Zhongzhu (Charlie) Zhou

Senior Research Scientist, Turbo & Kernel, Together AI
2 Townsend St., San Francisco, CA 94107 · (+1) 4153109490
zhongzhu.zhou@sydney.edu.au · zzz879978@outlook.com
Homepage · Google Scholar · GitHub · LinkedIn · DBLP · November 9, 2026

Research Highlights

- **EFFICIENT ML ARCHITECTURE:** efficient speculator design; efficient attention design
- **EFFICIENT POST-TRAINING:** efficient loss design; quantization
- **EFFICIENT POST-TRAINING ENGINES:** post-training engines; inference engines
- **AGENTS:** coding agents; AI for science

Work

- **Together.AI** — May. 2024 - Present
  Senior Research Scientist, Turbo & Kernel — *San Francisco, United States*
- **FutureMLS — Nonprofit Research Lab** — 2026 - Present
  Founder & Principal Investigator (PI) · [*futuremls.org*] — *San Francisco, United States*
  - *Mission:* make ML efficient and practical at scale via open-source algorithm–system co-design.
- **Dolby** — Mar. 2024 - Sep. 2024
  Research Intern — *Sydney, Australia*
- **DeepSpeed Team, Microsoft** — Mar. 2023 - Feb. 2024
  Research Intern — *Sydney, Australia*
- **WeChat, Tencent & UIUC** — Jul. 2018 - Jul. 2020
  Research Intern, Technical Architecture — *Champaign, United States + Guangzhou, China*
- **Microsoft (China) Co., Ltd., Guangzhou Branch** — Sep. 2018 - Feb. 2019
  Project Assistant to Senior Cloud Architect — *Guangzhou, China*

Education

- **The University of Sydney (USYD)** — Sydney, Australia
  Doctor of Philosophy (Ph.D.) — *Oct. 2022 - Feb. 2026*
  - Accumulated GPA: **4.0/4.0, High Distinction (HD)**
- **The University of Sydney (USYD)** — Sydney, Australia
  Visiting Scholar — *Mar. 2022 - Oct. 2022*
- **Sun Yat-sen University (SYSU)** — Guangzhou, China
  Research Associate — *Sep. 2019 - Mar. 2022*
  - Accumulated GPA: **3.41/4.0**
- **University of Illinois Urbana-Champaign (UIUC)** — Champaign, United States + Guangzhou, China
  Summer Session Student — *Jun. 2018 - Sep. 2018*
- **Sun Yat-sen University (SYSU)** — Guangzhou, China
  Bachelor of Engineering in Computer Science and Technology — *Sep. 2015 - Jun. 2019*
  - Overall GPA: **3.9/4.0**
- **Changjun High School** — Changsha, China
  Senior Student — *Sep. 2012 - Jun. 2015*

Professional Service

- **Member:** IEEE (ID: 97841404); ACM (ID: 6708618); China Computer Federation (CCF; ID: B8293G)
- Reviewer for Conferences
  - IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2027)
  - The Fifteenth International Conference on Learning Representations (ICLR 2027)
  - Fortieth Conference on Neural Information Processing Systems (NeurIPS 2026)
  - The 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP 2026)
  - The 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)
  - The 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)
  - Thirty-ninth Conference on Neural Information Processing Systems (NeurIPS 2025)
  - Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)
- Reviewer for Journal
  - IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
  - IEEE Transactions on Computers
- Program Committee
  - 1st International Workshop on Sustainable and Efficient Language, Vision, and Action Models (SELVA 2026) at ACL 2026
  - Computer Science Research Methods (CSRM 2023) (INFO5993/ INFO4990 in University of Sydney)
  - ACM International Conference on Architectural Support for Programming Languages and Operating Systems Artifact Evaluation Committee (ASPLOS'24 AEC)
- Web Chair
  - 32nd IEEE International Symposium on High-Performance Computer Architecture (HPCA 2026)

Publications

Book

- ***C Language Programming in Chinese*** — **Tianjin University Press** · *ISBN: 9787561847251* · [Book]

