# 刘源昊

Tel:188-0039-2632 Email:13365551665@163.com Wechat:lyh_5480

## 工作经历

**腾讯** 2024.05 – 2025.08

运营开发工程师 IEG 互动娱乐事业群-增值服务部 上海徐汇

- 负责腾讯游戏运营活动的全栈开发，后端 Golang/PHP/低代码平台，前端 Vue/JS/jQuery 等;
- 支持**王者荣耀**、QQ 炫舞等 20+ 业务的**营销活动开发**。累计交付需求 70+，服务用户**千万级**，无线上事故。

**字节跳动** 2025.09 – 2026.01

服务端研发工程师 生活服务-营销基础 上海徐汇

- 负责字节生活服务运营基础能力的 Golang 服务端开发，高可用、高并发优化，支撑多场大型活动顺利落地;
- 支持营销工具券减类的 B、C 端服务端开发。累计交付需求 30+，服务用户**千万级**，无线上事故。

## Part-Time 经历

**香港浸会大学深圳研究院** 2025.12 – 至 今

大模型研究员 广东深圳

- 负责大语言模型幻觉治理及后训练算法研发，设计涵盖事实准确率、幻觉率等多维度的量化评测体系;
- 基于自研评测指标优化奖励函数，指导 Agent 底座模型定向微调，增强 Agent 在复杂任务中的执行可靠性。

**青岛朝乾禾熹文化传媒有限公司** 2025.08 – 2025.11

全栈开发工程师 速凌电竞项目组 Remote

- 独立完成陪玩准入考核平台的产品原型设计、前后端开发及架构搭建，实现进店考核全流程自动化管理;
- 独立研发陪玩考核专用反作弊桌面端工具，落地陪玩历史游戏违规行为检测能力，相关成果已成功申请软著。

## 项目经历

**王者荣耀 KPL 年总选手投票** 腾讯

- 作为**项目负责人**，设计实现全流程投票系统，支撑单日 **1.6 亿 PV/8000 万 UV** 高并发，保障投票公平、稳定;
- 基于 Redis incrby 原子操作实现毫秒级票数更新与实时榜单展示；通过「缓存预热 + 热点 Key 永不过期 + 主动更新 + Redisson 分布式锁解决缓存重建并发冲突」，有效缓解缓存击穿问题，数据库峰值 QPS 下降 **73%**;
- 设计写时分片架构，基于 Redis Cluster 分布式集群，将用户 ID 哈希分片至 16 个子 Key，实现 28 万 + QPS 极端写入场景下的集群负载均衡，读时异步聚合 + 缓存兜底保证全局票数最终一致性。

**炫舞手游跨系统角色转移** 腾讯

- 作为**项目负责人**，设计并实现跨系统角色转移完整解决方案，支撑单日 **49 万 PV/23 万 UV**;
- 搭建完整角色转移流程，实现从账号冻结、数据迁移、状态同步到激活校验的端到端链路，通过游戏侧 IDIP 原子操作保证**数据一致性**，设计多级异常监控与自动回滚机制，使异常操作自动回滚率提升至99.99%;
- 设计基于 Redis Lua 脚本的原子化库存管理方案，通过「同步预扣库存 + 异步补偿兜底（针对分布式事务异常）」解决高并发下超卖/少卖问题；基于 Redis Lua 脚本 INCRBY+EXPIRE 原子操作实现每日转区名额精准控制，核心操作引入版本号乐观锁机制，库存数据准确率从 97% 提升至 **99.99%**，保障零数据错乱。

**Golang 入参治理通用中间件** 字节

- 针对 Golang 服务入参校验碎片化、代码冗余、校验逻辑不统一的痛点，研发高性能通用中间件，利用反射机制解析结构体标签，设计结构体元信息服务启动时一次性预加载策略，结合 **pprof** 采集火焰图定位内存分配瓶颈并引入 sync.Pool 复用解析对象，将单接口解析耗时从 200μs 降至 30μs，达到微秒级性能;
- 采用**适配器模式**屏蔽框架、打点中间件等不同版本间的差异，实现一套代码无缝兼容业务线中存量的旧版框架与新版微服务，解决了底层依赖升级带来的侵入性修改难题。基于**责任链模式**解耦参数解析、格式校验、业务准入与链路打点流程，支持通过**策略模式**自定义业务逻辑，实现了治理策略的配置热更新与动态拦截;

- 集成 Prometheus + Grafana + Jaeger 搭建全链路可观测性体系，通过中间件在请求入口处自动埋点，实现非法入参的实时告警与全链路拓扑展示，配合输出标准化接入手册，推动该组件成为部门 Golang 开发基线。成功落地于部门 10+ 业务线，平均精简了业务侧 85% 的冗余校验代码，显著提升了系统的鲁棒性。

