

# Tony (Tongyun) Yang

+31-645-7703-56 • Amsterdam, NL • [tonyyunyang.github.io](https://github.com/tonyyunyang) • [tonyyunyang@outlook.com](mailto:tonyyunyang@outlook.com)

The limits of language mean the limits of the world?  
Perhaps, yet we dwell in more than we can name.

## Statement

---

I believe AI is moving from scaling race to systems race. Larger models will still matter, but they are not enough. What matters is how we train, route, compress, deploy, and apply them under real constraints.

My previous experience spans a wide spectrum of AI, from LLM routing to medical imaging. My interests have converged in two fields, 1) where data is sensitive and public data is scarce, and 2) applied systems in products that exploit full power of frontier models. I believe they pave the next paradigm of AI.

## Experience

---

### Independent AI Researcher

Mar 2026 – Present

#### Long-term Self-motivated Role

Amsterdam, The Netherlands / Remote

Collaborate with industry (Tencent, Gradient Network, Commonstack, etc.) and academia (McGill, Tsinghua, etc.), leading research spanning cost-efficient LLM, LLM for optimization, agentic harness frameworks and world models, etc.

### Marie Skłodowska-Curie Fellow

Jul 2025 – May 2026

#### IMDEA Networks

Madrid, Spain

Investigated privacy-preserving wireless sensing systems converting wireless signals into actionable insights for human/obstacle detection, with focus on representation learning and world models for wireless signals to enable feature extraction and dynamic predictive reasoning.

### AI Research Engineer

Jan 2025 – Jul 2025

#### TU Delft Imaging Physics Department

Delft, The Netherlands

Investigated weight redundancy in nnU-Net using unstructured pruning, followed by structured pruning. Demonstrated that nnU-Net can be structurally pruned to 99% sparsity with minimal performance degradation, leading to a 6× improvement in training efficiency and gain in inference speed, with results accepted for publication in MIDL'25.

### HCI Research Assistant

Jun 2023 – Oct 2024

#### TU Delft Embedded Systems Department

Delft, The Netherlands

Constructed the first large-scale public dataset for emotion recognition based solely on eye-tracking in immersive environments, covering seven discrete emotions. Developed an efficient and scalable recognition method, with results accepted for publication in IMWUT/UbiComp'25.

### Support Engineer

Dec 2020 – Jul 2021

#### NXP Semiconductors N.V.

Nijmegen, The Netherlands

Developed an automated test bench system for CAN chips, integrating comprehensive diagnostic protocols to enable efficient defect analysis and ensure robust post-production performance.

## Publications

---

1. Yao, Y., Wang, Z., **Yang, T.**, Chen, W., Han, Y., Xiao, J., Bao, C., Shi, T. (2026). MERA: Model Evolution and Routing with Skill Adaptation for Agentic Systems at Scale. In *ACM Conference on AI and Agentic Systems (ACM CAIS), RLEval Workshop*.
2. Yang, P.\*, Chen, W.\*, **Yang, T.**, Feng, P., Xing, J., Guo, W., Yao, Y., Han, Y., Li, H., Wang, X., Wang, Z., Xiao, J., Yang, A., Tian L., Ai, L., Yang, E., Shi, T. (2026). TwinRouterBench: Fast Static and Live Dynamic Evaluation for Realistic Agentic LLM Routing. In *ACM Conference on AI and Agentic Systems (ACM CAIS), RLEval Workshop*.
3. **Yang, T.**\*, Regmi, B.\*, Du, L., Bulling, A., Zhang, X., & Lan, G. (2025). Through the Eyes of Emotion: A Multi-faceted Eye Tracking Dataset for Emotion Recognition in Virtual Reality. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (UbiComp/ISWC, IMWUT)*, 9(3), 1-41. ACM, New York, NY, USA.
4. Zhao, Y., Kellman, P., Xue, H., **Yang, T.**, Zhang, Y., Han, Y., Simonetti, O., & Tao, Q. (2025). Reverse Imaging for Wide-spectrum Generalization of Cardiac MRI Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 15962, 555-565. Springer Nature Switzerland.
5. **Yang, T.**, Zhao, Y., & Tao, Q. (2025). Pruning nnU-Net with Minimal Performance Loss. In *Medical Imaging with Deep Learning (MIDL), Short Papers*.

