InfoQ 极客传媒

面向智能研发的知识引擎构建及业务应用

演讲人：吴锐

蚂蚁集团 / 高级算法工程师

AiCon

全球人工智能开发与应用大会

AiCon
全球人工智能开发与应用大会

目录

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

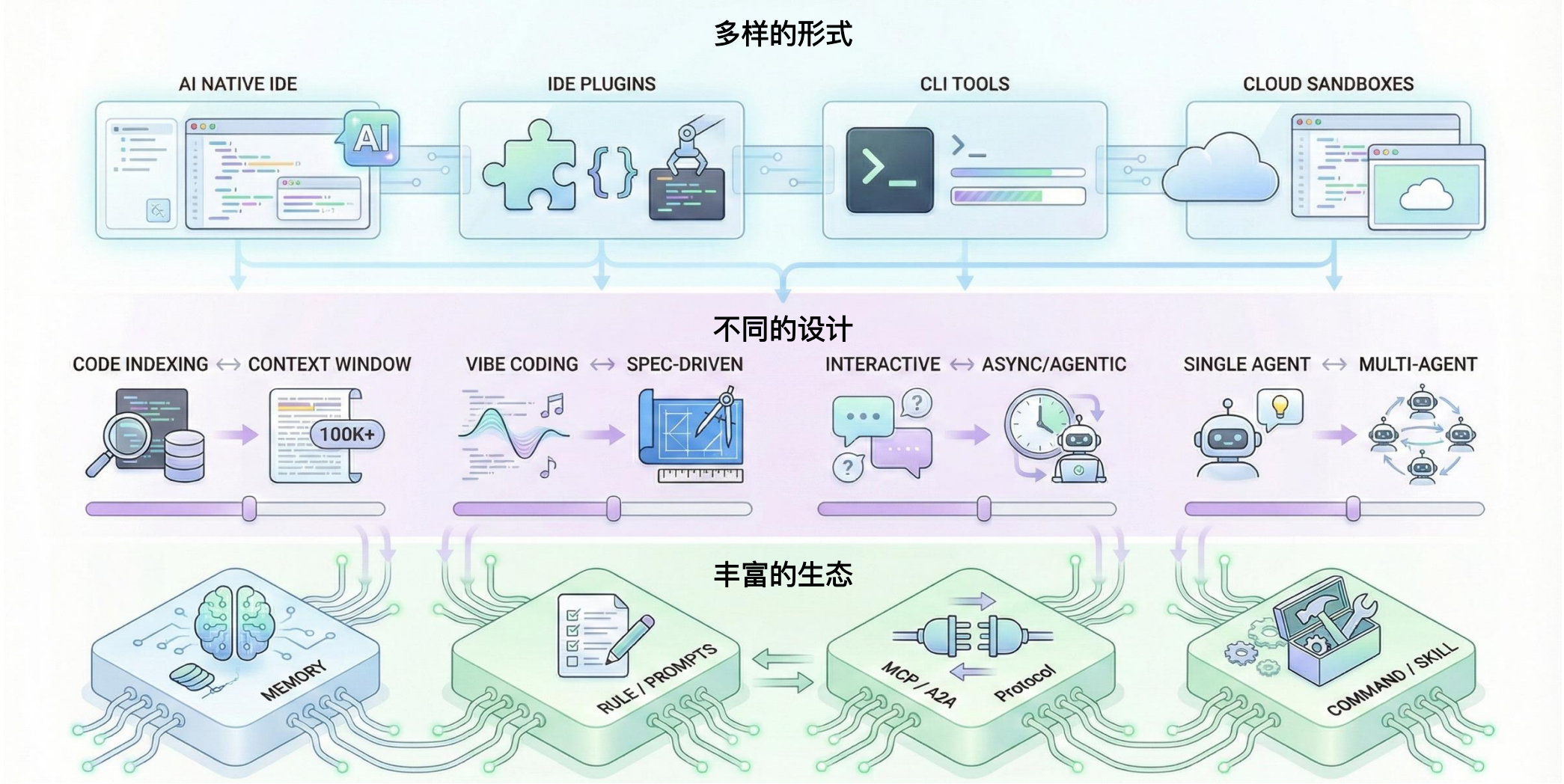
会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

