Sahel Sharifymoghaddam

Ph.D. Student, NLP/IR
University of Waterloo

☎ +1(647)771-3706
✉ sahel.sharifymoghaddam@uwaterloo.ca
in /sahelsharifymoghaddam
Sahel Sharifymoghaddam

Profile

Ph.D. student researching at the intersection of Natural Language Processing (NLP) and Information Retrieval (IR), with a focus on leveraging large language models (LLMs) for retrieval and ranking, grounding LLMs through retrieval-augmented generation (RAG), and their multi-modal extensions. Bringing several years of industry experience as a Machine Learning (ML) Engineer, leading cross-functional collaborations to design, implement, and launch Generative AI products on Google Cloud Infrastructure.

Education

Ph.D. in Computer Science (University of Waterloo, Waterloo, Ontario, Canada) **2023 - present**

- Research Interests: Use of LLMs for IR and RAG for grounding LLMs, multi-modal LLMs and retrievers.
- Supervisor: Prof. J. Lin (jimmylin@uwaterloo.ca)
- Coursework: Optimization for Data Science, Advanced Topics in AI (Trust, Explainablity and Social Media), Advanced Topics in AI (Foundation Models), Advanced Topics in AI (Diffusion Models), **GPA: 95.5/100**

M.A.Sc. in Computer Engineering (University of Toronto, Toronto, Ontario, Canada) **2013 - 2015**

- Thesis: Optimized Data Placement for In-memory Data Analytics (A+)
- Supervisor: Prof. C. Amza (amza@eecg.toronto.edu)
- Coursework: Advanced Operating Systems, Dependable Software Systems, Parallel Programming, Advanced Topics in Data Management Systems, Trends in Middle-ware Systems (Big Data), **GPA: 3.94/4**

B.Sc. in Computer Engineering (Sharif University of Technology, Tehran, Iran) **2009 - 2013**

- Selected Coursework: Object Oriented System Design, Data Structures and Algorithms, Design of Algorithms, Computer Architecture, Compiler Design, Computer Network, Operation Systems, Database Systems, Artificial Intelligence and 8 courses on various Math topics, **GPA: 18.56/20** (Department average for class of 2013 is 16.03.)

Selected Publications

- **Sharifymoghaddam, S.***, Upadhyay, S.*, Thakur, N.*, Pradeep, R., & Lin, J. (2025). **Chatbot Arena Meets Nuggets: Towards Explanations and Diagnostics in the Evaluation of LLM Responses.** *arXiv preprint arXiv:2504.20006.*
- **Sharifymoghaddam, S.***, Upadhyay, S.*, Chen, W., & Lin, J. (2025). **UniRAG: Universal Retrieval Augmentation for Large Vision Language Models.** *In Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 2026-2039).
- Pradeep, R., Thakur, N., **Sharifymoghaddam, S.**, Zhang, E., Nguyen, R., Campos, D., ... & Lin, J. (2025, April). **Ragnarök: A reusable RAG framework and baselines for TREC 2024 retrieval-augmented generation track.** *In European Conference on Information Retrieval* (pp. 132-148).
- Pradeep, R., **Sharifymoghaddam, S.**, & Lin, J. (2023). **RankZephyr: Effective and Robust Zero-Shot Listwise Reranking is a Breeze!** *arXiv preprint arXiv:2312.02724.*
- Pradeep, R.*, **Sharifymoghaddam, S.***, & Lin, J. (2023). **RankVicuna: Zero-shot listwise document reranking with open-source large language models.** *arXiv preprint arXiv:2309.15088.*

Selected Awards and Honors

- Recipient of the **Math Domestic Graduate** award, the **Dean of Math Excellence** scholarship and the **Go-Bell** scholarship awarded by University of Waterloo (2023 - present)
- Recipient of the **monthly scholarship for graduate studies** awarded by University of Toronto (2013-2015).
- **Achieved** 5th **highest GPA** among all computer engineering students (110+), Sharif University of Technology, class of 2013. Recipient of GPA-based **Honorary Admission for Graduate Study** (Declined).
- **Ranked** 4th in **National Scientific Olympiads** in Computer Engineering, Iran, 2012.
- **Ranked** 17th among **300,000+** participants in the nationwide university entrance exam, class of 2009. Recipient of the **4-year monthly undergraduate scholarship award** from the National Elite Foundation of Iran.

