Taeksang Peter Kim

tspeterkim3@gmail.com | 582-263-1348 | tspeterkim.github.io

EDUCATION

UIUC
MENG IN ELECTRICAL AND COMPUTER ENGINEERING
Dec 2025 | Champaign, IL

UCLA
BS + MS IN COMPUTER SCIENCE
Jun 2017 + Dec 2018 | Los Angeles, CA

INTL. SCHOOL OF BRUSSELS
VALEDICTORIAN
June 2013 | Brussels, Belgium

PROJECTS

MINIMAL FLASH ATTENTION
Flash Attention in ~100 lines of CUDA
900+ Stars on Github
Made the **front page** of Hacker News

ONE BILLION ROW CHALLENGE IN CUDA
First CUDA solution to the challenge
Runs in 17s on a V100
Made the **front page** of Hacker News

OUT-OF-ORDER RISC-V CPU
Speculative, RV32IM spec compliant CPU with Early Branch Recovery
Ranked 9th in class competition

LINKS

Github:// **tspeterkim**
LinkedIn:// **tspeterkim3**
X:// **@tspeterkim**

SKILLS

PROGRAMMING
Python • C++ • Go • Java • SQL • Shell

TOOLS
AWS • PyTorch • TensorFlow • Pandas
Ray • Spark • Jenkins • Airflow

EXPERIENCE

ANNAPURNA LABS (AWS) | SDE INTERN
May 2025 - Aug 2025 | Seattle, WA

- Achieved 2M-token context length for LLM inference on NeuronX by implementing Sliding Window Attention and Windowed Context Encoding.
- Extended FlashAttention NKI Kernel to support Sliding Window mask.

COUPANG | SENIOR ML ENGINEER
Aug 2022 - Aug 2024 | Seoul, South Korea

- Built the recommendation system for the largest South Korean streaming service by MAU.
- Trained a next video prediction model and productionized it as the default ranking algorithm, raising overall playback time by 20%.
- Applied Thompson Sampling to deliver artwork personalization, raising feed engagement by 14%.

BUZZVIL | ML ENGINEER (COMPULSORY MILITARY SERVICE)
Mar 2019 – Jul 2022 | Seoul, South Korea

- Improved e-commerce ads for Buzzvil, South Korea's 3rd largest mobile advertising platform (after Google and Kakao).
- Increased ad spending return by 30% through a user-action sequence based BERT model and an Item CF model.
- Designed a company-wide introductory ML course, and made it the most popular study in the company (15% of the entire company enrolled).

LINKEDIN | ML ENGINEERING INTERN
Jun 2018 – Sep 2018 | Mountain View, CA

- Trained job response prediction models for the search page that scaled up to 500M+ members.

PAYPAL | SOFTWARE ENGINEERING INTERN
Jun 2016 – Sep 2016 | San Jose, CA

- Deployed Apache Eagle on top of existing Hadoop clusters as the first real-time log monitoring tool deployed in production.

RESEARCH

ADAPT @ UIUC Jan 2025 – Current | Champaign, IL
Advised by Prof. Charith Mendis

- Wrote Triton kernel for fusing QKV projection with data movement operators for 4X faster LLM decoding over TensorRT.
- Optimized softmax CUDA kernel by using vectorized loads and avoiding atomic adds through local register adds.

PATENTS

[1] Jina Hwang and Taeksang Kim. Sports awareness, mobile feed video player, 2023. Korean Patent No. 10-0071581, Issued Feb 2024.

[2] Taeksang Kim. Next watch prediction using GPT, 2023. Korean Patent No. 10-0036443, Issued Dec 2023.