## Projects

---

### Diffusion Weights

Jan 2026

Investigated whether neural network fine-tuning can be reframed as a generative prediction problem rather than step-by-step SGD optimization. We used U-Net diffusion model to forecast future parameter states from earlier checkpoints. Predictions across short, medium, and long-term horizons were merged into the final model. Experiments showed 3.2x wall-clock acceleration on benchmarks including MMLU, OpenBookQA, and WMT16. Downstream performance remained comparable to full fine-tuning. The outcome of this project is submitted to ACL'26.

### LLM Router

Feb 2026

Collaborated with Gradient Networks to build a benchmark for evaluating LLM routing strategies. The project tested whether easy steps in complex tasks could be assigned to lower-cost models without sacrificing accuracy. Experiments covered coding, tool-use, etc. Results showed routing can preserve same performance while reducing cost by >90% compared with using a single SOTA model for every step. The outcome of this project is submitted to NeurIPS'26 and is currently under review.

### Cost-Adaptive LLM Routing with Specialist Models

Mar 2026

Extended the LLM router benchmark work by exploring how specialist models can further reduce inference cost. The project focused on repeated daily workflows, where tasks share similar patterns but differ in context. We collected trajectories from stronger models and used them to fine-tune smaller models for these scenarios. The central idea is that as more usage data is collected, the small model improves, and the overall cost of using the system decreases. The outcome of this project is in preparation for submission to EMNLP'26.

### Human Intent World Model

Mar 2026

Collaborated with MeetaVista to improve customer experience in AI-powered sales. We built a synthetic dataset for modeling customer intent, based on knowledge distilled from classical sales literature. The dataset includes visual and behavioral cues, intent labels, and reasoning traces linked to existing sales knowledge. We fine-tuned a vision-language model on this data so that it could infer customer intent from visual cues and respond using relevant sales reasoning. The model performed well in downstream real-world interaction tasks after deployment.

### Agentic Coding Bench

May 2026

Collaborated with Gradient Networks to construct a benchmark for agentic coding. Rather than testing isolated, single-function tasks like most existing benchmarks, we measure whether an AI agent can plan and build a complete software project from scratch. Each challenge is derived from a real open-source project and given to the agent only as a product brief, with the reference solution and grading hidden. The goal is to compare today's strongest coding agents on realistic, project-level work, and we plan to scale the benchmark from an initial proof-of-concept set toward a large public release. The outcome is in preparation for submission to a top-tier AI venue.

## Education

---

### MSc Computer Systems Engineering

TU Delft

Sep 2022 – Oct 2024  
Delft, The Netherlands

### BSc Electrical Engineering

Shanghai Maritime University & Exchange abroad in NL

Sep 2017 – Jul 2021  
China & The Netherlands

## Teaching

---

### Teaching Assistant & Mentor

TU Delft

Sep 2023 – Jan 2025  
Delft, The Netherlands

- ET 4310 Supercomputing for Big Data (2024/2025 Q1) / TA
- CESE 4030 Embedded Systems Lab (2023/24 Q3) / TA
- CESE 4000 Software Fundamentals (2023/24 Q1) / TA
- CESE 4010 Advanced Computing Systems (2023/24 Q1) / TA
- CESE MSc Programme Student Mentor (2023/24)

## Skills & Languages

---

*AI-Native:* Claude Code, Codex, OpenCode, pi, Hermes Agent

*Programming:* Python, C/C++, Rust, Bash

*ML/Frameworks:* PyTorch, HuggingFace, JAX

*Tooling:* Git, Docker, Linux, LaTeX

Mandarin (Native)  
English (Professional)

## Personal Interests

---

I run regularly, my personal best in a half marathon is 1:43:53. I body build through regular strength training, I also enjoy cooking (especially for family), hiking, and tennis.