Shan Lu

shanlu@google.com | (774) 567-5060 | **LinkedIn**: shan-lu-profile
Focus: Applied LLM, RAG, Agentic Systems, Evaluation

EXPERIENCE

Google, *Machine Learning Engineer*
Ads Insights Team, Ads Mar 2022 - Present | *New York, NY*

- Ads Advisor (Agentic RAG)
  - Designed and deployed Ads Advisor, a production **agentic RAG system** that performs **multi-turn reasoning** over complex **tabular datasets** to diagnose campaign performance.
  - Engineered a **hybrid retrieval mechanism** that **dynamically routes** queries between **semantic vector search** and **synthesized SQL generation**, minimizing hallucination rates on quantitative queries.
  - Implemented **System 2** reasoning patterns (Chain-of-Thought) to decompose high-level user intent into atomic sub-queries against the Ads database.
- Diagnostic Agent & Smart Summary (Eval)
  - Developed an **LLM-as-a-Judge** evaluation pipeline to autonomously score agent **factuality** and **instruction following**, reducing reliance on human raters and accelerating iteration cycles.
  - Implemented **constrained decoding** at inference time to enforce strict output schemas (JSON/Protobuf), coupled with self-correction loops to ensure 100% syntax compliance for downstream API execution.
  - Deployed generative summarization models for **time-series performance data**, engineering a **ground-truth entailment evaluation** framework to verify that generated summaries are strictly supported by the retrieved statistical context.
- Semantic Taxonomy Generation (Data Scale)
  - Scaled the semantic taxonomy for Consumer Interest Insights from 3k to 4M+ categories by engineering a high-throughput **n-gram labeling pipeline**.
  - Optimized dense retrieval performance for advertiser-facing search query analysis, achieving a 28% increase in data coverage by fine-tuning embedding models on domain-specific query logs.

Personal Health Record Team, Health Aug 2020 - Mar 2022 | *Palo Alto, CA*

- Engineered the **query understanding** backend for medical records search, implementing **query expansion** and **spell-correction** logic for a feature serving **1.5M daily impressions**.

Waybak Team, Research (Intern) May 2019 - Aug 2019 | *New York, NY*

- Conducted **Computer Vision** research on **clustering** and **recognition** of facade images, acquiring **94.3%** accuracy.
- Integrated 2D=>3D reconstruction **pipeline** including segmentation, rectification, inpainting, and embedding.
- Built the inference module of a freeform image inpainter based on **gated convolution** and **SN-PatchGAN** loss using TensorFlow for model training and C++ for inference.

MIT CSAIL, *Computer Vision Undergraduate Researcher*
Geometric Data Processing Group Sep 2019 - Jan 2020 | *Cambridge, MA*

- Developed a **deep supervised learning** model to learn feature representations for **3D scene view synthesis**.
- Optimized 3D scene structure and camera motion from multi-view images using **feature-metric bundle adjustment**, generating novel views from a latent **3D voxel representation**.

Computational Perception & Cognition Lab Sep 2018 - Aug 2019 | *Cambridge, MA*

- Trained ML models to predict **natural zoom levels** from crowdsourced mobile interaction data, generating **attention maps.**

EDUCATION

Wellesley College, Wellesley, MA Sep 2016 - May 2020
B.A. in Computer Science and Mathematics (Magna Cum Laude)

- **Massachusetts Institute of Technology (MIT)**: Completed advanced Computer Science core via cross-registration

SKILLS

- **Programming:** Python, C++, SQL, Java
- **Machine Learning:** TensorFlow, PyTorch, RAG (Retrieval-Augmented Generation), Agentic Workflows, Prompt Engineering, Vector Search, Evaluation/Reward Modeling