你好，我是

赖正麟

Lai Zhenglin



全栈工程师 · 算法工程师

me@laizhenglin.com　github.com/Bearisbug　linkedin.com/in/laizhenglin

关于

全栈与算法工程师，主导过大模型应用的整套架构，也做 AI 安全研究：两篇 NeurIPS 论文的共同一作（SAFEx、ICAT）。

在字节跳动做过 SRE Agent，在华为补全过 MindSpore 框架；业余写开源工具，比如 Quilt。

currently: 主导 Kando AI 的整套架构

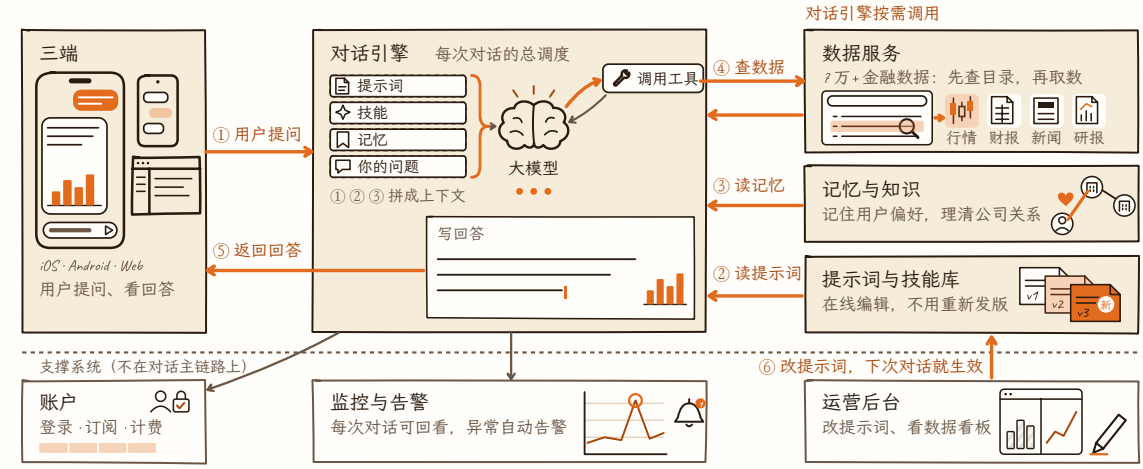
经历

2026.07 – 至今

核心技术负责人 · Kando AI

- 主导 Kando AI（面向个人投资者的 AI 投研对话产品）整套架构：三端和运营后台，后端 9 个服务部署在 AWS，配监控告警；5 万轮历史对话迁移零失败
- 设计对话引擎：每次对话按需读取提示词、技能和用户记忆，大模型通过 MCP 调用工具查数据、写回答；提示词和技能可在线编辑，改完下次对话即生效，不用重新发版
- 搭建金融数据服务：把行情、财报、新闻、研报等 7 万+ 数据项整理成一份目录，模型先查目录、再取数，两个动作取代旧版 27 个手工接口

数据、三端、运维，我主导搭起来



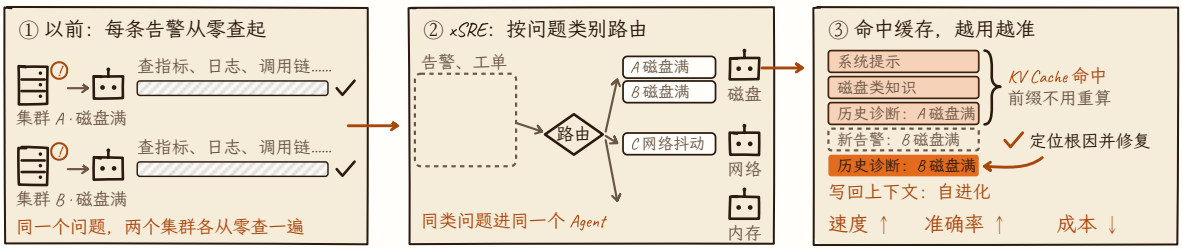
按编号看：① 提问 → ②③ 读提示词、技能、记忆 → ④ 调工具查数据 → ⑤ 回答；⑥ 改提示词，下次对话生效。