Conference Paper

- ***Did the Judge See the Answer? A Crossed Dose–Response Audit of LLM Relevance Judgments*** — **WSDM 2027 Main**
- ***A Fourier Analysis of Context Dependence in Retrieval-Augmented Generation*** — **WSDM 2027 Findings**
- ***You Cannot Retrieve Your Way Out of a Coarse Index: A Sketching Theory of Chunk Size*** — **WSDM 2027 Findings**
- ***On-Policy Residual Learning for Diffusion Language Models*** — **arXiv** + under review · [Paper], [Project], [Code], [Code (Engine)], [Hugging Face]
- ***When Does a Speech Model Know to Hand Off? Reading Competence from Hidden States for Real-Time Escalation*** — **NeurIPS 2026 Workshop RTCA** · [Paper]
- ***Orthros: Phase-Aware Heterogeneous Attention for Efficient Transformers*** — **NeurIPS 2026** · [Paper]
- ***Osprey: Target-agnostic Pre-training Makes Stronger Drafters in Speculative Decoding*** — **EMNLP 2026** · [Paper], [Code], [Hugging Face]

- *Tail-Likelihood Reinforcement Learning* — **arXiv** + under review · [Paper], [Project], [Code]
- *Massive Activations in Hybrid Linear Attention Large Language Models: Pre-Attention Spikes and Inter-Spike Plateaus* — **arXiv** + under review · [Paper], [Code], [Media Coverage]
- *OSCAR: Offline Spectral Covariance-Aware Rotation for 2-bit KV Cache Quantization* — **NeurIPS** (**Conference on Neural Information Processing Systems**) **2026** · [Paper], [Project], [Code], [Hugging Face]
- *Taylor-Calibrate: Principled Initialization for Hybrid Linear Attention Distillation* — **arXiv** + under review · [Paper], [Code]
- *I-DLM: Introspective Diffusion Language Models* — **COLM** (**The Third Conference on Language Modeling**) **2026** · [Paper], [Project], [Code], [Hugging Face]
- *Squeeze Evolve: Unified Multi-Model Orchestration for Verifier-Free Evolution* — **COLM** (**The Third Conference on Language Modeling**) **2026** · [Paper], [Project], [Code]
- *Compute Allocation in Bilinear Contrastive Representation Learning via Scaling Laws* — **COLM 2026 Workshop on Efficient Reasoning** (**Spotlight**) · [Paper], [Workshop]
- *From One-Pass SGD to Data Reuse: Mini-Batch Scaling Laws in Sketched Linear Regression* — **arXiv** + under review · [Paper]
- *SAW-INT4: System-Aware 4-Bit KV-Cache Quantization for Real-World LLM Serving* — **arXiv** + under review · [Paper], [Code]
- *CARE: Covariance-Aware and Rank-Enhanced Decomposition for Enabling Multi-Head Latent Attention* — **ICLR** (**The Fourteenth International Conference on Learning Representations**) **2026** · [Paper], [Project], [Code]
- *Aurora: When RL Meets Adaptive Speculative Training – A Unified Training-Serving System* — **ICML Forty-third International Conference on Machine Learning 2026** · [Paper], [Project], [Project], [Code]
- *KITTY: Accurate and Efficient 2-bit KV Cache Quantization with Dynamic Channel-wise Precision Boost* — **MLSys** (**Ninth Annual Conference on Machine Learning and Systems**) **2026** · [Paper], [Code]
- *CoderForge-Preview: SOTA open dataset for training efficient coding agents* — **Together AI Blog** · [Project], [Hugging Face]
- *Efficient Compression Algorithm-System Co-Design for Large-Scale Model Training and Inference* — under review · **Ph.D. Thesis**, University of Sydney
- *Understanding and Steering the Cognitive Behaviors of Reasoning Models at Test-Time* — **NeurIPS 2025 Workshop on Efficient Reasoning** · [Paper], [Project (Workshop)], [Code]
- *Ladder-Residual: Parallelism-Aware Architecture for Accelerating Large Model Inference with Communication Overlapping* — **ICML Forty-second International Conference on Machine Learning 2025** · [Paper], [Code]
- *Imitate Optimal Policy: Prevail and Induce Action Collapse in Policy Gradient* — **arXiv** + under review · [Paper]
- *CorDA: Context-Oriented Decomposition Adaptation of Large Language Models for Task-Aware Parameter-Efficient Fine-tuning* — **NeurIPS** (**Annual Conference on Neural Information Processing Systems**) **2024** · [Paper], [Code], [Hugging Face]
- *Quant-LLM: Accelerating the Serving of Large Language Models via FP6-Centric Algorithm-System Co-Design on Modern GPU* — **ATC** (**USENIX Annual Technical Conference**) **2024** · [Paper], [Code]
- *Flash-LLM: Enabling Cost-Effective and Highly-Efficient Large Generative Model Inference with Unstructured Sparsity* — **VLDB** (**International Conference on Very Large Databases**) **2024** · [Paper], [Code]
- *JSidentify: A Hybrid Framework for Detecting Plagiarism Among JavaScript Code in Online Mini Games* — **ICSE** (**International Conference on Software Engineering**) **2020** · [Paper (DOI)], [Paper], [Talk]

