通义灵码技术解析，打造 AI 原生开发新范式

通义灵码产品技术负责人 / 陈鑫

InfoQ 极客传媒

已结束 北京站 2000 人

QCon 全球软件开发大会 暨智能软件开发生态展

架构演进 | AIGC
研发效能 | 向量数据库

6 月 14–15 日 深圳站 1000 人

ArchSummit 全球架构师峰会 暨智能软件开发生态展

业务架构 | 高可用
AI 架构 | 架构师成长

10 月 18–19 日 上海站 1500 人

QCon 全球软件开发大会 暨智能软件开发生态展

AIGC | AI 原生应用
软件工程 | 金融科技

4月 5月 6月 8月 10月 12月

5 月 17–18 日 北京站 2000 人

AiCon 全球人工智能开发与应用大会 暨大模型应用生态展

通用大模型 | AI Agent
LLMOps | AI Infra

8 月 16–17 日 上海站 1000 人

FCon 全球金融科技大会 暨智能软件开发生态展

金融大模型 | 核心系统
数字化转型 | 数据安全

12 月 20–21 日 北京站 1000 人

ArchSummit 全球架构师峰会 暨智能软件开发生态展

未来架构 | 高可用高并发
AI 应用 | 架构师思维

极客邦科技 2024 年会议规划

促进软件开发及相关领域知识与创新的传播

访问大会官网

参会咨询

Contents
目录

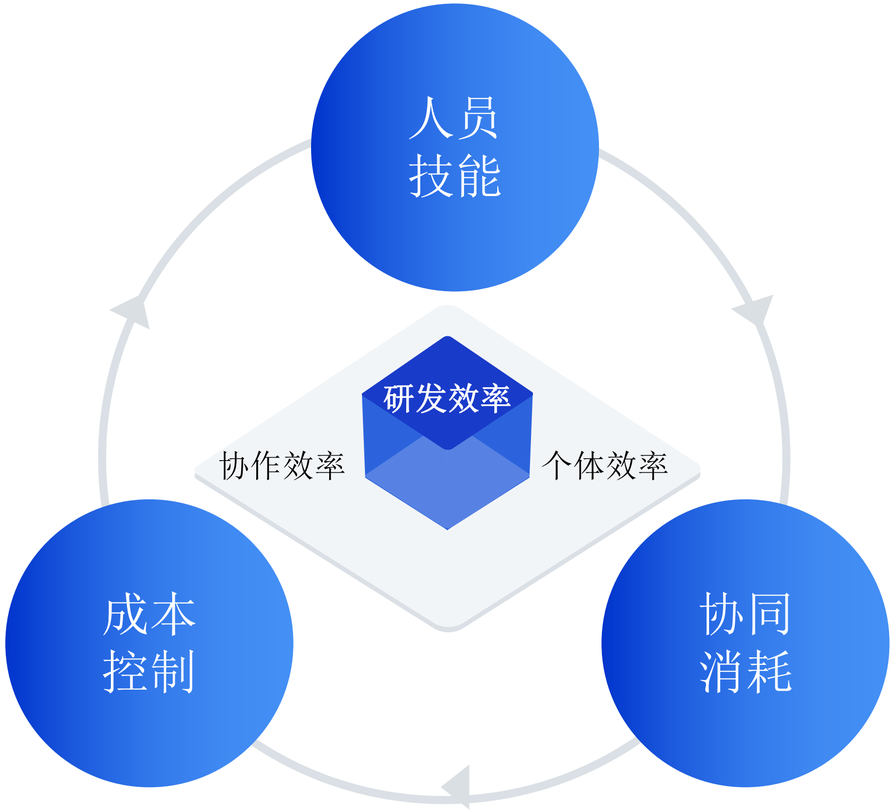
AIGC 对软件研发的根本性影响

01

AIGC对企业研发效能核心因素的影响

人员技能是效能的基石，也是效能破局点

能力提升 弥补能力短板



成本是效能优化的目的，同时也是约束条件

工具赋能 事务性工作替代

软件架构和组织复杂度正相关，并决定协同消耗的大小

流程规范 打造超级个体

企业软件研发的挑战及智能化的机会

1 个体效率	研发人员重复性工作，简单工作，沟通的工作特别多，浪费时间。	GPT Copilot and Agent 提升研发的一致性
2 协作效率	研发管理流程化，缺乏灵活性，组织容易产生效率竖井，响应能力弱。	LLM 简化流程 提升应对可变性能力
3 研发体验	现有工具散乱，操作不统一，学习成本高，切换代价大。	对话方式，统一入口 降低研发的复杂度
4 数字资产	研发知识缺乏沉淀，资产价值没有发挥出来。大部分都是负债，资产积累少。	SFT, RAG 增强 LLM 隐性知识显性化

