Fei Fang

Education

Stanford University *Jan 2022*
Master of Science in Computer Science (Specialization in Artificial Intelligence) Stanford, CA

Stanford University *Jan 2022*
Bachelor of Science in Mathematics, with Distinction Stanford, CA

Employment

Cresta Intelligence — Conversational AI for contact centers *May 2024 - Present*
Machine Learning Engineer Palo Alto, CA

- Finetuned 7B-parameter large language model (LLM) using knowledge distillation for a domain-specific question answering (QA) system backed by retrieval-augmented generation (RAG), outperforming GPT-4o in internal benchmarks.
- Deployed the optimized LLM to enhance real-time customer support efficiency at major enterprises such as United Airlines and Square.

Glean Technologies — Enterprise-grade AI-driven search engine and assistant *Jan 2022 - April 2024*
Machine Learning Engineer Palo Alto, CA

- Led the development of a multilingual RAG-based LLM assistant for Japanese and English, deployed to 10+ major Japanese corporations including Toyota.
- Productionized the ColBERT model for reranking relevant contexts in RAG, demonstrating a 7-point increase in Mean Average Precision (MAP).
- Developed a novel algorithm inspired by ColBERT which efficiently probes a finetuned semantic bi-encoder for word-level measures of relevance at query time, making the retrieval system more interpretable and easier to debug.
- Defined quantitative metrics to measure the impact of semantic search on the hybrid retrieval system in production traffic, enabling A/B testing for all modeling experiments.
- Optimized serving efficiency across the vector search stack, resulting in a 14% reduction in median latency and an 11% reduction in 90th percentile latency.

Publications

[1] E. Kreiss, **F. Fang**, N. D. Goodman, and C. Potts. "Concadia: Towards Image-Based Text Generation with a Purpose". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022, pp. 4667–4684.

[2] **F. Fang**, K. Sinha, N. D. Goodman, C. Potts, and E. Kreiss. "Color Overmodification Emerges from Data-Driven Learning and Pragmatic Reasoning". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 44. 2022.

Teaching

CS221: Artificial Intelligence, Stanford University *Fall 2021*
Co-Head Course Assistant, Student Liaison

CS103: Mathematical Foundations of Computing, Stanford University *Summer 2021*
Co-Lecturer

CS103: Mathematical Foundations of Computing, Stanford University *Spring 2021*
Tutorial Leader

CS221: Artificial Intelligence, Stanford University *Winter 2021*
Head Course Assistant

CS103: Mathematical Foundations of Computing, Stanford University *Spring & Fall 2019; Winter & Spring 2020*
Course Assistant

Service

Lemontree (foodhelpline.org) — SMS-based helpline connecting people to local food banks *Oct 2024 - Present*
AI Strategy Advisor

Awards, Grants, Fellowships

Computer Science Teaching Fellow *Jan 2021*
Appointed as summer session instructor for a core requirement course of the Computer Science department Stanford University

Kung-Yi Kao Prize for Outstanding Progress in the Study of Japanese *Jun 2019*
Awarded to an outstanding undergraduate in an East Asian Language Study Stanford University

East Asia Undergraduate Summer Language Study Grant *Jun 2017*
Received funding to undertake intensive Japanese study during the summer session Stanford University

Talks and Posters

Wikimedia Research Showcase, virtual *Apr 2024*
Talk on "Concadia: Towards Image-Based Text Generation with a Purpose"
E. Kreiss (presenting), F. Fang, N. D. Goodman, C. Potts.

EMNLP 2022, virtual *Dec 2022*
Poster on "Concadia: Towards Image-Based Text Generation with a Purpose"
E. Kreiss (presenting), F. Fang, N. D. Goodman, C. Potts.

CogSci 2022, virtual *Jul 2022*
Talk on "Color Overmodification Emerges from Data-Driven Learning and Pragmatic Reasoning"
F. Fang & E. Kreiss (presenting jointly), K. Sinha, N. D. Goodman, C. Potts.

Languages

Listed in chronological order of acquisition: **Mandarin** (native), **Cantonese** (native), **English** (native proficiency), **Spanish** (professional proficiency), **French** (elementary proficiency), **Japanese** (professional proficiency)