AiDD AI+研发数字峰会
AI+ Development Digital summit
第5届 5

探索工程智能体和RAG建设

汪晟杰 | 腾讯云

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 · 敦煌站
K+ 思考周®研习社
时间：2025.08.29-30

K+峰会 · 上海站
K+ 金融专场
时间：2025.10.17-18

K+峰会 · 香港站
K+ 思考周®研习社
时间：2025.11.25-26

K+峰会详情

AiDD峰会

AiDD峰会 · 上海站
AI+研发数字峰会
时间：2025.05.17-18

AiDD峰会 · 北京站
AI+研发数字峰会
时间：2025.08.08-09

AiDD峰会 · 深圳站
AI+研发数字峰会
时间：2025.11.28-29

AiDD峰会详情

演讲嘉宾

汪晟杰

腾讯云资深技术产品专家

腾讯资深技术产品专家，20年工作经验，负责腾讯云开发者AI代码助手产品，十多年协作SaaS、Teambition，和 SAP 云平台 SuccessFactors HCM、Sybase 数据库、PowerDesigner 等产品的开发经理，在软件架构设计、产品管理和项目工程管理、团队敏捷、AI研发提效等方面拥有丰富的行业经验。

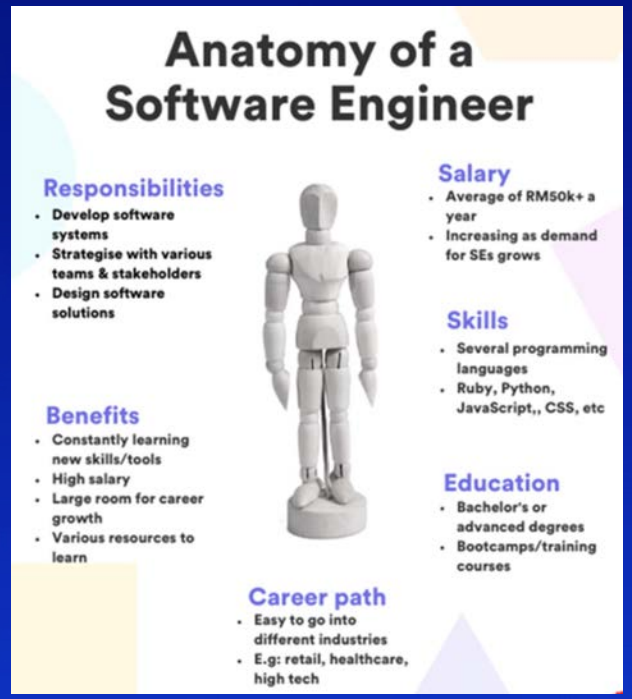
目录
CONTENTS

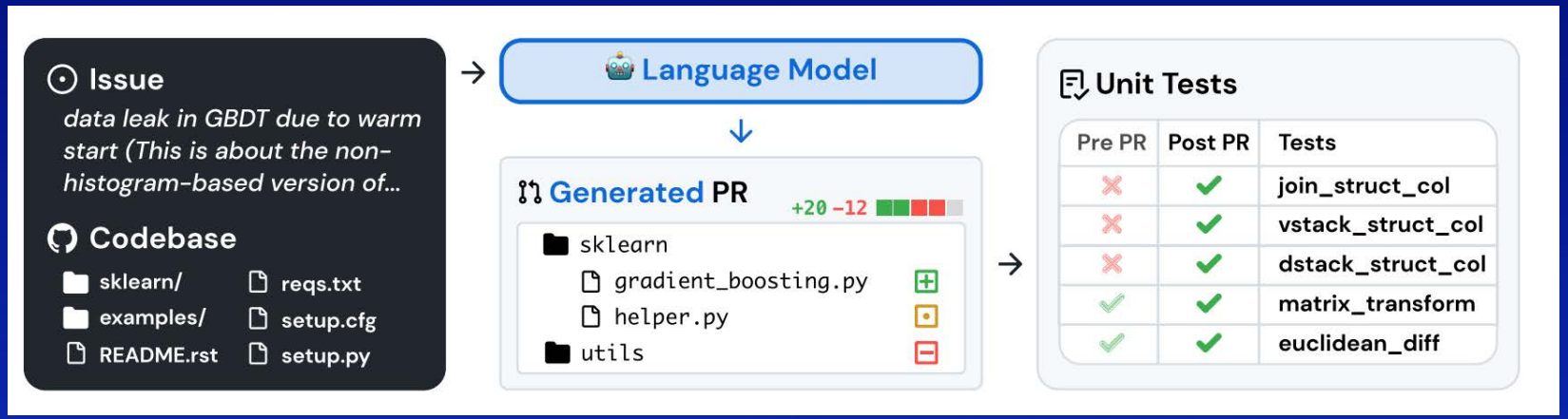
1. 建设工程智能体的背景
2. 问题/痛点
3. 解决思路/整体方案
4. 架构实现/技术实践
5. 总结与展望