人工智能带来的新的人机协同模式

LLM as Copilot	LLM as Agent	LLM as Facilitator
人机对话能力	自主完成任务能力	跨领域复合型整合能力
不改变软件工程的专业分工，但增强每个专业技术，基于AI的研发工具平台辅助人完成任务。影响个体工作。	单一领域专家，能够自主使用工具并完成预定任务。多个Agent之间可以互相协作完成复杂工作。影响角色互动。	影响软件研发流程的角色分工，基于AI的研发工具平台辅助决策。辅助计划、预测发现和协调的工作。影响组织决策。
（解决单点事务性工作效率问题）	（解决复杂任务协同效率问题）	（解决信息整合、分析、决策问题）
工具：专业增强 人：见多识广，提升Prompt能力	工具：多领域专家 人：给定上下文，完成知识对齐	工具：跨领域会诊 人：负责创意、纠偏

研发知识传递形态的改变

企业一线开发者

任务协同 | 架构设计 | 代码编写 | 软件测试 | 问题排查 | 软件发布 | 日常运维 | 知识查询

智能研发工具

智能编码 | 智能评审 | 智能项管 | 平台工程 | 智能问答 | AutoDev | 个人助理 | 智能洞察

LLM 智能大脑

自定义Prompt | RAG 检索增强 | Agent 平台

基础模型能力

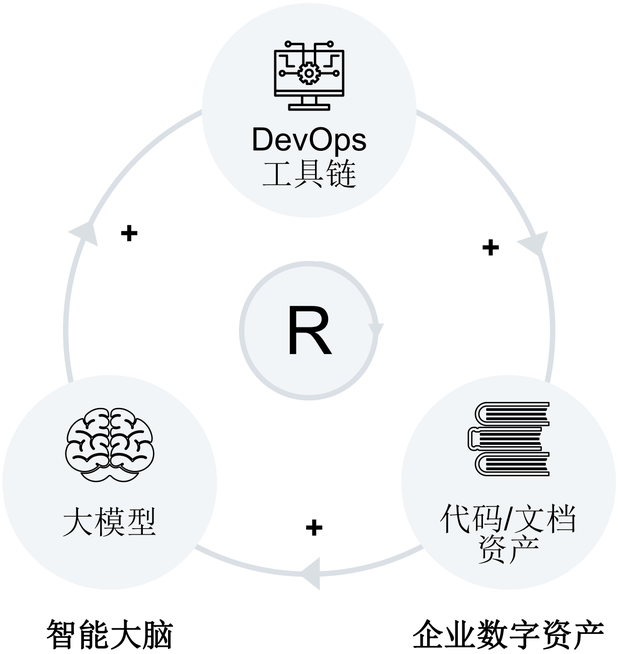
代码补全模型 | 研发问答模型 | 企业专属模型 | 模型训练SFT

实时增强

知识梳理

- 确定优化目标
- 梳理资产
- 数据清洗
- 安全与隐私

软件研发全生命周期



打造最佳 Copilot 人机协同模式

02

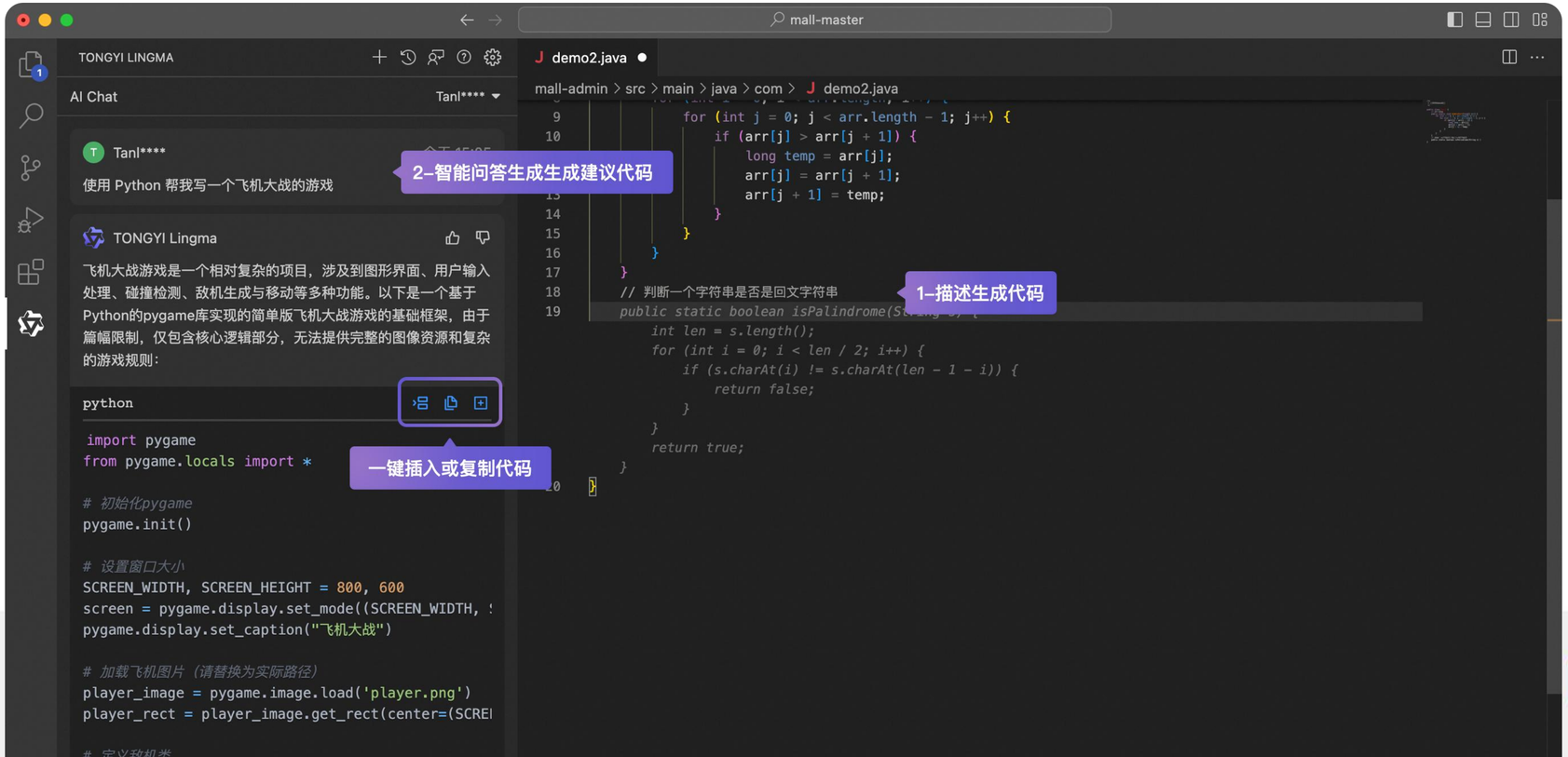
代码开发人机协同的 Copilot 模式

解决小任务
上下文宽度限制

人工确认采纳
模型幻觉问题

高频次
准确率有限

短输出
推理成本与性能



什么是开发者最喜爱的 Copilot

高频
刚需

触手
可及

知我
所想

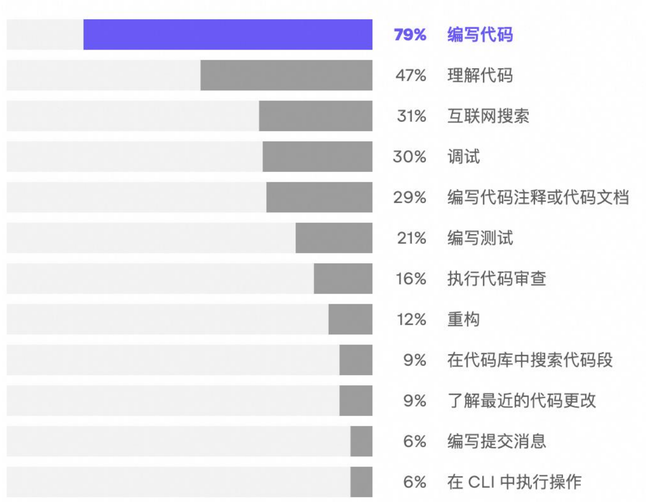
唯我
专属

解决开发者最高频刚需场景

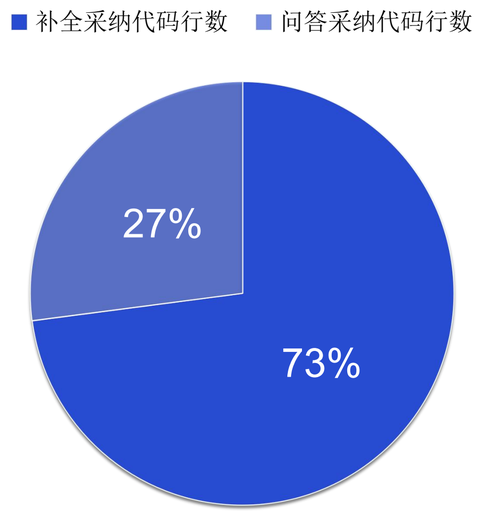
高频刚需

触手可及

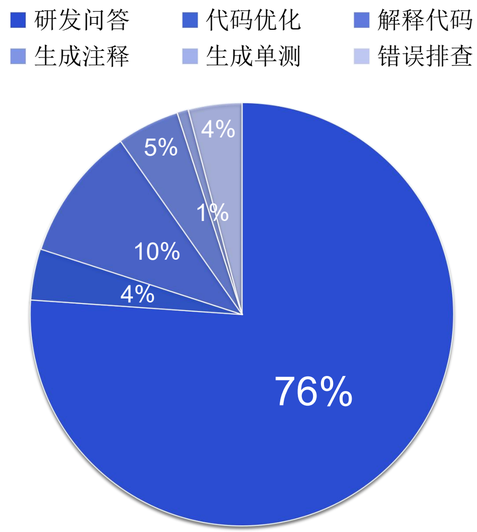
开发者最耗时的活动



补全与问答代码采纳分布



问答功能使用分布



注：「开发者最耗时的活动」数据来源 JetBrains 2023 年开发者生态报告

打造沉浸式编程体验

高频刚需

触手可及

知我所想

AI for Developer

代码补全任务是**性能敏感型**任务。使用专门训练的**小参数代码模型**，实现代码生成效率与质量的平衡。

代码补全任务 codeqwen2 模型

在**中等参数模型规模**下，提供代码解释、注释生成、单元测试、代码优化、运行错误修复、提交信息生成、重构建议等**7项代码技能**。

代码专项技能 qwen-plus 模型

研发问答任务对模型知识面、编程能力、推理能力有更高要求。需要最**大参数模型并叠加RAG技术**，大幅消除模型幻觉，提升回答质量。

研发自由问答 qwen-max 模型

插件侧跨 IDE 端架构设计

高频刚需

触手可及

知我所想

本地 Agent 服务

