# 杨童耘
# Tony

+86-15692192736 | +31-0645770356
tonyyunyang@outlook.com
阿姆斯特丹(现居 时差晚-6h) | 上海 | 深圳

## 研究领域

我的研究兴趣围绕人工智能系统如何理解世界、参与世界。具体方向包括世界模型、大语言模型、计算机视觉、以及医疗 AI 等。我关注模型能力如何被放入真实环境中，与人的需求、社会结构和具体任务发生连接。
我尤其关注两个问题。**第一**，AI 如何从语言和文本中的智能，进一步走向对现实世界的感知、建模和行动。**第二**，AI 如何成为扩展人类能力的工具，并使这种能力以更公平、可及的方式影响社会。

## 工作经历

**2026年3月 — 至今** **独立研究与产学研合作** **独立研究员**

- 与工业界和学术界团队合作，包括 Tencent（腾讯）、Gradient Network、MeetaVista（米塔视界）等公司，以及 McGill（加拿大麦吉尔大学）、清华大学等学术机构。研究内容涵盖低成本高效 LLM及其面向复杂任务的路由系统、面向组合优化的LLM推理框架、智能体评测与运行框架、世界模型等方向。

**2025年7月 — 2026年5月** **IMDEA Networks | 马德里高等研究院** **玛丽居里博士学者**

- 研究隐私保护型无线感知系统，探索如何将无线信号转化为可用于人和障碍物检测的有效信息。研究重点包括无线信号的表征学习，以及面向无线信号的世界模型，以支持特征提取和动态预测推理。

**2025年1月 — 2025年7月** **TU Delft | 荷兰代尔夫特理工大学** **医疗图像AI研究员**

- 研究 nnU-Net 中的权重冗余问题，先后采用非结构化剪枝和结构化剪枝方法提升模型效率。实验表明，nnU-Net 可以在结构化剪枝后达到 99% 稀疏度，同时仅产生很小的性能损失。该方法带来了约 6 倍训练效率提升，并提高了推理速度，相关成果已被 MIDL 2025 接收发表。

**2023年6月 — 2024年10月** **TU Delft | 荷兰代尔夫特理工大学** **情感计算AI研究助理**

- 构建了首个仅基于沉浸式环境中眼动数据的大规模公开情绪识别数据集，覆盖七类离散情绪。同时开发了高效、可扩展的识别方法。相关成果已被 IMWUT / UbiComp 2025 接收发表。

**2020年12月 — 2021年7月** **NXP | 恩智浦半导体** **芯片技术支持工程师**

- 开发用于 CAN 芯片的自动化测试平台，集成完整的诊断流程，用于提升缺陷分析效率，并支持芯片量产后的稳定性验证。

## 论文发表

- **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.

- **Yang, T.**, Zhao, Y., & Tao, Q. (2025). Pruning nnU-Net with Minimal Performance Loss. In Medical Imaging with Deep Learning (MIDL).

- 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.

## 项目经历

2026年2月 **LLM Router | 复杂任务的低成本大语言模型路由系统**

- 与 Gradient Networks 合作构建 LLM 路由评测基准，用于评估不同 LLM 路由策略的有效性。项目重点考察复杂任务中的简单步骤是否可以交由成本更低的模型完成，同时不牺牲整体准确率。实验覆盖了 SWE-bench、BFCL 等近期常用基准。结果显示，与在所有步骤中都使用单一 SOTA 模型相比，合理的路由策略可以在保持相同性能的同时，将成本降低超过 90%。
- 项目链接：https://github.com/CommonstackAI/CommonRouterBench

2026年3月 **Cost-Adaptive LLM Routing | 成本自适应的大语言模型路由**

- 在 LLM Router 项目的基础上，进一步探索如何通过专用小模型降低推理成本。项目主要面向重复出现的日常工作流：这些任务在结构上相似，但具体上下文不同。我们收集强模型生成的轨迹，并用这些数据微调较小模型，使其适应特定场景。项目的核心思路是：随着系统使用次数增加，可用于训练的数据不断积累，小模型能力逐步提升，系统整体使用成本也随之下降。该项目成果正在准备投稿至 NeurIPS 2025。

2026年3月 **Human Intent World Model | 意图理解的视觉语言世界模型**

- 与 MeetaVista 合作，探索如何提升 AI 销售场景中的客户体验。我们基于经典销售学文献中的知识，构建了一个用于建模客户意图的合成数据集。数据包括视觉与行为线索、意图标签，以及与销售知识相连接的推理过程。随后，我们基于该数据微调视觉语言模型，使其能够从视觉线索中推断客户意图，并结合相关销售逻辑做出回应。模型部署后，在真实交互任务中取得了良好表现。该项目成果正在准备投稿至顶级 AI 会议。

2026年4月 **LLM for Optimization | 组合优化的大语言模型推理框架**

- 与 Tencent 合作开发一个将 LLM 应用于优化问题的框架，例如旅行商问题等。该框架通过 optimization harness 引导模型进行结构化搜索，并结合反馈与记忆检索，使模型能够更高效地接近最优解，而不是仅依赖直接生成答案。项目目前仍在进行中，后续计划投稿至顶级 AI 会议。

## 教育经历

2022年9月-2024年10月 TUDelft | 荷兰代尔夫特理工大学 计算机工程 | 硕士

2017年9月-2021年7月 上海海事大学（1.5年荷兰海外交换） 电气工程及其自动化 | 学士

## 教学经历

2023年9月 —— 2025年1月 **TU Delft | 荷兰代尔夫特理工大学** **助教与学生导师**

- ET 4310 Supercomputing for Big Data，助教，2024/2025 Q1
- CESE 4030 Embedded Systems Lab，助教，2023/2024 Q3
- CESE 4000 Software Fundamentals，助教，2023/2024 Q1
- CESE 4010 Advanced Computing Systems，助教，2023/2024 Q1
- CESE 硕士项目学生导师，2023/2024

## 专业技能

AI-Native 工具：Claude Code, Codex, 及其他开源框架如Openclaw, pi, Hermer-Agent等
编程语言：Python, C/C++, Rust, Bash
机器学习框架：PyTorch, HuggingFace, JAX
开发工具：Git, Docker, Linux, LaTeX

## 语言水平

中文： 母语
英文： 专业学术/工作流利 | 海外6年 | 雅思8.0

## 兴趣爱好

跑步，半程马拉松 PB 1:43:53
健身 | 做饭 | 徒步 | 网球

# Tony (Tongyun) Yang

Amsterdam (Currently Based) | Shanghai | Shenzhen
+31-(0)645-7703-56 | +86-156-9219-2756 • 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 to enhance efficiency. 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

### 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 recent benchmarks such as SWE-bench for coding and BFCL for tool use, etc. Results showed that routing can preserve same performance while reducing cost by >90% compared with using a single SOTA model for every step. (https://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 NeurIPS'25.

### 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, such as the Traveling Salesman Problem, etc. 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 a top-tier AI venue.

## 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.