Tony (Tongyun) Yang

Amsterdam (Currently Based) | Shanghai | Shenzhen
+31-(0)645-7703-56 | +86-156-9219-2756 • tonyyunyang.github.io • tonyyunyang@outlook.com

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

Research Area & Interest

My research interest spans the spectrum of AI, including world models, LLMs, computer vision, AI safety and AI for medical purposes, etc. Broadly, I am interested in understanding how intelligent systems can be built, aligned, and deployed across diverse settings.
In particular, I am motivated by 1) improving how AI can pervade not only language, but all aspects of the world more effectively and meaningfully, and 2) enhancing human capabilities through AI in ways that promote fairness and access.

Experience

Independent AI Researcher — Mar 2026 – Present
Long-term Self-motivated Role — **Amsterdam, The Netherlands / Remote**
Collaborate with industry (Tencent, Gradient Network, MeetaVista, 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.

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. **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.*

2. 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.*

3. **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. (github.com/CommonstackAI/CommonRouterBench)

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. The outcome of this project is in preparation for a submission to a top-tier AI venue.

LLM for Optimization **Apr 2026**

Collaborated with Tencent to develop a framework for applying LLMs to optimization problems. The framework uses an optimization harness to guide the model toward better solutions through structured search, feedback, and memory retrival. The goal is to help the model approach optimal answers more efficiently, rather than relying on direct generation alone. The project is still on-going, and the outcome is planned for a submission to AAAI'26.

Education

MSc Computer & Embedded Systems Engineering **Sep 2022 – Oct 2024**

TU Delft **Delft, The Netherlands**

BSc Electrical Engineering & Mechatronics **Sep 2017 – Jul 2021**

Shanghai Maritime University & Exchange abroad in NL **China & The Netherlands**

Teaching

Teaching Assistant & Mentor **Sep 2023 – Jan 2025**

TU Delft **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.