Ke Wan

Sherman Oaks, CA | (408)666-2648 | kwa@microsoft.com | linkedin.com/in/kewanucsd

Publications

- **Cai, Z., Xiao, W., Sun, H., Luo, C., Zhang, Y., Wan, K., Li, Y., Zhou, Y., Chang, L.-W., Gu, J., Dong, Z., Anandkumar, A., Asi, A., Hu, J.** (2025).
  R-KV: Redundancy-aware KV Cache Compression for Reasoning Models.
  The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS), peer-reviewed
  Conference poster: https://neurips.cc/virtual/2025/loc/san-diego/poster/120110

- **He, Y., Chen, Q., Wan, K., Nabijiang, A., Cao, Y., Liu, B.** (2025).
  A Multi-Stage Machine Learning Pipeline for Automated Bowel Preparation Scale Assessment in Colonoscopy Videos.
  Proceedings of the International Conference on Machine Learning and Applications (ICMLA), peer-reviewed
  Accepted paper list: https://www.icmla-conference.org/icmla25/regularpapers.pdf
  It shows the beneficiary's paper (Submission number is 300) listed under Regular Papers, confirming acceptance after peer review.

- **Cai, Z., Qiu, H., Zhao, H., Wan, K., Li, J., Gu, J., Xiao, W., Peng, N., Hu, J.** (2025).
  From Preferences to Prejudice: The Role of Alignment Tuning in Shaping Social Bias in Video Diffusion Models.
  Transactions on Machine Learning Research (TMLR), under review.
  Open Review URL: https://openreview.net/forum?id=C0yxuS6jty

Peer Review Service

- Neural Networks (Elsevier, Q1) — **9** completed manuscript reviews (2025)
- IEEE Transactions on Image Processing (TIP) — **6** completed manuscript reviews (2025)

Selected Research Contributions and Impact

- R-KV: Redundancy-aware KV Cache Compression for Reasoning Models
  - Published at NeurIPS 2025
  - Adopted by Qualcomm AI Research (KAVA, 2025)
  - Baseline in NVIDIA Research (ThinKV, 2025)
  - 1,100+ stars, 180+ forks on GitHub
  - One of the top open-source repositories in LLM inference optimization, with broad community adoption
  - 17 academic citations within months

Experience

Software Engineer II — Oct. 2025 – Present
Microsoft — *Sherman Oaks, CA*

- Work on the design and development of **large-scale LLM inference platforms and AI agent serving systems**, supporting **AI intelligence** and **high-volume user-facing workloads** in production environments using **Python**, **Flask**, and **Kubernetes**.
- Support deployment of **flexible fine-tuned models with specialized knowledge**, addressing system complexity related to orchestration, performance stability, and scalability.
- Contribute to **SGLang-based LLM serving integration and inference orchestration**, focusing on scalability, reliability, and efficient runtime behavior under real-world operational constraints.
- Develop **LLM serving backends and internal APIs**, enabling robust serving orchestration and runtime management for internal AI platforms.

Software Engineer Jun. 2023 – Aug. 2025

Activision Blizzard Inc. *Sherman Oaks, CA*

- Contributed to Atlas, an anti-cheat **monorepo** for Game *Call of Duty: Black Ops 6* in 2024 with **Python** and **Golang**, processing **4.4M+** cheating events/hour at peak, **50%** improvement over 2023 platform.
- Built a **Kafka** consumer with **Golang** to fetch events with **Protobuf**, routing 25K messages/sec to Dead Letter Queues and API servers.
- Set up a **Flask** API server with 19 endpoints for querying events, deployed via **Gunicorn** on **Kubernetes** with **Horizontal Pod Autoscaling**, handling **20k** HTTP requests/sec.
- Developed a multi-threaded **AWS S3** Manager handling **6k** jobs/sec for persistent event storage, implemented metadata storage service for production use with **SQLAlchemy** across two databases.
- Refactored unit test modules using **Pytest**, achieving 100% coverage, enhanced **GitHub workflows** for pull-based **CI** unit testing and maintained **Docker** images for project's **CD** pipeline.

Software Engineer Intern Jun. 2022 – Sept. 2022

Activision Blizzard Inc. *Santa Monica, CA*

- Contributed to **Segmentation 2.0**, a real-time matchmaking system that grouped players by skill in predefined ratios with **Python**, improving matchmaking balance and game fairness over 1.0 version.
- Designed a calculation pipeline: produced user update events to **Kafka**, buffered via **RabbitMQ**, processed events and executed calculation with **Celery workers**, and stored results in **Redis**, handling **1.2M** events/day.
- Created 2 **RESTful APIs** in **Tornado** to get results from calculation pipeline, supporting **350k** requests/day.
- Self-driven a log monitoring and alerting system based on **Demonata** framework to gather, reformat, and transfer logs from matchmaking system to **Kibana** log platform, handing 17M logs per day, crafted **Grafana** dashboards to visualize log metrics and alerts.

Software Engineer Intern Mar. 2021 – Jul. 2021

ByteDance Ltd. *Beijing, China*

- Launched a reusable UI widget library with **Flutter** for ByteDance, migrating UI elements to a remote bucket, reducing app size by **20%**.
- Designed an app update SDK with ByteDance components to unify the codebase post-acquisition and enhance maintainability.
- Enforced multiple update services in **Flutter**, supporting multiple update modes including manual, automatic, and debug updates.
- Created 2 **gRPC APIs** using **Go** and **Protobuf** for fetching update package profiles and posting metrics/issues, integrated **Redis** to cache frequently used package metadata with 91% hit rate.

Education

University of California San Diego Sep. 2021 – Mar. 2023

Master of Science in Computer Science | GPA: 3.97/4.0 *La Jolla, CA*

Beijing Jiaotong University Sep. 2016 – Jul. 2020

Bachelor of Science in Computer Science | GPA: 3.79/4.0 | ranking: 1/219 *Beijing, China*

Technical Skills

Languages: Python, Go, Java, C/C++, SQL, Shell Script, Dart
Frameworks: Flask, Gunicorn, Pytest, PubSub, Kafka, Protobuf, Spark, Hadoop, Flutter, RabbitMQ, Tornado
Tools: Docker, Redis, GCP, Kubernetes, Git, GitHub Workflow, MySQL, Terraform, AWS