Cole McIntosh

AI and Data Engineer

Email | LinkedIn | GitHub | Website | Hugging Face

Foreword

Innovative AI and Data Engineer specializing in developing cutting-edge AI systems and data solutions. With deep expertise in **Generative AI**, **Large Language Models (LLMs)** and **AI Agents**, I create sophisticated, scalable systems that leverage the power of generative AI for intelligent, autonomous decision-making and problem-solving. By combining technical prowess with a strategic mindset, I deliver AI and data solutions that address complex challenges and drive future innovations.

Skills

- **AI / Machine Learning:**
  - **LLM Expertise:** Fine-tuning (LoRA, QLoRA, RoPE Scaling), RAG Architecture & Optimization (incl. Reranking), Prompt Engineering, LLM Evaluation, Agentic Systems Design, Vector Databases (FAISS), Multimodal AI, Chain-of-Thought Reasoning, Structured Output Generation
  - **Frameworks/Libraries:** Langchain, Huggingface (Transformers, TRL, Datasets), PyTorch, Unsloth, PyCaret (AutoML), Cohere Rerank, Groq API
  - **Concepts:** Natural Language Processing (NLP), Information Retrieval, Knowledge Systems, Predictive Modeling, Generative AI
- **Data Engineering & Cloud:**
  - **Platforms:** AWS, Google Cloud, Vercel, Databricks, Azure
  - **Data Tools:** Data Pipelines (Python, SQL), ETL Solutions, Business Intelligence Reporting, Data Modeling (SQL)
  - **Integrations:** Salesforce CRM, Slack Automations, Jira
- **Programming Languages:** Python, SQL, Typescript
- **Web & Tools:** NextJS, React, Streamlit, Git

Experience

Pay Ready - AI/Data Engineer (*2023 - Present*)

- **Developed** agentic AI systems, fully automating complex business processes, reducing associated manual effort by **60%**.
- **Engineered** scalable, AI-driven knowledge systems, enhancing company-wide information access for **60+** users and improving AI assistant evaluation accuracy by **35%**.
- **Designed and implemented** end-to-end generative AI solutions using LLMs to automate customer support ticket handling and report generation, enabling **2x faster** delivery of actionable insights.
- **Implemented** multimodal AI systems analyzing text and document images concurrently, improving decision-making accuracy in document verification by **15%**.
- **Created** a retrieval augmented generative AI Slack bot querying internal documentation, reducing average information retrieval time by **75%** and boosting productivity for **50+** team members across multiple departments.
- **Built** a full suite of in-house AI/automation tools with SSO, evolving into a client portal servicing clients.

Otter - Scaled Tooling Analyst (*2022 - 2023*)

- **Built** scalable Python data pipelines processing **~500 GB** of data weekly for cross-team analytics, incorporating generative AI based scoring that empowered the account management team to upsell and manage clients proactively and more successfully.
- **Implemented** internal generative AI tools automating support ticket categorization and routing, enhancing operational efficiency and saving an estimated **10 hours/week** of manual effort.
- **Developed** a generative AI, retrieval augmented (RAG) chatbot integrating Jira team spaces, enhancing cross-team project visibility and reducing cross-functional query resolution time by **30%**.

RealPage - Business Analyst (*2021 - 2022*)

- **Designed and maintained** comprehensive business intelligence reporting systems using Tableau, supporting **75+** users across **5** departments.
- **Developed** complex SQL data models enabling advanced customer segmentation analytics and automated reporting for **50+** key clients.
- **Created** automation pipelines using SQL and Python scripts, streamlining month-end reporting processes and reducing associated manual workload by **25%**.

Projects

(This section highlights key personal projects and significant open-source contributions demonstrating core competencies)

Chain of Thought Reranking - *Advanced LLM Reasoning Optimization*

- Developed a novel technique and system to optimize LLM responses by generating detailed reasoning (CoT), segmenting the chain, reranking segments for relevance (using Cohere Rerank), and generating a refined answer, demonstrably improving accuracy on complex tasks.
- **Blog:** www.colemcintosh.io/blog/chain-of-thought-reranking
- **Code:** https://github.com/colesmcintosh/chain-of-thought-reranking

Entropy-Based CoT Injection - *Dynamic LLM Prompting*

- Created an adaptive approach to enhance LLM reasoning by monitoring model uncertainty (entropy) during generation and dynamically injecting Chain-of-Thought prompts only when a critical uncertainty threshold is reached.
- **Blog:** www.colemcintosh.io/blog/entropy-based-chain-of-thought-injection
- **Code:** https://github.com/colesmcintosh/entropy-injection-cot

