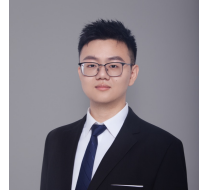


Xiao Wang

wangxiao02131@gmail.com | +49 17675790197
linkedin.com/in/xiaowang131 | xiaowang.lol | github.com/lololo-xiao
Saarbrücken, Saarland (open to relocation, available now)



Skills

Languages: Python, SQL, Bash

LLM / ML: PyTorch, Hugging Face Transformers, Post-training (SFT, RLHF, DPO), LLM/Agent Evaluation

Agents & LLMOps: Agent Orchestration, Function-calling, MCP, Arize Phoenix, OpenTelemetry tracing

Cloud & Infra: Google Cloud (Vertex AI, Cloud Run, Cloud Build), Docker, Git

Spoken: Chinese (native), English (C1), German (A1)

Experience

Research Assistant (HiWi) | Saarland University

01.2026 – 06.2026

KV-cache compression and memory-efficient fine-tuning for long-context LLMs (PyTorch, Hugging Face).

- Unified several KV-cache compression methods (H2O, InfLLM, attention-sink / sliding-window) behind one DynamicCache-compatible interface using PyTorch forward hooks and instance-level monkey-patching, with no subclassing or library forks.
- Built a compress-then-distill fine-tuning pipeline (LoRA plus knowledge distillation) that compresses a long-context KV cache $\sim 9\times$ while recovering full-context accuracy on a held-out probe; also resolved a gradient-checkpointing and KV-cache-reuse incompatibility in Hugging Face to enable memory-efficient training.

LLM Engineer Intern | Meituan (Shanghai, China)

05.2024 – 08.2024

LLM-powered conversational AI for a children's smartwatch app (K-12).

- **Data:** Built SFT and preference datasets for child-safety and ethics with a hybrid rule-based and model-based cleaning pipeline, cutting the bad-case rate by $\sim 30\%$.
- **Training:** Ran post-training alignment (SFT, online RLHF, DPO) on a distributed Megatron-LM stack with tensor and data parallelism across multiple A100 nodes, improving response safety and coherence.
- **Evaluation:** Designed an LLM-as-judge evaluation system and a child-value benchmark using rubric-based, chain-of-thought, few-shot prompts; key alignment metrics improved $\sim 15\%$.
- **Deployment:** Ran stress tests, supported model rollout, and set up performance monitoring to keep the product stable in production on Meituan's internal cloud.

Projects

CareerPilot | Google Cloud Rapid Agent Hackathon, Arize track

2026

Autonomous, self-improving job-matching agent for international graduates in Germany.

- **Agent loop:** Built a Python agent that plans, ranks jobs against a candidate, rewrites its own ranking rubric when quality is low, and stops autonomously at a quality threshold; orchestration is hand-built on Gemini (Google Gen AI SDK, Vertex AI) with function-calling for tool selection.
- **Evaluation:** Added an LLM-as-judge scoring fit, risk detection, and actionability, driving a self-improvement loop plus an automatic retry that re-prompts the model on schema-validation failures.
- **Observability and memory:** Used the Arize Phoenix MCP server as cross-run memory (versioned rubric prompts) and OpenTelemetry tracing to emit a span for every step, making the full reasoning path observable.
- **Deployment:** Packaged Python and Node in one container so the MCP server runs as a subprocess, and deployed to Google Cloud Run with secrets in Secret Manager via Cloud Build.

Publications

MultiJustice: A Chinese Dataset for Multi-Party, Multi-Charge Legal Prediction | First author, NLPCC 2025

First Chinese benchmark for multi-defendant, multi-charge legal prediction; built the data pipeline and four 5,000-case datasets and benchmarked BERT, T5, LawFormer, and InternLM2.

InternLM-Law: An Open-Source Chinese Legal LLM | Co-author, COLING 2025

Built and ran the model evaluation pipelines for legal reasoning and QA tasks.

Education

Saarland University – MSc Language Science and Technology

10.2022 – 05.2026

Wuhan University – BSc Information Management and Information System

09.2018 – 06.2022