

小驹智能 XiaoJu AI

小驹智能技术白皮书

开箱即用的买断制 AI 知识库问答系统

文档版本：V2.0（图表增强版）

发布日期：2026 年 6 月

官网：www.xiaojusmart.com.cn

联系：sales@xiaojusmart.com

“一分钟理解，三分钟部署，一次性买断永久使用”

目录

摘要 (Executive Summary)

核心价值 | 定价一览 | 行动号召

第一章：为什么需要轻量化 AI 知识库？

第二章：小驹智能架构设计

架构概览 | 单体 vs 微服务 | 零外部依赖 | 架构演进 | LLM 双模式 | 数据库设计

第三章：RAG 引擎详解

完整流程 | Hybrid 检索 | 性能对比 | 上下文管理 | 高级特性

第四章：部署与升级

部署流程 | 时间对比 | 配置分离 | 五步升级闭环

第五章：安全与授权

授权校验 | 金字塔分层 | 防白嫖机制

第六章：客户案例

指标对比 | 成本对比

第七章：定价与商业模式

四套餐 | 功能矩阵 | TCO 对比

第八章：技术路线图

第九章：前端交互设计

第十章：竞品对比

全景图 | 核心维度 | 差异化优势 | 适用场景

第十一章：结论与行动号召

摘要

Executive Summary

核心价值：开箱即用的买断制 AI 知识库问答系统，3 分钟部署，永久使用。



图 1：小驹 AI「三个一」核心价值主张 — 一分钟理解、三分钟部署、一次性买断

为什么需要小驹智能？

企业知识库当前面临三大痛点，传统方案（SaaS 订阅 & 开源自建）各有利弊，市场亟需一个轻量化方案：



定价逻辑：让客户 3 年内回本

小驹定价 = 竞品部署成本 × 13% ~ 66%



入门版 Starter	标准版 Standard	专业版 Pro	企业版 Enterprise
¥9,800 买断制 · 永久有效 RAG 问答 + 管理后台 内置 Ollama 开箱即用	¥19,800 买断制 · 永久有效 + H5 嵌入网页 + 系统监控告警	¥29,800 买断制 · 永久有效 + 微信小程序 + 高级RAG 品牌定制	¥59,800 买断制 · 永久有效 + 多租户架构 + 专属定制SLA保障

图 3：定价逻辑推导 — 小驹定价 = 竞品部署成本 × 13%~66%

买断制 vs SaaS 订阅：3 年 TCO 对比

以小驹AI 专业版 (¥29,800 买断) vs SaaS 专业版 (¥30,000/年) 为例



图 4：买断制 vs SaaS 订阅 3 年 TCO 对比

行动号召



图 5：从了解到达成合作只需三步

第一章：为什么需要"轻量化 AI 知识库"？

核心论点：现有方案太重，中小企业需要"买断制 + 零外部依赖"的轻量化方案

1.1 当前市场的两个极端

AI 知识库赛道呈现两个极端——SaaS 订阅（数据上云、订阅费贵）和开源自建（部署复杂、需要专职 IT）。市场需要一个中间方案。

SaaS 订阅方案

Coze 企业版 ¥30,000/年
数据上云，合规风险高
长期成本不可控

开源自建方案

Dify / MaxKB
需专职 IT 部署维护
7~8 个容器，部署半天起

小驹智能

3 分钟部署，3 个容器
买断制，永久使用
零外部依赖，数据私有化

1.2 目标客户的真实痛点

- **中小企业 (50-500人)**：无专职 IT，3 分钟部署 + 买断制
- **政府机构**：数据不能上云，完全私有化 + 符合政府采购习惯
- **教育机构**：预算有限，¥9,800 起买断制、无年费
- **医院/金融机构**：合规要求私有化，零外部依赖降低故障点

1.3 小结

小驹智能的定位：不是 SaaS 的替代品，不是开源方案的简化版，而是两者之间的轻量化方案——私有化部署 + 零外部依赖 + 买断制。

第二章：小驹智能架构设计

架构设计的核心矛盾是“简单性 vs 扩展性”。V1.0 选择单体架构，是因为目标客户不需要分布式。

2.1 架构概览

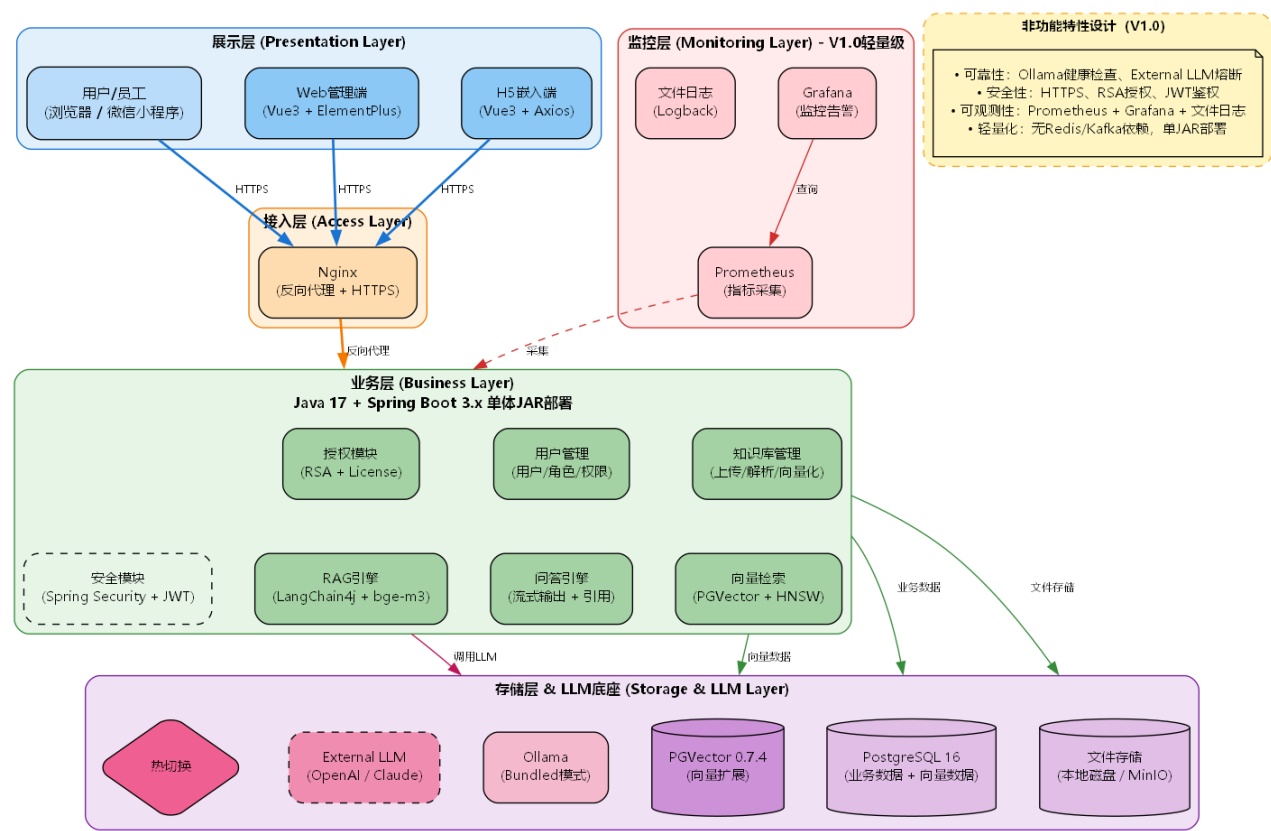
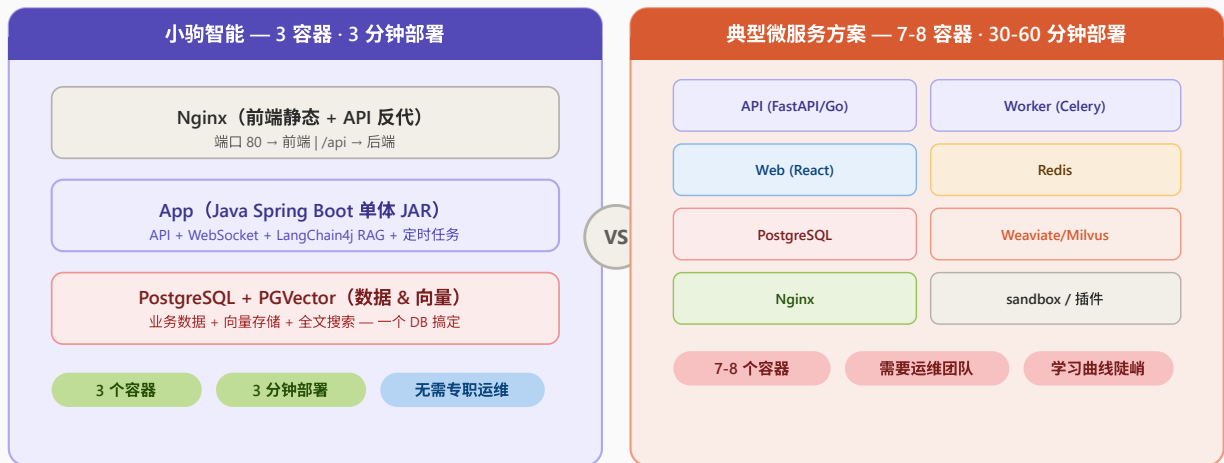