Work Experience at Google

Senior Machine Learning Engineer
Google Cloud - Applied AI **Apr. 2021 - Apr. 2025**

Led multiple projects in the Sales AI and Healthcare AI teams, driving efforts across design, implementation, testing, and launch. Built and deployed Generative AI applications on Google Cloud Platform (GCP), delivering end-to-end solutions for real-world enterprise use cases. Prior to working with Generative AI, focused on fine-tuning domain-specific language models for core NLP tasks such as Named Entity Recognition (NER), Relation Extraction, and Entity Linking/Disambiguation.

Selected Projects

- In Sales AI, worked on diverse data processing pipelines using Generative AI and prompt engineering techniques. Implemented prompt optimization, LLM-as-a-judge evaluation, synthetic data generation, grounded generation with Google Search and citation insertion, and ensured security and privacy compliance for the serving infrastructure.
- Led six feature launches for the Cloud Healthcare NLP API that processes medical notes for Entity Extraction, Relation Extraction, and Entity Linking. Responsibilities spanned design, implementation, and deployment on Vertex AI. Launches included integrating new Entity Extraction and Entity Linking models, and upgrading model serving infrastructure for performance and scalability.
- Developed GCP DocumentAI processors for Healthcare and Contract AI use cases. Trained custom entity extraction models, designed deployment architectures, scoped the projects into incremental, deliverable features, and led cross-functional execution involving nine engineers across three teams.

Software Engineer
Chromium **Apr. 2016 - Apr. 2021**

Worked end-to-end on multiple projects within the Web Payments and Input & Scrolling teams, leading efforts in design, implementation, launch, and public documentation. Actively guided cross-team collaborations, addressed high-priority feature requests, and contributed to W3C standards.

Selected Projects

- Led GPay's adoption of the Web Payments API, coordinating cross-team efforts, triaging feature requests, and analyzing the impact of Chrome's privacy sandbox on GPay APIs.
- Defined the Shipping Option and Address Change events in the W3C standards for the Payment Handler API, implemented them, and launched the features in Chromium—enabling real-time transaction updates for merchants.
- Improved scrolling smoothness and latency across all platforms by making document-level wheel event listeners passive by default and implementing scroll latching and asynchronous wheel events. Enhanced autoscroll latency and unified fling handling logic by migrating touchscreen, touchpad, and autoscroll fling events from the renderer process to the browser process, resolving multiple outstanding issues.

Selected Teaching Experience

- Teaching Assistant, University of Waterloo
  - **Computational Linguistics (Graduate Course)**, with Prof. F. Shi (Winter 2025)
  - **Object-Oriented Software Development**, with Caroline Kierstead (Fall 2023, Winter 2024)
- Teaching Assistant, University of Toronto
  - **Foundations of Computing**, with Dr. D. C. Drosu (Winter 2014, Winter 2015)
  - **Parallel Programming (Graduate Course)**, with Prof. C. Amza (Fall 2014)
  - **Operating Systems**, with Prof. A. Goel (Fall 2014)

Computer Skills

- **Machine Learning:** Deep learning, Natural Language Processing, TensorFlow and Keras, Pytorch, Pandas, Numpy, LangChain, LlamaIndex, Weights and Biases
- **Programming Languages:** Proficient in Python, C++, Go, Java, and JavaScript; submitted and reviewed hundreds of thousands of lines of production code across 10+ languages at Google.
- **Google Cloud Basics:** Compute Engine, IAM, Spanner, Cloud Storage, Vertex AI, and Familiar with GCP Deployment Manager, Cloud Infrastructure, Docker and more.
- **Other Qualifications:** System Design, API Design (gRPC with protocol buffers, and W3C standards), Relational Databases (MySQL), Operating Systems, Network, Compilers, Computer Architecture, Data Structures and Algorithms, Git