

# 200+ AI 系统交付经验

AI delivery architect with 200+ system deployments across healthcare, manufacturing, finance, government & enterprise. Sizing → build → ship → SLA.

私有 LLM (AW36) · RAG · Agent · Multi-model · AIOps · Enterprise Delivery

200+

AI  
systems 系统

6

capability  
lanes 方向

20

case  
studies 案例

10+

industries  
行业

## 六大能力方向

Six capability lanes

### ① 私有 LLM 部署

Private Deployment

- vLLM · AW36 · Milvus
- A800 · 4090 · H20
- 内网合规 · 多租户

3 cases

### ② RAG 知识库

Retrieval-Augmented

- bge · Milvus · rerank
- 三库合一 · 引用溯源
- 语音 · 混合检索

3 cases

### ③ Agent workflows

Agent Workflows

- 多智能体 · 人机审核
- 招投标 · 合同 · CRM
- 长时任务 · 幂等

3 cases

### ④ 多模型调度

Multi-model Routing

- cost-aware router
- 混合云 · 地域 · 敏感度
- fallback · 灰度

3 cases

### ⑤ AI + DevOps

AIops / MLOps

- K8s + Slurm 混合调度
- MTTR 根因 · Trace 融合
- CI/CD code-review

3 cases

### ⑥ 政企场景

Government / Enterprise

- 招投标 · 视觉质检
- 集团合同 · 医保智审
- 国资 · 央企 · 医疗

5 cases

## 20 代表案例 · 硬指标

20 flagship cases with hard metrics

#### 01 · 1 三甲医院 AW36-72B

查指南从 8-15 分钟 → 平均 45 秒

#### 01 · 2 车间断网离线大模型

事故排查平均耗时 -45% (从 SOP 查阅到…)

#### 01 · 3 集团多租户 LLM 网关

集团 GPU 采购降 60% (从各线自建改为…)

#### 02 · 1 律所三库合一 RAG

合同起草模板复用率 从 40% → 82% (律…)

#### 02 · 2 电网现场语音 RAG

现场事故排查平均耗时 -40% (电网侧 6…)

#### 02 · 3 药械厂客服 RAG

平均处理时长 (MTTR) -55% (从 22 分钟…)

#### 03 · 1 招投标应答 Agent

中标率 +25% (从 38% → 47%, 12 个月对…)

#### 03 · 2 政采合同审核 Agent

法务人员审核工时 -60% (从平均 4 人天/…)

#### 03 · 3 CRM 线索跟进 Agent

MQL → SQL 转化率 +42% (从 22% → 31…)

#### 04 · 1 出海 SaaS 三网关调度

LLM 总成本 -38% (从 ARR 12% → 7.5%)

#### 04 · 2 医院公私混合云调度

私有集群利用率从 45% → 68% (更多诊…)

#### 04 · 3 内容工厂多模型 A/B

客户采纳率 +25% (从 62% → 78%, 减少…)

#### 05 · 1 GPU 集群 K8s+Slurm

GPU 整体利用率: 32% → 71% (季度均值)

#### 05 · 2 AIOps 告警根因

MTTR: 45 分钟 → 12 分钟 (73% 降)

#### 05 · 3 CI/CD AI 代码审查

defect escape rate -35% (季度事故 3…)

#### 06 · 1 政企招标全周期

采购平均周期 -35% (从 45 天 → 29 天)

#### 06 · 2 制造视觉质检

defect escape rate -74% (从 每月 3-5…)

#### 06 · 3 集团法务智审

合同事故 -80% (从每年 3-5 起 → 每年 0…)

#### 06 · 4 央企投标全周期

投标周期: 45 天 → 25 天 (-45%)

#### 06 · 5 医院医保智审

医保拒付率: 8-12% → 6-8% (相对下降 3…)

## 合作模式 · Delivery Modes

### POC

POC 落地

4-8 周验证效果

### BUILD

私有部署

硬件 sizing + 全栈交付

### TEAM

驻场组队

带队 + 培训 + 交付

### RESCUE

应急救火

接手救回 stuck 项目

### SLA

长期驻场

顾问 + 迭代 + 复盘

## 联系 · Contact

william.yangshun@gmail.com

抖音 92454365424 · TikTok @ailaopao

ailaopao-geo.pages.dev · 20 案例详版 → INDEX.md

私有部署 · POC 落地 · 长期驻场

Private LLM · RAG · Agents · Multi-model routing · AIOps

Sizing → Build → Ship → SLA · 全栈交付