2025.10 – 2026.06

AI Agent 研发实习生 · 字节跳动

- 优化 SRE Agent 的执行速度和根因诊断准确率（告警或工单进来后，Agent 自动定位根因、诊断并修复）
- 提出自进化的 xSRE 集群架构：按问题类别把告警路由给专属 Agent，同类问题高频命中 KV Cache、复用历史诊断，同时优化速度、准确率和成本

我做的：让同类问题越查越快、越准



例子：同一天，集群 A 和 B 都出现磁盘满。

2024.09 – 2025.04

算法框架实习生 · 华为

– MindSpore 框架补全：对照 PyTorch 的接口和昇腾的算子，补齐 MindSpore 缺的代码

作品

ICAT

Incident-Case–Grounded Adaptive Testing for Physical-Risk Prediction in Embodied World Models

Zhenglin Lai*, Sirui Huang*, Yuteng Li, Changxin Huang, Jianqiang Li, Bingzhe Wu

视频世界模型能预见危险后果吗？ICAT 从事故报告和安全标准出发构造 909 个测试，按「危险情境—危险动作—严重后果」整条风险链打分：即使场景保住、动作也做了，最强的系统也只在 23.5% 的测试里画出了有记载的后果。

*NeurIPS 2026 · Poster · *共同一作*



NeurIPS 2026 海报 ↗

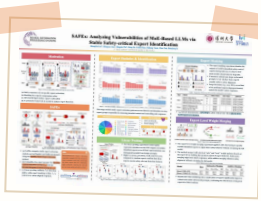
SAFEx

Analyzing Vulnerabilities of MoE-Based LLMs via Stable Safety-critical Expert Identification

Zhenglin Lai*, Mengyao Liao*, Bingzhe Wu, Dong Xu, Zebin Zhao, Zhihang Yuan, Chao Fan, Jianqiang Li

MoE 大模型会不会拒绝有害请求，主要取决于少数几个「安全关键专家」。SAFEx 用稳定性筛选把它们找出来，分成识别有害内容、控制拒答两组；在 Qwen3-30B-A3B 的 6144 个专家里只屏蔽 12 个，拒答率就下降 22%。

*NeurIPS 2025 · Poster · 代码 · *共同一作*



NeurIPS 2025 海报 ↗

Imbo

AI 编程助手之间的即时通信：Claude Code、Codex 等六种工具的会话可以互发消息，发给自己别的电脑上的会话，也能发给别人的 Agent；消息在对方干活的间隙送达，空闲的会话会被唤醒。个人独立开发并开源，约 2.3 万行 Rust。

Rust · Tokio · Axum · PostgreSQL · MCP · 141 stars

Quilt

AI 原生的无限画布设计工具：一句话描述 App，生成风格一致的 4–6 屏 HTML 设计稿铺在画布上，可直接点击当原型、导出成单文件；本机的编码 Agent 通过 MCP 双向读写画布。个人独立开发并开源，约 1.7 万行 TypeScript。

TypeScript · React 19 · Hono · MCP · 106 stars

技术栈

语言	TypeScript · Python · Swift · Kotlin
前端	React · Next.js · Vite · Tailwind CSS · shadcn/ui · TanStack Query
移动端	SwiftUI · React Native · Expo
后端与数据	FastAPI · NestJS · PostgreSQL · Redis · Drizzle · DuckDB
AI 与 Agent	MCP · Claude Agent SDK · Anthropic / OpenAI API
模型与视觉	PyTorch · vLLM · LoRA · YOLO · OpenCV
基础设施	Docker · AWS · Terraform · GitHub Actions · Cloudflare · Playwright

圈起来的每天都在用

教育

深圳大学 · 计算机科学与技术

2022.09 – 2026.07