Hanxiang Qin

(410)-258-9449 | hancsane@gmail.com | LinkedIn | GitHub | Personal Website

Research expertise in NLP, Multimodality, and efficient ML, combined with industrial experience engineering platforms serving hundreds of millions of users. Passionate about advancing state-of-the-art research and building reliable, high-performance production systems.

Education

Johns Hopkins University — Baltimore, MD
MS in Computer Science, ***GPA: 3.97 / 4.00*** — *Expected 05/2026*

Zhejiang University — Hangzhou, China
Dual Degree: B.Eng. in Computer Science and Technology & BS in Horticulture — *07/2020*

- **Academic Awards**: Outstanding Graduation Thesis

Research Experience

Multi-Vector Index Compression in Any Modality — **SIGIR 2026**
Advisors: Alexander Martin, Prof. Benjamin Van Durme — *Under Review*

- **Publication:** H. Qin, A. Martin, R. Jha, C. Zuo, R. Kriz, B. Van Durme. preprint arXiv:2602.21202. [arXiv] [Code].
- Exploited the sparse utilization of multi-vector indices to develop an efficient compression framework to make late-interaction models scalable for token-heavy multimodal content.
- Proposed a hybrid attention-similarity method, Attention-Guided Clustering (AGC), for dimensionality reduction. Introduced learnable **Universal Query Tokens** to identify critical signals as anchors, then applied **salience-based weighted aggregation** to reduce redundancy.
- Retained ∼ **97**% of the uncompressed model's performance while utilizing only a **2.4**% – **4.9**% Token Budget, preventing the representation collapse or model failure observed in Soft Pooling

Efficient Fine-tuning of Pre-trained Language Models for Commonsense Causal Reasoning — *12/2019*
Undergraduate Researcher | Advisor: Prof. Ming Cai

Achieved a state-of-the-art **88.4%** accuracy on the COPA dataset by optimizing RoBERTa for extremely efficient fine-tuning via selective unfreezing, task-specific heads, multi-task learning, and data augmentation.

Professional Experience

Pinduoduo Inc. (Branded as Temu in the US) — Shanghai, China
Software Engineer — *07/2020 – 06/2023*

- Coordinated a team to develop core features across the **primary user journey**, including product, promotion, checkout, and review flows, partnering with product and design teams to support large-scale user growth.
- Optimized application **performance** and trimmed binary package **size** through structural framework refactoring, improving code maintainability.
- Drove complex **feature distribution** and **personalization strategies**, designing and verifying rigorous **A/B tests** to ensure increased sales conversions and improved UX.
- Achieved **zero production incidents** over 3 years for the platform's highest-traffic page, delivering highly reliable functionality across iOS, Android, Web for **hundreds of millions** of DAUs.

Code Efficiency and Package Quality Analysis Platform

- Developed and coordinated with other teams to deploy an analytics infrastructure combining static code analysis and dynamic runtime tracking to automatically aggregate, process, and visualize code inefficiencies.
- Mentored peer engineers by delivering tech talks and authoring documentation, empowering the team to analyze and optimize their own codebases.
- Drove a team-wide increase in average code usage rate (**79%** → **96%**) and cut application package size in half (**178MB** → **89.5MB**) using insights from this platform.

Product Page Performance Optimization

- Decreased page load time by **34% (420ms** → **276ms)**, directly contributing to a **12%** increase in user retention.

- Resolved system bottlenecks, identified through event tracking and performance profiling, by reallocating thread workloads, implementing data model warm-up, and reprioritizing network and microservice requests.
- Leveraged cross-page caching to achieve **near-zero** Largest Contentful Paint (LCP), and implemented progressive data loading to drastically reduce Visually Complete (VC) time.

Server-Driven UI (SDUI) for Product Pages

- Designed a unified Server-Driven UI protocol enabling dynamic layout adjustments and content additions without requiring client App Store updates.
- Cut development costs and shaved **4 weeks** off iteration cycles, accelerating cross-platform A/B testing and feature rollouts to amplify impact of positive features.
- Ensured a **zero downtime** and continuous experience during this massive architectural transition on the highest-traffic page.