01 智能研发现状

AI Coding 百花齐放



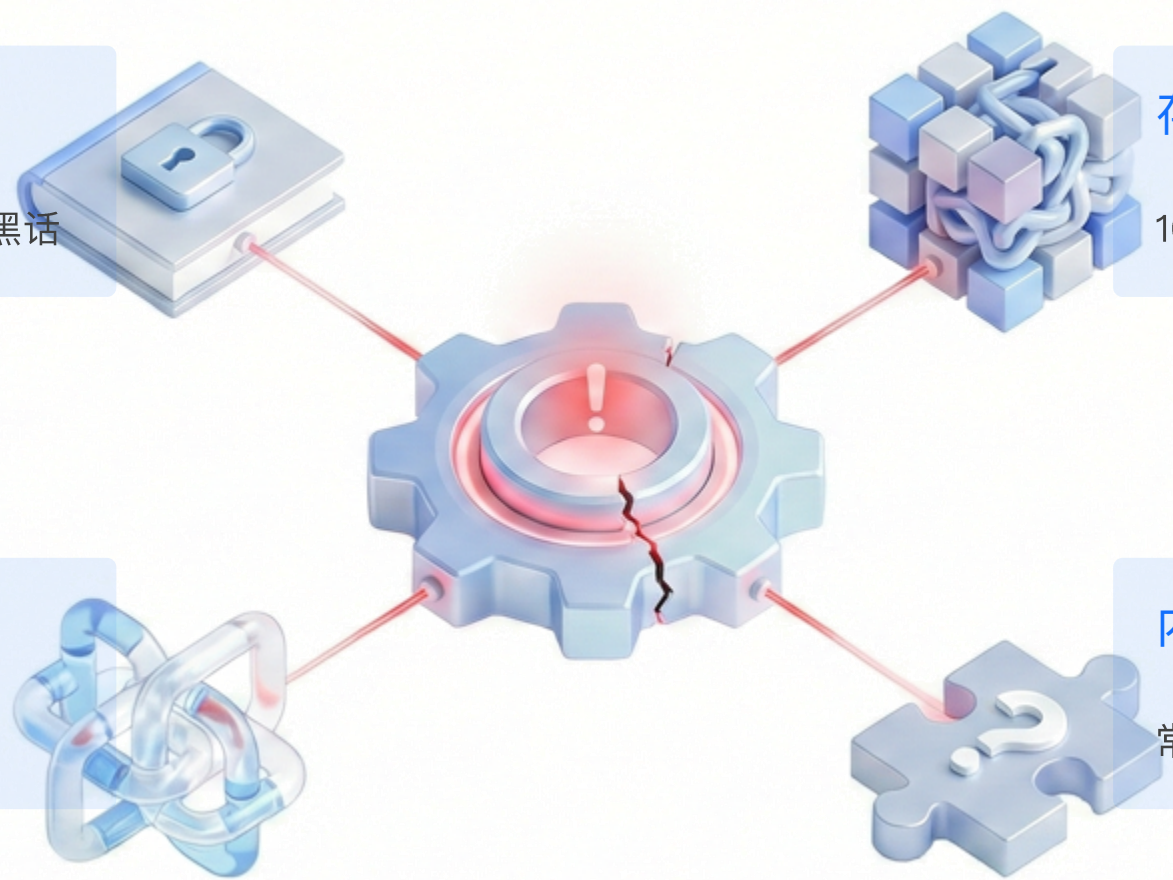
业务落地的“最后一公里”

