# Agent面试题整理

## Agent 面试题整理

整合两篇面经，按主题分类去重。纯前端八股（CSS 动画、Vue 响应式等）和纯 HR 问题已剔除；与 Agent 工程实践相关的问题保留。

---

### 一、Agent 架构设计

1. 你做的 AI 系统主要是做什么的？整体功能架构是怎样的？
2. 从架构角度看，AI 系统应该分成哪些层？
3. 假设现在让你做一个某门店的 AI 数字化系统（或通用型智能客服），你会怎么设计？技术选型怎么做？
4. 抽象通用 AI 能力模块做了什么？
5. 从 0 到 1 搭建一个 AI 项目，怎么设计架构？怎么拆模块、数据怎么流转？
6. 传统业务系统的后端分几层？你搭框架的整体思路是什么？

### 二、多 Agent（Multi-Agent）

1. 什么是多 Agent？为什么需要多 Agent？
2. 单个 Agent 也可以分配多个职能，为什么还要多 Agent？多 Agent 主要解决什么问题？
3. 什么场景下需要多 Agent 架构？有没有构建过 Multi-Agent 系统？
4. 多 Agent 到底是如何决策的？比如中途拉一个新的 Agent 进来，为什么做这样的决策？
5. 多 Agent 协作时，单 Agent 交互协议有什么坑？注意的点？
6. 主子模式下，子 Agent 产生幻觉怎么办？
7. 如何让多 Agent 自由协同而不出现"瞎聊停不下来"的情况？
8. 没有用强状态机，如何保证任务流转不偏移？
9. 动态路由如何根据条件分配任务分支？
10. Shared State / Agent Team / 主子结构之间的区别是什么？

### 三、记忆与上下文管理

1. 短期记忆和长期记忆如何区分管理？
2. 上下文压缩的策略是什么？压缩时如果大模型能力不足导致质量下降，怎么解决？
3. 上下文压缩的技术细节？
4. 记忆系统如何设计？为什么分五层（或多层）？
5. 记忆的治理：如何做记忆合并？出现冲突怎么办？

6. 低权威的记忆如何晋级？聊天里的消息可能变成类似架构决策吗？
7. 记忆是本地文件方案还是在线方案？为什么这样选？
8. Session 隔离如何做？其他项目的记忆会不会搞乱当前项目？
9. 长程任务中，Agent 做着做着压缩了上下文丢失了很多细节，怎么处理？
10. 多轮对话中（比如五个人聊了一百多轮），如何管理上下文？摘要只在结束时做会不会失忆？
11. 每个 Agent 只关注自己相关的上下文，每一轮聊天记录拼入聊天，有什么优化方案？

## 四、RAG 与知识库

1. 简单介绍一下 RAG 的流程。
2. 知识库的检索是怎么做的？如何加速检索过程，有什么策略？
3. 知识库的数据源来自哪里？文档、PDF、课件是如何切分、入库、人工审核的？
4. ES 的切片方式有哪些？ES 有哪些分词机制？
5. 使用什么向量库？选型依据是什么？向量库目前的数据量？
6. 向量检索使用了哪些工具和模型？如何实现、效果怎样、如何优化？
7. Elasticsearch + 向量库怎么做混合检索？整体流程？
8. 混合检索已经做了分组、排序、去重，为什么还需要再次精排（Re-rank）？
9. 项目中有没有使用领域词典？
10. grep（全文检索）和向量检索到底如何结合？BM25 + 向量？
11. 什么是查询改写？作用是什么？
12. 政策文件场景下，怎么确认召回数据的准确性？
13. AI 回答效果差的时候，从哪些环节调整优化？

## 五、大模型与 Prompt

1. 大模型的选型逻辑是什么？
2. 有没有做过国内外 AI 工具或模型的对比？同一个模型不同工具，或同一个工具不同模型，效果差异研究过吗？
3. 每个模型的优势和劣势是什么？你对每个模型的 taste 是怎样的？
4. 大模型有没有做过微调？是直接接入第三方 API 还是私有化部署？
5. 如果政府项目限制使用国外模型，完全使用国内工具有没有影响？
6. 切换国产模型的过程中，有没有遇到 harness 要调整的情况？如何调整？
7. 用较弱模型（如 GLM-4）时 tools 调用不准确，如何解决？
8. Prompt 提示词是你编写还是专人负责？你会怎么撰写需求，交给 AI 生成提示词？
9. AI 输出如何保证序列化 / 结构化？为什么选择 Markdown 而不是 JSON？