图 6：小驹智能架构总览（Bundled & External 两种模式，三大容器：App + PostgreSQL/PGVector + Ollama）

2.2 为什么选择单体架构而不是微服务？

单体式 vs 微服务：为什么我们选择简化架构

容器越少 = 部署越快 = 运维越简单



简化的代价：单点故障风险 (V2.0 解决) | 增益：部署 60x 更快，运维成本降低 80%

图 7：小驹 3 容器 vs 典型微服务 7-8 容器 — 部署速度 60× 更快

2.3 零外部依赖设计哲学

零外部依赖设计哲学

Dify 的 7 项外部依赖 vs 小驹的 2 项 — 减少的不是服务，是心智负担

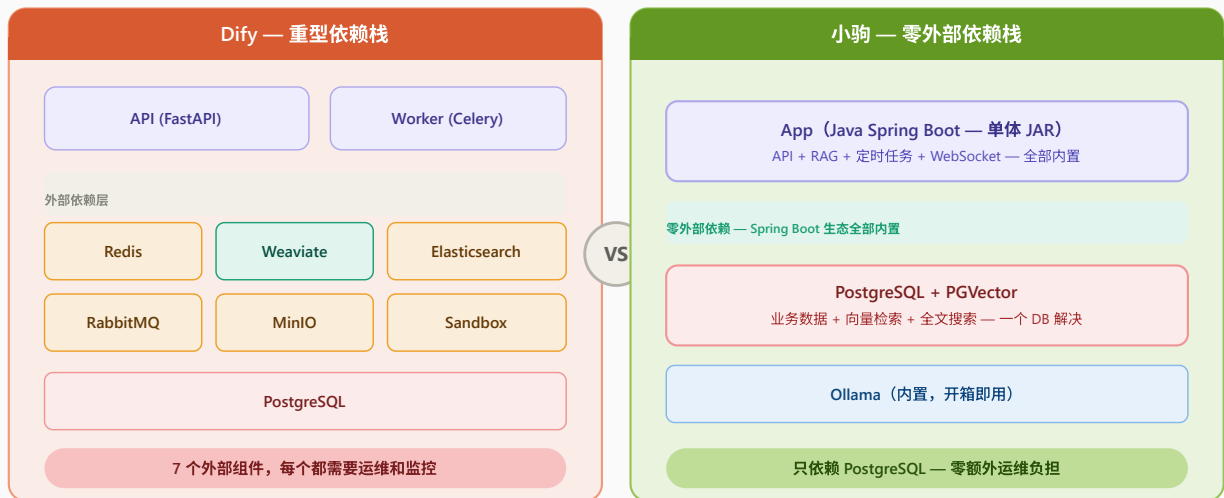


图 8：Dify 的 7 项外部依赖 vs 小驹的零外部依赖

核心原则：

- 无 Redis：会话状态存 PostgreSQL
- 无 Kafka：异步任务用 DB 状态表 + 定时轮询
- 无 MinIO：文档存本地磁盘
- 单库解决：PostgreSQL + PGVector，一库多用