**天龙八部智能 AI 问答系统** 腾讯

- 构建端到端 RAG 数据管道，针对游戏静态知识，通过游戏官方数据源爬虫抓取，经清洗/术语归一化、去重、实体链接、质量打分完成数据处理并分层入库；采用 WeData 实现批处理，结合 VectorDB 向量库 + BM25 混合索引、Redis 热表优化检索能力；实现增量 embedding 微批更新延迟 ≤ 2h，设计版本化元数据与 diff 回滚机制，检索层引入基于指数衰减的时效性排序，有效降低过时答案率至3%-；
- 主导研发基于**规划 - 执行**模式的异构多智能体协同系统，利用 **Router Agent** 实现“门派技能 + 宝石属性”多维收益计算等复杂游戏逻辑任务的动态分发与原子化拆解，系统支撑日均 5 万+次高并发请求，将长链路推理任务的逻辑完备性与执行成功率提升 **35%+**；
- 针对游戏术语稀疏性核心痛点，构建基于 **Domain-Specific Embedding** 与双路召回的增强检索体系，引入 **Graph-RAG** 技术建立游戏世界观实体关联，并结合游戏领域知识库设计基于 Self-Reflection 的幻觉校验机制，确保回答 92%+ 遵循游戏世界观约束，同时知识召回准确率达 96%+；
- 基于 DDD 领域驱动设计理念打造高度解耦的 Agent 插件化架构，后端采用 Gin 框架搭建高可用服务，通过关键词提取+上下文裁剪与分级缓存技术将系统平均响应时延降低 **42%**，并有效节省约 **30%** 的推理 Token 成本；该技术方案已成功申请专利，且被采纳为部门 AI 项目开发基准模板。

**智能需求文档生成与对齐平台** 腾讯

- 针对部门内 PRD 不规范、开发测试难对齐的痛点，定义需求抽取字段规范、验收准则模板及标准化可机读交付文档，将生成结果转化为可执行契约与测试用例，接入 CI/CD 流水线，使需求澄清平均时间降低 30%+；
- 将整体流程拆分为提取、上下文检索、合成、合规规则、验证 5 个专责智能体，实现多智能体核心引擎；在生成链路中引入自定义置信度评分机制，有效抑制模型幻觉，支持生成结果逐段回溯，提升可靠性与可追溯性；
- 优化向量库与检索层性能，实现热点缓存与异步预热；在大模型推理环节加入域外内容检测及不确定性阈值控制，触发人工复核，降低失误率；将复核界面用户操作回流至主动学习与标注流水线，用于模型持续微调与业务适配更新，使业务适配性提升 40%，生成结果可追溯率达 100%，开发、产品接受率提升至 92%+。

**基于 RL 后训练的智能 Agent 大模型底座优化** 浸会研究院

- 作为 **技术负责人**，主导基于 RL 后训练的大模型底座优化与 Agent 应用架构设计，缓解 Agent 在复杂业务流中的“指令漂移”与逻辑幻觉难题。通过建立多场景任务执行的质量评测体系，实现了模型能力与业务需求的深度对齐，显著提升了 Agent 在自动化工作流中的执行可靠性与多轮交互的一致性；
- 落地了融合 **RLHF 与 RLAIF 的指令遵循增强方案**，采用“先 RLAIF 生成伪偏好数据，再小批量 RLHF 精调”的融合方式，针对 Agent 开发中的工具调用精度与长链任务规划能力进行定向优化；引入过程监督机制（按 Agent 规划-工具调用-结果验证划分步骤，设计步骤级奖励函数），解决 Agent 在闭环任务中“理解易、执行难”的痛点，使模型幻觉率降低 35%+，确保 Agent 输出在工业级应用中的合规性与逻辑完备性；
- 搭建标准化数据工程与效能评估体系，采用合成数据生成（种子任务+领域知识库引导）+人工纠偏相结合的方案，沉淀数万组涵盖“多步规划-工具调用-结果自检”全链路的高质量偏好样本；针对 Agent 长链推理导致的响应延迟瓶颈，引入基于 Page Attention 的分页式 KV-Cache 优化 + 多轮对话 KV-Cache 复用优化，结合流式采样技术，将系统平均首字时延降低 50%+，显著提升用户体验。

## 教育经历

**华东师范大学 985** 软件工程 硕士 软件工程学院；GPA: 3.56/4.0，方向：可信人工智能 2021 – 2024

**合肥工业大学 211** 数学与应用数学 本科 数学学院；Score: **89.98**/100，RANK: **5**/75(6%) 2017 – 2021

## 其他

- 曾于唯品会运营中心（策略算法工程师）实习、华为 ICT - 5G 开发部（通用软件开发工程师）实习；
- 拥有全栈开发能力，可从 0-1 搭建产品，具备大模型幻觉治理、Agent 架构研发、RL 后训练对齐的 AI 研究经验，以及算法工程化、高并发 AI 系统落地的工程实践经验；重度 AI 使用者，擅长运用 AI 为工作全流程提效。