作为插件端与远程服务的桥梁，为不同插件端共享核心业务/算法逻辑，降低插件端的实现成本。并且能将用户数据存储在本地，从而保障用户的隐私安全。

本地离线模型服务

提供离线的微型补全模型，满足无网络环境场景的使用诉求。

基于语义理解的自适应生成粒度决策

触手可及

知我所想

唯我专属

生成单行代码：无法构建完整的函数或模块

代码块的不同位置提供不同生成规则：准确度低

通义灵码基于代码的语义信息，充分让模型理解不同场景下所需的生成粒度，从而让模型能够根据当前正在编写的代码位置，**模型自适应决策应该生成的代码粒度**。

生成粒度决策准确率，Java 语言从**47%提升到56%**，Python 语言从**26%提升到44%**，其他语言均有较大提升

```java
package com.alibaba.force;

import com.aliyun.odps.udf.UDF;
import com.google.googlejavaformat.java.Formatter;
import com.google.googlejavaformat.java.JavaFormatterOptions;
import org.apache.commons.lang3.StringUtils;

/**
 * @Description 格式化Java代码的UDF函数
 * @Author bogw.wbg
 * @Date 2023/6/30
 **/
public class CodeDataFormat extends UDF {
    public String evaluate(String language, String content) {
        // 如果代码为空则直接返回
        if (StringUtils.isBlank(content)) {
            return content;
        }
        // 判断是java代码并执行格式化处理
        if (language.equalsIgnoreCase("java")) {
            try {
                // 调用Java Formatter完成代码格式化
                return new Formatter(JavaFormatterOptions
                                     .builder()
                                     .style(JavaFormatterOptions.Style.AOSP)
                                     .build()
                                    ).formatSourceAndFixImports(content);
            } catch (Exception e) {
                // 如果发生错误则返回原始代码
                e.printStackTrace();
                return content;
            }
        }
        return content;
    }
}
```