Journal Paper

- *RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Models* — **TPAMI** (**IEEE Transactions on Pattern Analysis and Machine Intelligence**) · Track: *Survey Papers* · [Paper]
- *Binary Neural Network for Automated Visual Surface Defect Detection* — **Sensors MDPI** (**Multidisciplinary Digital Publishing Institute**) · Track: *Special Issue Intelligent Sensing and Monitoring for Industrial Process* · [Paper]

Preprint Paper

- *FlashState: Unlocking Flexible Hybrid State Management for Agentic LLM Serving* — **alphaXiv 2026** · [Paper], [Project], [Code]
- *SinkState: On-the-Fly SVD of Recurrent States for Efficient Hybrid LLM Serving* — **alphaXiv 2026** · [Paper]
- *QATFactory: A Versatile, Deployment-Aligned Framework for Quantization-aware Training and Distillation of LLMs* — **arXiv 2026** · [Paper], [Code]
- *Linear RNN Scaling Laws: When Longer Sequences Beat More Sequences* — **arXiv 2026** · [Paper]
- *DeepSpeed-Chat: Easy, Fast and Affordable RLHF Training of ChatGPT-like Models at All Scales* — **arXiv 2023** · [Paper], [Project], [Code]
- *DeepSpeed4Science Initiative: Enabling Large-Scale Scientific Discovery through Sophisticated AI System Technologies* — **arXiv 2023** · [Paper], [Project], [Project], [Code]

Talks

- **Beyond the Paper: Bringing 2-bit KV Cache to Production** — Global C++ and System Software Technology Conference, Beijing, November 20–21, 2026 (upcoming) [Conference]
- **Efficient AI That Keeps Learning and Improving - Architectures, Post-Training, and Adaptive Inference** — GlobalAI-2026; The University of Sydney, October 5, 2026 [Program]
- **Toward AGI: Efficient Compression across the Data–PostTraining–Inference Loop** — Invited Virtual Tech Talk, Pluralis Research, September 30, 2026
- **Aurora: Adaptive Speculative Training as Online Reinforcement Learning** — Cornell RL Seminar Weekly, August 1, 2026
- **Compression-Centric Systems for Large-Model Post Training & Inference** — Invited Talk, Amazon, July 22, 2026
- **Frontier Research: Scaling Agentic RL & Verified Reasoning** — AI-tonomy Summit (Models & Agents), AI Researcher Forum, Plug and Play Tech Center, Silicon Valley, June 5, 2026
- **I-DLM: Introspective Diffusion Language Models** — Ant Group (蚂蚁集团), April 2026
- **CARE: Covariance-Aware and Rank-Enhanced Decomposition for Enabling Multi-Head Latent Attention** — ICLR (International Conference on Learning Representations) 2026
- **Panel Discussion: AI in Cross-Border Digital Technologies Infrastructure Platforms & Global Launchpads** — AI in the Cross-Border Digital Technologies Ecosystem: Infrastructure, Platforms & Global Launchpads, Nov, 20, 2025
- **JSidentify: A Hybrid Framework for Detecting Plagiarism Among JavaScript Code in Online Mini Games** — ICSE (International Conference on Software Engineering), Jul, 11, 2020

