

Byungkyu (Jay) Kang

Senior Staff AI Scientist (ML/NLP/LLMs)

homepage — CV — San Diego, CA — Google Scholar

penguinkang@gmail.com

+1-805-455-9856

Summary

AI Architect and Principal-level Machine Learning Scientist with 10+ years shipping production ML/NLP systems (large and small language models, recommendation, anomaly detection) for a Fortune-500 financial-software company; models in production serve **50M+** consumers. Leads cross-functional teams (PM/Eng/Design), mentors scientists and engineers, and owns technical roadmap and reviews end to end. **16 US patent applications filed (5 granted)** in LLM/NLP; Ph.D. in Computer Science.

Core Skills

ML/AI: LLM/SLM fine-tuning & alignment, RAG, RecSys, anomaly detection, XAI, graph/sequence models

Data/Platforms: PyTorch, TensorFlow, scikit-learn, Spark, AWS, Neo4j, OpenSearch, SQL/NoSQL

Prod/Infra: Batch/real-time pipelines, A/B testing, monitoring, AutoML, model eval/guardrails

Lang/Tools: Python, Scala, Java, JS, Docker, Linux/macOS/Windows

Experience

Intuit — Intuit AI, San Diego, CA

Jun 2018 – Present

Senior Staff AI Scientist; Staff (2021–2025); Senior (2018–2021)

- Led the company's GenAI conversational assistant for customer-support experts (**Intuit Assist**) and a real-time recommendation model routing customers to the right support channel; earlier built an automatic call/case-note summarizer cutting post-call agent documentation work.
- Led a systematic **24-wave architecture-search/fine-tuning study** for real-time speech-intent classification: fusing lightweight frozen audio+text encoders – replacing a costly frontier multimodal LLM entirely – hits the production target (**macro F1 0.746**) at **~1,000x fewer params** and a fraction of the latency.
- Delivered fine-tuned production models: an SLM replacing a frontier-LLM call in a real-time memory/context-extraction service (**~8x lower latency**, 97%+ reliability), and a conversational-coaching model for live sales/support agents. Also built a model confidence scorer (**0.80 Pearson / 0.88 AUROC** vs. LLM-judge) for quality-gating.
- Own **production ML infrastructure** across cloud ML platforms (AWS SageMaker, Kubernetes) – training recipes, distributed RL/eval orchestration; built a paved-path fine-tuning/RL library (auto-sft) with an automated-research-agent feature; fixed a GPU/storage bottleneck blocking 20+ eval configs.
- Shipped a 5-quarter production knowledge-base recommendation feature with measurable handle-time/satisfaction impact; led development of a proprietary RAG-based GenAI agent platform (PSA & HelpGPT); founding member of the company's first data-science team – built call-routing classification and a domain embedding model (**87% → 94%** query coverage).
- **16 US patent applications** filed; **5 granted**. Own on-call/RCA for the GenAI assistant service.

Nokia Technologies — CTO Analytics, Sunnyvale, CA

Sep 2016 – Mar 2018

Data Scientist

- Built large-scale analytics for device/content signals using Spark/Scala and ML for recommendation and search ranking.

IBM TJ Watson Research (Network Science & MR), Yorktown Heights, NY

Jun 2016 – Sep 2016

Visiting Scholar

- Graph mining + reasoning prototypes; interactive visual analytics for model insights.

Education

Ph.D., Computer Science, University of California, Santa Barbara

Mar 2016

Dissertation: *Information Reliability on the Social Web—Models and Applications in Intelligent User Interfaces*

B.Eng., Electrical & Electronics Engineering, Chung-ang University, Seoul, Korea

Feb 2007

Selected Patents (US)

- **11.907,657** — Dynamically Extracting N-grams for Automated Vocabulary Updates (2024)
- **11.941,144** — Methods and Systems for Automatic Call Data Generation (2024)
- **12.067,976** — Automated Search and Presentation Computing System (2024)
- Multiple applications filed: query augmentation, response synthesis, framework-agnostic summarization, context-specific generative answers, anomaly prediction, abandonment behavior, etc.

Selected Publications

- **Kang, J.**, “From Multimodal LLMs to Lightweight Classifiers: When Bypassing the Language Model Improves Speech Intent Detection” (working paper, 2026).
- Sadeq *et al.*, “Mitigating Hallucination in Fictional Character Role-Play” *EMNLP*, 2024.
- Sikdar, **Kang**, *et al.*, “Understanding Information Credibility on Twitter” *ASE/IEEE SocialCom*, 2013 (**Best Paper**).