LangChain Salesforce - *Enterprise LLM-CRM Integration*

- Authored and maintain the `langchain-salesforce` package, a robust LangChain integration enabling seamless, secure interaction (CRUD, SOQL, Schema) with Salesforce CRM data within LLM applications and agents.
- **Blog:** www.colemcintosh.io/blog/langchain-salesforce
- **Code:** https://github.com/colesmcintosh/langchain-salesforce
- **Docs:** python.langchain.com/docs/integrations/tools/salesforce

Llama Kernel - *Memory-Efficient PyTorch Inference*

- Developed a memory-efficient PyTorch kernel for Llama 3.2 inference, implementing 4-bit quantization and advanced optimization techniques to reduce resource consumption.
- **Code:** https://github.com/colesmcintosh/llama-kernel

AWS Knowledge Base RAG - *Bedrock RAG Implementation*

- Developed a practical guide and implementation demonstrating how to query an AWS Bedrock Knowledge Base using the RetrieveAndGenerate API for effective RAG applications.
- **Code:** https://github.com/colesmcintosh/aws-knowledge-base-rag

NumPy MCP Server - *LLM Numerical Computation via MCP*

- Built a Model Context Protocol (MCP) server using FastMCP, enabling LLMs (like Claude Desktop) to perform complex numerical computations (linear algebra, stats) by interfacing directly with NumPy.
- **Blog:** www.colemcintosh.io/blog/numpy-mcp-server
- **Code:** https://github.com/colesmcintosh/numpy-mcp

cursor.solutions - *Open Source AI Documentation Site*

- Developed and maintain the open-source documentation and resource website for Cursor AI, built with NextJS/React.
- **Link:** https://cursor.solutions/

Llama 3.2 1B Mango - *Optimized CoT Fine-Tune*

- Fine-tuned a Llama 3.2 1B model optimized for chain-of-thought reasoning, achieving 2x faster training using Unsloth, TRL, LoRA/QLoRA/RoPE. Trained on SkunkworksAI/reasoning-0.01 (29.9k examples).
- **Link:** https://huggingface.co/colesmcintosh/Llama-3.2-1B-Instruct-Mango

Structured Output with Multimodal Agents - *Advanced Task Routing*

- Developed a system demonstrating structured output generation via multimodal agents, showcasing advanced AI integration and intelligent task routing capabilities based on diverse inputs.
- **Code:** https://github.com/colesmcintosh/structured-output-with-multimodal-agents

Smol Vision - *Resource-Efficient Local Vision Pipeline*

- Built a local LLM vision pipeline for efficient image analysis *without* dedicated GPU requirements, demonstrating expertise in optimizing AI models for resource-constrained environments.
- **Code:** https://github.com/colesmcintosh/smol-vision

Reranked RAG Demo - *Personal Website Q&A App*

- Developed a Streamlit application showcasing RAG with reranking (Cohere). Answers questions about my work using scraped website content, LangChain, Groq's LLM, and FAISS.
- **Link:** https://cole-mcintosh-rag.streamlit.app/

Chain of Thought Visualization - *Real-time Reasoning Demo*

- Developed a front-end project (NextJS/Vercel) demonstrating CoT reasoning with structured outputs using Mistral's Ministral 3B. Streams responses to visualize the reasoning process, showcasing efficient use of smaller models.
- **Link:** https://cot-with-structured-output.vercel.app/

Finance Fine Tuned Mistral 7B - *Domain-Specific Financial LLM*

- Fine-tuned a Mistral 7B model tailored for finance tasks using the alpaca finance dataset, enhancing capabilities in financial analysis, prediction, and domain-specific understanding.
- **Link:** https://huggingface.co/colesmcintosh/mistral_7b_finance_finetuned

(Note: Active contributor to open-source projects including LangChain, Block's Goose, Huggingface Smol Agents, and LiteLLM - see GitHub for details)

Education

- **AWS Cloud Technical Essentials** - Amazon
- **Google Cloud Fundamentals: Core Infrastructure** - Google
- **Python Specialization** - University of Michigan
- **SQL for Data Science** - University of California, Davis

Afterword

Thank you for reviewing my resume. I am passionate about the possibilities of AI and data technology and eager to connect with like-minded individuals and innovative teams. Please feel free to reach out!