Konstantin Golobokov

kogolobo@uw.edu ● Seattle, WA ● (610) 322-4132 ● GitHub ● LinkedIn Profile ● Personal Website

As an experienced applied machine learning researcher pursuing a Ph.D. in Applied Mathematics, I am seeking a **summer internship position** focused on developing **efficient deep learning methods**. With a strong foundation in applied machine learning, bolstered by impactful contributions at Microsoft and rigorous academic training, **I aim to advance the field through innovative research**.

Skills

- **Machine Learning:** Natural Language Generation, Representation Learning, Few-Shot Learning, Semantic Parsing
- **Coursework:** CUDA Programming, Convex Optimization, Numerical Linear Algebra, Statistical Learning
- **Computing:** C++ (CUDA, OpenMPI, OpenCV), Python (PyTorch, HuggingFace, pySpark), MATLAB, Slurm

Experience

Research Assistant, University of Washington, IFML, Seattle, WA — Jul 2024 – Present

- Researched **foundations of NLP** in **reasoning** and robust low-resource language **summarization**
- Used **efficient adaptation** of **LLMs** for reasoning graph generation and evaluate **out-of-domain generalization**
- **Collected 1.1 Tb of data** in 5 low-resource languages, trained robust **paraphrase** and **style transfer** models

Senior Applied Researcher, Microsoft Corporation, Azure AI, Redmond, WA — Sept 2023 – Feb 2024

- Researched **parameter-efficient training** for **LLMs** on multi-GPU clusters to reduce training cost
- Benchmarked open-source models on language modeling tasks, provided **finetuning quality comparison** across LLama and OpenAI models, reduced uncertainty for client teams and senior leadership
- Built quality monitoring infrastructure to **detect finetuning quality regressions**

Applied Researcher, Microsoft Corporation, Azure AI, Bellevue, WA — Aug 2022 – Sept 2023

- Built a research demo of **ChatGPT** augmented with domain-specific knowledge in-context; empowered **70+ customer teams** to onboard product scenarios, and recorded **10,000 users in 2 months**
- **Researched semantic parsing** approaches for code generation, focusing on low-domain programming languages; produced **1 publication** and **1 patent** application, **3 production launches**
- Led coordination with product teams, sourced feedback, and **decided project direction.**

Machine Learning Scientist, Microsoft Corporation, Bing Ads, Bellevue, WA — Sept 2018 – Aug 2022

- **Led 3 literature review and research planning** sessions, developed cutting-edge models for **controlled text generation** and unsupervised **representation learning, produced 2 publications**
- **Pitched research ideas** to the leadership team in **7 marketplace review meetings**, presented research results in **3 conferences** and **3 technical talks**
- Owned **6 production launches** in BingAds marketplace, producing up to **+10.65% ads revenue gain**

Algorithm R&D Intern, Lyrical Labs, Chicago, IL — May 2017 – Aug 2017

- Used a **random forest classifier** to detect salient regions in a video frame, **improved video encoding quality** on 5 customer video clips
- Wrote machine learning code in C++ OpenCV, optimized for performance, and integrated it into production

Education

B.Sc.Eng. Computer Science, University of Michigan – (3.8/4.0) — Aug 2018
Summa Cum Laude Honors, Varsity Wrestling Team Letter Winner

M.S. Applied & Computational Mathematics, University of Washington – (3.9/4.0) — Jun 2024

Ph.D. Applied Mathematics, University of Washington — Jun 2029

Selected Publications

- **Golobokov, K.**, Chai, J., Dong, V.Y., Gu, M., Chi, B., Cao, J., Yan, Y., Liu Y., 2022. DeepGen: Diverse Search Ad Generation and Real-Time Customization. *EMNLP 2022*
- Chai, J., Pryzant, R., Dong, V.Y., **Golobokov, K.**, Zhu, C., Liu Y., 2022. FAST: Improving Controllability for Text Generation with Feedback Aware Self-Training. *Preprint*

Project

System Design in C++ Search Engine Project, Ann Arbor, MI — Jan 2018 – Apr 2018

- Wrote a functioning search engine from scratch in C++; designed own data structures. Wrote 3000 lines of code.
- Optimized code performance with OS primitives. Crawled 8,000+ pages of Wikipedia.