Honor and Awards

- **[2024] APR Intern Program Scholarship** (**SC3600**), University of Sydney
- **[2023–2025]** Progress Evaluation: satisfactory or excellent, University of Sydney
- **[2022] Jingdong Technology** (**JD**) **Co Ltd Research Scholarship in Artificial Intelligence**, University of Sydney
- **[2021]** SYSU Overseas Visiting and Collaborative Research Program Funding Plan, Sun Yat-sen University
- **[2020–2022]** The Third Class Scholarship ×3, Sun Yat-sen University
- **[2019] Research Honor Degree**, Sun Yat-sen University
- **[2019]** The Second Class Scholarship (Top 15% of the major), SYSU
- **[2018] The First Class Scholarship** (**Top 5%** of the major), SYSU × 2 (2015–16 & 2017–18)
- **[2018]** Illinois Computer Science Summer Research Program, UIUC; The Third Prize, Microsoft Hackathon, South China

- **[2017]** Meritorious Winner, COMAP's Mathematical Contest in Modeling, United States; The Second Prize, Student Innovation Software Development Competition, SYSU; The Third Prize, ACM-ICPC, SYSU; The Third Prize, The Chinese Mathematics Competitions
- **[2016]** **National Scholarship** (**Top 1** of the major), China
- **[2016]** The Second Prize, The Chinese Mathematics Competitions; The Second Class Scholarship (Top 15% of the major), SYSU (2016–2017)
- **[2013]** The Second Prize, National Olympiad in Informatics in Provinces (NOIP), Changjun High School

Teaching

- **COMP3520: Operating Systems Internals**
  Fall 2023, Tutor, The University of Sydney