2.4 架构演进路径

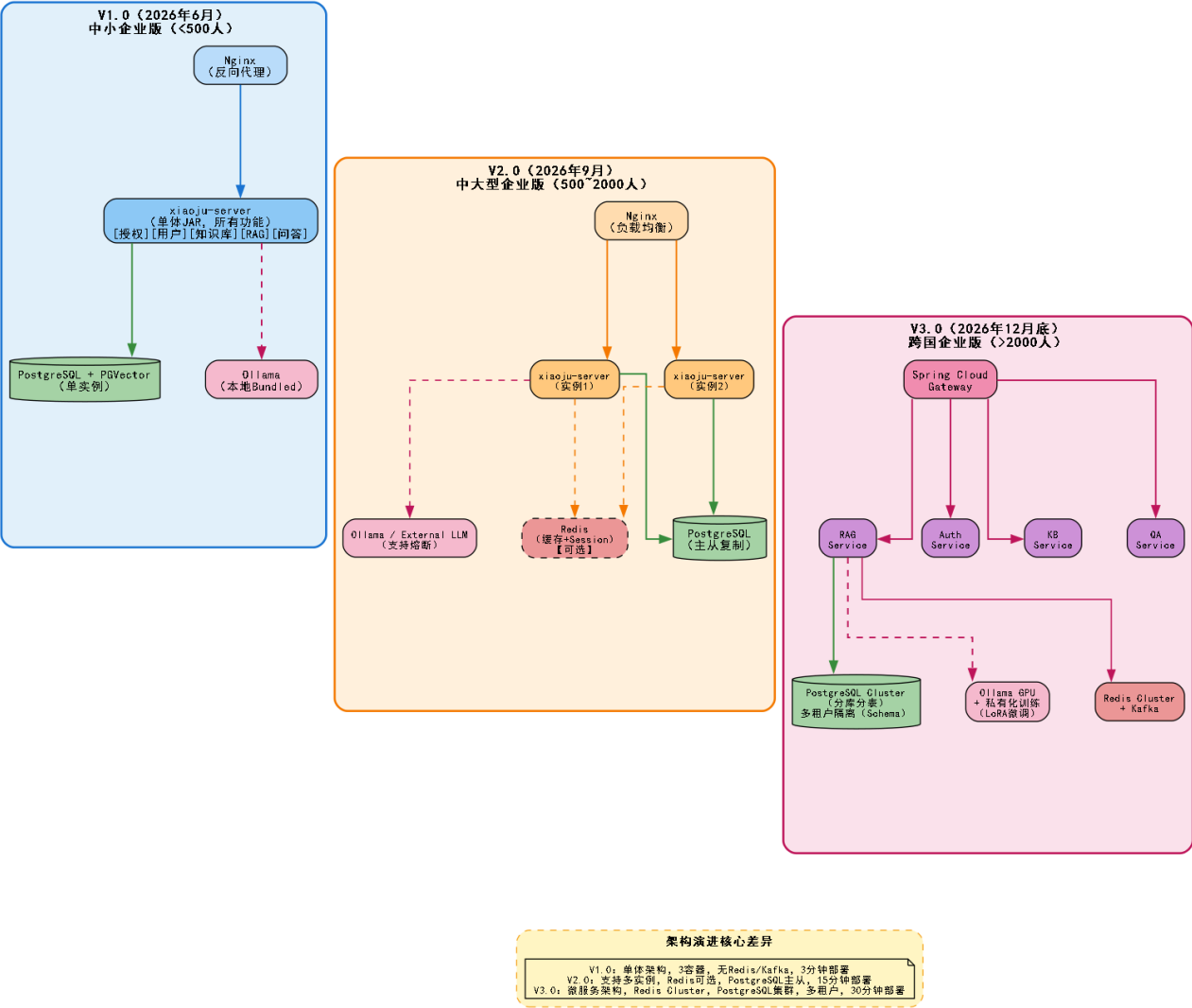


图 9: V1.0 → V2.0 → V3.0 三阶段演进路径

2.5 版本选择决策树

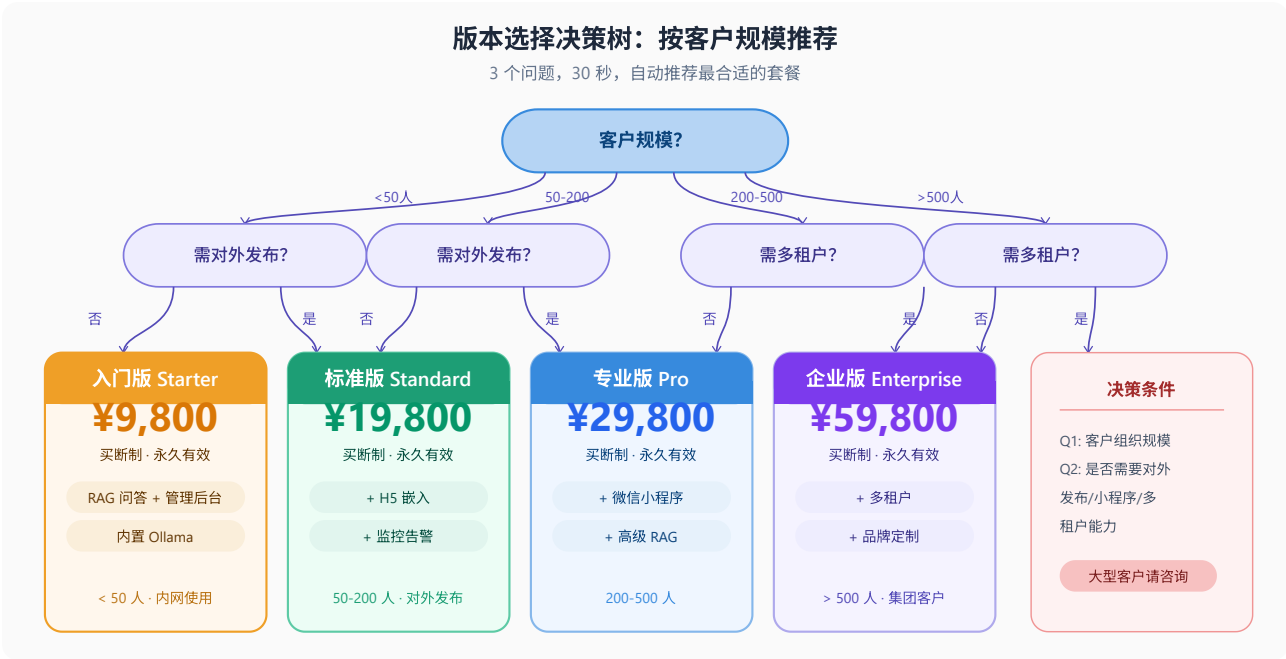


图 10：按客户规模自动推荐最佳套餐

2.6 LLM 部署双模式

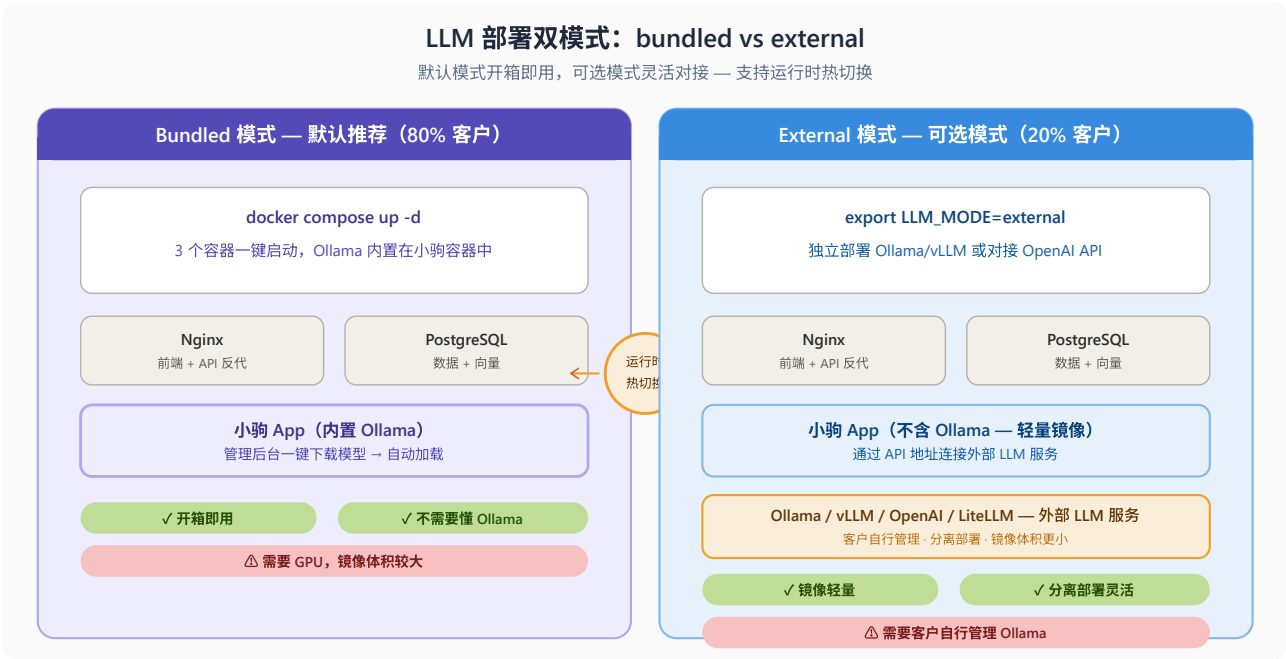


图 11：Bundled（内置 Ollama，开箱即用）vs External（对接外部 LLM） — 支持运行时热切换

2.7 数据库设计

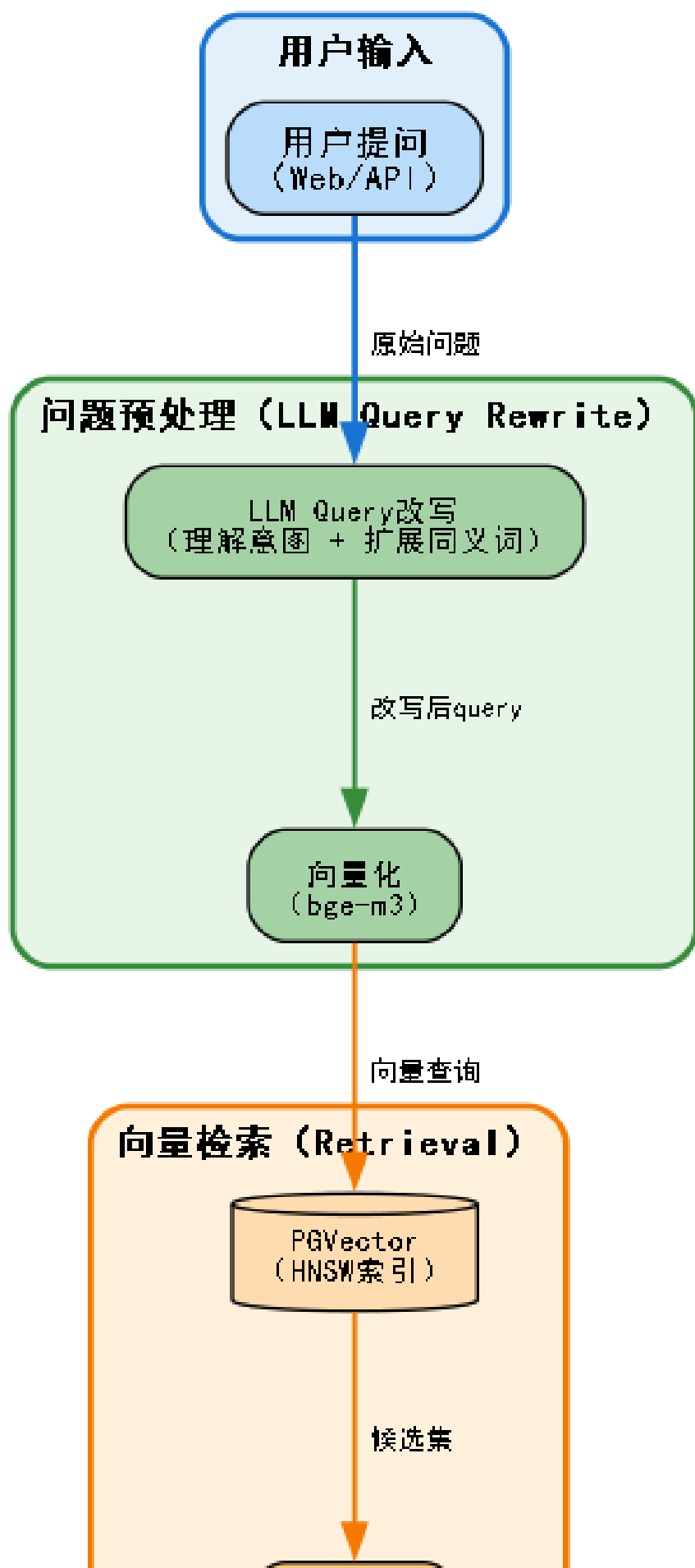
为什么用 PGVector（而不是独立向量数据库）？

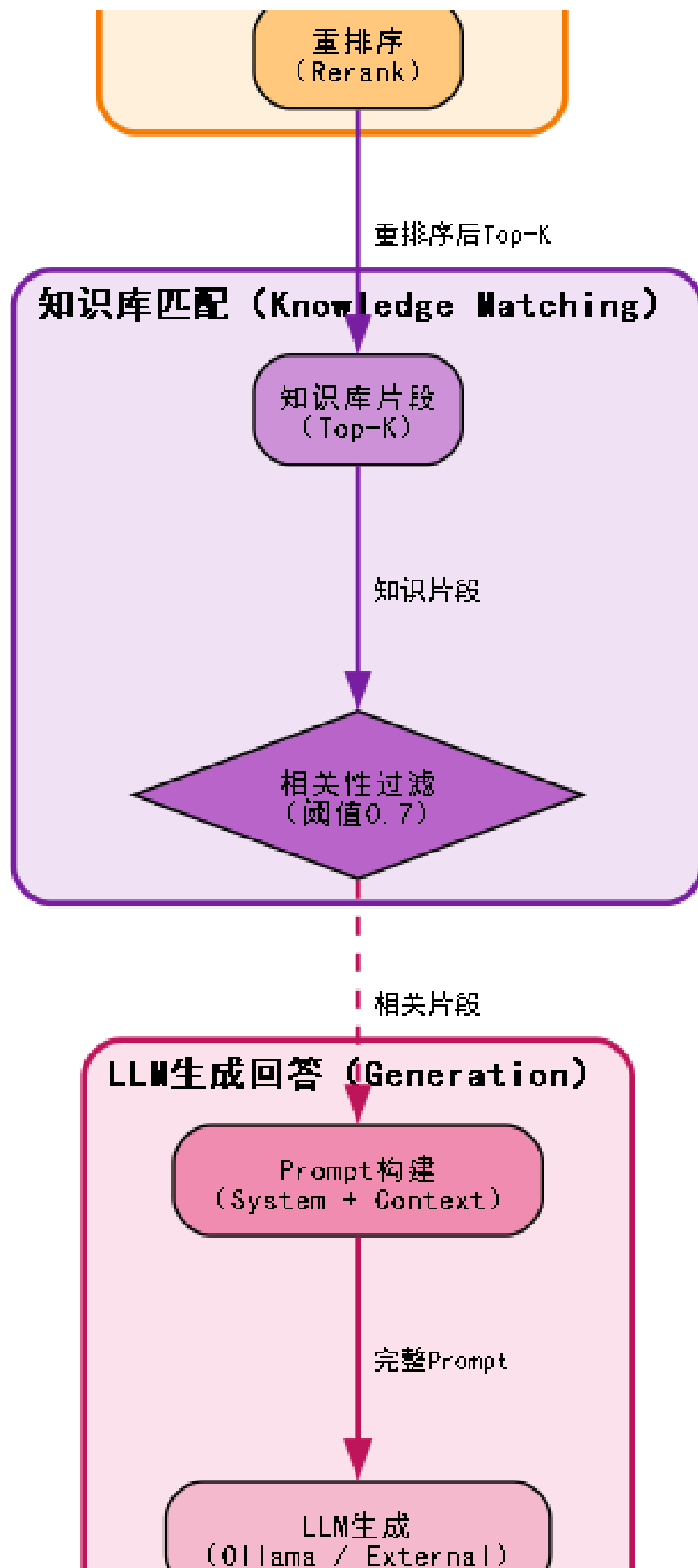
对比项	PGVector	Milvus	Pinecone
部署复杂度	简单 PostgreSQL 自带	复杂 7 个容器	需云服务
成本	免费	需独立服务器	按向量数收费
适用场景	中小企业 (<100万文档)	大型企业 (>1000万)	云原生 SaaS

