Jiang Yu

CONTACT	7675 Palmilla Dr Apt 6405, San Diego, CA 92122 *Email:* me@thejiangyu.com *Mobile:* 917-215-1187

SUMMARY

Engineering Manager (Machine Learning) with 20 years building and leading large-scale Search/Ranking and Recommender Systems. Currently lead Search Ranking & Agentic AI at eBay, delivering $1B+ revenue growth through modern multi-stage ranking, constrained optimization/allocation, and large-model innovation. Proven track record building high-performing ML teams shipping end-to-end systems (modeling, experimentation, and production serving). Recipient of eBay's inaugural **AI Business Impact Award (2025)** (presented by CEO Jamie Iannone). Dual PhDs.

QUALIFICATIONS

- **Leadership:** People management; org alignment; execution; cross-functional influence (Product/Eng/Sci/Analytics/Exec)
- **ML Systems:** Multi-stage retrieval & ranking; reranking/blending; constrained optimization & allocation; calibration; exploration–exploitation; A/B experimentation
- **AI/ML:** Learning-to-rank/IR; recommender systems; deep learning; NLP/NLU; Generative AI; RL
- **Stack:** Python, Java, Scala; PyTorch, TensorFlow, XGBoost; AWS; distributed training/serving; logging & analytics pipelines

PROFESSIONAL EXPERIENCE

eBay — Manager, Applied Research (Search & Recommendations) Jun 2023 – Present
Lead a cross-functional ML team (Applied Researchers, ML Engineers, Software Engineers) owning core Search Ranking, Organic Reranking, Allocation Frameworks, and Monetization optimization across the multi-stage stack (retrieval → ranker → reranker/blender).
Execution & Influence: Sustained 80%+ quarterly goal completion across 10 consecutive quarters; aligned Search/Marketplace/Response Prediction roadmaps; quantified business tradeoffs with auditable budget frameworks.
Business Impact: Helped drive the Search & Monetization org beyond the **$1B 2024** revenue growth target ahead of schedule; delivered **86%** of 2025 revenue goals in H1; received eBay's **AI Business Impact Award (2025)**.
Technical Leadership & Innovation:

- Architected a **Ranking-as-Allocation** framework enabling programmable constraints in organic ranking via primal–dual optimization and auditable budgets.
- Built the next-gen ranking training platform: **20× faster** iteration, **20× larger** models/data; reduced training latency **2 days → 5 hours**; eliminated distributed-training instability.
- Led ranking innovations (Mixture-of-Ranking-Experts, RL-based reranking, slate-aware optimization), generating **$300M+** incremental annual revenue.
- Founded and led the **Generative AI Working Group** for Search & Ads to accelerate adoption and de-risk production rollouts.

Amazon — Applied Scientist (Lead), Alexa AI   Jun 2020 – Jun 2023
Led 7+ scientists and engineers building multilingual conversational AI systems across ASR, MT, and NLU.
Product & Team Leadership: Drove Conversational Language Learning from PR/FAQ → funding → prototype → beta; authored a 20+ page runbook that standardized onboarding and scaled delivery across orgs.
Impact: Reduced false rejections **10%** and repeat failures **15%** in Alexa Translation; improved ASR WER **5%** for Live Translation.
Research & Modeling:

- Led formality-control MT research; won **1st place** at **IWSLT 2022**; co–first author on a published paper.
- Built GPT-based paraphrasing/data augmentation pipeline; improved conversational model accuracy by **+4%**.
- Built multilingual language ID for low-latency spoken queries; delivered **50%** error reduction.
- Published work on knowledge distillation for multilingual NLU (AMLC 2022).

Memorial Sloan Kettering Cancer Center (MSKCC) — Tech Lead, Software Engineer (ML in Healthcare)   Apr 2019 – May 2020
Built deep learning models for clinical sequence prediction using diagnosis & lab data; designed and shipped a clinical research database web app; built ETL pipelines and Tableau dashboards for large-scale medical data visualization.

Selected Earlier Roles:
Intel (2018): Scaled Transformer/GNMT across GPU/CPU clusters; optimized MoE via dynamic layer skipping.
Co-Founder, G-wear (2017–2018): Built IoT+ML safety monitoring for elderly care; **1st place**, NYU Healthcare Makerthon.

Tencent WeChat AI Lab (2017): Built medical chatbot and knowledge-graph pipelines.
Twitter (2016): Improved recommendation RCE **+33%**; launched personalization features (Insta-Stories).
AWS (2016): Built global call-center cloud integration, later launched as **Amazon Connect**.
AIG (2015): Built fall detection & human activity recognition models.
Adjunct Professor, NYU (2015): Instructor, Python Programming.
Founder, Rental Sharing Service (2006–2008): Built and scaled rental platform to **100K RMB** annual revenue and thousands of transactions.
R&D Lead, Nanjing Institute of Communications Engineering (2006–2013): Led software & networking projects; managed engineers and researchers.

PUBLICATIONS & PATENTS	Author of 20+ publications and patents (including two doctorate dissertations).
HONORS & AWARDS	**AI Business Impact Award**, eBay (2025)
	1st Place, IWSLT Machine Translation Formality Control Challenge (2022)
	1st Place, NYU Healthcare Makerthon (2017)
	Best Remote Office Award, Twitter Hack Week (2016)
	Graduate Innovation Fellowship, NYU (2013–2015)
	Top Rank, China National College Entrance Exam (school-wide, 1st/500)

EDUCATION	**Stanford University Graduate School of Business**	
	Executive Education LEAD Program	2025 – 2026
	New York University	
	Ph.D. Computer Science (part time)	2013 – 2023
	Dissertation: Topics in Interactive Search with Large Language Models	
	Nanjing Institute of Communications Engineering	
	Ph.D. Cooperative Information Theory	2006 – 2013
	B.S. Electrical Engineering **Rank: 5th/178**	2001 – 2005