GEORGE HALAL | Personal Site: https://georgehalal.github.io | georgehalal@alumni.stanford.edu | +1 (650) 422-9033

Stanford PhD turned AI engineer with expertise in agentic search, synthetic data generation, and LLM finetuning.
Most recently, I trained and open-sourced rerankers on the cost/performance frontier across public and customer benchmarks.

EDUCATION

Stanford University | Ph.D. Physics | GPA: 4.00/4.00 | June 2019–July 2024
Lehigh University | B.S. Physics & Minor in Applied Mathematics | GPA: 3.97/4.00 | Aug. 2015–May 2019

EXPERIENCE

Member of Technical Staff | Contextual AI, Mountain View, CA | July 2024—Present

Best-performing, Most Efficient, Instruction-Following LLM Rerankers | v2 Blog | v1 Blog | Snowflake Blog | OS models | OS evals

- Built a synthetic data pipeline to generate contrastive data covering desired behaviors and diverse domains.
- Experimented with QAT, distillation, reinforcement learning, and curriculum learning among other techniques.
- Achieved SOTA performance on instruction following, question answering, multilinguality, product search/recommendation systems, and real world use cases.
- Selected as the default reranker for Snowflake Cortex AI among other companies.

Agentic Search Tool Use Optimization

- Optimized the type, number, cost, and latency of knowledge search and navigation tools an agent uses during rollouts.

Graph-based Search (Graph RAG) | Paper in Prep

- Developed an LLM-based pipeline to turn documents into knowledge graphs for efficient retrieval at query time.
- Shipped to production as part of a mixture of retrievers for answering certain queries.
- Separately, mentored a Stanford CS student on his master's thesis, "End-to-End Retrieval on Black-Box Knowledge Graphs."

Graduate Student Researcher | Stanford University, Stanford, CA | June 2019–July 2024

Transformer-Based Super-Resolution for Dust Polarization Images | GitHub Link

- Trained a multi-image encoder, a transformer-based fusion module, and a decoder to increase the image resolutions by 4x.

Causal Inference for Modeling the Effects of the Nearby Dust Geometry on Magnetic Fields | Paper Link

Spherical Harmonic Convolutional Hough Transform | GitHub Link | Paper Link | Invited Talk Link

- Achieved 3000x speedup and 5x memory reduction over the previous SOTA for modeling the structure of interstellar gas.

Modeling the Foreground Obscuring Radiation from the Early Universe | Paper Link | Award Link | Invited Talks: Harvard, Spain, S4

- Applied computer vision and Bayesian inference for quantifying this signal, setting new limits on early universe expansion.

Conditional Wasserstein Generative Adversarial Network with Gradient Penalty for Generating Observed Galaxy Properties

Deep Learning for Modeling the Transfer Function of Galaxy Detection, achieving an ROC-AUC score of 0.95 | GitHub Link

Deep Learning for Searching for 2-ν Double-β Decay of ^{136}Xe to the Excited State of ^{136}Ba in EXO-200 Data | Poster Link

- Developed a data acquisition pipeline and an LSTM-based model, achieving an ROC-AUC score of 0.98.

Data Scientist Intern | Alife Health, San Francisco, CA | June 2023—Sept. 2023

Causal Inference, A/B Testing, and Machine Learning for IVF Intracycle Dose Adjustments

Undergraduate Student Researcher | Yale University and Lehigh University | Nov. 2016–May 2019

Deep Learning for Heavy-Flavor Jet Classification at RHIC | Report Link | Talk Link

Deep Learning for Collision Geometry Determination

SKILLS

Python • PyTorch • WandB • Pandas • vLLM • Hugging Face (transformers, accelerate, peft, trl) • NumPy • asyncio • threading • OpenAI • OpenAI Agents SDK • Pydantic • Statsmodels • SciPy • Seaborn • Xgboost • Scikit-learn • Matplotlib • requests • LaTeX • SQL • SLURM

PUBLICATIONS | 15+ peer-reviewed (1,253+ citations) including 3 first/corresponding-author in top astrophysics journal