类级别

函数级别

逻辑块级别

行级别

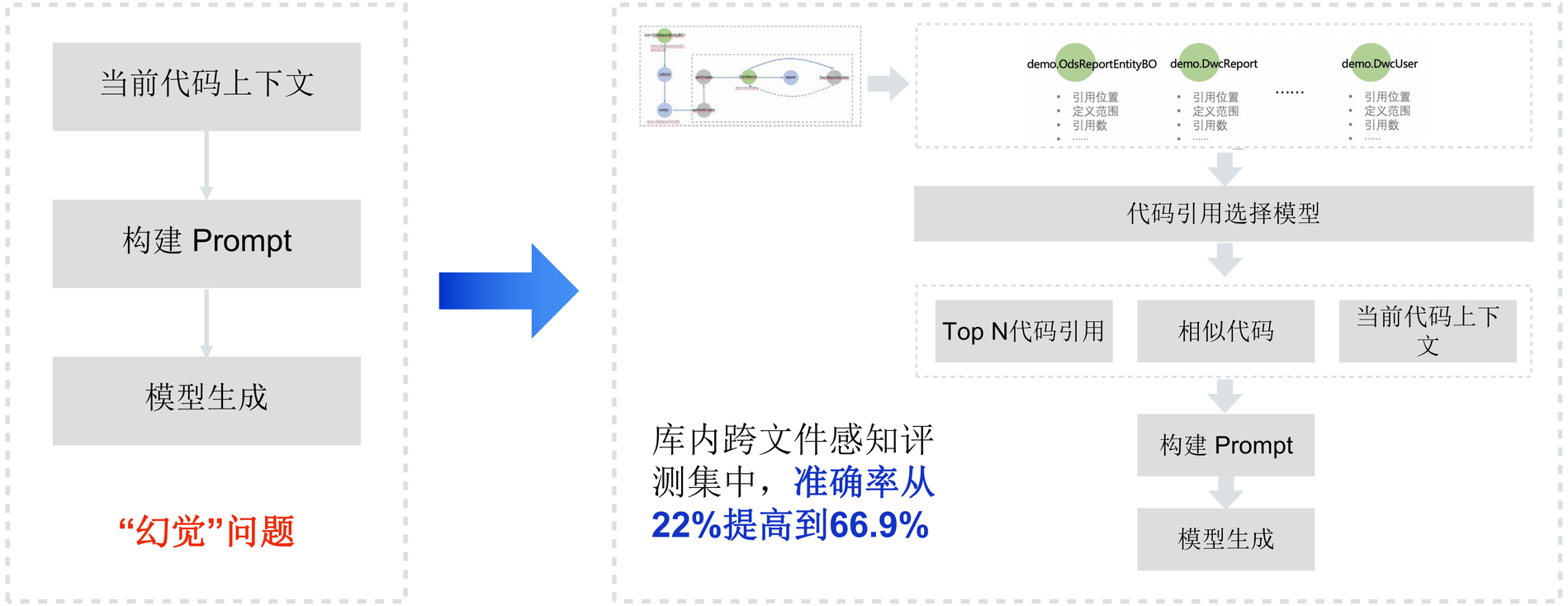
基于库内感知的代码生成及问答

触手可及

知我所想

唯我专属

通义灵码通过先进的端侧实时代码语义分析技术，实时分析当前正在编写的代码，并基于代码语义分析、代码引用链路跟踪、动态语言类型推导、相似代码分析等先进的技术方法获取所需的相关代码引用、相似的代码片段等语义信息，弥补单纯关注当前代码文件所需的不足，避免在生成的代码中引用了代码库内不存在的API等大模型常见的“幻觉”问题。