隐藏在代码外的业务知识

业务专家经验、业务规则、业务黑话

存量代码的复杂性

10w+行的代码仓库、复杂设计模式



复杂的系统间依赖

跨系统调用、复杂的包依赖

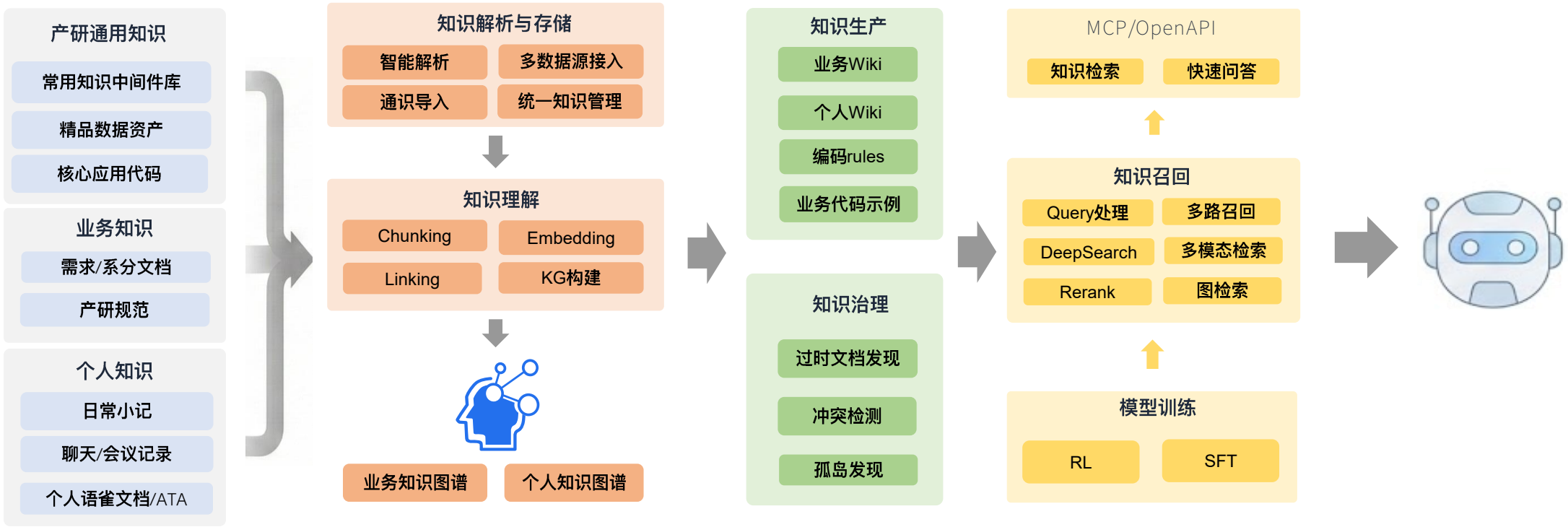
内部研发知识的匮乏

常见的中间件的接入方式

02

构建研发知识引擎

架构设计



知识引擎的必要性

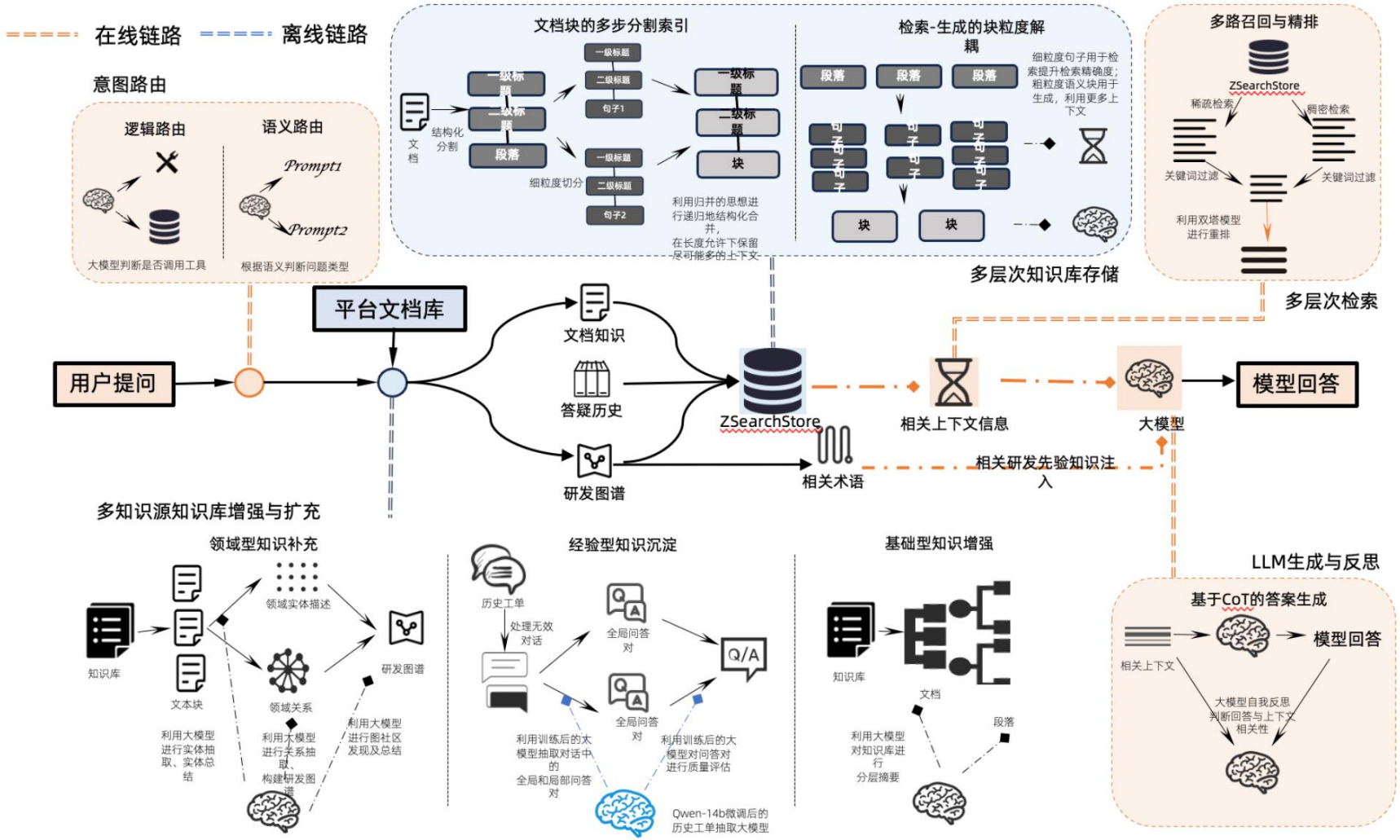
1. 深度理解、高效管理存量业务知识，让AI读懂现有业务，降低生成代码的幻觉率
2. 挂载企业内部私有的中间件文档与规范，填补通用模型的认知空白
3. 提供准确、快速的检索能力，在有限上下文内提供关键信息，降低模型理解门槛

