Shengjia Zhang

shengjia.zhang@zju.edu.cn | https://cynthia-shengjia.github.io

Education

Zhejiang University, College of Computer Science and Technology — Hangzhou, China
M.Eng. in Computer Science and Technology; Advisor: Jiawei Chen — *09/2024 – Present*

Chongqing University, College of Big Data and Software — Chongqing, China
B.Eng. in Data Science and Big Data Technology; GPA: 3.87/4.00 (Avg: 91.84/100); Rank: 5/256 — *09/2020 – 06/2024*

Publications (* denotes equal contribution)

Talos: Optimizing Top-K Accuracy in Recommender Systems
Shengjia Zhang*, Weiqin Yang*, Jiawei Chen, Peng Wu, Yuegang Sun, Gang Wang, Qihao Shi, Can Wang
WWW 2026 (CCF-A)

Advancing Loss Functions in Recommender Systems: A Comparative Study with a Rényi Divergence-Based Solution
Shengjia Zhang, Jiawei Chen, Changdong Li, Sheng Zhou, Qihao Shi, Yan Feng, Chun Chen, Can Wang
AAAI 2025 (CCF-A)

Breaking the Top-K Barrier: Advancing Top-K Ranking Metrics Optimization in Recommender Systems
Weiqin Yang, Jiawei Chen, **Shengjia Zhang**, Peng Wu, Yuegang Sun, Yan Feng, Chun Chen, Can Wang
KDD 2025 (CCF-A)

OThink-R1: Intrinsic Fast/Slow Thinking Mode Switching for Over-Reasoning Mitigation
Shengjia Zhang*, Junjie Wu*, Jiawei Chen, Changwang Zhang, Zhe Li, Xingyu Lou, Wangchunshu Zhou, Sheng Zhou, Can Wang, Jun Wang
Arxiv 2025

Selected Awards

National Scholarship, **Zhejiang University** — 2025
Huawei Scholarship (2/10 Candidates), **Chongqing University** — 2023
Outstanding Winner & **Frank Giordano Award** (1/15105), Mathematical Contest in Modeling (MCM) [Certificate] — 2022

Skills

Languages: C++, Python, CUDA
Frameworks & Tools: PyTorch, Triton

Research Experience

Ranking Optimization in Recommender Systems — 09/2023 – 03/2025
(Distributionally Robustness Optimization → Top-K NDCG Optimization → Top-K Accuracy Optimization)

- **Talos: Optimizing Top-K Accuracy in Recommender Systems**
  (1) Developed Talos, a differentiable loss that serves as a tight upper bound for Top-K accuracy, resolving the inconsistency between the optimization objective and evaluation metrics.
  (2) Reformulated the expensive global sorting ($O(N \log N)$) into comparisons between prediction scores and a learnable Top-K threshold ($O(N)$), employing a novel unbiased quantile regression technique for precise Top-K threshhold estimation.
  (3) Established theoretical guarantees: (a) equivalence to *Distributionally Robust Optimization (DRO)* for stability against distribution shifts; (b) the convergence guarantee the proposed optimization algorithm.
  Contributions: idea, implementation & experiments, proof, writing.
- **DrRL: The *Distributionally Robustness Optimization* based Loss Function in Recommender Systems**
  (1) Proposed a generalized robust loss framework by incorporating **Rényi Divergence** into Distributionally Robust Optimization (DRO), theoretically unifying Softmax Loss and Cosine Contrastive Loss as special cases.
  (2) Derived a novel loss function where the robustness intensity is flexibly controlled by the Rényi order parameter α, enabling adaptive defense against diverse distribution shifts and achieving superior performance.
  Contributions: idea, implementation & experiments, proof, writing.

- **SL@K: Optimizing NDCG@K in Recommender Systems**
  (1) Developed SL@K, a differentiable surrogate loss that serves as a tight upper bound for NDCG@K, by reformulating the Top-K truncation into efficient score-threshold comparisons.
  (2) Adopted a Monte-Carlo method for efficient Top-K threshold estimation.
  Contributions: idea, implementation on Transformer-based recommender

Adaptive Fast/Slow Thinking for Large Reasoning Models 03/2025 – 08/2025

- **OThink-R1: Enabling Large Reasoning Models to Adaptively Switch between Fast and Slow Thinking**
  (1) Proposed OThink-R1, a hybrid reasoning framework that enables a single model to autonomously select fast thinking (direct response) or slow thinking (chain-of-thought reasoning) based on problem difficulty.
  (2) Identified three essential and three redundant reasoning patterns from model trajectories, and designed an LLM-based judge to classify reasoning necessity for constructing a hybrid fine-tuning dataset.
  (3) Introduced a dual KL-divergence constraint anchoring to both a reasoning model and a non-reasoning model, preventing mode collapse during training.
  Contributions: idea, implementation & writing, experiments

Research Interests

I am interested in *efficient machine learning*, with a focus on designing hardware-aware algorithms to accelerate the training and inference of large-scale models. **I am deeply familiar with FlashAttention**, and have implemented it from scratch in Triton. I also have hands-on experience with CUDA optimization techniques (shared memory tiling, register blocking, and vectorized memory access). I am particularly drawn to efficient attention mechanisms and their integration into scalable serving and training systems.

Industrial Experience

OPPO Research Institute Shenzhen, China
Towards optimizing the large reasoning model by dynamically allocating reasoning depth and computational resources based on task complexity—fast responses for simple queries, deep reasoning for complex problems. *03/2025 – 08/2025*