Gefei Gu

412-670-1512 gefeig@andrew.cmu.edu linkedin.com/in/gefeigu https://frankgu3528.github.io/

Education

Carnegie Mellon University **Sep. 2025 – Dec 2026**
M.S. in Intelligent Information Systems (NLP Research Master), School of Computer Sciences *Pittsburgh, PA*

Zhejiang University **Sep. 2021 – Jun 2025**
B.S. in Statistics, School of Mathematical Sciences *Hangzhou, China*

- Awards: First Level Scholarship of Zhejiang University (Top 3%)

Technical Skills

Programming Languages: Python, C++, R, SQL, HTML/CSS, Shell
Machine Learning Tools: PyTorch, HuggingFace, Transformers, Numpy, Pandas, Sklearn, NLTK
Frameworks & Tools: Linux, Verl, Git, Docker, Jupyter, LaTeX, vLLM, AWS

Work Experience

Alibaba **Mar 2025 – July 2025**
Enhancing E-commerce DeepSearch Agents via Reinforcement Learning *(Research Intern)* *Hangzhou, China*

- Designed and implemented Product-Searcher, an end-to-end reinforcement learning approach to optimize deep product search on real-world e-commerce platform. Built an offline training environment for scalable training and generalization.
- Trained Qwen-7B using GRPO to achieve a 53% improvement over baselines in offline evaluations, subsequently outperforming GPT-4o by 6.7% in online evaluation on Alibaba's e-commerce search engine.
- Successfully deployed trained model to production, serving live traffic on Alibaba's e-commerce platform.
- Authored a first-author paper introducing our work, submitted to EMNLP Industry track.

Projects

An Automatic Realistic Framework for Long-Context LLM Evaluation **Yale University**
Research Assistant, Supervisor: Prof. Arman Cohan *Apr 2024 – Aug 2024*

- Developed an automatic evaluation framework for long-context LLMs that generates realistic benchmarks at varying document lengths and evidence depths without inserting artificial information.
- Implemented quality control mechanisms including LLM-based QA generation, quality checking modules, and RAG-based filterer to ensure high-quality benchmark questions with 96% human validation accuracy.
- Constructed a cross-domain benchmark spanning finance, legal, patent, and scientific literature with 8k-128k token lengths and 5-95% evidence depths. Evaluated and visualized 10 LLMs on our benchmark and observed significant performance degradation as context length increased.

Benchmarking the Long Context Referencing Ability of Long-context LLMs **Yale University**
Research Assistant, Supervisor: Prof. Arman Cohan *Aug 2024 - Nov 2024*

- Designed and constructed Ref-Long, a novel benchmark with three distinct datasets (synthetic, semi-realistic, and real-world academic papers) to systematically evaluate the long-context referencing capabilities of LLMs.
- Evaluated 13 LLMs on their ability to locate and attribute specific information within contexts up to 75K tokens. Observed critical performance decline as input length increases and conducted error analysis to categorize failure modes.
- Finetuned Llama-3.1 using a synthetic long-context dataset and observed limited performance improvements.

Precise Scientific Retrieval Through Dynamic Reasoning and Self-Reflection **Yale University**
Research Assistant, Supervisor: Prof. Arman Cohan *Dec 2024 - Feb 2025*

- Designed a multi-round Reflect-Refine algorithm that optimizes queries through iterative document reflection, improving retrieval performance. The method is applicable to a wide range of search engines and RAG systems.
- Evaluated on information-retrieval benchmarks including Bright and MultiHop-RAG, demonstrating 5% average improvement in retrieval effectiveness (ndcg@10) over baselines. Integrated the approach into a full RAG pipeline, achieving a 2.3% improvement in question-answering results.

Publications

1. **Gefei Gu**, Wenhui Chen, Qiankun Shi, et al. **Product-Searcher: Real-World E-commerce Product DeepSearch via Reinforcement Learning**. [Under Review] (EMNLP 2025 Industry Track)
2. Junjie Wu*, **Gefei Gu***, Yanan Zheng, Dit-Yan Yeung, Arman Cohan. **Ref-Long: Benchmarking the Long-context Referencing Capability of Long-context Language Models**. (ACL 2025, Main)
3. **Gefei Gu**, Yilun Zhao, Ruoxi Ning, Yanan Zheng, Arman Cohan. **TAIL: A Toolkit for Automatic and Realistic Long-Context Large Language Model Evaluation**. (EMNLP 2024, System Demonstrations)
4. Zhizhang Yuan, Fanqi Shen, Meng Li, Yuguo Yu, **Gefei Gu**, Fei Wu, Chenhao Tan, Yang Yang. **BrainWave: A Brain Signal Foundation Model for Clinical Applications**. [Under Review] (Nature Machine Intelligence)