Selected Professional Experience

- **Turbo & Kernel, Together.AI** May. 2024 - Present
  Senior Research Scientist *San Francisco, United States*
  - **Together Post-Training:** RL/SFT service with custom training loops, multi-LoRA and full-weight training. [Project], [Code (Cookbook)]
  - **Dissecting NVIDIA Blackwell and Blackwell Ultra: a multi-level microbenchmark study**
  - **Differentiable harness**
  - **Learning token-population geometry for sparse mixture-of-experts**
  - **QATFactory: a versatile, deployment-aligned framework for quantization-aware training and distillation of LLMs** [Paper], [Code]
  - **QuantProxy: evaluating cheap proxies for quantized LLMs**
  - **Hadamard rotation for NVFP4 KV caches: no measurable gain and a zero-code penalty**
  - **When does a speech model know to hand off? Reading competence from hidden states for real-time escalation (NeurIPS 2026 Workshop RTCA)** [Paper]
  - **ColdGPTQ: document-diverse calibration and count-adaptive damping for MXFP4 MoE quantization**
  - **On-policy residual learning for diffusion language models (ORL):** Co-advised on-policy residual learning and entropy contraction for faster diffusion decoding. [Paper], [Project], [Code], [Code (Engine)], [Hugging Face]
  - **Orthus + Loop model**
  - **FoldDecoding: accelerating decode attention over 2- and 4-bit KV caches through affine folding**
  - **Tail-likelihood reinforcement learning (TailRL):** Developing maximum-likelihood-style RL objectives that use pass@k likelihood expansions to reweight low-success tail prompts and preserve solution diversity. [Paper], [Project], [Code]
  - **Mix architecture speculator**
  - **Scaling law of speculators**
  - **LCFS:** Contributed to design discussions around hierarchical fairness, vruntime-style accounting, and weight partitioning across distributed serving instances.
  - **AgentGo:** Helped build the staged system path from telemetry and shadow prediction to offline replay, observability, and config-gated runtime integration for prediction-aware scheduling.
  - **Taylor-Calibrate:** Led the Taylor-guided initialization method for converting pretrained Transformers into hybrid linear-attention students by calibrating value projections, recurrent timescales, write gates, and output gates. [Paper], [Code]
  - **Agent Evolve:** Built a bio-inspired LLM multiagent framework with pheromone-style memory, evolutionary division of labor, and skill inheritance for open-ended population adaptation.
  - **OSCAR:** Led the design of an offline spectral covariance-aware rotation method for 2-bit KV-cache quantization that calibrates attention-aware key/value rotations and clipping thresholds. [Paper], [Project], [Code], [Hugging Face]
  - **SmartKV**
  - **I-DLM:** Helped build an inference path for introspective diffusion language models that reuses AR serving optimizations while supporting introspective strided decoding. [Paper], [Project], [Code], [Hugging Face]
  - **Squeeze Evolve:** Owned coding-benchmark execution and evaluation pipelines, especially for LiveCodeBench V6, and supported ablations on routing thresholds and aggregation behavior across AIME 2025 and HMMT 2025. [Paper], [Project], [Code]
  - **SAW-INT4:** Supported fused rotation-quantization kernel integration for paged KV-cache serving, matching plain INT4 throughput while recovering most accuracy lost by naive INT4. [Paper], [Code]
  - **Osprey:** Coauthored Osprey, which combines target-agnostic pre-training with target-specific adaptation to turn pretrained small language models into efficient, transferable drafters. [Paper], [Code], [Hugging Face]
  - **XoRL:** Built much of the training-side RL framework, including agent PPO trainers, asynchronous rollout and pipeline-training paths, and the execution flow that converts multi-turn agent-environment interaction into PPO and GRPO training batches. [Project (EN)], [Project (ZH)], [Project (Docs)], [Code]
  - **CoderForge:** Led the training pipeline for OpenHands R2E-Gym & SWE-Bench-scale data: curated high-signal SWE-smith / Rebench examples and fixed attention-mask plus position-ID issues in XoRL. [Project], [Hugging Face]
  - **CREST:** Co-developed the core idea of a training-free test-time steering method that identifies and modulates cognitive attention heads, improving accuracy by up to 17.5% and reducing token usage by 37.6% across reasoning benchmarks. [Paper], [Project (Workshop)], [Code]
  - **Ladder-Residual:** Implemented and optimized the `gpt-fast` inference path with CUDA Graphs and PyTorch compile ("reduce-overhead") for large-model serving. [Paper], [Code]
  - **CARE:** Ran the full experimental suite and carried out `vLLM` integration and theoretical analysis. [Paper], [Project], [Code]
  - **LEXICO** [Paper], [Code]
  - **Pulsar / SGLang:** Designed and implemented KV-cache prompt caching for the Phoenix speculator in Pulsar, stabilizing acceptance rates and reducing end-to-end latency. [Project (SGLang Docs)], [Code (SGLang)]
  - **Axolotl:** Integrated long-context attention (Ulysses, Ring Attention) into Axolotl and supported SFT data flow from successful trajectories to extend supervised post-training to larger context windows. [Project (Docs)], [Code]
- **Dolby** Mar. 2024 - Sep.2024
  Research Intern *Sydney Australia*
  - **Extreme Efficient Video Coding System:** Designed Optimal Brain Stride-wise Quantization (OBSQ), a domain-specific quantization methodology that selectively compresses neural network weights based on content type (e.g., video conferencing, gaming), enabling real-time 1080p performance with minimal quality loss.
- **DeepSpeed Team, Microsoft** Mar. 2023 - Feb. 2024
  Research Intern *Sydney Australia*
  - **DeepSpeed-Chat:** Apply INT4 and INT8 quantization to the RLHF pipeline, increase the batch size and improve the speed of the training and generation phases of RLHF without significantly compromising accuracy. [Paper], [Project], [Code]
  - **DeepSpeed4Science:** Revised the blog content, font size, technical research architecture, and code related to GenSLMs-'Megatron-DeepSpeed for Large-Scale AI4Science Model Training'. [Paper], [Project], [Project], [Code]

