Bryce Kan

925-548-3667 | brycekan@usc.edu | https://www.linkedin.com/in/brycekan/ | https://brycekan123.github.io/

Education

University of Southern California Aug. 2024 – Dec. 2026
Masters of Science in Computer Science
Coursework: Algorithms, Machine Learning, Natural Language Processing, Trustworthy LLMs, Hardware for ML

University of Washington Aug. 2018 – Dec. 2021
Bachelor of Science in Molecular and Cell Biology

Technical Skills

Languages: Python, C++/C, CUDA, Java, SQL
Frameworks/Libraries: PyTorch, sentence-transformers, Scikit-learn, Langgraph, Pandas, Matplotlib, Numpy,
Developer Tools: Git, Linux, Jupyter, VS Code, SLURM, SSH, Conda

Experience

SWE - Agentic AI Master's Intern May. 2026 – August. 2026
Honeywell *Atlanta, GA*

- Built **agentic evaluation judge framework** across task completion, context retention, and error recovery scenarios
- Routed LLM inference across Anthropic and OpenAI via **OpenAI-compatible layer**; built FastAPI-backed data visualizer
- Dockerized and deployed eval pipeline; integrated **Databricks** model-serving endpoints, tracking per-call cost and latency

Graduate Research Assistant Mar. 2025 – Present
USC Machine Learning Center *Los Angeles, CA*
RaMem: Contextual Reinstatement for Long-Term Agentic Memory

- Built **RAG-based agentic memory pipeline** enabling LLM agents to recall evidence across multi-session conversations.
- Engineered **temporal evidence filtering** with session metadata to remove semantically related but invalid memories.
- Implemented hybrid retrieval with **Reciprocal Rank Fusion (RRF)** over semantic search and BM25, improving F1 by **11.6 points** across GPT-4o, GPT-4.1-mini, and Qwen3-8B.
- Reduced prompt context length by **50%** while preserving performance against state-of-the-art agent memory baselines.

Student-Job Recommendation System

- Led end-to-end development as **1st author**, mentoring **2 Master's students** and collaborating with PhD candidates
- Tackled **cold-start scenarios** using synthetic explanation generation via **GPT-4o-mini distillation** into **Llama-3-8B-Instruct** for student-teacher learning
- Implemented **multi-step fine-tuning** via **dual optimization framework** with LoRA, simultaneously 2 objectives training for ranking (**InfoNCE contrastive loss**)
- Achieved **up to 2.6× (Recall@5) and over 3× (NDCG@5) improvement** over LLM baselines

Research Associate Jun. 2023 – Jul. 2024
Bruker *San Jose, CA*

- Developed custom **Python pipelines** to automate experimental data analysis, adapting to evolving scientific demands
- Designed interactive **Plotly dashboards** to visualize complex biological datasets, enabling faster data-driven insights

Process Engineer Jan. 2022 – Jun. 2023
Applied Materials *Santa Clara, CA*

- Launched Lab Inventory Database as final product, delivering full workflow traceability and reducing processing time by **80**%
- Built automated data visualization pipelines using Pandas and Plotly to extract insights from **terabyte-sized datasets**

Projects

Multimodal Cancer Prediction with Whole-Slide Histopathology | *PyTorch, OpenCV, Segmentation*

- Integrated gigapixel WSI and clinical-text features through patch-level segmentation and attention-based clustering analysis.

MCP-Powered GitHub Repository Analytics [Link] | *Python, FastMCP, SQLite, GitHub API*

- Custom MCP server with **10 tools**, storing repository data in a SQL database to generate insights while minimizing API calls.

MLP Optimization via Kernel Fusion and Vectorization [Link] | *CUDA, cuBlAS, GPU Optimization*

- Implemented cuBLAS-based 2-layer MLP with fused kernels, reducing kernel launches by **33%** and achieving **1.14× speedup.**

Publications

- Yang, W., **Kan, B.**, et al. (2026). *RaMem: Contextual Reinstatement for Long-Term Agentic Memory.* arXiv. [Link]
- **Kan, B.** et al. (2026). *De-conflating Preference and Qualification: Constrained Dual-Perspective Reasoning for Job Recommendation with Large Language Models.* arXiv. [Link]
- Seo, D.-O., Zhang, E. T., Piantadosi, S. C., Marcus, D. J., Motard, L. E., **Kan, B. K.**, et al. (2021). *A locus coeruleus to dentate gyrus noradrenergic circuit modulates aversive contextual processing.* Neuron. [Link]