Review Gallery Page Architecture Refactoring

- Refactored a heavily coupled legacy codebase into a generic, modular container architecture for multiple flows.
- Integrated **VIPER** architecture and Aspect-Oriented Programming (**AOP**) paradigms to decouple business logic, largely reducing maintenance overhead amidst rapidly scaling product requirements.

Projects

Diffusion Language Model Optimization: Efficient Inference and RL Stability 11/2025

- Explored acceleration schemes targeting the inference overhead bottleneck of Diffusion Language Models (DLMs), integrating Self-Speculative Decoding (SSD) and DPad Suffix Dropout mechanisms to reduce computational complexity, reusing pruned attention contexts during the draft and verification stages.
- Investigated Reinforcement Learning stability in Diffusion LMs. Experimented with a PPO-style Actor-Critic architecture with token-level value optimization on GSM8K dataset to mitigate training instability from denoising randomness. Identified reward hacking as the core problem in this setting.

Kaggle AI Mathematical Olympiad 2 (Silver Medal, Top 1.2%) 03/2025

- Developed an efficient pipeline for solving mathematical problems under tight time and resource constraints through agent workflows, prompt engineering, hyperparameter tuning, majority voting, and quantization.
- Explored verifiable reward functions for Reinforcement Learning with GRPO, incorporating length-regularized rewards and explicit penalties for incorrect solutions to improve reasoning precision and conciseness.

Retrieval-Augmented Generation Argumentation Agent 12/2024

- Developed a RAG argumentation agent featuring a two-stage BM25 and re-ranker retrieval system to ensure highly accurate, evidence-grounded responses from the Kialo Dataset.
- Integrated context-aware LLM query expansion, prompt engineering, and turn-weighting heuristics to maintain conversational relevance and prevent topic drift.

Foundational Natural Language Processing Models 11/2024

- Developed core NLP architectures including N-gram Language Models, Earley Parsers, BiRNN-CRFs for POS tagging, and Seq2Seq translation models; designed a lightweight speech recognition model based on the CTC architecture integrating adapted Res2Net, ResidualCNN, and BiGRU layers; systematically explored the methodological evolution and architectural implementation of foundational techniques.

Development of an Epidemic Service Platform 05/2020

- Developed a full-stack application in response to COVID-19 featuring daily health check-ins, group medicine orders, and neighborhood forums.
- Built cross-platform frontend with *React Native (Expo)* and backend with *Express.js*, combining with agile methodologies to accelerate development and team collaboration.

Foundational Computer Science Projects 2018 – 2020

- **Vulnerability Exploitation & CTFs** Solved CTF challenges by reverse engineering and exploiting vulnerabilities including stack overflow, ROP, Ret2libc, format string, Set-UID, Android repackaging, Meltdown, and Specter.
- **Tiger Language Compiler** Built a Tiger programming language compiler from scratch by designing tokenizer, parser, and implementing code for LLVM IR generation.
- **MiniSQL Database Engine** Engineered a complete C++ MiniSQL database backend from scratch to process CRUD operations, with performance optimization using custom B+ tree indexing and memory buffering.

Qualifications

Languages: Python (5 yrs.), Objective-C (3 yrs.), JavaScript (3 yrs.), TypeScript (3 yrs.), C/C++ (3 yrs.), Java (2 yrs.), Swift, Shell, SQL, Verilog HDL, Ruby

Frameworks & Libraries: PyTorch, Transformers, Ray, vLLM, DeepSpeed, verl, NumPy, NLTK, Scikit-learn, SciPy, Pandas, Expo, React Native, Express.js, Node.js, Spring, LLVM, UIKit, SDWebImage, Masonry, IGListKit

Tools & Environments: Git, Slurm, LLDB, CocoaPods, MachOView, MySQL, pwntools, IDA, Jenkins, SonarQube, Kibana, Docker, OCLint, Jira, Linux

Competencies: Cross-platform, Full-stack, iOS, Agile Development, Natural Language Processing, Retrieval-Augmented Generation, Large Language Models, Multimodal Modeling, Performance Optimization, Reinforcement Learning, Code Quality Management, RESTful API, A/B Testing