

Zihan Li

Email: zihanl13@illinois.edu

GitHub: <https://github.com/Prime-ppzi>

Education

University of Illinois at Urbana-Champaign

Expected May 2028

Master of Computer Science

University of Illinois at Urbana-Champaign

Graduated May 2026

Bachelor of Science, Mathematics and Computer Science

GPA: 3.98/4.00

- Selected Coursework: Statistics and Probability (A+), Numerical Analysis (A+), Data Structures (A), Computer Architecture (A+), Algorithms and Models of Computation (A), Machine Learning (A+)
- Graduate-level Coursework: Deep Learning Theory (A), Statistical Reinforcement Learning (A+)
- Dean's List: Fall 2022 - Fall 2024

Technical Skills

Programming Languages: Python, SQL, C++, C, Java, Assembly

Frameworks/Tools: PyTorch, DeepSpeed, DDP, FSDP, Accelerate, Weights & Biases, Slurm, Git, Docker

Data: Relational data modeling, multi-source integration, temporal joins, feature engineering, model evaluation

Languages: Mandarin (Native), English (Fluent)

Work & Research Experience

Generative AI Research — Diffusion LLM

Champaign, IL

Deep Learning Research Assistant (Advised by Prof. Alexander Schwing)

October 2025 - Present

- Co-authored paper on Diffusion LLMs (dLLMs), currently under review.
- Pioneered a novel architecture transitioning from discrete token sequences to continuous diffusion processes, significantly improving semantic coherence and sampling efficiency for generative modeling.
- Designed and engineered the end-to-end training pipeline from scratch using PyTorch, optimizing data loading, forward/backward passes, and model checkpointing.
- Scaled model training across multi-node Slurm GPU clusters using DDP and HuggingFace Accelerate.
- Integrated DeepSpeed and FlashAttention to overcome memory bottlenecks during large-scale model training, and utilized Weights & Biases for rigorous experiment tracking and hyperparameter tuning.

Optimizing a Generalized Moment of Inertia Amongst Tetrahedra

Champaign, IL

Research Assistant (Advised by Prof. Carrie Clark)

August 2024 - December 2024

- Applied Monte Carlo methods to explore trends in moment of inertia under geometric and mass variations.
- Designed and implemented an efficient algorithm in Python to compute exact solutions to the nonlinear optimization problem, reducing computation time by 23%.
- Validated theoretical conjectures through algorithmic simulation and identified hypotheses extensible to higher-dimensional spaces.
- Authored a comprehensive research report summarizing findings and future directions.

Personal Project Highlights

Flight Delay Analysis and Prediction Web Application

Fall 2026 - Present

- Processed and integrated 300,000+ U.S. flight records with historical NOAA weather observations using SQL for route, airline, and delay analysis.
- Built a leakage-safe evaluation pipeline with chronological splits to compare Logistic Regression with tree-based classifiers using precision, recall, F1, and ROC-AUC.
- Deployed the selected classifier through a backend prediction endpoint integrated with a relational database.

Sudoku Solver Using BERT-Like Transformer Architecture

June 2025

- Designed a lightweight 6-layer BERT-style transformer with 3D embeddings; optimized model architecture for structured reasoning, achieving 96% cell-level accuracy within only 5 epochs.
- Demonstrated the effectiveness of multi-dimensional embedding for lightweight transformers in structured reasoning tasks.

Infrastructure & DevOps Experience

Online

Linux Server Administrator

August 2019 - August 2024

- Engineered and maintained a high-availability Linux server infrastructure, supporting real-time multiplayer applications and high-bandwidth media streaming for a global community.
- Developed custom Python plugins and Bash automation scripts for remote system monitoring, resource allocation, and auto-recovery, effectively eliminating 90% of manual maintenance overhead.
- Optimized backend performance to ensure low-latency connectivity and stable routing for 100+ daily active users across multiple geographic regions.
- Collaborated with a distributed, cross-functional team of administrators to troubleshoot server anomalies and deploy seamless updates with zero downtime.