Selin Yıldırım

PhD Student in Computer Science

University of Illinois at Urbana-Champaign

+1447902-2251 | seliny2@illinois.edu | /selinyldrm | selinnilesy | https://selinyldrm.github.io/

Summary

4th year PhD student in Computer Science with a passion in producing impactful academic works at the intersection of Artifical Intelligence and High Performance Computing. Taking a role in a brilliant academic research group where cutting-edge innovations and team collaboration are basic priorities. An excellent communicator and team player with the ability to meet deadlines and provide deliverables.

Experience

- RESEARCH ASSOCIATE

California, US

Institution: Advanced Micro Devices

05/2025 - 08/2025

Speculative Decoding for Image Generation: Developed a novel speculative decoding algorithm for resolving the challenges in accelerating text-to-image generation. Achieved speedup up to 3.8x over vanilla image generation without quality loss.

- RESEARCH INTERN

Remote

Institution: NVIDIA

08/2024 - 12/2024

Towards Memory-Efficient Decoding on Constrained GPUs: Introduced a hardware-aware, memory-efficient and fast speculative decoding approach that leverages smart memory techniques to enhance text inference performance on consumer-grade GPUs.

- DEEP LEARNING PERFORMANCE INFERENCE INTERN

California, US

Institution: NVIDIA

05/2024 - 08/24

TensorRT-WinJIT/WinAI: Contibuted to TensorRT JIT compiler stack by implementing efficient deep learning compiler algorithms for Windows. Worked on a novel training-free KV cache compression algorithm for TensorRT-LLM.

- GRADUATE RESEARCH ASSISTANT

Illinois, US

Institution: Coordinated Science Laboratory, University of Illinois at Urbana-Champaign

05/2023 - Current

Efficient AI Algorithms and Systems: Improving speculative-decoding based generative AI techniques to provide faster and resource-friendly inference. Designed a resource efficient scheduling system for co-locating various deep learning jobs on NVIDIA GPUs with resource isolation. Worked hands-on with Transformers models, hypervisors, GPU drivers, Nvidia MPS, Nvidia MIG, CUDA, ROCM, Kubernetes and Docker.

- GRADUATE RESEARCH ASSISTANT

Illinois, US

Institution: Parasol Laboratory, University of Illinois at Urbana-Champaign

08/2022 - 05/2023

Exoplanet: Achieved upto 14x acceleration of NASA's exoplanet search algorithm on distributed GPUs. Worked hands-on with parallel graph algorithms as well as dynamic data-scheduling strategies.

- UNDERGRADUATE RESEARCH ASSISTANT

Turkiye

Institution: Department of Computer Engineering, Middle East Technical University

10/2021 - 08/2022

Sparsity Optimizations: Implemented a novel data compression algorithm for eliminating the sparsity in Skew-Symmetric Sparse Matrix-Vector Multiplication kernel. Our work was awarded the 2nd place in an academic research competition.

- HIGH PERFORMANCE COMPUTING INTERN, ERASMUS+

Germany

Institution: High Performance Computing Stuttgart (HLRS), University of Stuttgart

07/2021 - 10/2021

Scientific Kernel Optimizations: Worked on European Union's Horizon 2020 research project, namely *ExaHyPE*, as part of SPMT team. Optimized intermediate I/O operations in the kernel by using parallelism. Optimized ExaHyPE's Compressible Navier-Stokes with asynchronous tasks. Side project as paralell Sieve of Eratosthenes Prime Finder Algorithm with Intel TBB.

- COMPUTER VISION INTERN

Turkiye

Company: Argos AI, Middle East Technical University Technopark

10/2020 - 3/2021

Undergraduate Internship: Engineered computer vision in C++ to control airport cameras and capture point of views. Performed foreign object detection on image samples to eliminate foreign objects in airports.

- MACHINE LEARNING INTERN

Turkiye

Company: Otsimo, Middle East Technical University Technopark

06/2020-10/2020

Undergraduate Internship: Worked on a speech recognition algorithm inspired by Apple Siri. Trained models and optimized supervised learning.