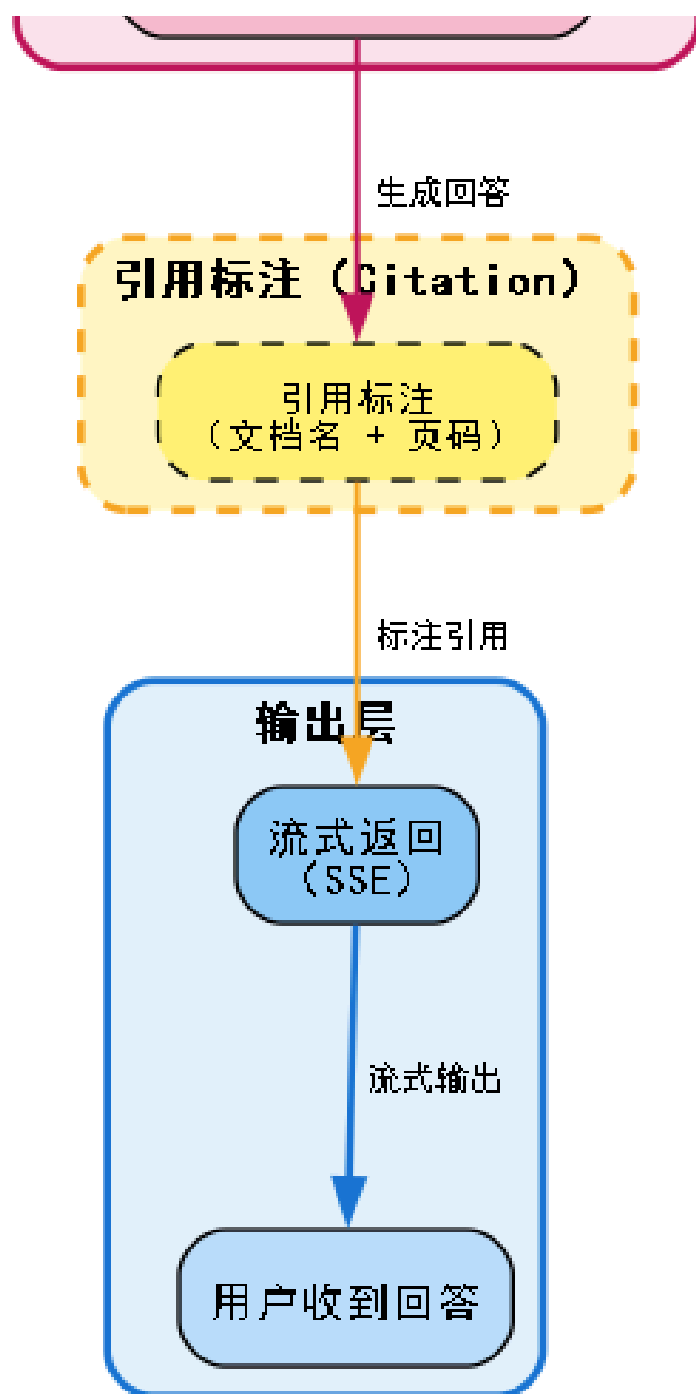
第三章：RAG 引擎详解

RAG（检索增强生成）是小狗智能的核心引擎，V1.0 支持离线和在线两阶段处理。

3.1 RAG 完整流程







RAG核心特性 (LLM Query Rewrite 方案)

- Query改写: LLM理解用户意图, 扩展同义词, 提升召回率
 - 向量模型: bge-m3 (1024维)
 - 检索算法: HNSW (效率 vs 精度平衡)
 - 重排序: Rerank模型提升相关性
 - 引用标注: 可追溯至原始文档
 - 流式输出: SSE技术, 延迟<500ms

