VICTOR MAY

Brooklyn, NY
Homepage: mrcabbage972.github.io · Google Scholar

SUMMARY

I'm an engineer with end-to-end experience in building ML-centric products—from rapid prototyping to production-grade systems and low-level optimization. I've led teams in both industry and the open-source community, and have worked at companies ranging from early-stage startups to Google. Recently, I've developed a deeper focus on research and am actively building my publication record.

WORK EXPERIENCE

- **Staff ML Engineer** New York City
  Google *June 2024-Now*
  - **Gemini CLI (June 2025 - Present):** Research and development on agent quality for Gemini CLI, a popular open-source SWE agent with 70k+ Github stars.
    - ∗ Achieved a 58% solve rate on the SWE-bench benchmark using the Gemini 2.5 Pro model, with a fast single shot method requiring no code execution.
    - ∗ Spearheading efforts to enable full repository understanding via neural memory modules (Titans) in collaboration with a Google Research team.
    - ∗ Developing model routing strategies for tool-calling agents in collaboration with a Google Research team.
  - **Code Modernization Agent (June 2024 - February 2025):**
    - ∗ Proposed and implemented a dynamic, tool-using autonomous LLM agent to upgrade Java projects from JDK8 to JDK17, achieving a 60% success rate.
    - ∗ Developed a full-stack solution for the agent featuring a GCP backend and a GitHub app, which was deployed to Trusted Partners via an Experimental Launch.
    - ∗ Led the development of the evaluation framework and dataset curation for benchmarking Java migration capabilities, resulting in a submission to a top-tier software engineering conference.
- **Advisory Board Member** New York City
  Ontocord.ai *January 2024-Now*
  - Advising on tech roadmap, team building and fundraising.
  - Led on the **Aurora-M** large language model continual pretraining training project and paper.
  - Leading on **Mixture Vitae**, an upcoming permissively licensed LLM pretraining dataset and paper.
- **ML Engineering Manager** New York City
  Chegg *June 2023 - May 2024*
  - Established and led an R&D team to prove the feasibility of a **Scientific Diagram-aware Q&A AI agent**.
  - Trained Vision-Language models for diagram understanding and generation, leveraging continual pretraining on science textbook data, RAG, prompt engineering, synthetic data generation, and constrained generation to build a product-ready end-to-end pipeline.
  - Responsible for the planning of quarterly goals, resources and high-level technical roadmaps.
  - People manager of 5 MLE's, responsible for career growth, performance reviews and compensation.
- **Staff ML Engineer** New York City
  Chegg *May 2022 - June 2023*

  - Initiated and led an effort to fine-tune domain-specific LLMs using Chegg's proprietary dataset of QnA pairs with detailed and structured step-by-step reasoning, building out data preparation pipelines, training infra and evaluation processes. This work served as a basis for a company-wide project on replacing OpenAI models in the **Chegg Study** product in with fine-tuned permissively-licensed models, cutting inference costs in half.
  - Built a streaming LLM routing and caching service that was adopted by 5 engineering teams.
  - As a member in Chegg's **Technical Architecture Group**, co-designed the architecture for **CheggMate**, the company's LLM-powered AI tutoring service for students and worked with a range of engineering teams to design their API contracts to implement the system design.

- **Senior ML Engineer, NLP** Tel-Aviv, New York City
  Taboola *March 2020 - April 2022*
  - End-to-end development of a system for extracting high-level concepts from multilingual short texts. The project included research and prototyping to meet the desired quality goals, building multiple data pipelines and micro-services and optimizing GPU utilization via CUDA optimization. This system and also the one described below are processing millions of texts per day in production.
  - Designed and implemented a data pipeline that orchestrates all of Taboola's NLP processing.
  - Developed a proof-of-concept of a highly accurate multilingual Named Entity Recognition model using stacked PEFT adapters for transformers and multi-task learning. Demonstrated that the NER task can be decoupled from language, enabling improved performance in long-tail languages using unlabeled data and multilingual transfer.
- **Senior ML Engineer, Recommender Systems** Tel-Aviv, New York City
  Taboola *March 2015 - March 2020*
  - Designed and implemented Taboola's first real-time personalized recommender system for ad ranking, including the MLOps infrastructure and the first models running on production traffic. The system was rolled out on all of Taboola's traffic, serving around 20B pageviews per day.
  - Built a system for driving the subscription rates of web publishers by a modeling-based personalization of a user's browsing experience, using recommenders and multi-arm bandits. The system was shipped and a dozen B2B customers were onboarded.
- **Computer Vision Algorithm Developer** Tel-Aviv
  Extreme Reality 3D *April 2012 - March 2015*
  - Developed algorithms for real-time human pose estimation and gesture recognition in video. Our technology had been licensed by the likes of Samsung and SEGA games.

PUBLICATIONS

- FreshBrew: A Benchmark for Evaluating AI Agents on Java Code Migration, Under Review (2025)
- GitChameleon: Evaluating AI Code Generation Against Python Library Version Incompatibilities, Under Review (2025)
- Aurora-M: The First Open Source Multilingual Language Model Red-teamed according to the U.S. Executive Order, Accepted to COLING 2025 Industry Track.
- An algorithm for improving Non-Local Means operators via low-rank approximation, Accepted to IEEE Transactions on Image Processing (Volume:25 , Issue: 3, 2015).

EDUCATION

- **M.Sc., Applied Mathematics, Magna Cum Laude**
  Tel-Aviv University

- **B.Sc., Computer Sciences and Mathematics**
  Bar-Ilan University

COMMUNITY INVOLVEMENT

- Contributed to developing the LAION OpenAssistant model.
- 2x Kaggle Expert, winner of 3 medals in NLP competitions.