Skills & Experience

Generative AI & Machine Learning (3 years): In-depth experience with deep learning frameworks and generative AI performance. Hands-on experience with autoregressive text and image models with continuous and discrete representations, as well as diffusion models.

Parallel Programming (5 years): Expertise in hand-written CUDA kernels, MPI, OpenMP, STAPL, Intel TBB for high-performance computing. Proficient in C++, Python, Java, Assembly, ANSI C, Golang, Haskell, Prolog, with a focus on performance-critical applications.

Architecture & Compilers (3 years): Advanced understanding of computer organization, instruction sets, and architectures (X86, RISC, CISC). Knowledge and experience in compilers such as Torch Inductor, JAX, TVM, XLA, LLVM and MLIR for efficient machine learning execution.

Systems Performance (5 years): Experience with Nsight Systems, Nsight Compute, Nvidia-smi, NVtop and Prometheus for in-depth performance profiling and tuning. Strong background in low-level performance analysis and improvements.

Education

University of Illinois at Urbana-Champaign — *Illinois, US*
Doctor of Philosophy in Computer Science — *2022 - Current*
Compiler, Architecture and Parallel Computing research track. **GPA:3.9**

Middle East Technical University — *Turkiye*
Bachelor of Science in Computer Science — *2017 - 2022*
Top ranked CS department with the highest research output in Turkey. Top-interests were Parallel Programming, Guided Research, Computer Architecture, Operating Systems and Quantum Computing.

Teaching

University of Illinois at Urbana-Champaign — *Illinois, US*
CS521 - Machine Learning and Compilers — *8/2025 - 12/2025*
Teaching TensorRT, JAX, TVM, and several popular compiler frameworks for machine learning to graduate students.

Publications and Conferences

11/2025	**Under Review,** Context Aware Speculative Decoding
10/2025	**ICCAD'25,** Invited Paper: Hardware-Software Co-Design for Highly Optimized, Customized, and Reliable AI Systems
05/2025	**ArXiv (Under Revision),** *SpecMemo: Speculative Decoding is in Your Pocket.*
07/2024	**SIAM'24,** *Contributed talk: PARS3: Parallel Sparse Skew-Symmetric Matrix-Vector Multiplication with Reverse Cuthill-McKee Reordering*
09/2021	**Euro MPI 2021,** Attendee

Leadership & Projects

2026-2027	**CSLSC Conference Chair,** Co-organizing a student-organized conference (CSLSC) at UIUC to bring together scientists from academia and industry to discuss cutting-edge research across multiple disciplines. Particularly serving as AI/ML Track session chair.
10/2021 - 07/2022	**Wanderlust,** Senior undergraduate project: A mobile application planning holistic touristic trips for deriving maximum travel satisfaction. Employs a novel, Multi-Variate, Traveling Salesman algorithm powered by AI-based combinatorics to obtain the best travel itinerary.

Certifications

10/2020	**Machine Learning, Stanford University, Grade:98.82/100,** *(Certificate)*
09/2020	**Problem Solving,** *(Certificate)*
09/2020	**C++,** *(Certificate)*
09/2020	**Intermediate C,** *(Certificate)*
11/2019	**Extracurricular : Work and Travel Exchange Program,** *(Certificate)*

Achievements

University

07/2022	***Second Best Research Award***, Guided Research, Department of Computer Engineering, METU	*Turkiye*
2022-	**Graduate Fellowship**, University of Illinois at Urbana-Champaign	*IL, USA*
2017-2022	**Undergraduate Fellowship**, *Sema Yazar Youth Foundation*	*Turkiye*
2018-2022	**High Honor Student (5x)**, Department of Computer Engineering, METU	*Turkiye*

Exams

2017	**Top 0.05% among 2.162.895 participants**, National University Entrance Exam (LYS)	*Turkiye*
2013	**Top 0.1% among 1.112.604 participants**, National High School Entrance Examination (SBS)	*Turkiye*

Languages

English	IELTS: 7.5 (C1) , European Online Linguistic Support: C1
German	A2
Latin	A1
Turkish	Native Language