知识索引与检索

基础检索能力：朴素RAG

部分关键优化点：

- Chunking结构化分割
- 索引–检索解耦
- 知识粒度对齐（HyDE）
- QA挖掘
- 文档标签与优先级



知识索引与检索

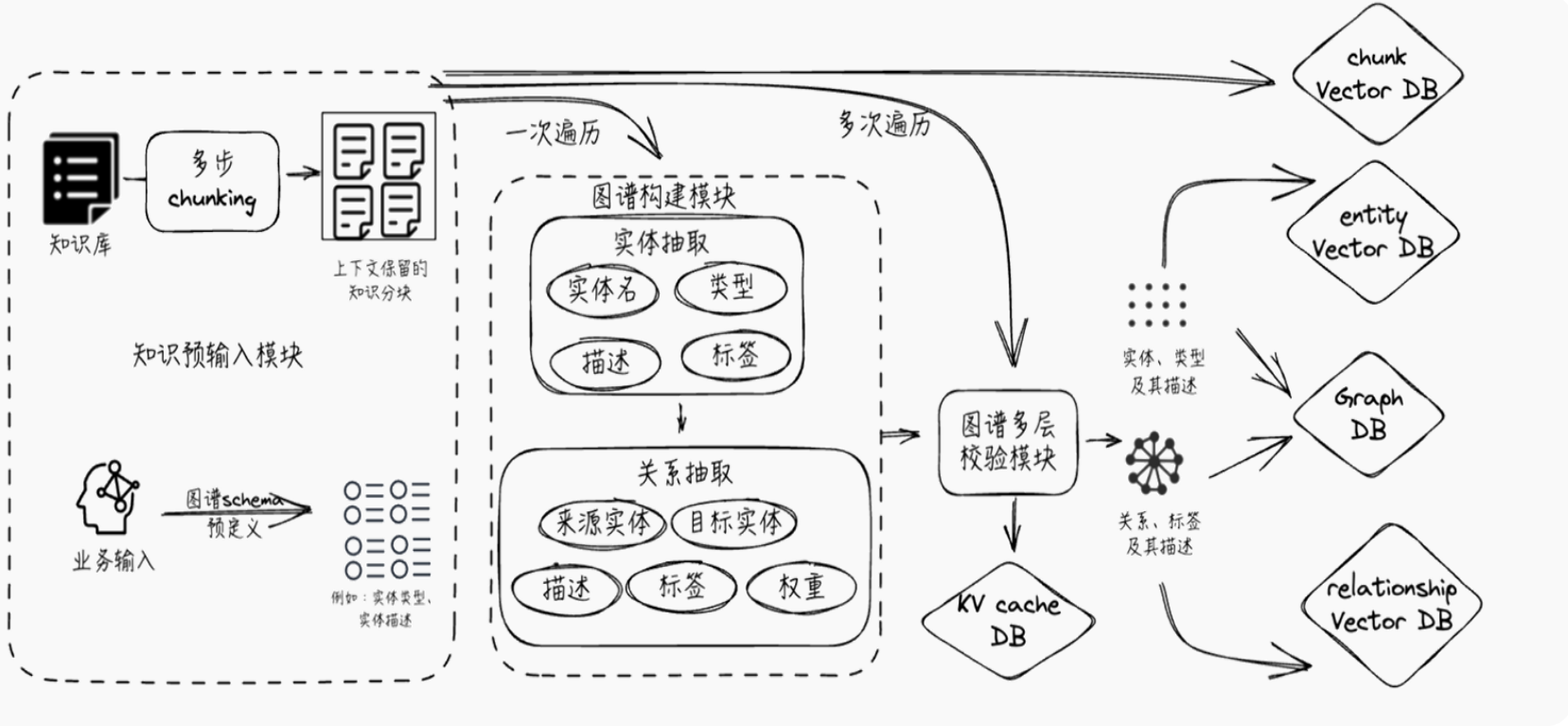
进阶检索能力：GraphRAG

优势：

1. 依赖Node和Edge进行结构化的信息关联，有效地扩大检索范围，提升召回率
2. 前置进行了信息的归纳与总结，面对全局性问题效果更好

劣势：

1. 依赖模型抽取三元组，导致索引成本高、耗时长
2. 工程复杂度高，不易于维护