本地库内检索增强

触手可及

知我所想

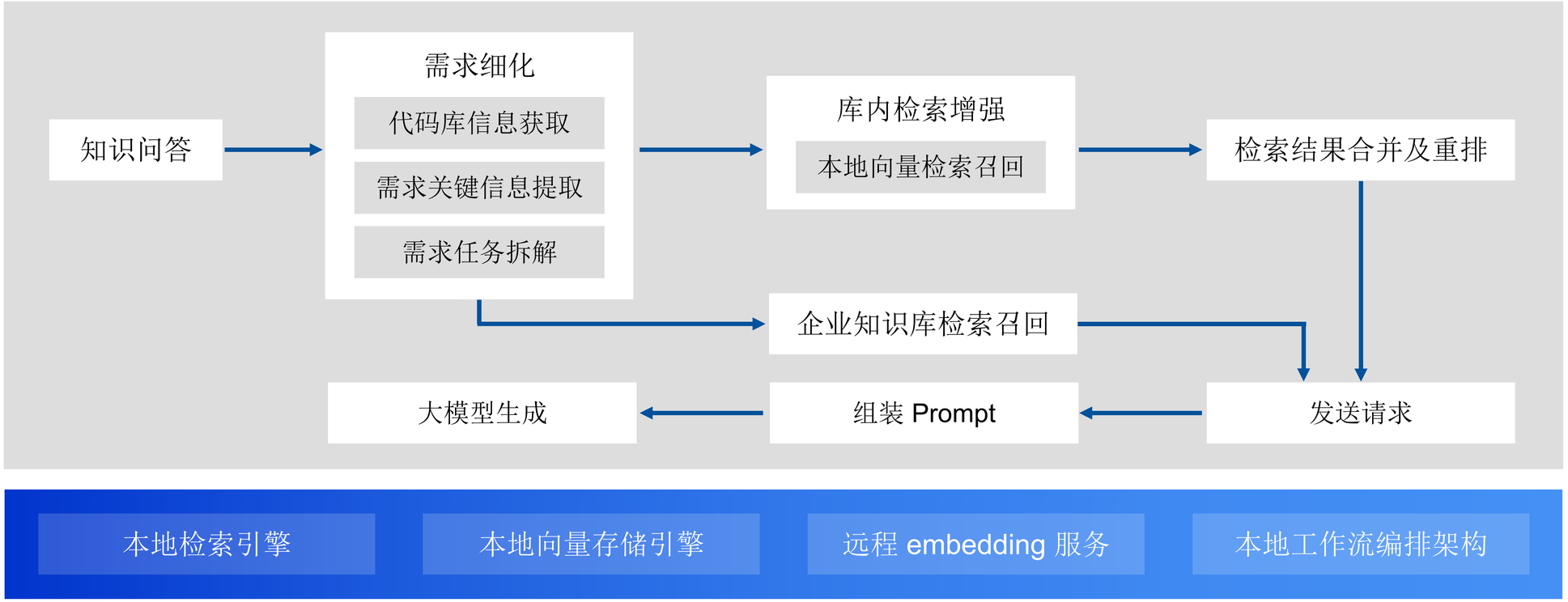
唯我专属

本地库内检索增强服务

通过感知本地工作空间中源文件进行预处理，建立在用户本地的向量化索引，基于本地工作流编排引擎，完成多阶段任务。

安全性高

基于本地的解析和索引等服务，保障用户的代码都存在本地，保障用户代码安全和隐私。



企业数据个性化场景

知我所想

唯我专属

项目管理	开发	测试	运维
• 所在行业存在较多专有词汇 • 对需求/任务/缺陷的内容及格式有固定的规范/要求 • 需要学习已有的项目管理策略/经验	• 编码需要符合企业制定规范 • 需要引用企业内的二方包 • 需要调用企业内的API接口 • 代码的业务逻辑较复杂，存在较少的通用代码 • 适配企业内已有的数据库表结构，并学习SQL相关逻辑 • 企业内通常使用自研开发框架，如前端框架、组件库等	• 