

Yiyang Wei

🌐 [Homepage](#) | [LinkedIn](#) | 📞 +86 18078212148 | ✉️ 2002weiyiyang@gmail.com

EDUCATION

Zhejiang University

B.Eng in Information Engineering

Hangzhou, China

Sep. 2020 – July 2024

- Major GPA(Third Year): 3.72/4
- Selected Courses: Calculus(A+), Linear Algebra(A+), C Programming(A+), Lectures on Programming(A+), Fundamental Data Structures(A), Operating System(A+)

PUBLICATIONS

A Survey of Reasoning-Intensive Retrieval: Progress and Challenges

[ACL 2026 \(Main\)](#)

Yiyang Wei, Tingyu Song, Siyue Zhang, and Yilun Zhao.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, pp. 42101–42122. [DOI](#)

Can Retrievers Find the Same Paper from Different Aspects? A Multi-Aspect Full-Paper Scientific Retrieval Benchmark

[KDD 2027 \(Under Review\)](#)

Yiyang Wei, Fang Guo, Qiji Zhou, Zhizhang Fu, Mengru Ding, Kai Yang, and Yue Zhang.

arXiv preprint, [arXiv:2608.15624](#)

SELECTED WORK EXPERIENCE

Westlake University

WestlakeNLP Lab — Research Assistant

April 2026 – Now

Hangzhou, China

- **Multi-Aspect Scientific IR Benchmark (KDD 2027, Under Review):** Led the development of MAPLE, an expert-validated full-paper retrieval benchmark with 2,095 text- and multimodal-referenced queries, and MAPLE-Synth, an OpenReview-grounded retrieval-based ICL pipeline for paper search query generation; evaluated diverse retrievers and identified an 82.4-point gap between AnyAspect@20 (98.1%) and AllAspect@20 (15.7%), highlighting limitations in holistic full-paper representation.

Yale University

Research Collaborator

Sep. 2025 – March 2026

Remote

- **First-Authored Survey(ACL 2026(Main))** A comprehensive survey on Reasoning-Intensive Retrieval (RIR), covering benchmarks, methods across the retrieval pipeline, applications, and open challenges.
- **Synthetic Training Data for Retrievers:** Designed a persona-driven pipeline to construct retriever training data by generating realistic, instruction-rich queries across diverse difficulty levels and styles and mining reasoning-based hard negatives; fine-tuned Qwen3-Embedding on the resulting dataset, improving overall recall by 12% on BRIGHT and achieving state-of-the-art performance on the coding subtask.

Alibaba

DingTalk Wukong(Now [Qwen ToB](#)) Group — AI Engineer, Full-Time

July 2024 – July 2025

Hangzhou, China

- **DingTalk AI Image Search:** Fully engineered the foundational online search and offline index construction pipelines. Subsequently optimized the system by deploying a closeness-aware online ranking module and a fine-tuned CLIP offline indexing pipeline, resulting in an 18.9% increase in CTR, a 5% boost in recall, and a 95% reduction in indexing costs.
- **Large-Scale RAG-Based AI Search System:** Led the design of intent recognition and parameter extraction modules, and built **Text-to-SQL**-based user behavior(including send, receive, etc.) retrieval and indexing pipelines, supporting 2M+ users and 100k daily active users, enabling personalized and context-aware search.
- **RAG Evaluation System:** Developed an end-to-end RAG evaluation pipeline that quantified hallucination risk, faithfulness, response-level relevance for (query, retrieval, answer) core triplets in AI Search System.

Zhejiang University

Research Intern

Nov. 2023 – May 2024

Hangzhou, China

- Developed a dual-layer federated/pipelined learning framework for LLM fine-tuning on edge devices, reducing accuracy drop to 0.63% (vs. 1.44% in traditional federated learning, a 56% improvement).
- Upgraded communication to gRPC, 2x faster on edge transmission and 40% bandwidth reduction vs HTTP/1.1.

Trip.com

April 2023 – July 2023

Intelligence Infrastructure Group — AI Engineer Intern

Shanghai, China

- Deployed a GPT API service platform handling over 100K daily requests using Spring Boot, SOA, Elastic Search, CAT, and Hystrix for API integration, logging, monitoring, alarming, rate limiting, and load balancing.
- Built an Email-GPT assistant service with Retrieval Augmented Generation (RAG), LangChain, Gradio, and PostgreSQL to parse, translate, and facilitate interaction with international emails, providing writing suggestions

SKILLS

Programming: Python, C/C++, Java, Go, SQL

IR Techniques: Retrieval-Augmented Generation (RAG), Dense Retrieval (BGE-embedding, HNSW, K-means), Sparse Retrieval (BM25, Inverted Index), Multi-modal Search (CLIP, Vision Transformer (ViT))

Framework: Pytorch, MS-SWIFT, MTEB, LangChain, Flask, Springboot, Gin, Gorm, gRPC

Middlewares: PostgreSQL, MySQL, Redis, Message queue, Elastic Search, RPC, Sentinel, Hystrix

Developer Tools: Codex, Claude Code, Git, LaTeX, Docker, PyCharm, VS Code, IntelliJ, Clion, Goland, Vim, PostMan

Languages: Chinese Mandarin (Native), English (Fluent, TOEFL iBT 101)