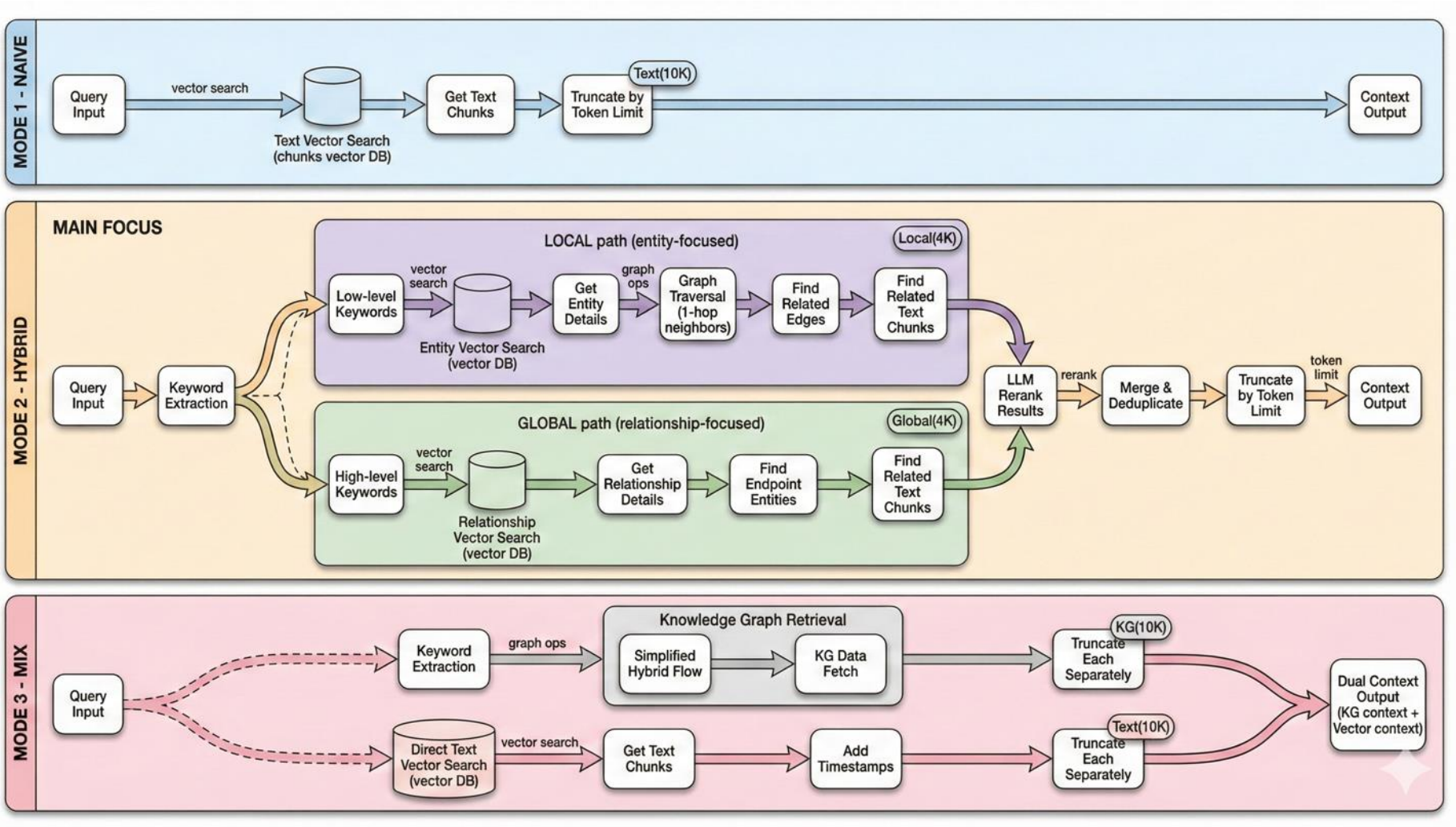
知识索引与检索

综合多路检索

Chunk检索确保简单问题能够快速定位到答案

图谱检索提升复杂Query的召回率：

1. 实体/关系分别进行关键词抽取与检索，确保明确/抽象描述均能召回
2. 在Graph上进行基于点/边关系的扩展，提高召回内容
3.重排确保相关性



知识索引与检索

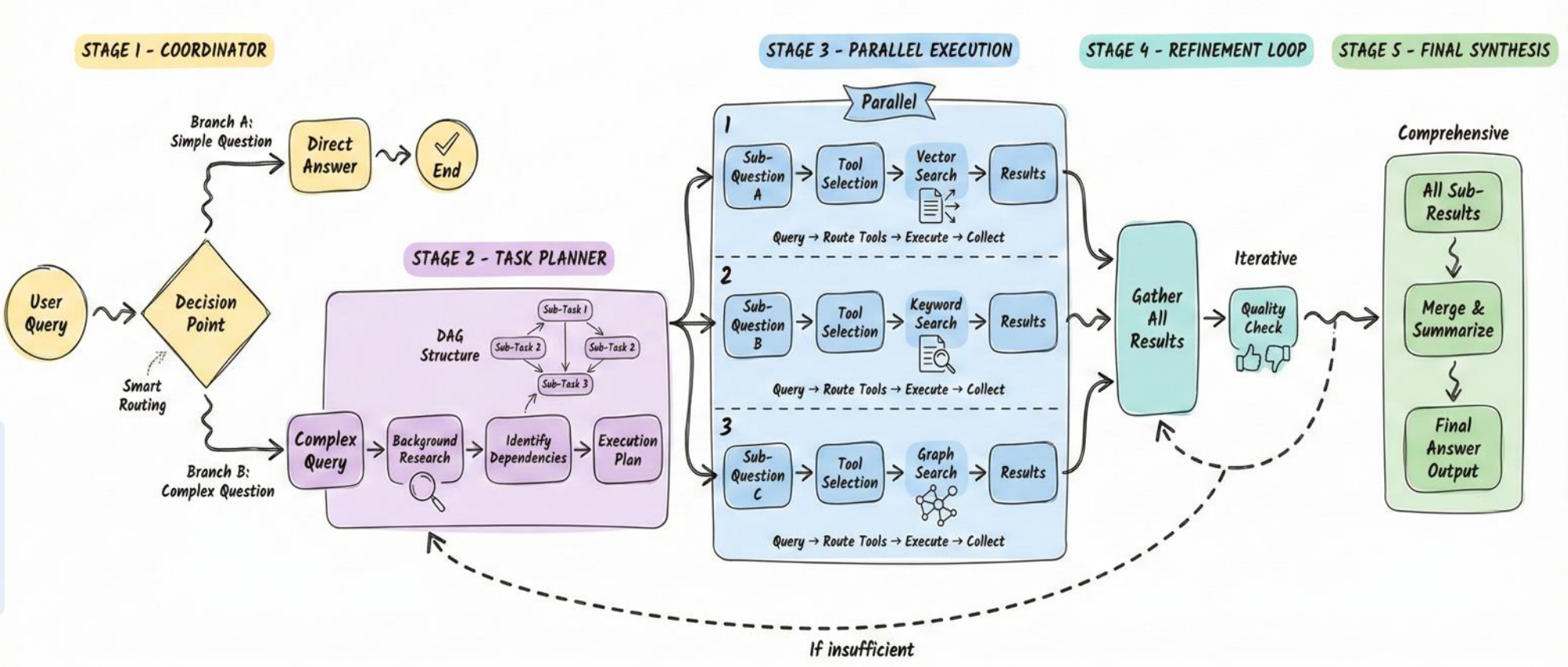
Search Agent： 依靠Agent能力，在搜索、阅读、推理中不断循环往复，直到总结出最优答案。