需要符合企业内指定的测试规范 • 所在行业的业务逻辑较复杂 • 企业内通常使用自研的测试框架	• 需要复用企业内的运维手册 • 运维人员需要学习企业内的大量运维脚本/知识 • 需要快速获取企业内的运维接口/API

企业数据个性化流程

知我所想

唯我专属

企业代码/知识库

数据处理

模型微调/RAG索引

效果评测

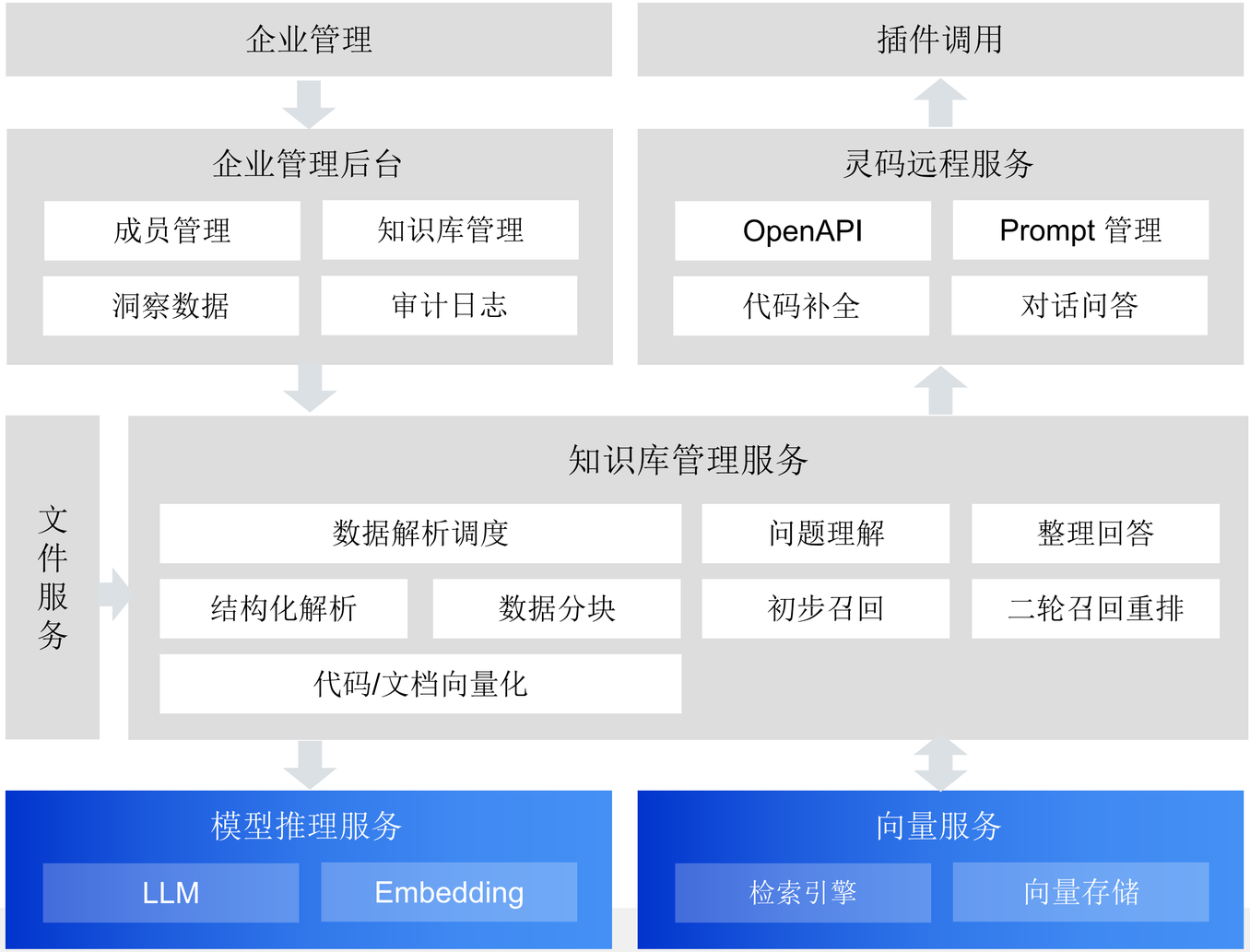
- 代码数据处理
  - 过滤过小或过大的文件
  - 过滤文件行数大于xxxx的文件
  - 过滤掉注释比例大于 xx% 的文件
  - 过滤掉反编译产出的文件
  - ……
- 文档数据处理
  - 各种文档类型转换为markdown格式
  - 根据标题段落构建文档结构树
  - 使用大模型、规则等策略抽取QA问答对
  - 使用大模型、规则等策略拆分文档chunk
  - 使用大模型摘要、扩展文档及QA对
  - ……