- **Future System Architecture (FSA) Lab, The University of Sydney (USYD)** Mar. 2022 - Feb. 2026
  Visiting Scholar, Ph.D. student *Sydney Australia*
  - **CorDA:** Designed the experimental methodology for evaluating task-aware parameter-efficient fine-tuning (dataset selection, baselines, and metrics across math/code/instruction-following). [Paper], [Code], [Hugging Face]
  - **Reinforcement Learning:** Conducted experiments applying ETF classifiers to 5+ neural networks in 10+ discrete-action reinforcement learning environments (e.g., Atari, Gym Classic) [Paper]
- **School of Computer Science and Engineering, SYSU** Sep. 2019 - Mar. 2022
  Research Associate *Guangzhou, China*
  Advisor: *Dan Huang* (Professor, SYSU)
  - **Pre-Expedite:** Utilized POSIX to create ZERO file blocks (Loop Device). Established a VFS within the ZERO file blocks, allowing each user to store small files in their designated ZERO file blocks. [Code]
- **Weixin Group, Tencent Holdings Ltd. & Dep. of CS, UIUC** Jul. 2018 - Jul. 2020
  Research Intern, Testing, Technical-Architecture Department *Champaign, United States + Guangzhou, China*
  - **JSidentify: Developed an edit distance estimation and network flow algorithm** to measure similarity in bytecode generated by Ignition, TurboFan Interpreter. [Paper], [Paper (DOI)], [Talk]

Selected Other Projects

- **LeetCode Record** (Jun. 2017 - Present) — Utilized languages such as C, CPP, Python3, Java, and Go to solve LeetCode algorithm questions based on my preference. [Code]
- **System Related Conference Papers Crawler** (Jun. 2021 - Present) — Leveraged Python, BeautifulSoup4, and Requests to scrape papers and crucial deadlines for major computer system conferences. [Code]
- **DDLs** (Dec. 2017 - May. 2018) — Developed DDLs, an Android application for personal deadline management, using Java and Android Studio for the front-end, incorporating MVC architecture, and NodeJS with Express.js for the back-end RESTful API. [Code (Backend)], [Code (Frontend)]
- **ChainLoveHelp** (May. 2018 - May. 2018) — ChainLoveHelp is dedicated to providing a peer-to-peer platform for university task posting and processing based on blockchain technology.

Conference Participation

• COLM 2026; ICLR 2025/26; HPCA, PPoPP, CGO & CC 2026; GTC 2025; ATC 2024; ASPLOS 2023; SC / ISCA 2021; ICSE 2020.

Certification

• Pearson Test of English: Overall: 67 (March, 21, 2024); IELTS Test Scores: Overall: 6.5 (October, 08, 2020); Coursera: Sample-based Learning Methods and Fundamentals of Reinforcement Learning (2021).

