Sagi Shaier

sagishaier@gmail.com | Linkedin.com/in/Shaier | Publications | GitHub.com/Shaier | Website

SUMMARY

I'm an AI researcher focused on foundational problems – efficiency, sparsity, reasoning, transfer learning, and continual learning – where progress ripples across the entire field rather than solving one narrow task. I draw on insights from biological systems to discover simple, generalizable ideas. I care about both the science and its real-world impact, and enjoy building research agendas from the ground up while collaborating across disciplines to tackle problems that matter at scale.

PROFESSIONAL EXPERIENCE

Johannes Gutenberg University of Mainz · Visiting Professor April 2026-Present

● Designed and taught graduate & undergraduate courses spanning NLP applications and continual learning.

Aleph Alpha Research · AI Researcher Oct 2025-April 2026

● Invented Excitation (patent-pending), a lightweight optimizer-agnostic framework that **sharpens expert specialization in MoEs** via utilization-driven competitive update dynamics, **accelerating convergence and improving performance.**

● **Led development of a production-grade framework** for citation generation and hallucination mitigation.

Cohere · Intern of Technical Staff May 2025-Sep 2025

● Collaborated with Cohere researchers and fellow interns to **develop an agentic search system**, contributing to dataset construction, data annotation, and evaluation frameworks.

● **Co-developed a multimodal, multilingual document parsing pipeline** using smart page segmentation, designing an interactive annotation pipeline to enable scalable model training.

University of Colorado Boulder · Doctoral Researcher Aug 2020-May 2025

● **Awarded the prestigious Social Impact Award** out of 363 international candidates for identifying fairness concerns in biomedical QA systems, paving the way for more reliable and fair AI in high-stakes healthcare applications.

● **Designed and led a research mentorship program for 12 groups** of graduate students over 3 years across NLP projects including information retrieval, multihop QA, knowledge graphs, multilingual AI, and LLMs — with outcomes including peer-reviewed publications, research positions at academic labs, and engineering roles at FAANG companies.

● **Initiated and led international research collaborations** across institutions and disciplines, including government organizations, Google DeepMind, medical schools, and universities, spanning NLP, biomedical AI, and computer vision.

● **Minimized inference requirements by over 90% in "foundational" models** using biologically-inspired algorithms, accelerating progress towards artificial general intelligence (AGI) and highly-scalable generative AI.

● Developed a framework for **building more accurate and trustworthy language models** that prioritize factuality and mitigate hallucinations by grounding their generation with citations.

● **Increased accuracy by up to 42%** in large language models (LLMs) knowledge assessment through the development of a novel evaluation method, which also **reduces language generation redundancy by up to 40%**.

● **Improved accuracy by up to 36% in retrieval augmented generation (RAG)** models using attention-based strategies.

National Institute of Mental Health (NIMH) · ML Researcher (Volunteer) May 2023-May 2025

● **Designed a biologically-inspired mixture-of-experts algorithm** to induce sparsity and modularity in any neural system, enhancing model efficiency, performance, and transfer learning capabilities.

● Collaborated with several researchers teams in **developing multimodal NLP applications**, such as knowledge representation and QA systems in the biomedical domain using both structured and unstructured data.

● Produced intuitive visualizations that **simplify complex algorithms**, making them accessible to a broader audience.

● Developed **biologically-inspired continual learning algorithms** for computer vision and language models.

Oracle · Research Intern Jan 2024-April 2024

● Designed and implemented scalable multi-GPU machine learning systems for large scale training and inference, which **supports 70B parameter models** in both supervised and unsupervised settings.

● **Developed 5 novel datasets** to assess language models' ability to answer complex, ambiguous questions with citations, spanning multiple domains and requiring challenging multihop reasoning.

● **Increased LLMs' question answering accuracy by 19.4%** and **citation generation accuracy by 86.7%** through innovative prompt engineering and fine-tuning techniques, yielding more accurate, trustworthy, and reliable AI.

Pacific Northwest National Laboratory (PNNL) · National Security Research Intern May 2021-October 2021

● Applied topological methods to high-dimensional text embeddings to **improve factual knowledge representation.**

Quantum Metric · Data Scientist (Research Team) Dec 2019-May 2020

● Implemented real-time anomaly detection systems that reduced service disruptions, **saving millions annually**.

Welocalize · Machine Learning Intern May 2019-Aug 2019

● Built predictive models to **surface organizational bottlenecks** and optimize project tracking workflows.

