Improving RAG Applications

For MIT Generative AI Course

Jason Liu

February 25, 2025

Independent Consultant

Table of Contents

About Me

What are we doing here?

Retrieval & Recommendation Systems

Effective Use of AI in Business

Individual Contributor Career Paths

About Me

About Me

Goal: To showcase my diverse experience – feel free to ask questions about any area during Q&A

- Independent Consultant & Staff-level ML Engineer & Educator
  - Meta, Stitchfix, NYU from (2016-2023)
- University of Waterloo
  - B.Math in Mathematical Physics & Computational Mathematics
  - Minor in Statistics (Class of 2017)
- Creator of Instructor - Python library for structured LLM outputs
  - 9500+ GitHub stars
  - 1.5M+ monthly downloads
  - Cited by OpenAI as inspiration for structured output feature
  - Popular 'Pydantic is all you need' Talk from AI Conference
- a16z Scout & Angel Investor & Startup Advisor

My Journey

- Transitioned to independent consulting due to RSI (2022)
- Do I want to get 20% of my coding back, or find 100x more leverage?
- Now focused on:
    - Teaching teams to work with AI and be quantitative
    - Writing more 'popular ai' content
    - Advisory work for early stage startups
    - Open source projects, independent research

Work & Consulting & Advisory Engagements

Client	Contact	Industry
Zapier	VP of Product	Automation
HubSpot	GM	Sales & Marketing
Enterpret	CTO	Analytics
Tensorlake	CEO	Data
Limitless AI	CTO	AI
Trunk Tools	VP Eng	Construction
Naro	CTO	Sales & Marketing

Additionally, I've worked with innovative startups including New Computer, Sandbar, Dunbar, Bytebot, Kay.ai, Raycast, Weights & Biases, Modal Labs, Timescale, and Pydantic on various technical and strategic initiatives.

Where My Students Come From

Company	Industry
OpenAI	AI Research & Development
Anthropic	AI Research & Development
Google	Search Engine
Salesforce	Customer Relationship Management Software
Microsoft	Software, Cloud Computing
Amazon	E-commerce, Cloud Computing
Zapier	Automation Software
Adobe	Software, Creative Tools
Accenture	Consulting, Technology Services
McKinsey & Company	Management Consulting
Bain & Company	Consulting
PwC	Professional Services
Cisco	Networking Technology
Electronic Arts	Gaming
Shopify	E-commerce Platform

What are we doing here?

Course Context & Goals

- Learning Objectives
  - Developing durable AI knowledge that outlasts specific technical implementations
  - Understanding ML systems as continuously evolving products rather than ”deploy once and forget”
  - Recognizing the parallels between recommendation systems and retrieval systems
  - Identifying valuable business applications through effective data analysis
  - Mastering the key skills for building successful AI systems in the era of democratized tools

Setting the Context

- Why This Matters Now
  - Democratization of AI tools means the competitive advantage comes from thinking deeply
  - Growing gap between research capabilities and business implementation
  - Science now drives product development, reversing traditional patterns
  - Opportunity for individual contributors to have outsized impact through thoughtful implementation
- Interactive Format
  - This group is quite diverse, so I'll try to keep it broad
  - The goal is to seed you with good questions, rather than dump information
  - We'll leave plenty of time for questions about AI, Business, and Career paths

Key Questions to Consider

- How has machine learning research and implementation evolved from 2015 to 2025?
- What behavioral practices should teams adopt when working with AI systems?
- How do we identify economically valuable AI applications?
- What's the right balance between research exploration and product implementation?
- How do we design systems that can evolve effectively over time?
- When should you join established labs versus work independently?
- How can individuals and small teams achieve leverage without large resources?
- What skills matter most in the AI era? (Hint: thinking ¿ coding)

Three Key Arcs We'll Explore

- Technical: From Recommendation to Retrieval Systems
  - The surprising similarities in architecture and challenges
  - Why understanding these parallels helps build better systems
- Organizational: Effective AI Implementation
  - The importance of observability and measurement
  - Balancing unified systems vs. specialized subsystems
- Personal: Career Considerations in AI
  - Information synthesis as a durable skill
  - How AI changes team dynamics and individual contributions

Retrieval & Recommendation Systems

Evolution of Retrieval Systems (2015-2025)

- Historical Context
  - From recommendation systems to modern retrieval
  - Parallels with fraud detection & triage systems
- Key Developments
  - Semantic search improvements
  - Integration with LLMs
  - Hybrid approaches
- TODO: Add specific examples and metrics

System Architecture Evolution

- Traditional Systems
  - Collaborative filtering
  - Content-based filtering
  - Hybrid systems
- Modern Approaches
  - Dense retrievers
  - Cross-encoders
  - Hybrid search systems
- TODO: Add architecture diagrams

Effective Use of AI in Business

Common Implementation Challenges

- Current Business Struggles
  - Over-engineering solutions
  - Misalignment of resources
  - Lack of data analysis focus
- Key Insight

  "If a manager says the agent needs more complex reasoning, it means you haven't thought about the problem yourself"

Data Analysis vs Coding

- Critical Skills
  - Rapid data analysis capabilities
  - Pattern recognition in large datasets
  - Business context understanding
- Key Insight

  "Since thinking has been democratized, any thinking on your part will be above average"
- TODO: Add real-world examples

Individual Contributor Career Paths

Research & Industry Landscape

- Current State
  - Top researchers staying at large firms
  - Resource access disparities
  - Adaptation strategies for smaller businesses
- Career Decisions
  - Lab vs. Independent work tradeoffs
  - Resource accessibility considerations
  - Impact potential analysis

Modern Business Models

- Revenue Metrics
  - Revenue per employee analysis
  - Solo entrepreneurship opportunities
  - Scaling considerations
- Future Trends
  - AI-enabled business models
  - Consulting evolution
  - TODO: Add specific metrics and examples