优势：

1. 依赖大模型自主决策，极大地提升了检索的灵活性，检索效果有质的飞跃
2. 更加易于扩展，只要接入新的tool即可扩大检索范围

劣势：

依赖大模型服务的反复调用，检索成本高、耗时（特别）长



知识索引与检索

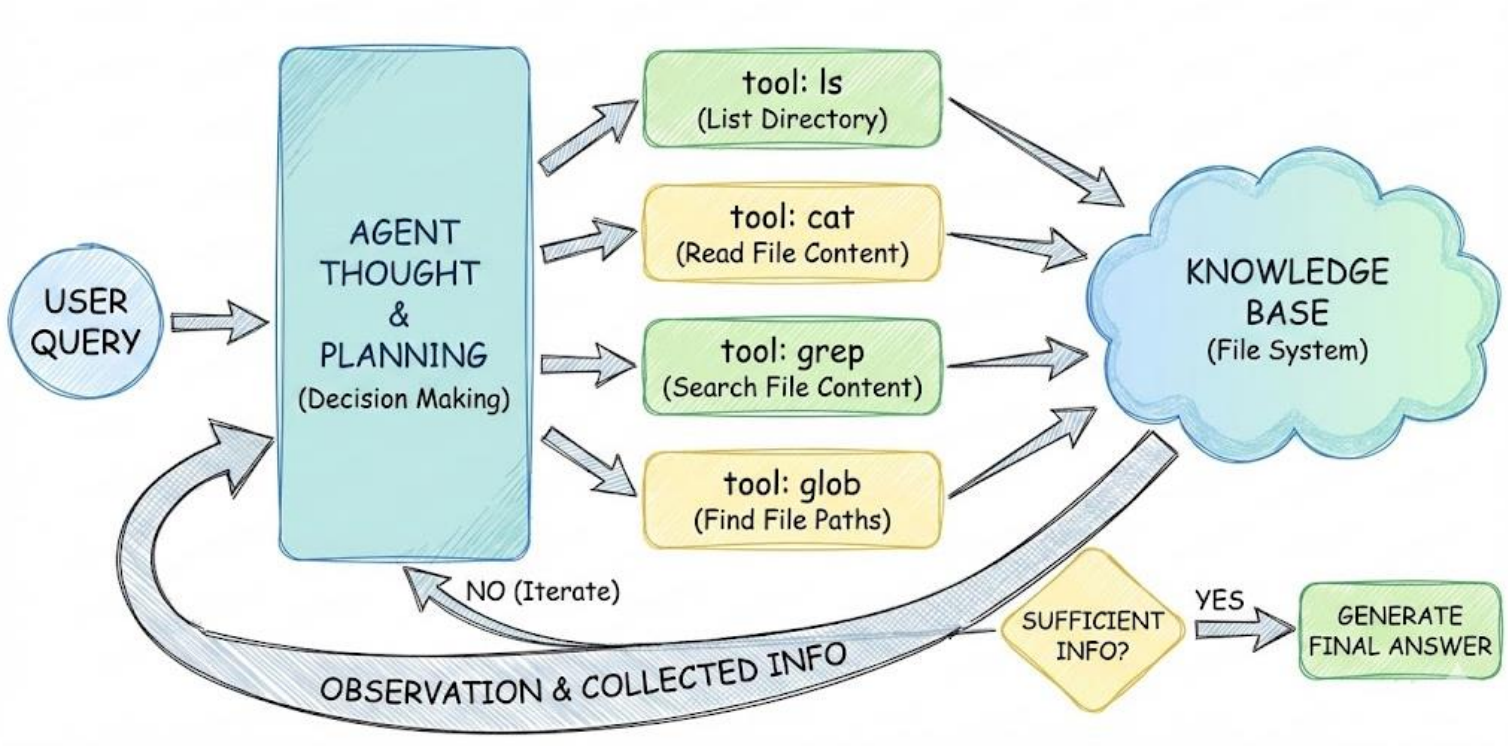
“File System”：将知识组织成一个文件系统

> Why Cline Doesn't Index Your Codebase

– 当你在chunk时，你实际上在破坏其内部的逻辑

– 索引总是会存在负担：存储成本、更新成本

当文档本身存在良好的目录结构/标题时，给 Agent 配置 ls/grep等类命令行工具，让他像人一样去「找」文档，比检索更为准确、有效。



知识压缩与生产

业务Wiki构建

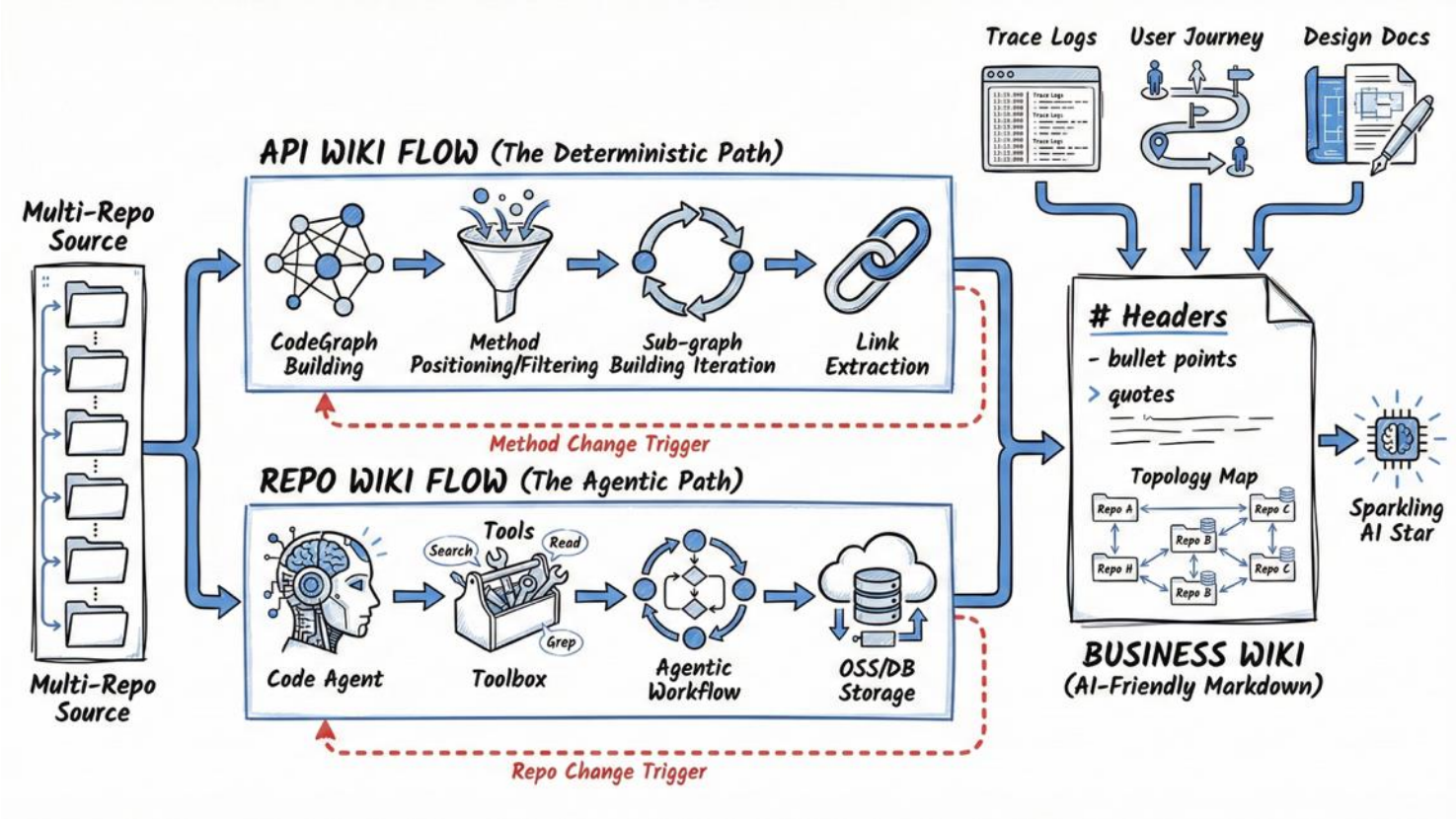
AI 在生成代码时缺乏对业务背景的深入理解，可能导致实际开发偏离真实需求。

——结合代码、文档、用户动线等多种类型数据，产出业务全貌文档，让模型真正「懂业务」。

Wiki 都有什么内容?