Hack Oregon · Data Scientist (Volunteer) Feb 2019-Sep 2019

● Created a spatial casualty model for Cascadia Earthquake to **inform medical response strategies**.

EDUCATION

University of Colorado Boulder · PhD Computer Science

Dissertation: "Factual Knowledge-enhanced Question Answering in Dynamic Environments".

University of Colorado Boulder · MS Computer Science

Kennesaw State University · BS Computational and Applied Mathematics,
Concentration in Epidemiology, Minor in Statistics, Pre-Med.

PUBLICATIONS

[1] **S. Shaier**, F. Pereira, K. von der Wense, L. Hunter, and M. Jones. More Experts Than Galaxies: Conditionally-overlapping Experts With Biologically-inspired Fixed Routing (**ICLR**) 2025.

[2] **S. Shaier**, G. Baker, C. Sridhar, K. von der Wense, and L. Hunter. MALAMUTE: A Multilingual, Highly-granular, Template-free, Education-based Probing Dataset (**ACL Findings**) 2025.

[3] **S. Shaier**, A. Kobren, and P. Ogren. Adaptive Question Answering: Enhancing Language Model Proficiency for Addressing Knowledge Conflicts with Source Citations. Empirical Methods in Natural Language Processing (**EMNLP**) 2024.

[4] **S. Shaier**, L. Hunter, and K. von der Wense. It Is Not About What You Say, It Is About How You Say It: A Surprisingly Simple Approach For Improving Reading Comprehension. Association for Computational Linguistics (**ACL Findings**) 2024.

[5] **S. Shaier**, L. Hunter, and K. von der Wense. Desiderata For The Context Use Of Question Answering Systems. European Chapter of the Association for Computational Linguistics (**EACL**) 2024.

[6] **S. Shaier**, K. Bennett, L. Hunter, and K. von der Wense. Comparing Template-based And Template-free Language Model Probing. European Chapter of the Association for Computational Linguistics (**EACL**) 2024.

[7] **S. Shaier**, L. Hunter, and K. von der Wense. Who Are All The Stochastic Parrots Imitating? They Should Tell Us!. Asia-Pacific Chapter of the Association for Computational Linguistics (**AACL**) 2023.

[8] **S. Shaier**, K. Bennett, L. Hunter, and K. von der Wense. Emerging Challenges In Personalized Medicine: Assessing Demographic Effects On Biomedical Question Answering Systems. Asia-Pacific Chapter of the Association for Computational Linguistics (**AACL**) 2023. **Won Social Impact Award.**

[9] **S. Shaier**, M. Raissi, and P. Seshaiyer. Data-driven approaches for predicting spread of infectious diseases through DINNs: Disease Informed Neural Networks (**Letters in Biomathematics**) 2022.

[10] **S. Shaier**, M. Burke. A Mathematical Model for the Effect of Domestic Animals on Human African Trypanosomiasis (Sleeping Sickness) (**KJUR**) 2019.

PREPRINTS

[1] **S. Shaier**. Excitation: Momentum For Experts (**arXiv**) 2026.

[2] **S. Shaier**, M. Guerrero, K. von der Wense. Asking Again and Again: Exploring LLM Robustness to Repeated Questions (**arXiv**) 2025.

[3] G. Baker, A. Raut, **S. Shaier,** L. Hunter, K. von der Wense. Lost in the Middle, and In-Between: Enhancing Language Models' Ability to Reason Over Long Contexts in Multi-Hop QA (**arXiv**) 2024.

HONORS & AWARDS

Outstanding Research & Paper Awards	· **University of Colorado Boulder**	2024-2025
Outstanding Student & Service Awards	· **University of Colorado Boulder**	2023-2024
Publication Recognition Award (5x)	· **University of Colorado Boulder**	2023-2024
James H. Martin Graduate Award	· **Nelson A. Prager Family Fund**	Nov 2023
Social Impact Award	· **AACL (1 of 363 international candidates)**	Aug 2023

SKILLS

Programming: Python, Bash, SQL, Java, MATLAB
Machine Learning: PyTorch, Torchtitan, HuggingFace, Large-scale training
MLOps: Distributed Systems, SLURM, Docker, Determined AI, UV, CI/CD, Git, Linux, OpenSearch, MCP, Jira
Project Management: strategic planning, budgeting, goal posting, delegation and supervision
Communication: scientific and analytical writing, public speaking and presenting, teaching and training