (Jun 2025) — standards/competitiveness framing replaces "safety"
State preemption	Implicit deference; state experimentation encouraged	Active preemption: Dec 2025 EO + DOJ AI Litigation Task Force (legal force contested)
Critical infra AI	DHS/CISA April 2024 guidelines under EO 14110	CISA Dec 2025 OT principles (joint w/ allies); 2024 doc orphaned but still hosted

What stayed: NIST AI RMF + GenAI Profile, the institute itself (rebranded), sector regulators, and a rapidly growing body of state AI laws.

SINGLE BEST REFERENCE

Stanford HAI — *2025 AI Index, Chapter 6: Policy and Governance* (balanced, footnoted, covers both administrations + state activity): hai.stanford.edu/...chapter6_final.pdf

13.2 — LIVE POLICY DEBATES

Three debates that will shape the regulatory environment your career will live in.

Open Model Release

Transparency vs. proliferation

Foundation Model Regulation

Compute thresholds, evals, audits

Governance Evolution

Voluntary → binding, national → international

13.2A — OPEN MODEL RELEASE

Benefits of open weights

Research access, competition, privacy — cited and substantiated.

Marginal misuse risks

CBRN, cyber, CSAM, disinformation — framed as *marginal* risk over the closed-model baseline, not absolute.

Recommended posture

Active monitoring + evidence collection rather than preemptive restriction; preserve flexibility for future regulatory pivots.

SINGLE BEST REFERENCE

NTIA — *Dual-Use Foundation Models with Widely Available Model Weights* (July 2024, response to EO 14110): ntia.gov/programs-and-initiatives/artificial-intelligence/open-model-weights-report

13.2B — FOUNDATION MODEL REGULATION

An IPCC-style consensus document for AI

Multilateral: 30 governments + UN, EU, OECD; chaired by Yoshua Bengio; ~100 nominated experts.

Surveys the regulatory toolkit

Capability evaluations, red-teaming, transparency mandates, third-party audits — and the limits of compute thresholds as a policy lever.

Frames the 2025–2026 trajectory

EU AI Act GPAI tier · the AISI network (UK, US, plus follow-ons) · emerging frontier safety frameworks.

SINGLE BEST REFERENCE

International AI Safety Report 2025 (UK AISI host, chair: Bengio; ~300 pages, ongoing 2026 updates): internationalaisafetyreport.org/publication/international-ai-safety-report-2025

13.2C — AI GOVERNANCE EVOLUTION

Voluntary → binding

NIST AI RMF and OECD AI Principles giving way to enforceable EU AI Act, US executive orders, and state laws.

National → multilateral

AI Safety Institutes proliferating since Bletchley 2023; international AISI network formalized at Seoul 2024 and Paris 2025.

Velocity

Legislative AI mentions up **21.3%** across 75 countries in 2024; US state AI laws went from **49 to 131**.

SINGLE BEST REFERENCE

Stanford HAI — *AI Index 2025, Chapter 6: Policy and Governance*: hai.stanford.edu/ai-index/2025-ai-index-report/policy-and-governance

13.3 CURRENT LANDSCAPE & 13.4 EMERGING TECHNOLOGIES

Top student-vote topics — deeper treatment

13.3 — WHERE THE ATTACKS ACTUALLY LIVE (2025–2026)

Real production incidents have stopped looking like academic adversarial-example papers. The frontier moved.

Indirect prompt injection

EchoLeak (CVE-2025-32711) was the first zero-click LLM exploit at production scale — emails → Copilot → data exfil. The pattern repeats: attacker text reaches the context window through any retrieval channel.

Tool-call abuse / excessive agency

Agents wired into email, calendar, repos, and browsers gain real-world leverage. 2025–2026 incidents centered on agent tool-misuse, not on "the model said something bad."

Supply-chain attacks on models

Pickle-deserialization RCE in HF model files (PyTorch / TensorFlow loaders); typosquatted model names; poisoned LoRA adapters distributed via community hubs.

RAG-corpus poisoning

Wiki, Drive, ticket-system content treated as authoritative once retrieved — the trust boundary is set at indexing time, not query time. (See Pres 20.)

REFERENCE

OWASP — *Top 10 for LLM Applications & Generative AI*: genai.owasp.org/llm-top-10

AI Incident Database (curated catalog of real-world incidents): incidentdatabase.ai

13.3 — STATE OF THE DEFENDERS

AI Safety Institutes

UK AISI, US AISI, Japan AISI, Singapore AI Safety Centre, EU AI Office — doing pre-deployment evals on frontier models. Network formalized at Seoul 2024, Paris 2025.

