ETHAN F. COOLEY

AI & Machine Learning Engineer | Charlotte, NC | linkedin.com/in/ethancooley | ethancooley.com | github.com/ethancooley

SUMMARY

Healthcare AI engineer with 6 years shipping production AI tools that live inside clinicians' daily work. Deep domain expertise in EHR data and HCC risk adjustment, the CMS model behind Medicare Advantage and PACE payment. Owns tools end to end, from data layer and integrations to agent and prompt design, structured evaluation, interface, monitoring, and documentation. Builds 0-to-1 prototypes personally, works directly with clinical and product teams, and holds to HIPAA standards. Currently building agentic AI on AWS Bedrock; ships daily with Claude Code and GitHub Copilot. M.Eng. in AI for Product Innovation, Duke University (May 2027).

PROFESSIONAL EXPERIENCE

Stanson Health · a division of Premier Inc. *Charlotte, NC*

Senior Machine Learning Engineer *July 2025 – Present*

- Building an agentic AI system on **AWS Bedrock** that automates guideline chart review nurses and reviewers do by hand, cutting review from hours to minutes for hundreds to thousands of patients a day: multi-step agents across Amazon Nova, Claude Haiku and Sonnet, and Qwen read patient records and determine whether each patient meets clinical guideline criteria.
- Lead production AI that reviews EHR documentation for HCC risk adjustment, flagging diagnoses the record does not support and surfacing undocumented conditions so care teams capture the full picture of patient complexity.
- Embed with the clinical product team to study how expert reviewers annotate HCC cases, then translate their decision patterns into model improvements so the AI decides the way the people who use it do.
- Run structured evaluations: architected an automated test suite that checks every LLM update against expected clinical behavior and catches failure modes before they reach users.
- Own the data layer the team's AI tools are built on: technical lead for the data engine on **AWS OpenSearch**, from ingestion and curation to retrieval.
- Build the integrations that put AI in front of clinical users: REST APIs linking front-end applications to models on **AWS SageMaker**.
- Cut LLM inference latency and cost with kernel caching and batched sentence embeddings; ship within HIPAA standards.

Machine Learning Engineer *Sept 2023 – July 2025*

- Observed clinicians and reviewers at work, then designed and built a tool around how they review and label cases, raising the quality and consistency of the data behind production models.
- Monitored production tools with **AWS CloudWatch**, catching and fixing issues before users noticed.
- Established technical foundations: team-wide code standards, architecture and API docs, and onboarding for new engineers.

Data Scientist *June 2021 – Sept 2023*

- Redesigned the clinical review workflow with active learning, routing the most uncertain cases to human reviewers.
- Automated repetitive annotation and case-preparation work with DAG-based pipelines, cutting manual effort.

Data Science Intern *Jan 2021 – May 2021*

- Iterated on production models with senior data scientists; built an app to compare predictions across model versions.

0-TO-1 PRODUCTS

IDT Prep Copilot · Agentic AI for PACE care teams, built solo (idt.ethancooley.com) *2026*

- Built an agentic copilot that preps PACE interdisciplinary team (IDT) meetings from synthetic charts: a **LangGraph** agent plans, gathers evidence through chart tools, drafts, verifies, and finalizes each participant brief, with every line citing its evidence.
- Owned it end to end: Python agent on Claude Haiku and Sonnet, Postgres with pgvector, a React dashboard that streams the agent's steps live, and AWS Lambda, S3 and CloudFront deployed with SAM through GitHub Actions CI/CD.
- Added answer-key evaluation, run-trace replay, and an admin console that sets per-user run and spending limits.

Groundmark · AI product on the Chrome Web Store, built solo (groundmark.dev) *2026*

- Took an AI tool from idea to public launch with 500+ installs and a 5-star rating, owning architecture, prompt design, evaluation, and interface: a Chrome extension that checks AI claims against their sources, on-device with **Gemini Nano**.

PlainMed · Patient-Facing LLM *2026*

- LoRA fine-tuned Qwen2.5-1.5B to rewrite clinical text in plain language for patients; shipped on Hugging Face.

TECHNICAL SKILLS

AI Tooling: Agentic AI on AWS Bedrock (Nova, Claude Haiku and Sonnet, Qwen), LangGraph, MCP, tool use, prompt design, LLM evaluation, RAG, embeddings, copilots and automations, LoRA fine-tuning, Claude Code, GitHub Copilot
Full-Stack Engineering: Python, SQL, JavaScript, React, Node.js, FastAPI, Flask, REST APIs, OAuth 2.0, Postgres and pgvector, Docker, Git, GitHub Actions; AWS Lambda, SageMaker, OpenSearch, S3, CloudFront, CloudWatch, SAM; Google Cloud
Healthcare: EHR data, HCC and CMS risk adjustment, ICD-10, PACE, clinical documentation, clinical decision support, HIPAA
Product & Delivery: 0-to-1 prototyping, discovery with clinical users, workflow automation, success metrics, user docs

EDUCATION

Duke University · M.Eng., Artificial Intelligence for Product Innovation *Expected May 2027*
Wake Forest University · B.S. in Engineering, Cum Laude *May 2021*

- Team manager, best senior capstone: CNN classifying cancer histopathology at 85% accuracy vs. 52% for pathologists.

Beyond Work: Ultramarathon running (48th, Run Rabbit Run 50-miler) · Triathlon · Backcountry skiing (AIARE 1) · Guitar (10+ yrs)