- 模型微调
  - 需要加入开放域数据及私域数据混合训练，如2:1比例混合
  - 如果企业内GPU资源不足，可以采用LoRA/QLoRA的方式，并且采用较小的alpha配置
  - 训练数据较小时，需要避免过拟合
- 检索增强
  - 采用关键词+向量混合检索的方式比仅用向量检索效果会更好，检索后进行重排能进一步提升召回
  - 如果数据量较少，需要尽可能抽取问答对，或使用大模型扩充内容，提高数据的泛化能力
  - 词嵌入模型对没有见过的数据，泛化性较差

企业级检索增强方案

知我所想

唯我专属



RAG 检索服务

负责RAG知识的处理与检索业务逻辑。主要分为：数据处理、检索召回。数据处理支持主流的文档和代码语言。检索召回分别为问答和补全提供生成参考。

嵌入服务

提供通用文本数据的向量化。

向量服务

提供向量数据的存储及索引服务。

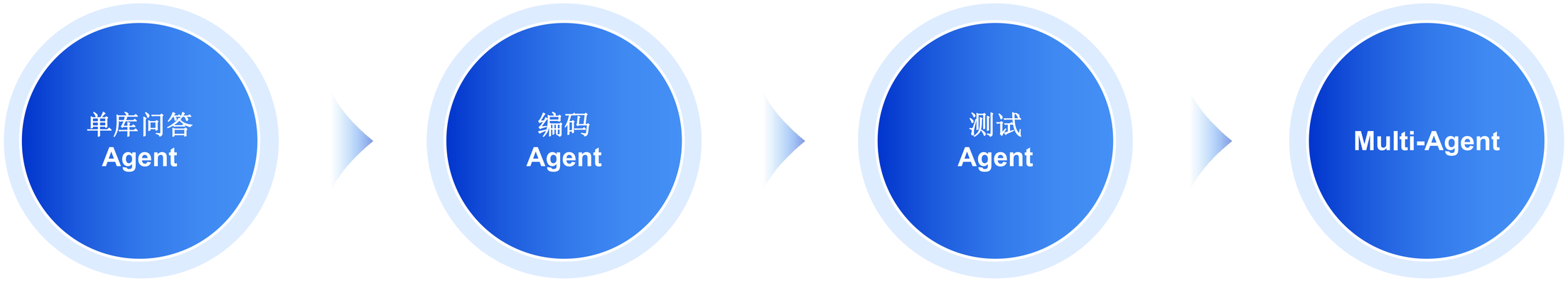
未来的软件研发 Agent 产品演进

03

研发领域多智能体协同


用户需求
产品经理 Agent
细化需求
用户故事
需求优先级
…
架构师 Agent
需求澄清
领域模型
API 时序流
…
项目经理 Agent
细分任务
任务优先级
任务约束
…
工程师 Agent
生成代码
代码评审
修复缺陷
…
测试 Agent
生成单测
单测执行
收集缺陷
…
运行环境
全局知识/工具
企业私域检索
开放域检索
企业 API 调用
工具调用
知识查询
公共消息队列
消息1
消息2
…
消息来源：项目经理 Agent
消息目标：开发工程师
消息内容：任务1：实现模块
人工干预：React or Vue
信息提取
功能需求
待完成任务
架构设计
代码工程信息

