

ZHACHORY VOLKER

hello@zhach.me | (940) 315-4580

Github: @Zhachory1

LinkedIn: zhachory1

Website: zhach.me

Staff Software Engineer with 10+ years of professional full-stack and Machine Learning experience. Proven track record of designing scalable Search and Recommendation stacks, optimizing robust data infrastructure, and drastically reducing compute overhead. Specialized in building high-impact ML ranking architectures at massive scale and leveraging modern AI development tools to accelerate delivery and business growth.

Experience

Staff Software Engineer - Rokr

October 2025 - Present | New York, New York

- Reduced machine learning audience generation pipeline costs by 75% (saving ~\$750k annually) by migrating to a highly stable, efficient two-tower infrastructure and implementing a robust data refresh system.
- Architected and deployed a dynamic ranking layer for on-the-fly audience generation, optimizing for CVR and scaling revenue from generated audiences by 120%.
- Led technical architecture and hands-on engineering for the Audience and Bidding domain, delivering a 40% uplift in CVR for ML-generated audiences by building independent, scalable systems.
- Spearheaded an org-wide AI integration initiative, upskilling engineers to utilize AI tooling (Claude, Cursor, OpenCode, Codex) for coding, documentation, system management, and debugging, improving overall Developer Satisfaction by 25%.

Senior Software Engineer - YouTube Search

March 2018 - September 2025 | New York, New York

- Led a resource optimization project achieving a 60% reduction in compute usage through efficient metric analysis and ranking model refactoring (estimated annual savings: \$1.8M).
- Designed and implemented a unified ranking framework to evaluate diverse content types (shelves, playlists, channels, articles) within the same scoring space on the YouTube Search results page, increasing overall Search Active Users (SAUs) by 2% and increasing CTR by an average of 25% for multiple features in a single launch.
- Created the YouTube Search News experience by implementing clustering algorithms, NLP models, and scalable infrastructure, resulting in a 38% increase in News Click-Through Rate (CTR) and a 10% rise (~8 Million queries) in News Searches per day.
- Launched a critical feature surfacing authoritative news articles for breaking news queries to combat misinformation, collaborating with Google News and improving news user satisfaction by 13%.

Software Engineering Resident - Google Research

March 2017 - March 2018 | New York, New York

- Applied NLP and Machine Learning to 10,000+ articles (RNNs, CNNs, Decision Trees) for associating user questions with relevant documents in the fact-checking corpus, which can be seen on the first page of Google current day.
- Created and maintained MapReduce pipelines (C++) for managing the size and processing of the fact-checking article corpus which was used by 6 Large News Publishers with XXM sized audiences.
- Developed a internal research tool using Java, TypeScript, and Angular to facilitate discovery of fact-checking articles still used by 100s of journalists today.
- Implemented robust evaluation functions for effective model training based on a published research paper's ("Watch Your Step: Learning Node Embeddings via Graph Attention") internal implementation.

Skills

- **Machine Learning & AI:** TensorFlow, PyTorch, Deep Learning, NLP, Transformers, RL, Genetic Algorithms, Clustering, Semi-Supervised Learning, Graph Mining, Embeddings, Ranking Systems, Fine-Tuning, RAG, Evals
- **Agentic AI:** Claude Code, Codex, Cursor, OpenCode, Coding Agents, Prompt Engineering, Loop & Context Engineering, MCP Servers & Tooling, Subagents, Skills & Workflows, Agent Evals
- **Software Development & Infrastructure:** Scalable Infrastructure Design, System Monitoring, CI/CD, Docker, Kubeflow, Buildkite, Object-Oriented Programming (OOP), Advanced Algorithms, Data Structures, Code Health
- **Soft Skills:** Leadership & Technical Mentorship, Collaboration & Teamwork, Communication, Problem-Solving, Conflict Resolution, Analytical Thinking, Adaptability, Continuous Learning

Open Source

- [Me Write Code](#) — Token-efficient terminal coding agent. 20+ providers, plan mode, subagents, MCP, and hooks.
- [Autopraxis](#) — Self-improving workflow skills for AI agents, turning messy goals into grounded briefs, reviewed plans, shipped work, and measured learning loops.
- [agent-fleet](#) — A portable council of specialist review personas for high-stakes engineering decisions.

Education

University of North Texas - Bachelors of Computer Science

August 2013 - May 2018 | GPA: 3.443