图 12: RAG 离线入库 + 在线问答双流程

3.2 Hybrid 混合检索原理

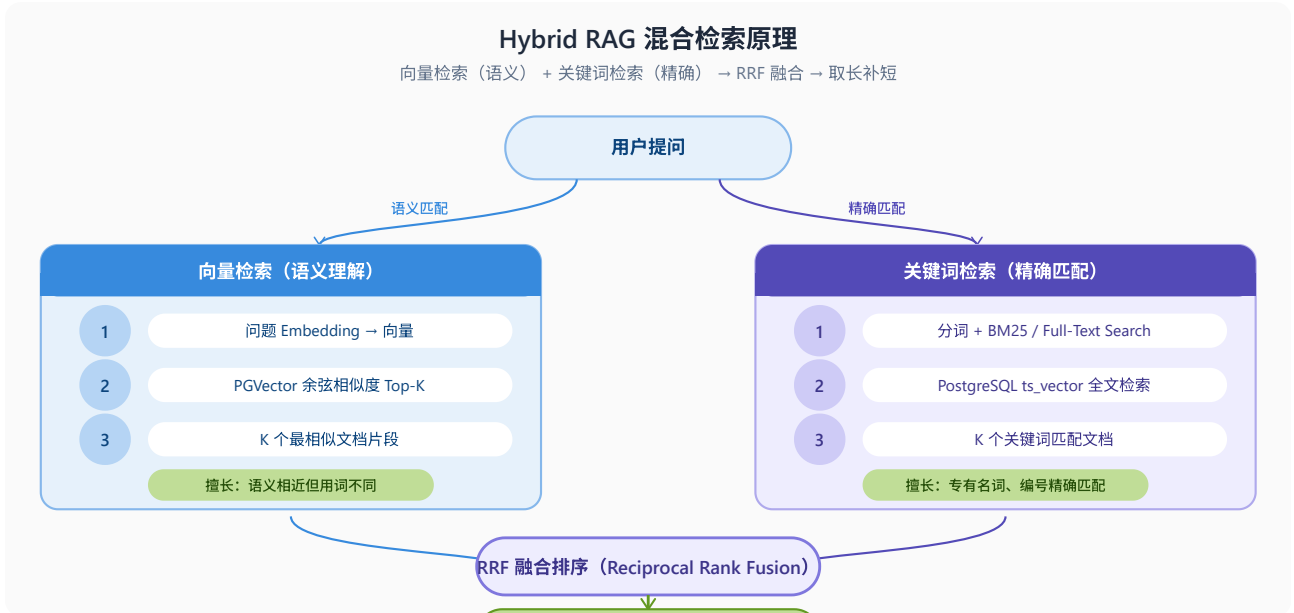


图 13: 向量检索（语义） + 关键词检索（精确） → RRF 融合 → 最佳结果

3.3 检索性能对比

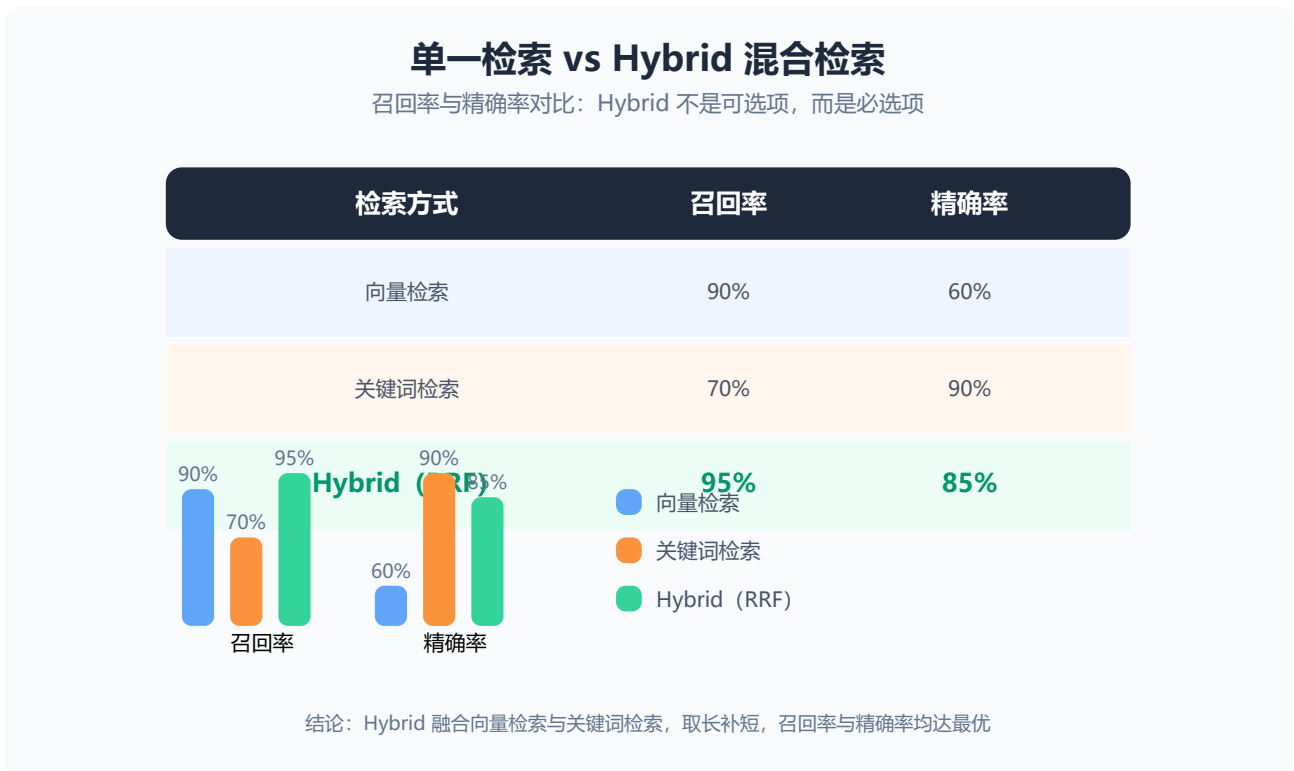


图 14: 向量/关键词/Hybrid 三种方式的召回率与精确率对比

3.4 多轮对话上下文管理

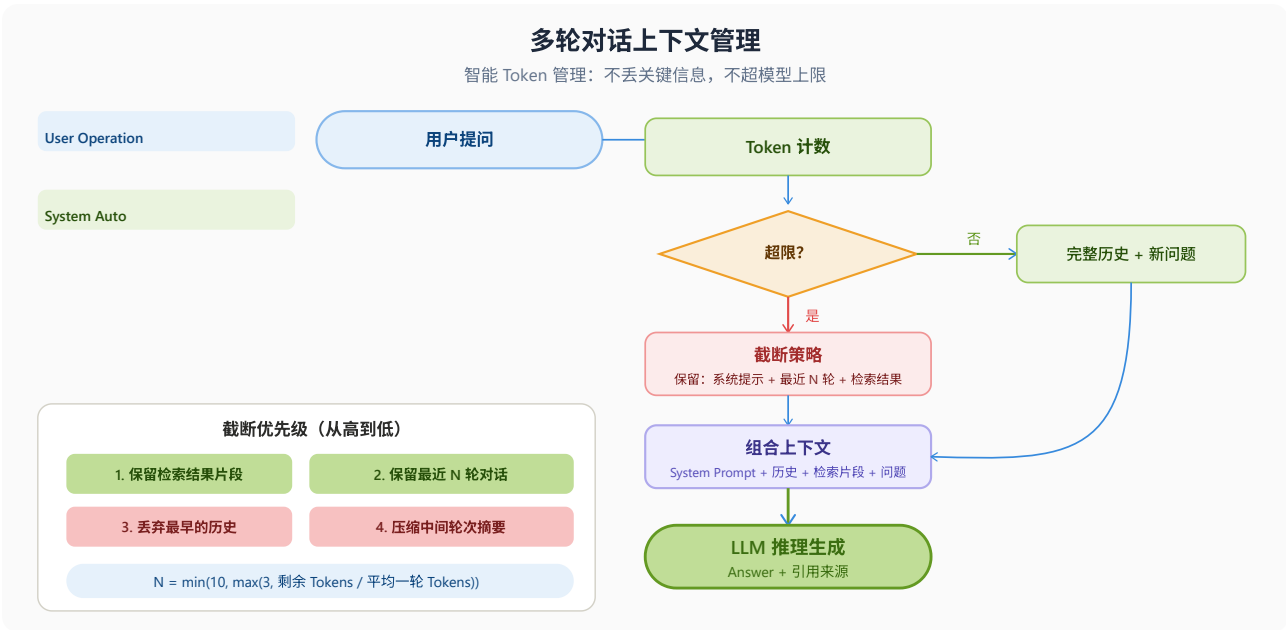


图 15：智能 Token 管理 — 不丢关键信息，不超模型上限

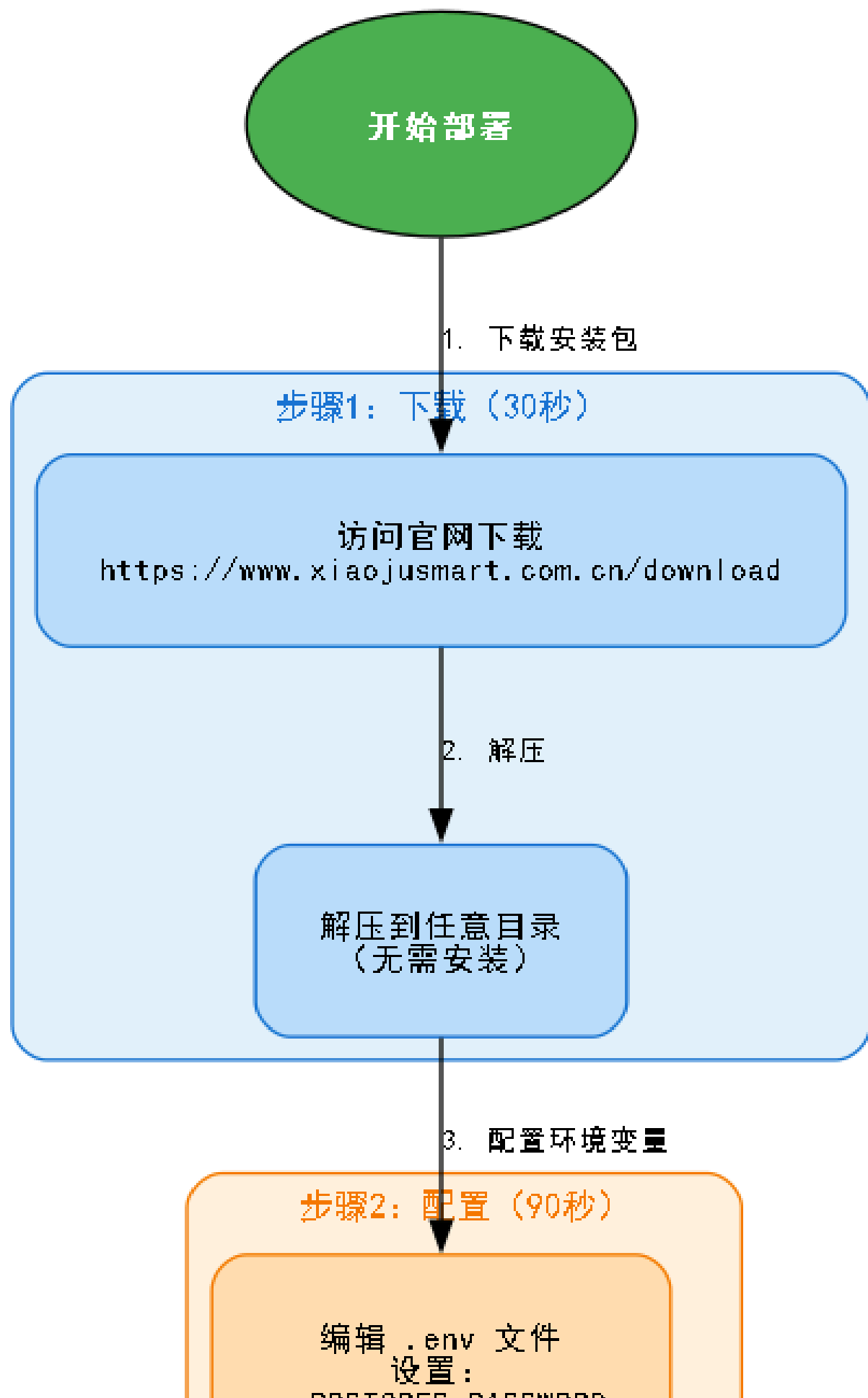
3.5 RAG 高级特性

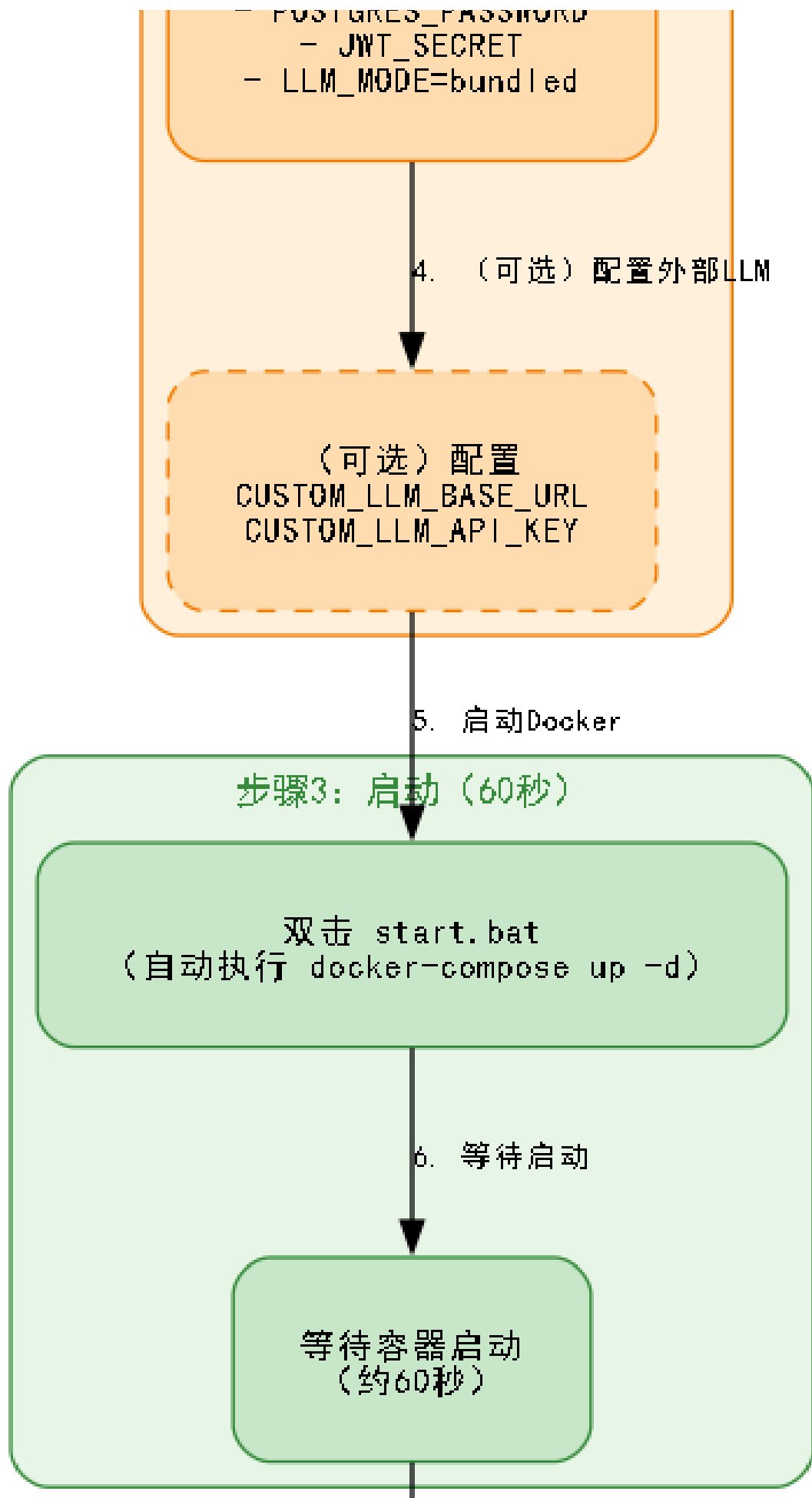


第四章：部署与升级

部署哲学：让客户 3 分钟上线，升级像"撤销"一样简单。