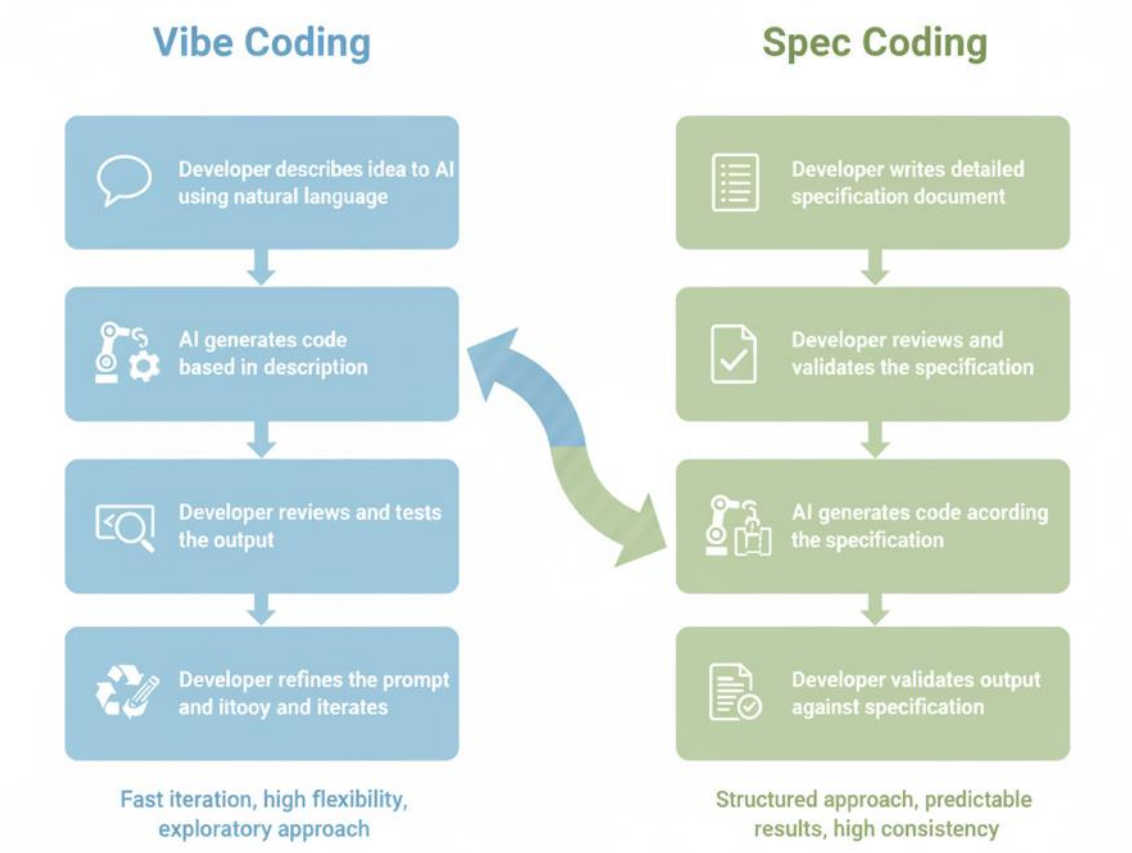
- 接口功能与链路描述
- 技术栈详解
- 项目模块概述、模块间关系
- 业务功能列表
- 业务功能与接口的关系
- 用户主要动线

