George Saad

647-544-5877 georgekysaad@gmail.com linkedin.com/in/gkysaad github.com/gkysaad georgesaad.ca

Experience

Amazon | *Applied Scientist* *New York City, NY | Mar. 2025 – Present*

- Built and productionized a **large-scale RAG-based, agentic LLM system**, emphasizing scalable alignment methods (human-in-the-loop prompting, generate-and-rank) deployed across 50M+ products and 100M+ daily searches
- Designed **post-training preference alignment pipelines** using SFT and custom multi-objective loss functions to optimize generation for purchase intent, relevance & brand sentiment, directly driving $500M+ in annualized revenue
- Incorporated and operationalized LLM evaluation frameworks in production, leveraging **LLM-as-a-judge**, automated sentiment/factuality/relevance guardrail models, and calibrated human evaluation to monitor hallucinations, bias, and alignment regressions at scale
- Authored **2 research papers** and **1 patent submission** on scalable LLM alignment, evaluation, and controllable generation for advertising systems

University of Toronto | *Research Assistant* *Toronto, ON | May 2023 – Feb. 2025*

- Conducted research on **LLM-integrated conversational recommender systems**, combining retrieval, query reformulation, and generation to improve recommendation quality
- Led projects resulting in **peer-reviewed publications at ACL, SIGIR, and RecSys**, focusing on LLM-based retrieval, reasoning, and evaluation

Amazon | *Applied Scientist Intern* *Seattle, WA | Jun. 2024 – Sep. 2024*

- Developed an **agentic, multi-stage LLM-based semantic disambiguation pipeline** that filtered data early and applied deeper LLM reasoning only to ambiguous cases, maintaining inference efficiency while improving product search ranking by 29% across 8.5M+ weekly queries
- Evaluated failure modes of **embedding-only semantic methods** and identified ambiguity regimes where LLM-based reasoning materially improved clustering and classification quality
- Implemented **production-ready modeling, evaluation**, and **logging code**, enabling integration with Amazon's search stack and reliable offline/online validation
- Conducted systematic prototyping, literature review, and empirical evaluation, presenting results and trade-offs to 50+ engineers and scientists

Vector Institute | *Applied Machine Learning Intern* *Toronto, ON | Jan. 2023 – Sep. 2023*

- Evaluated deep causal inference models (**TARNet, DragonNet**) on observational datasets, studying sensitivity to confounding and lack of ground-truth counterfactuals
- Implemented evaluation pipelines to compare causal estimators on synthetic and real-world data under distributional shift
- Presented findings through technical workshops for **200+ researchers and industry practitioners** across major sponsors (RBC, Deloitte, Shopify, etc.)

Meta (Instagram) | *Software Engineer Intern* *Menlo Park, CA | May 2022 – Jul. 2022*

- Built a **high-throughput Thrift service** to ingest and track 260M+ post-click external interactions daily, enabling richer feedback signals for personalized ad delivery and ranking on Instagram
- Developed app-level detection logic (Swift/Kotlin) to surface cross-app signals used for more accurate ad targeting, expanding delivery scope and increasing monetizable impressions by 120% for previously unlinked users
- Optimized Instagram Ads serving and data ingestion endpoints (Hack/PHP, Python Django), **reducing runtime and memory overhead** to support low-latency, large-scale personalization pipelines

Publications / Research

Q-STRUM Debate: Query-Driven Contrastive Summarization for Recommendation Comparison | Paper
G. Saad, S. Sanner - *Findings of ACL 2025*

Elaborative Subtopic Query Reformulation for Broad and Indirect Queries in Travel Dest. Rec. | Paper
Q. Wen*, Y. Liu*, J. Zhang*, **G. Saad**, A. Korikov, Y. Sambale, S. Sanner - *ROEGEN @ RecSys 2024*

Multi-Aspect Reviewed-Item Retrieval via LLM Query Decomposition and Aspect Fusion | Paper
A. Korikov*, **G. Saad***, E. Baron, M. Khan, M. Shah, S. Sanner - *IR-RAG @ SIGIR'24*

Education

University of Toronto *Sep. 2023 – Jan. 2025*

- MASc (Thesis-based), Research in LLMs & Recommendation supervised by Prof. Scott Sanner, GPA: 4.0 (out of 4.0)

University of Toronto *Sep. 2019 – May 2023*

- BASc in Engineering Science, Major in Machine Intelligence, Certificate in Engineering Business, GPA: 3.76 (out of 4.0)

Technical Skills

Languages: Python, Java, JavaScript, C/C++, PHP/Hack, HTML/CSS, YAML, PostgreSQL, MATLAB
Technologies: PyTorch, JAX, Numpy, Pandas, Scikit Learn, Matplotlib, Spring Boot, React, Node, Flask, Django