软件研发领域 Agent 可行性路径



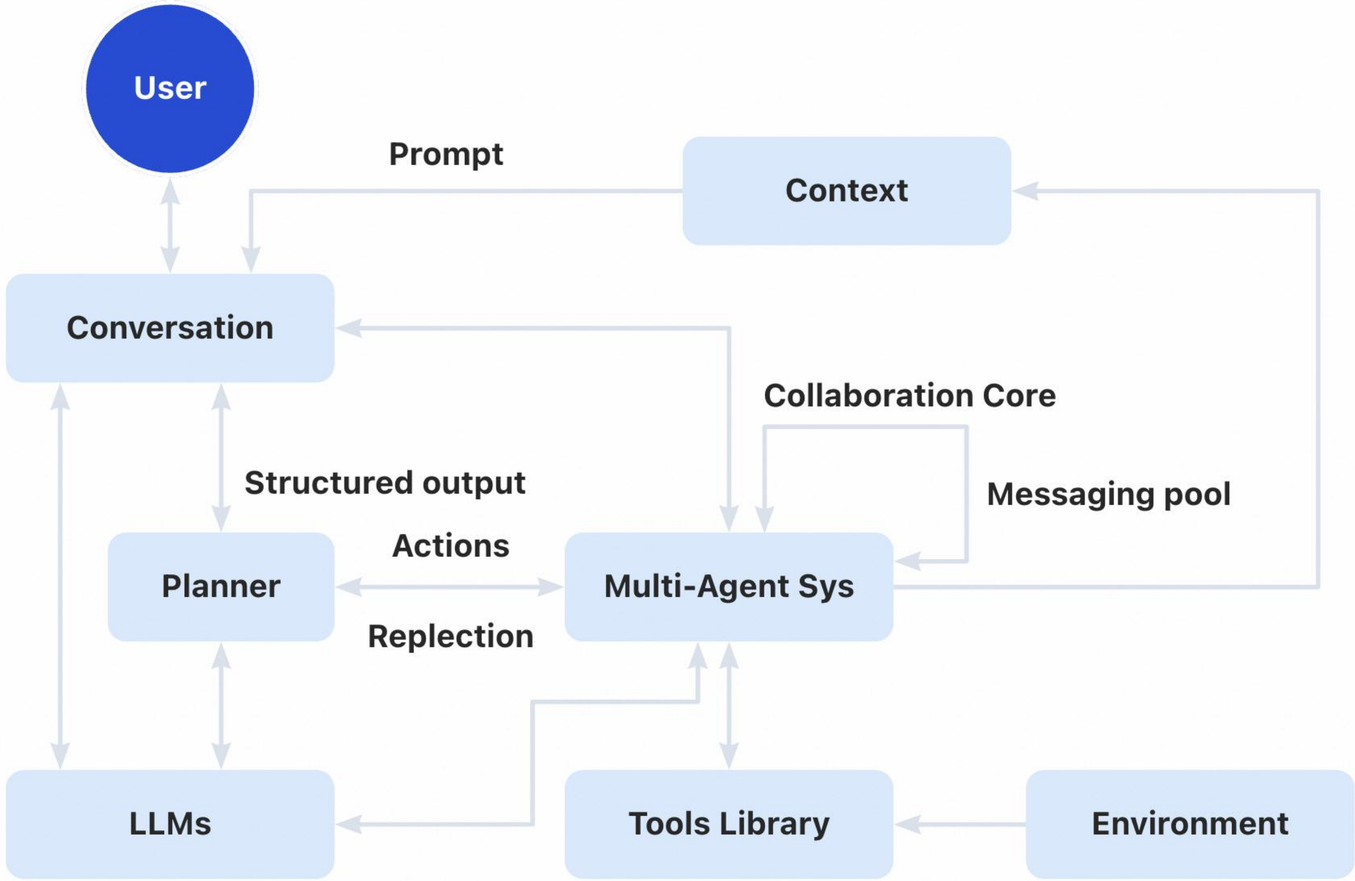
基于单库 RAG 技术，以及固定步骤实现对单库范围内的简单编码任务。

具备一定自主任务规划能力，以及使用工具能力，可自主完成单库范围内的编码任务。

具备自主测试能力的Agent，可以理解任务需求，阅读代码，生成测试代码，并完成运行。

多 Agents 基于 AI 调度共同完成任务。实现从需求->代码->测试的全流程自主化。

Multi-Agent 概念架构



未来智能软件研发工具链形态

InfoQ 极客传媒

已结束 北京站 2000 人

QCon 全球软件开发大会 暨智能软件开发生态展

架构演进 | AIGC
研发效能 | 向量数据库

6 月 14–15 日 深圳站 1000 人

ArchSummit 全球架构师峰会 暨智能软件开发生态展

业务架构 | 高可用
AI 架构 | 架构师成长

10 月 18–19 日 上海站 1500 人

QCon 全球软件开发大会 暨智能软件开发生态展

AIGC | AI 原生应用
软件工程 | 金融科技

4月 5月 6月 8月 10月 12月

5 月 17–18 日 北京站 2000 人

AiCon 全球人工智能开发与应用大会 暨大模型应用生态展

通用大模型 | AI Agent
LLMOps | AI Infra

8 月 16–17 日 上海站 1000 人

FCon 全球金融科技大会 暨智能软件开发生态展

金融大模型 | 核心系统
数字化转型 | 数据安全

12 月 20–21 日 北京站 1000 人

ArchSummit 全球架构师峰会 暨智能软件开发生态展

未来架构 | 高可用高并发
AI 应用 | 架构师思维

极客邦科技 2024 年会议规划

促进软件开发及相关领域知识与创新的传播

访问大会官网

参会咨询

THANKS

大模型正在重新定义软件

Large Language Model Is Redefining The Software