

BENJAMIN V. ROZONoyer

✉ brozonoyer@umass.edu · [brozonoyer.github.io](https://github.com/brozonoyer) · [brozonoyer](https://www.linkedin.com/in/brozonoyer) · [Google Scholar](https://scholar.google.com/citations?user=brozonoyer) · [LinkedIn](https://www.linkedin.com/in/brozonoyer)

ML researcher (PhD candidate, UMass Amherst) specializing in **discrete diffusion language models**, efficient LLM inference, and generative information retrieval. Publications at ICML, NeurIPS, AISTATS, ACL, and EMNLP; research experience at Google and Oracle Labs.

EXPERIENCE

Google Research / DeepMind — Student Researcher

Feb 2025 – Jun 2025

- Generative ranking for information retrieval: proved expressivity advantages over dual encoders and developed SToICaL, a rank-aware loss for LLM finetuning; results published (FoGen @ ICML 2026) and presented in an invited talk at Google DeepMind.

Oracle Labs — Research Assistant, ML Research Group (two internships)

Summers 2024, 2025

- Inference optimization for mixture-of-experts LLMs (2025); LLM agents for automated program repair (2024).

Charles River Analytics — Scientist Intern

Jun 2023 – Aug 2023

- Contrastive authorship embeddings (IARPA HIATUS); advanced CRA from last to 2nd place in program evaluation.

Raytheon BBN Technologies — Engineer II

Jun 2020 – Aug 2022

- Multilingual information extraction (DARPA AIDA); built full NLP stack for a COVID-19 machine reading system (IARPA); maintained internal NLP/deep learning libraries.

EDUCATION

PhD, Computer Science — University of Massachusetts Amherst

2022 – present

- Advisor: Andrew McCallum (IESL). Thesis research: discrete diffusion models for text generation.

MS, Computational Linguistics — Brandeis University

2020 – 2021

BS, Computer Science, Linguistics, and Mathematics — Brandeis University

2016 – 2020

SELECTED PUBLICATIONS

full list: brozonoyer.github.io/publications

- Rozonoyer, B.**, et al. [Learned Relay Representations for Forward-Thinking Discrete Diffusion Models](#). *SPIGM & FoGen @ ICML 2026*. Cut DLM inference latency by up to 32% while beating SFT on coding tasks.
- Rozonoyer, B.**, et al. [Autoregressive Ranking: Bridging the Gap Between Dual and Cross Encoders](#). *FoGen @ ICML 2026* (Google internship work).
- Patel, D.*, **Rozonoyer, B.***, et al. [Insertion Based Sequence Generation with Learnable Order Dynamics](#). *ICML 2026*.
- Patel, D., **Rozonoyer, B.**, et al. [A Continuous Time Markov Chain Framework for Insertion Language Models](#). *AISTATS 2026* ([Spotlight](#)).
- Rozonoyer, B.**, et al. [Learning Representations for Hierarchies with Minimal Support](#). *NeurIPS 2024*.

SKILLS

ML	PyTorch, JAX, Transformers, diffusion models, LLM training and inference, CUDA (NVIDIA-certified, 2024)
NLP / IR	Information extraction, retrieval, multilingual NLP, computational linguistics
Programming	Python, Julia, Java, CUDA C++ , SPARQL
Honors	ICML 2026 Silver Reviewer Award; Joshi Award for Outstanding Achievement in Computational Linguistics