4.1 部署流程





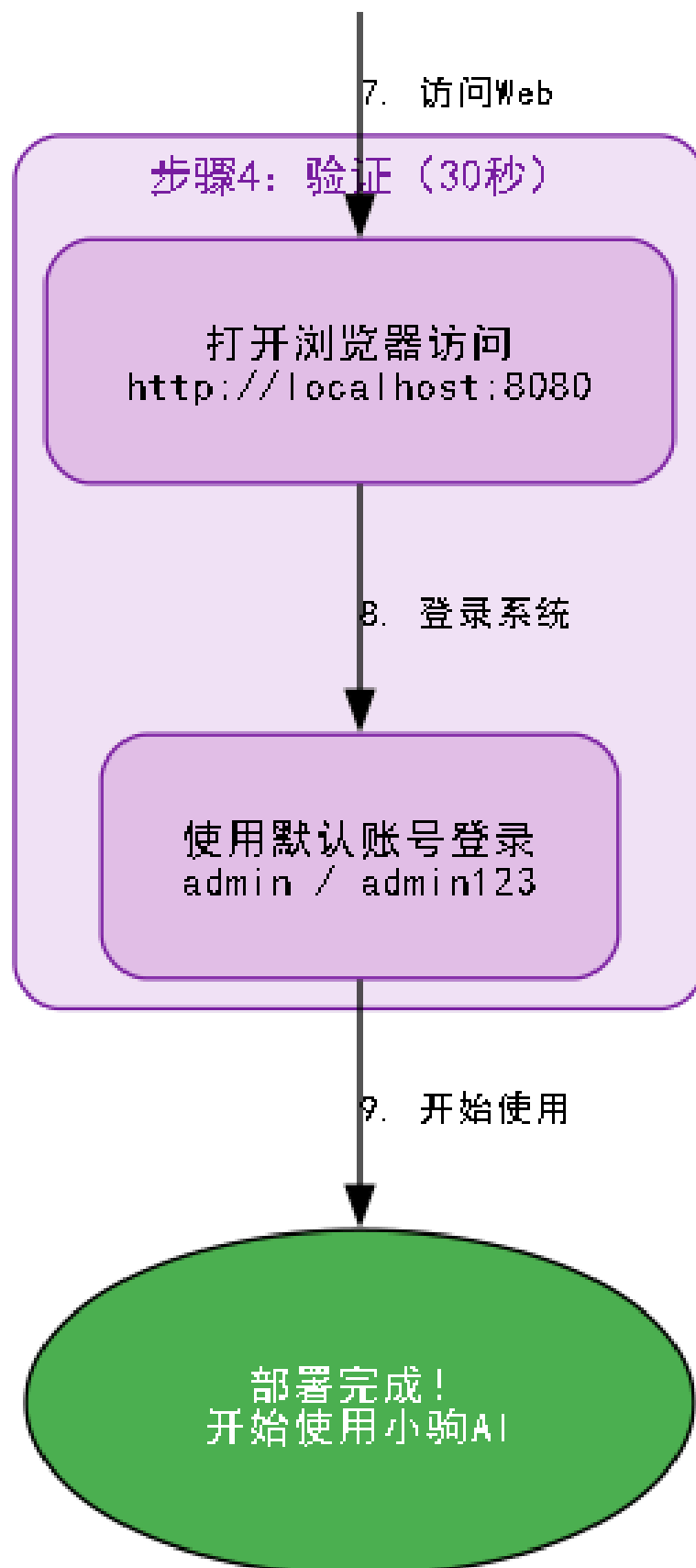


图 16: 6 步完成部署, 总耗时 < 3 分钟

4.2 部署时间对比

部署时间对比：为什么小驹只需 3 分钟？

零外部依赖是 3 分钟部署的核心原因

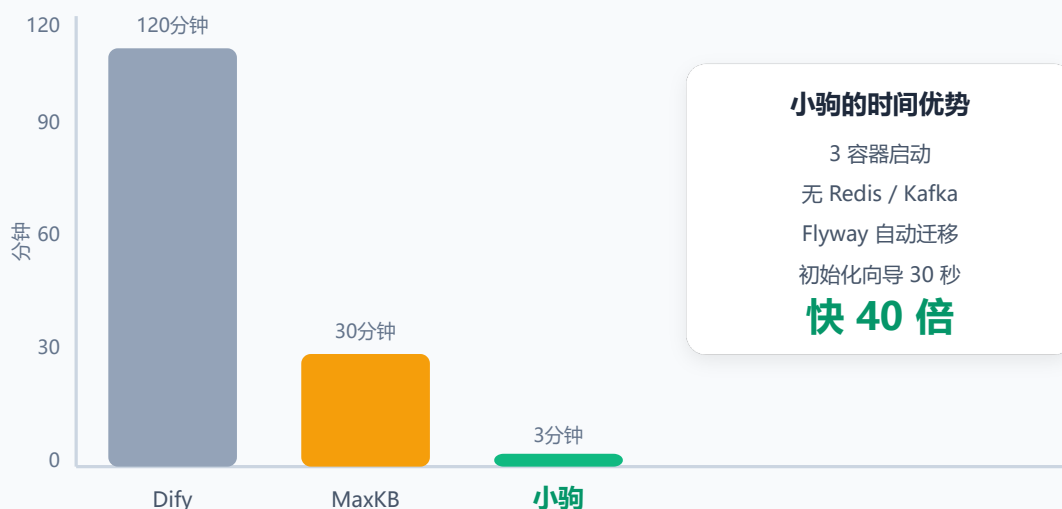


图 17：小驹 vs Dify vs MaxKB — 3 分钟 vs 1.5 小时 vs 半天

4.3 配置文件分离策略

配置文件分离策略：升级零覆盖

系统配置 vs 客户配置 — 分离即安全，升级不丢失



图 18：application.yml（系统配置，升级覆盖）vs application-deploy.yml（客户配置，升级保留）

4.4 五步升级闭环

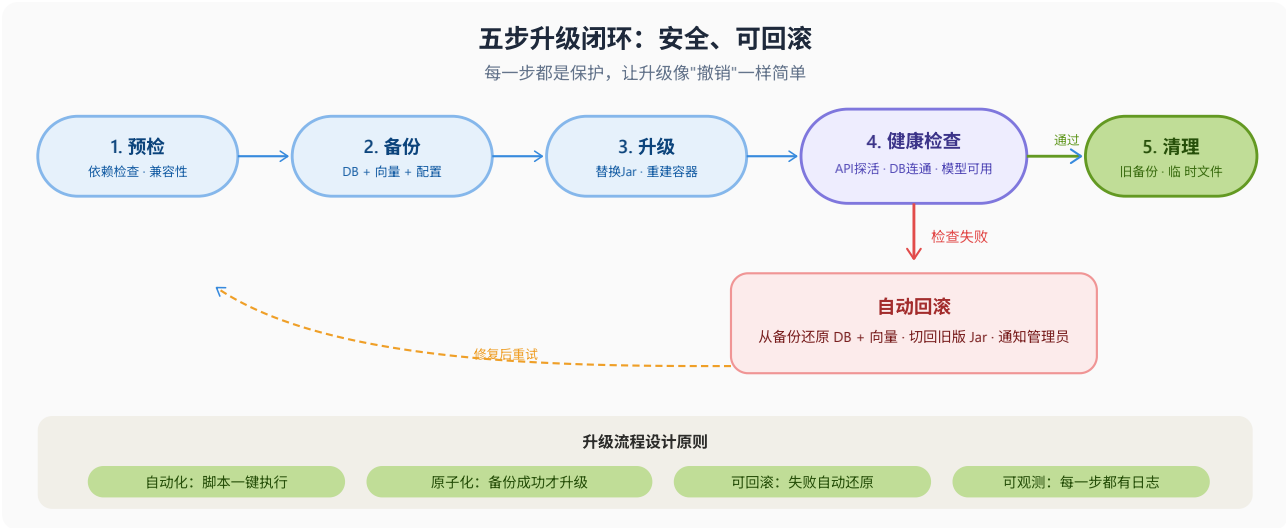
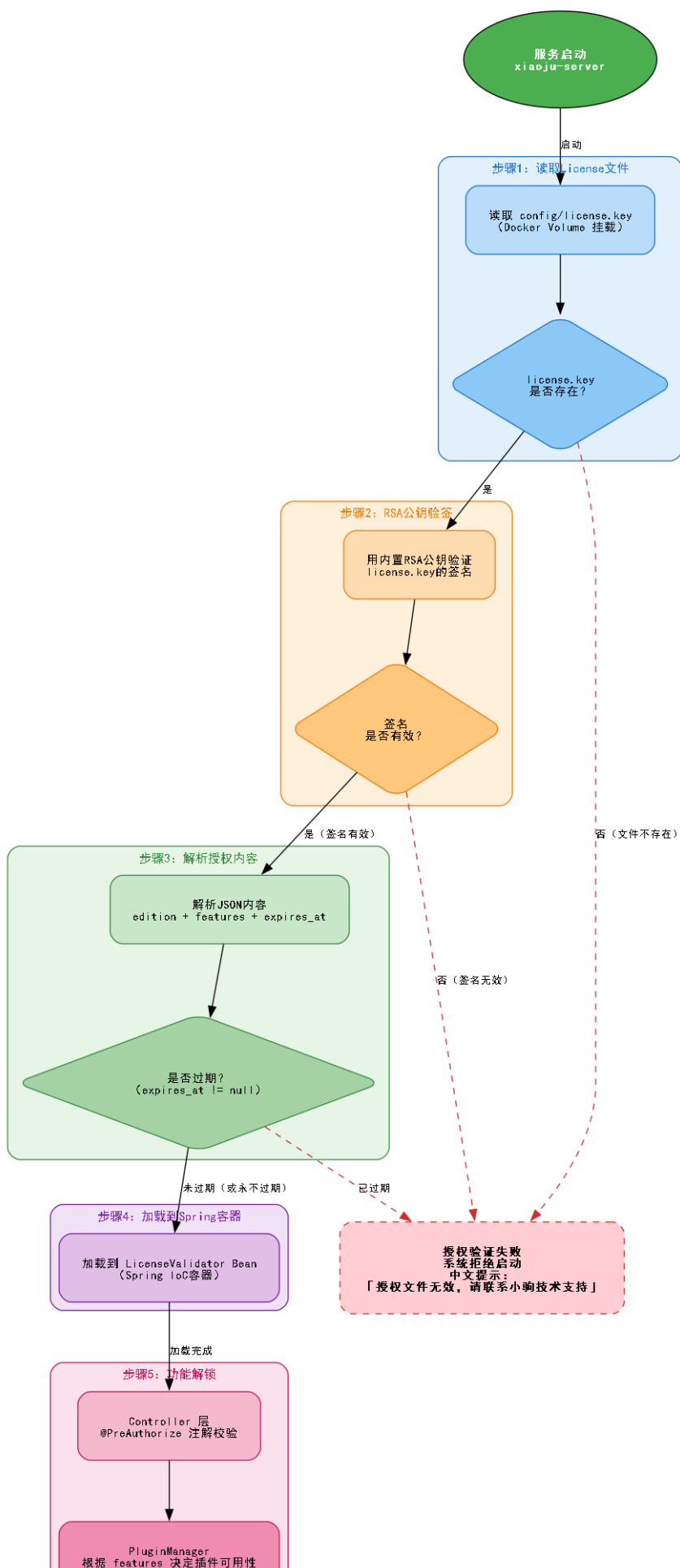


