WANG Xutao

+86 139-0302-3712 | wangxutao98@qq.com | https://github.com/WANGXutao98

Education

2020.07 - 2022.02
National University of Singapore
MSc in Computer Science
GPA: 3.6/4.0

2016.09 - 2020.06
Xi'an Jiaotong-Liverpool University
BSc in Information and Computing Science
GPA: 3.9/4.0 First-Class Honor

Skills

Programming Languages

Python, Golang, C++

Frameworks & Tools:

PyTorch, ONNX, vLLM, TensorRT-LLM

Databases & Ops

MySQL, Hadoop, MongoDB, Docker, Linux

Prototyping

Azure, Sketch

Awards

Hunyuan Annual Technical Breakthrough Award
Tencent
2024

Tencent H1-SEVP Outstanding Individual Award
2023

Tencent Individual Excellence Award
2023,2024

Tencent TEG-Open Source Collaboration Award
2023

Interests

Piano (Amateur Grade 10), Photography, Fitness, Basketball, Eletric Guitar

Publications

Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records
IEEE
Nov 2019

2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)

Inter-Personal Relation Extraction Model Based on Bidirectional GRU and Attention Mechanism
IEEE
Sep 2019

2019 IEEE 5th International Conference on Computer and Communications (ICCC)

Model Checking the Reliability of Data Center Network
IEEE
Jul 2018

2018 9th International Conference on Information Technology in Medicine and Education (ITME)

Languages

Chinese

English

Cantonese

Summary

With core AI R&D experience in Tencent's Hunyuan division and a proven track record in managing cutting-edge university research, I possess a full-stack, closed-loop capability spanning algorithms, engineering, and scientific investigation, underpinned by rigorous architectural thinking. I now aspire to leverage this foundation to probe the fundamental nature of technology and deliver pivotal innovations during my doctoral studies.

Experience

Shenzhen Loop Area Institute
AI Information Manager
Jul 2025 – Present

- Led tracking and analysis of AI research trends, key technologies, and disruptive innovations through in-depth review of top-tier conference papers and industry reports. Produced high-impact technical and strategic advisory reports to inform R&D planning.
- Established a high-level AI think tank, delivering research reports on Agent Memory, Multi-Agent Systems, LLM Hallucination, and Reasoning Uncertainty. Findings were adopted in research project guidelines.
- Directed the evaluation of large-scale computing cluster proposals, developing assessment frameworks for technical feasibility, efficiency, and sustainability to align resource allocation with research priorities.

Tencent Technology – TEG Multimodal Model Department
Algorithm Engineer
May 2024 – Jul 2025

Hunyuan-Tencent Yuanbao

- Engineered and productionized Strategy service for the Yuanbao APP and its WeChat contact mini-program; designed a high-availability architecture together with a cross-region, multi-active routing strategy that withstood traffic spikes of 20,000+ QPM while maintaining 100% service success rate.
- Leveraged user-behavior and system-performance telemetry from Yuanbao and WeChat to help build a closed-loop data flywheel.

Game for Peace AI Agent

- End-to-end owner of the "Gilly" AI Agent service in Game for Peace and developed the interaction pipeline and, with the team, proposed the MBA-RAG framework (Multi-arm Bandit-based Adaptive RAG); full-scale tests showed a ~20% reduction in retrieval steps versus Adaptive RAG while keeping relevance ≥90%.
- Built a streaming AI-Agent backend on trpc-Go coroutine scheduling plus vLLM Prefix-Caching & Continuous Batching, sustaining 305+ QPS with end-to-end P99 latency ≤2.6s; deployed a panoramic service-health dashboard using OpenTelemetry for real-time KPI visualization.
- Created a multi-model benchmark (DeepSeek-R1, GPT-4o, Claude-3.5, Qwen-2.5) covering response latency, long-context comprehension and multi-turn dialogue stability for streaming AI-Agent scenarios.

Tencent Technology – TEG AI Lab
Backend Development Engineer
Feb 2022 – May 2024

Virtual Human AI-NPC Engine Capabilities

- Led key technology research for AI-NPC engines in *YuanMeng* and *Honor of Kings: World*; developed and productionized intelligent expression & motion generation services plus centralized control middleware.
- Independently built a standardized FastAPI-based engineering framework supporting automatic service-log reporting, rapid API scaffolding, and containerized algorithm deployment. Established an SOP pipeline (requirement review → model adaptation → service rollout) that has enabled 10+ complex-scenario algorithm services to onboard quickly, boosting development efficiency by 40–50 % and cutting manpower costs by 70 %.
- In charge of inference optimization for in-game AI expression/motion algorithms: applied ONNX-Runtime operator fusion and model quantization to raise single-machine concurrency of the expression service by 250 % (30 → 75 QPS) and cut average latency by 55 % (220 ms → 98 ms); for motion generation, achieved a 400 % concurrency increase (25 → 100 QPS) with no accuracy loss while keeping average latency ≤ 150 ms—meeting players' stringent real-time interaction demands.

Virtual Human PaaS Platform

Headed backend development of the AI-Lab digital-human PaaS platform; designed and implemented a high-performance, modular Flask-based WebServer framework for rapid development and deployment. The framework is lightweight and extensible, encapsulating generic modules (DB operations, logging & monitoring) for plug-and-play usage, and supports automatic dynamic API documentation generation, improving development hand-off efficiency by 50 %.

National University of Singapore
Teaching Assistant – BT5153 Applied Machine Learning for Business Analytics
Dec 2020 – Jun 2021

- Instructed programming labs focused on data mining, covering data preprocessing, feature engineering, machine learning model implementation, and visualization using Python.
- Assisted in course material development, designed demo cases, and provided student support through office hours and feedback sessions.