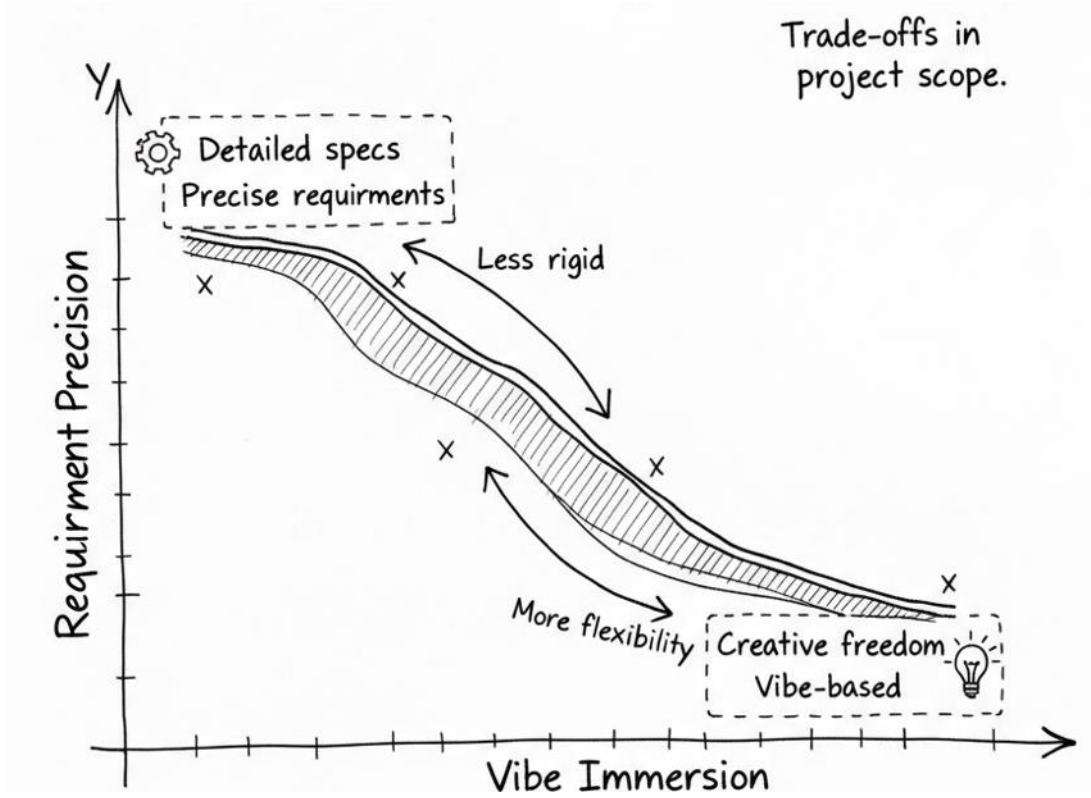
……



03 业务实践案例

Spec Driven Development

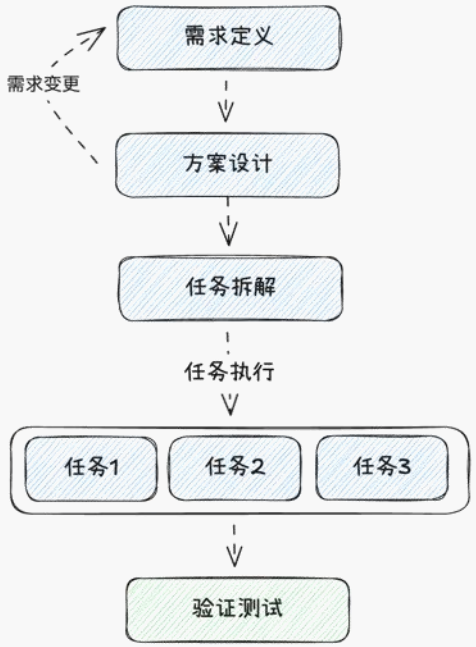




Specification：一份详细描述一个产品、系统、软件组件或项目应该如何工作、它具备什么功能、它的设计参数和限制的文件或集合，用来驱动 AI 生成代码。

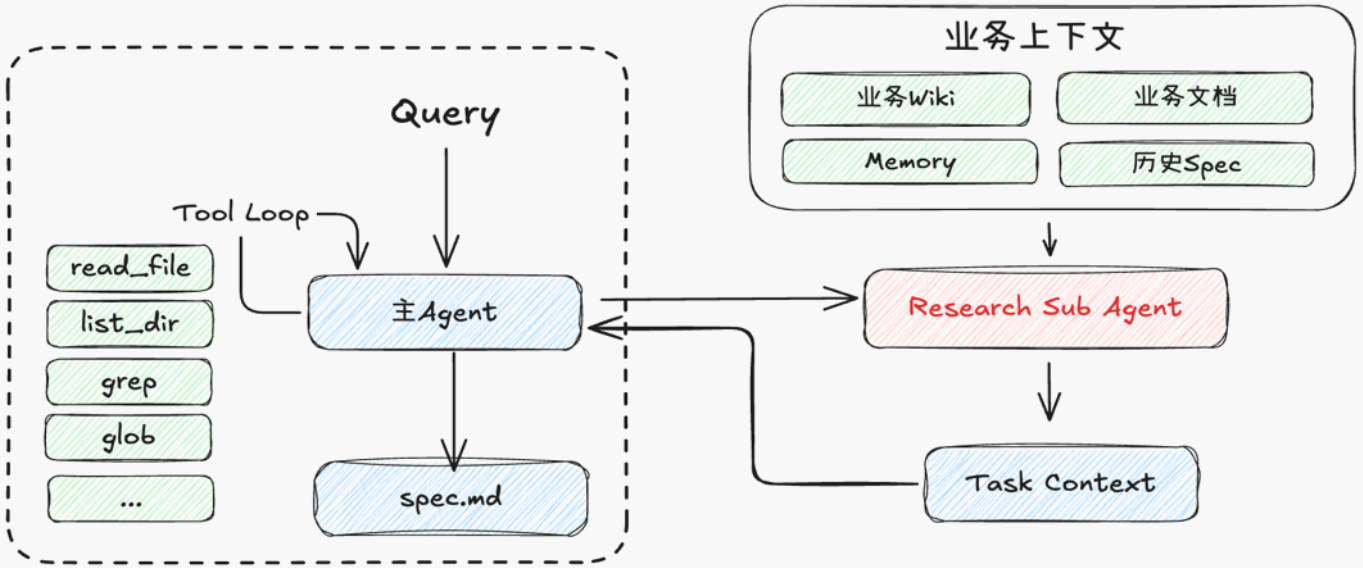
Spec Driven Development

SDD的典型流程



Agent实现

引入知识引擎



通过Sub Agent的方式引入上下文检索能力，确保单一职责，同时避免检索无关上下文污染主Agent链路

04 未来趋势判断

Context is All You Need

1. 从人的视角出发，良好的上下文管理是提升AI Coding效果的核心

2. 不止是coding，AI将参与到业务知识的主动维护中来

3. 现阶段AI在业务研发上仍然是辅助角色，架构设计、线上应急还是依赖人工

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

InfoQ 极客传媒

THANKS

探索 AI 应用边界

Explore the limits of AI applications

AiCon

全球人工智能开发与应用大会