图 19：预检 → 备份 → 升级 → 健康检查 → 失败回滚 — 安全可控

第五章：安全与授权

授权机制不是"防盗版"，而是"保护客户的正版权益"。

5.1 授权校验流程



设计原则:

1. 私钥永不离开开发环境 (离线存储)
2. 授权文件不写机器指纹 (降低运维成本)
3. 即使公钥泄露, 攻击者也只能验证签名, 无法伪造授权文件
4. 永久有效 (expires_at = null), 买断制不设过期时间

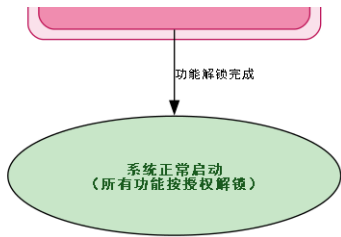


图 20：RSA 签名验证流程 — 启动时强制校验

5.2 金字塔安全分层策略



图 21：RSA 签名 → 硬件绑定 → 代码混淆 → 服务端校验，破解成本逐级递增

5.3 防"白嫖"机制



图 22：三种破解方式 vs 小驹防护措施 — 提高破解成本，让正版用户有面子

核心原则：

- 🚫 不联网：全离线校验，尊重隐私
- ♾️ 不限制：安装次数无限制，迁移/重装不受影响
- 👉 买断制是信任，不是监控

第六章：客户案例

三个真实场景，验证小驹智能在不同行业的落地效果。

6.1 客户案例指标对比

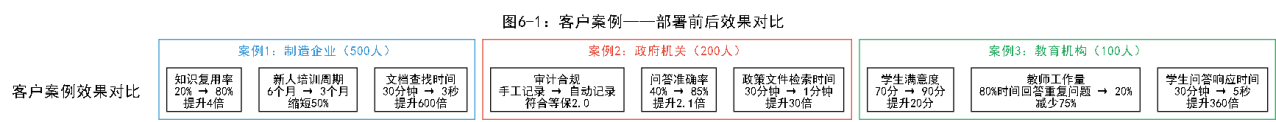


图 23：三个案例的部署前后关键指标对比

6.2 三年总成本对比



图 24：三个真实场景 — 小驹 AI 均显著降低 TCO

案例	客户规模	推荐套餐	Dify 3年成本	小驹成本	节省比例
制造企业	500 人	专业版 ¥29,800	¥225K	¥29.8K	Dify 的 13%
政府机关	200 人	入门版 ¥9,800	¥225K	¥9.8K	Dify 的 4%
教育机构	100 人	标准版 ¥19,800	¥225K	¥19.8K	Dify 的 9%

第七章：定价与商业模式

定价逻辑：让客户 3 年内回本。小驹定价 = 竞品部署成本 × 13%~66%。

7.1 四套餐定价



图 25：定价逻辑推导 — 让客户 3 年内回本

7.2 套餐功能矩阵

套餐功能矩阵：按需选择				
功能随版本递增，从基础问答到企业多租户				
功能	入门版	标准版	专业版	企业版
RAG 智能问答	✓	✓	✓	✓
管理后台	✓	✓	✓	✓
最大用户数	50人	500人	500人	无限制
H5 网页嵌入	✗	✓	✓	✓
监控告警	✗	✓	✓	✓
微信小程序	✗	✗	✓	✓
高级 RAG 策略	✗	✗	✓	✓
多租户架构	✗	✗	✗	✓
价格（买断制）	¥9,800	¥19,800	¥29,800	¥59,800

