ZHANGIR AZERBAYEV

zhangir-azerbayev.github.io ⋄ za2514@princeton.edu

EDUCATION

Princeton University *June 2023-Present*
Ph.D Computer Science, prospective
Research advisor: Jia Deng

Yale University *August 2019-May 2023*
B.S Mathematics
Research advisors: Dragomir Radev, Arman Cohan, Julian Jara-Ettinger

Stanford Online High School *August 2018-May 2019*
Part-time student. Courses: linear algebra, abstract algebra.

PAPERS

Publications:

Zhangir Azerbayev, Hailey Schoelkopf, et al. Llemma: An Open Language Model for Mathematics. *ICLR 2024*

Keiran Paster, Marco Dos Santos, **Zhangir Azerbayev**, Jimmy Ba. OpenWebMath: An Open Dataset of High-Quality Mathematical Web Text. *ICLR 2024*

Zhangir Azerbayev, Bartosz Piotrowski, Jeremy Avigad. ProofNet: A Benchmark for Formally Proving and Formalizing Undergraduate-level Mathematics. *MATH-AI workshop at NeurIPS 2022 contributed talk.*

Zhangir Azerbayev, Ansong Ni, Hailey Schoelkopf, Dragomir Radev. Explicit Knowledge Transfer for Weakly-supervised Code Generation. *Dl4Code Workshop at ICLR 2023.*

Ansong Ni, **Zhangir Azerbayev**, Mutethia Mutuma, Troy Feng, Yusen Zhang, Tao Yu, Ahmed Hassan Awadallah, and Dragomir Radev. SummerTime: Text Summarization Toolkit for Non-experts. *EMNLP 2021 System Demonstration.*

Preprints:

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, Jeremy Avigad. ProofNet: Autoformalizing and Formally Proving Undergraduate-Level Mathematics. *arXiv:2302.12433*

Marlene Berke, **Zhangir Azerbayev**, Mario Belledone, Zenna Tavares, Julian Jara-Ettinger. MetaCOG: Learning a metacognition to recover what objects are actually there. *arXiv:2110.03105*

OPEN SOURCE

Llemma

- Trained state-of-the-art open source language models for mathematics.
- 7B model. 34B model.

GPT-NeoX

- Contributions to one of the most popular language model pretraining libraries (6.1k stars).
- Fixed major numerical precision bug, contributed to FlashAttention-2 implementation.

proofGPT models

- Open-source pre-trained language models for mathematical tasks.
- Trained for 8 billion tokens on the proof-pile, a curated dataset of mathematical text.
- Huggingface: 1.3B weights., 6.7B weights.

Lean Chat

- A VS code extension for translating natural language theorem statements into Lean.
- Visual Studio Marketplace: `Lean Chat`
- Uses OpenAI's Codex API as a backend.

SummerTime: Toolkit for Automatic Text Summarization

- GitHub: `Yale-LILY/SummerTime`, 216 stars.
- One of four people responsible for the concept and early-stage design decisions.
- Changed 13,000 lines of code.

particle_filtering

- Github: `zhangir-azerbayev/particle_filtering`
- Raw numpy implementation of particle filtering for sequential Bayesian inference.

Lean Mathematical Library

- GitHub: `leanprover-community/mathlib`
- Contributed formalizations of alternating maps and exterior algebras.

WRITING

The Future of Interactive Theorem Proving?

- Guest post on Kevin Buzzard's Xena Project blog.
- Reached top 10 on Hacker News.