Lab-internal red teams

Anthropic, OpenAI, Google DeepMind, Meta — structured red-team programs publishing capability/safety reports per release. *Frontier safety frameworks* now standard.

Eval ecosystem

METR, Apollo Research, Pattern Labs, MLCommons AILuminate, plus academic shops (CHAI, MILA, FAR.AI). Evals are professionalizing — with all the methodology debates that implies.

Defensive frameworks landing

NIST AI RMF + GenAI Profile (AI 600-1), MITRE ATLAS, Meta's Agents Rule of Two, OWASP LLM Top 10. Different layers, complementary, none sufficient alone.

REFERENCES

UK AI Safety Institute: aisi.gov.uk · US AISI: nist.gov/aisi

MITRE ATLAS — AI threat matrix: atlas.mitre.org

METR — frontier model evaluation: metr.org

13.4 — EMERGING TECHNOLOGIES, IN THE SECURITY LENS

Three frontiers worth watching — each will reshape the attack surface in the next 12–24 months.

Agentic Systems

Tool use, MCP, computer use, multi-agent

Mechanistic Interpretability

Sparse autoencoders, circuits, probes

AI for AI Safety

Constitutional AI, debate, scalable oversight

13.4A — AGENTIC SYSTEMS

Model Context Protocol (MCP)

Anthropic's open standard for tool/server integration with LLM clients. Now the de-facto agent ↔ tool wiring layer (Claude, Cursor, Continue, ChatGPT, others).

Computer use / browser agents

Claude Computer Use, OpenAI ChatGPT Agent (Operator merged in Aug 2025), Google Gemini Agent (Project Mariner winds down May 4, 2026). Agents that drive a real browser or desktop — full web/app capability + every web/app vulnerability.

Multi-agent orchestration

Long-running agent crews (engineering, research, ops). New problems: cross-agent prompt injection, principal/delegate identity, audit trails across agent boundaries.

Reasoning + extended thinking

GPT-5.5 Thinking (o-series folded into GPT-5 line), Claude Opus 4.7 adaptive thinking, DeepSeek V4-Pro (R1 superseded; V4 collapses chat + reasoner). Models that *plan* before acting — longer attack chains.

REFERENCE (APRIL 2026)

Stanford HAI — *2026 AI Index Report* (released April 13, 2026; foregrounds agentic systems, security barriers, OSWorld / SWE-Bench / Cybench): hai.stanford.edu/ai-index/2026-ai-index-report

13.4B — MECHANISTIC INTERPRETABILITY

From "black box" to "we can name some circuits"

Sparse autoencoders identify human-interpretable features inside frontier models. Anthropic's *Scaling Monosemanticity* mapped millions of features in Claude 3 Sonnet; later work extends this to Claude 3.5+ and Llama-class models.

Why it matters for security

If we can detect deception features, sycophancy features, or jailbreak-trigger features *at activation time*, we get a defense layer that doesn't depend on prompt-text classification.

Where it's still pre-paradigmatic

Feature-finding works; *causally* intervening to suppress unsafe behavior at scale, in deployed systems, is research-grade. Don't ship interpretability-based safety as your only line.

REFERENCES

Anthropic — *Scaling Monosemanticity*: transformer-circuits.pub/2024/scaling-monosemanticity

Neel Nanda — *Mechanistic Interpretability Quickstart*: neelnanda.io/mechanistic-interpretability/quickstart

13.4C — AI FOR AI SAFETY

Scalable oversight

Use AI systems to help humans evaluate other AI systems on tasks humans cannot reliably grade alone. Constitutional AI, RLAIF, debate, recursive reward modeling.

Automated red-teaming

Models that generate jailbreaks, edge cases, and adversarial inputs against other models. Currently used in production at every frontier lab.

Verifiable / formal-method assists

LLMs as theorem-proving assistants (Lean, Coq), policy-as-code generators, and formal-spec drafters — bringing program-verification rigor into AI safety.

REFERENCES

Anthropic — *Constitutional AI*: anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback

OpenAI — *Weak-to-strong generalization*: openai.com/index/weak-to-strong-generalization

WHAT'S NEXT

WEDNESDAY, APRIL 29 — WEEK 14, SESSION 2

Unit 14 — **Career pathways**, professional development.

MONDAY, MAY 4 — FINAL CLASS SESSION

Synthesis & wrap-up + final exam review.

FINAL PROJECT

Due date **TBD** — will be announced once finalized.

FINAL EXAM TIME

I'll send out a **survey** — we get to choose. Watch your email.