图 26：四套餐 × 7 维度的功能覆盖对比

7.3 买断制 vs SaaS 订阅（3 年 TCO 对比）

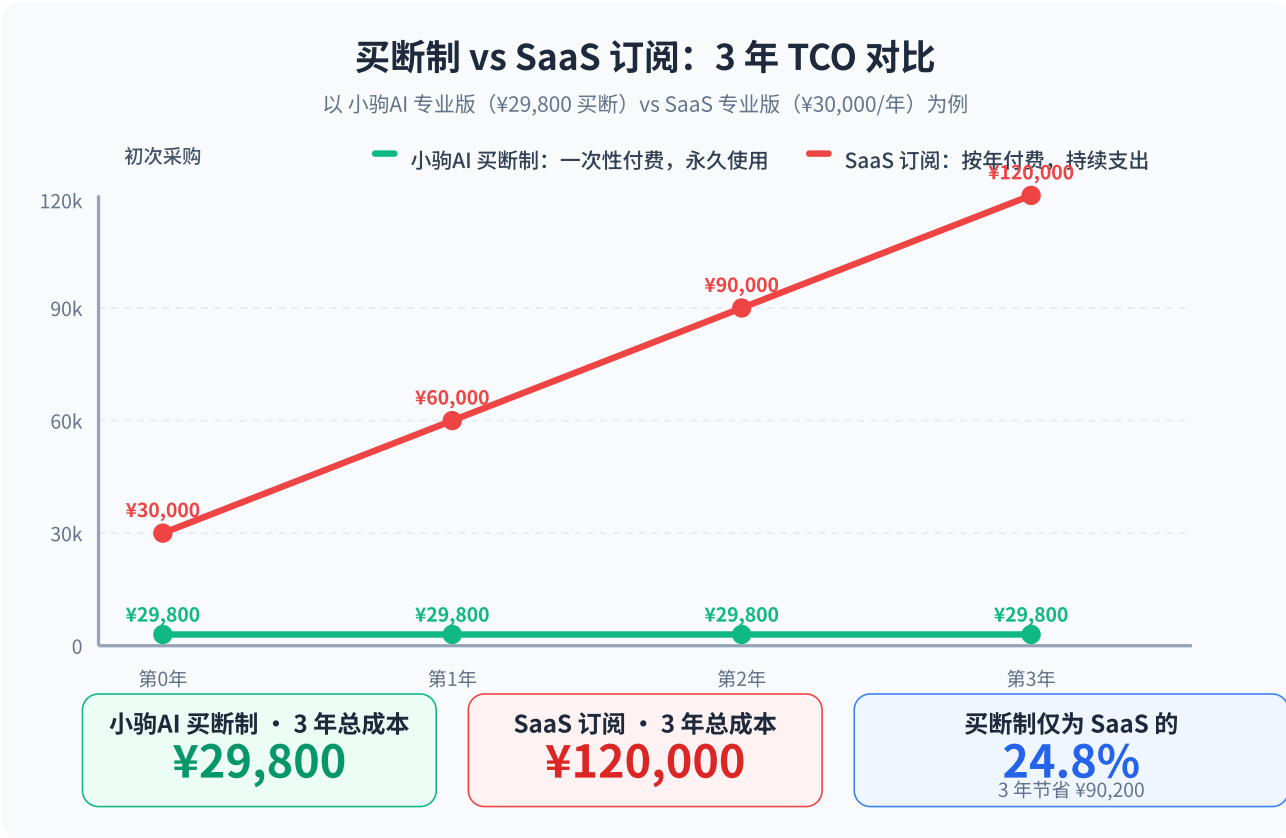


图 27：逐年成本累加折线图 — 清晰展示长期成本优势

第八章：技术路线图

技术路线图不是"画大饼"，而是"客户需求驱动的技术演进"。

8.1 版本路线图

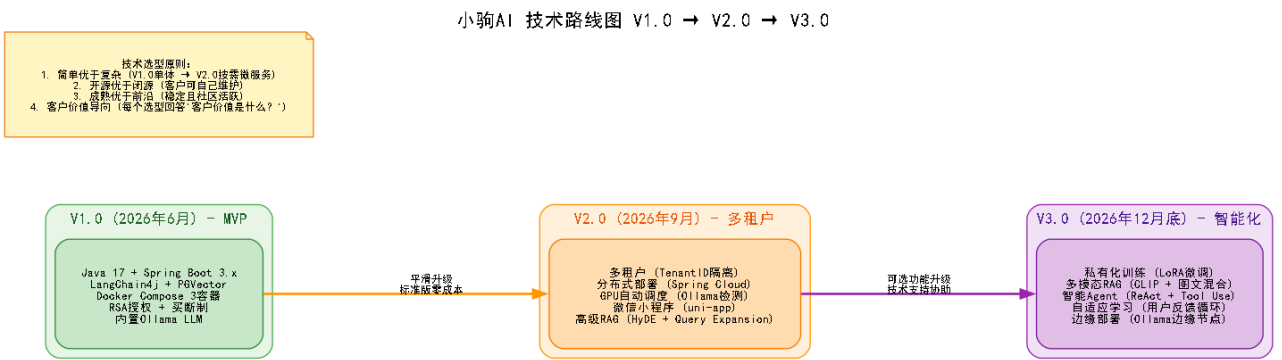


图 28: V1.0 (2026.06) → V2.0 (2026.09) → V3.0 (2026.12) 时间线及核心特性

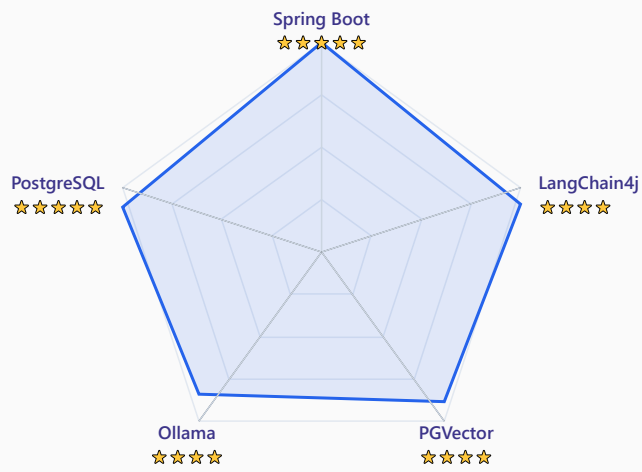
版本	发布时间	核心特性	目标客户
V1.0	2026 年 6 月	轻量化单体架构，3 分钟部署	中小企业 (<500 人)
V2.0	2026 年 9 月	微服务架构，多租户，GPU 调度	大型集团 (>500 人)
V3.0	2026 年 12 月	AI 原生，私有化训练，多模态 RAG	跨国企业 (>2000 人)

升级路径：V1.0 → V2.0 平滑升级，标准版零成本；V2.0 → V3.0 按需选择。

8.2 技术成熟度评估

技术成熟度评估

成熟优于前沿，稳定优于追新 — 每项技术都经过生产验证



技术债务评估

Spring Boot 3.x
企业级标准，20年生态

PostgreSQL 14+
最成熟的开源关系数据库

LangChain4j
Java生态首选，社区活跃

PGVector
PostgreSQL原生，零运维增量

Ollama
本地LLM事实标准，GitHub 100K+

无技术债务，全栈可维护

图 29：四维度技术成熟度评估 — 成熟优于前沿，稳定优于追新

第九章：前端交互设计

好的前端，让用户不用看说明书就能用。

9.1 管理后台

基于 Vue 3 + Element Plus 构建的响应式管理界面，支持：

- 知识库管理（创建/编辑/删除知识库）
- 文档上传与管理（批量上传、自动切片入库）
- 对话历史查看与反馈（👍/👎）
- 系统监控与审计日志

9.2 H5 嵌入控件

轻量级 H5 问答控件，可嵌入公司内网、公众号、第三方网站：

- 一键复制嵌入代码
- 支持自定义主题色
- 支持品牌 Logo 替换（专业版+）

9.3 微信小程序

专业版及以上支持微信小程序访问，员工通过微信即可使用 AI 知识库问答。

第十章：竞品对比

坦诚对比，不回避劣势，让客户做出明智选择。

10.1 竞品全景图



图 30：部署复杂度 vs 长期成本 — 小驹定位在左下角「理想区间」

10.2 核心维度对比

图1-1：传统知识库 vs 小驹AI——痛点与解决方案对比

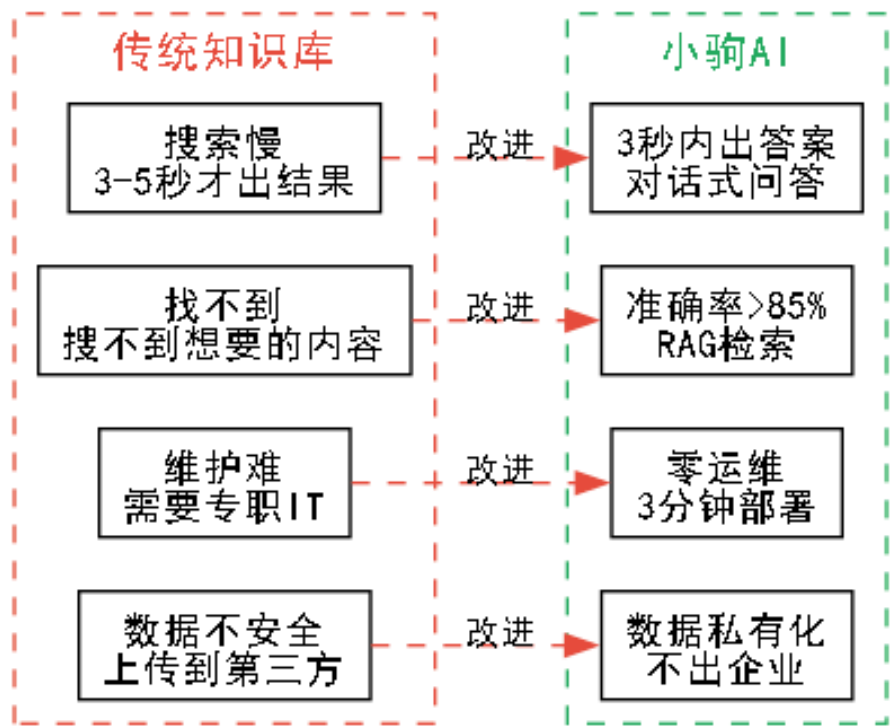


图 31：小驹 vs Dify vs MaxKB vs Coze 全维度对比

10.3 小驹智能差异化优势

- 1. 部署门槛最低：3 个容器、零外部依赖、3 分钟部署
- 2. 长期成本最优：买断制确保成本可控，3 年 TCO 是竞品的 8%~50%
- 3. 数据安全最高：完全私有化 + RSA 签名授权 + 零数据外传

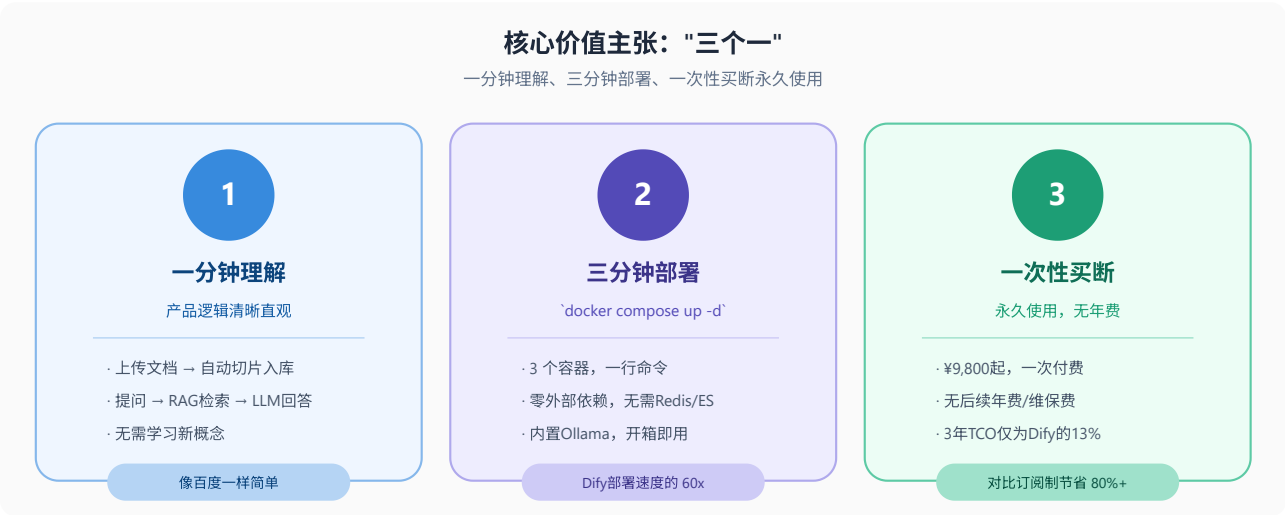
10.4 坦诚说明：小驹当前不适合的场景

场景	推荐产品	原因
需要 AI 工作流编排	Dify	小驹 V1.0 未实现，规划中 V2.0
需要多模态 RAG（图片/视频）	Dify	小驹 V3.0 规划中
需要大量社区插件	Dify	Dify 开源社区更活跃
需要云端免运维	Coze 企业版	小驹是私有化部署

第十一章：结论与行动号召

好的结论，不是"重复全文"，而是"强化核心价值"+"推动客户行动"。

11.1 全文总结



小驹智能的核心价值：「一分钟理解，三分钟部署，一次性买断永久使用。」

11.2 立即行动



图 32：从了解到达成合作只需三步

三步上手小驹 AI

第一步：下载试用

访问
www.xiaojusmart.com.cn/download
免费试用 7 天

第二步：预约演示

填写官网表单
技术团队 1v1 远程演示
(30 分钟)

第三步：采购部署

签署协议 → 获取
License
→ 部署到生产环境

联系方式

类型	邮箱	说明
销售咨询	sales@xiaojusmart.com	产品购买、方案咨询
技术支持	support@xiaojusmart.com	部署问题、技术答疑
商务合作	business@xiaojusmart.com	渠道合作、战略合作

小驹智能 XiaoJu AI

开箱即用的买断制 AI 知识库问答系统

文档版本：V2.0（图表增强版）

发布日期：2026 年 6 月

图表数量：31 张（SVG 矢量图 + PNG 渲染图）

总页数参考：~35 页（含图表）

"好的白皮书，不是「我有什么卖给你」，
而是「你的痛点，我有解药」。"