Skills

- *Efficient ML Architecture:* **Speculative decoding**; attention architecture conversion; parameter-efficient fine-tuning; **test-time reasoning**.
- *Quantization & Compression:* Low-bit weight and **KV-cache** quantization; mixed-precision policies; **rotation** and **calibration**; knowledge and activation **distillation**.
- *Post-Training & **Reinforcement Learning**:* **SFT**; **RLHF**; **PPO/GRPO**; likelihood-based RL objectives; advantage estimation; asynchronous rollouts.
- *Distributed Training & Inference:* Tensor and sequence parallelism; **FSDP/ZeRO**; **MoE-LoRA**; **vLLM/SGLang**; **KV-cache** management; serving scheduling.
- *GPU Kernel Optimization:* **CUDA/Triton**; **Tensor Core** sparse kernels; fused **rotation**-quantization kernels; **CUDA Graphs**; **torch.compile**.
- *Agents, Evaluation & Scientific ML:* Coding-agent training; **SWE-Bench/LiveCodeBench** evaluation; **Pass@k**; AI4Science data I/O; **diffusion models**; **neural video compression**.
- *HPC & Program Analysis:* Parallel/distributed file systems; heterogeneous resource scheduling; application profiling; bytecode analysis and code-clone detection.
- *Programming Languages:* Pascal (11 yrs), **C** (11 yrs), **C++** (11 yrs), **Python** (6 yrs), HTML, CSS, JavaScript (6 yrs), Java (6 yrs), SQL (6 yrs), Bash (6 yrs), LaTeX (5 yrs), Matlab (5 yrs), **CUDA** (5 yrs), R (4 yrs), Go (4 yrs), **Triton** (2 yrs)
- *Systems and Infrastructure:* ***MPI/OpenMPI*** (6 yrs), **Linux Kernel** (5 yrs), Distributed/Parallel File Systems - e.g., **Lustre**, **HDFS** (5 yrs), **Kubernetes**, **Kubernetes** Scheduler, **Kubernetes** SR-IOV (5 yrs), **Docker** (5 yrs), Hadoop (4 yrs), Spark (4 yrs), YARN (4 yrs), Mesos (4 yrs), **NVLink** (2 yrs), **NVshMem** (2 yrs), TensorCore, CudaCore Programming (2 yrs)
- *Machine Learning and AI:* **TensorFlow** (5 yrs), **PyTorch** (4 yrs), TorchServe (4 yrs), TensorBoard (4 yrs), Ray (4 yrs), **JAX** (2 yrs), **Triton** (2 yrs), **DeepSpeed** (1 yr), **HuggingFace** (1 yr), **Reinforcement Learning** (4 yrs), CNN, RNN, ResNet, Attention Block, UNet, **Transformer**, ViT (5 yrs), Neural Architecture Search (3 yrs), **Diffusion Models** (1 yr), GPT-2,3,4 (1 yr), **Reinforcement learning** from human feedback (1 yr), **VeOmni** (1 yr), **TorchTitan** (1 yr), **FSDP** (1 yr), **Zero** 1,2,3 (1 yr)
- *Databases and Storage:* MySQL (6 yrs), Oracle SQL (6 yrs), MongoDB (6 yrs), PostgreSQL (6 yrs), Redis (4 yrs), Hive SQL (3 yrs)
- *Front-end Development:* PHP (6 yrs), Vue.js (6 yrs), ReactJS (6 yrs), ASP.NET (6 yrs), jQuery (6 yrs), AngularJS (6 yrs), Apache (6 yrs), MeteorJS (6 yrs)
- *Back-end Development:* Spring Boot (6 yrs), Django (6 yrs), Flask (6 yrs), Node.js (6 yrs), Express (6 yrs), REST API Design (6 yrs), CI/CD
- *Mobile Development:* Android Studio (6 yrs, Java, Kotlin), XCode (6 yrs, Swift, Objective-**C**), React Native (6 yrs, Cross-platform), Flutter (6 yrs, Cross-platform)
- *Web Crawling & Testing:* Urllib (6 yrs), BeautifulSoup (6 yrs), Scrapy (6 yrs), Requests (6 yrs), JSON (6 yrs), Selenium (6 yrs), Pytest (6 yrs), JUnit (6 yrs)
- *Version Control & Build Systems:* **Git** (8 yrs), Gradle (6 yrs, Android, Java), Maven (6 yrs, Java), npm (6 yrs, JavaScript), pip (6 yrs, **Python**)
- *Development Tools & Libraries:* Airflow (2 yrs), Kafka (2 yrs), Elasticsearch (2 yrs), OpenCV (5 yrs), Pandas (5 yrs), **NumPy** (5 yrs), **SciPy** (5 yrs), NLTK (5 yrs), Matplotlib (5 yrs), Seaborn (5 yrs), Azure Data Factory (2 yrs), AWS (2 yrs), Google Cloud Platform (2 yrs)

Extracurricular

- **Fitness:** fencing, jogging, bodybuilding, table tennis and badminton
- **Leisure:** saxophone, web / mobile development, magic and video games
- **Volunteer:** HPCA 2026 volunteer; former SYSU Student Union Vice President; Changjun High School volunteer.