## 六、Agent 工作流与工具

1. 你日常的 AI 开发工作流程是怎么样的?
2. 有在使用什么 MCP 或 SKILL? 说说 MCP、Function、Skills 三者的区别。
3. AI 调用工具失败了怎么解决?
4. 有没有做基于 AI 的界面测试? AI 是否能覆盖前端点击、输入、提交等自动化测试?
5. 接到一个需求后，你会怎么做?
6. 只有口头需求、无完整页面设计时，如何借助 AI 输出页面原型与 UI/UX 方案?
7. 有没有 AI 编程规范? AI 写的代码相对不可控，有没有 code review 流程?
8. 没用 AI 和用了 AI 后，效率提升有多大? 按百分比说你觉得能提升多少?
9. 你会用什么 AI 工具来完成项目规划、草图、文档和开发?

**七、Agent 自进化与评测**

1. Agent 如何自进化? 要具备自我进化，你觉得重点是什么?
2. 自演进和自进化是如何做的?
3. 有没有做 benchmark? 如何测评? 修改完 harness 后如何评估效果?
4. eval / benchmark 相关：准备虚拟环境、真实工具，搭环境到 agent RL 的流程?
5. 如何横向对比不同模型的 thinking level ROI?
6. 如何对自己的记忆系统、A2A 协议做 eval?
7. 你觉得未来人和 Agent 的协作模式可能是什么?

**八、成本与性能优化**

1. Token 消耗量大概多少? 个人一个月 token 使用量是多少?
2. 有没有考虑过怎么节约 AI 成本? 做了哪些成本优化方案?
3. 任务重试会造成重复执行、重复调用大模型，如何做重复推理优化、降低成本?
4. 最终降了多少成本? AI 投入前后带来的收益有多少?
5. 在 Agent loop 过程中如何做 gate（参数 / 权限校验）来减少 token 消耗?
6. 用 Skills 承载真实执行，减少对 token 消耗敏感的场景怎么处理?

**九、工程实践与后端**

1. LangChain / LangGraph 主要用了什么 AI 技术? 为啥选它们? 选型依据是什么?
2. LangChain 不同版本 / 模块的区别?
3. 能讲讲 Claude Code 的多层架构吗? 核心循环流程说一下?
4. Claude Code 的上下文治理是怎么做的?
5. Claude Code 的架构设计和你的 multi-agent 有什么不同?
6. 有没有了解过像 Manus 这种通用 Agent? 通用 Agent 一般分几层功能?

7. SSE 和 WebSocket 的区别？各自适用场景？
8. SSE 断线续传怎么做？用户关掉窗口再打开，如何继续接收流？
9. 后端流式推送数据，前端如何实现边接收边流式渲染？
10. AGUI 协议动态渲染表单 / 卡片的整体实现流程？
11. 消息队列任务如何保证不丢失、保证任务一致性？任务中途中断如何实现断点续跑？
12. 如何保证数据库和缓存之间的数据一致性？
13. 多租户场景下 K8S 怎么部署的？
14. sandbox 云端方案如何设计？如何考虑轻量化的沙箱方案？
15. env 管理如何做？Agent 跑的时候生成的代码怎么管理？
16. 意图识别是如何实现的？使用了哪些模型？用户问题如何路由到 AI 系统？

## 十、产品与实战场景

1. 你的 AI 系统用户量有多少？日活有多少？Agent 的使用量是多少？
2. 如何推动 AI 系统落地的？花了多久落地？
3. 假设设计一个秒杀系统，你会怎么设计架构？高并发下如何避免超卖？
4. 支付接入流程怎么设计？如何保证用户支付后积分 / 虚拟产品到账？
5. 如果要做成云端版本 / 逻辑多租，你有设计吗？
6. 你的项目开源了吗？为什么公司会允许你开源？
7. 你觉得你做的项目和其他同类产品比的亮点是什么？