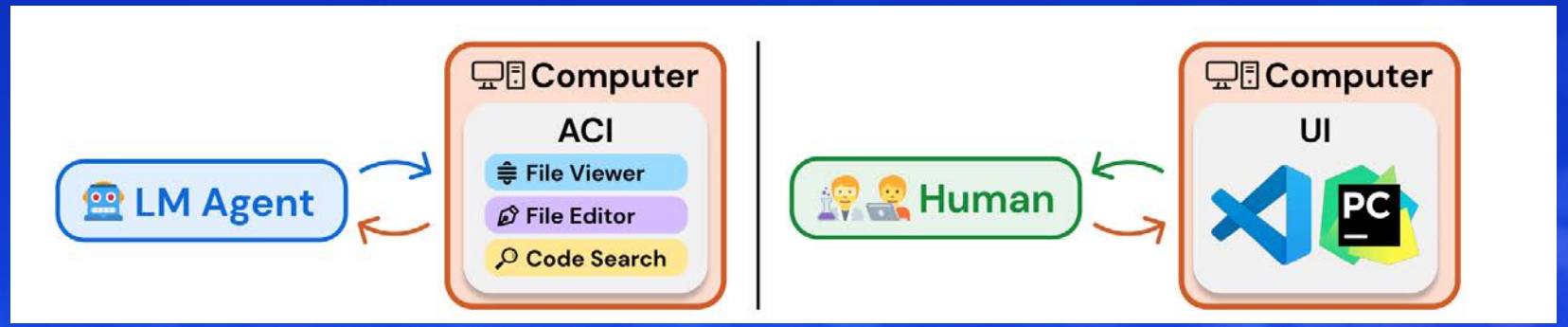
PART 01

工程智能体的背景

SWE Agent


Anatomy of a Software Engineer
Responsibilities
Develop software systems
Strategise with various teams & stakeholders
Design software solutions
Salary
Average of RM50k+ a year
Increasing as demand for SEs grows
Skills
Several programming languages
Ruby, Python, JavaScript,, CSS, etc
Benefits
Constantly learning new skills/tools
High salary
Large room for career growth
Various resources to learn
Education
Bachelor's or advanced degrees
Bootcamps/training courses
Career path
Easy to go into different industries
E.g: retail, healthcare, high tech


Issue
data leak in GBDT due to warm start (This is about the non-histogram-based version of...
Codebase
sklearn/
examples/
README.rst
reqs.txt
setup.cfg
setup.py
Language Model
Generated PR
+20 -12
sklearn
gradient_boosting.py
helper.py
utils
Unit Tests
Pre PR
Post PR
Tests
join_struct_col
vstack_struct_col
dstack_struct_col
matrix_transform
euclidean_diff


LM Agent
Computer
ACI
File Viewer
File Editor
Code Search
Human
Computer
UI
PC

工程智能体的场景

AISE 场景 | AI 理解/规划需求 | 需求生成设计稿 | 设计稿生成代码 | 需求生成测试用例 | 设计稿生成代码 | 测试用例生成测试代码

LLM 大模型– 基座模型

需求领域模型 | 测试领域模型 | 设计领域模型 | .....

需求理解Agent	测试用例Agent	页面编排 Agent	测试执行Agent	API
产品规划Agent	API 测试 Agent	序列对比 Agent	结果校验Agent	RAG
UI 组件Agent	UX 效果Agent	UI 测试Agent	安全测试Agent	Tools

Prompt 工程 | Agent 逻辑解析器 | 人工review 与交互

研发领域 LLM 多智能体

理解与决策

企业专属知识库

记忆与存储

业务知识库 | 需求知识库 | 代码知识库 | 测试知识库

自动化/平台工程

执行与反馈

需求自动化 | 代码仓库Hook | IAC 资源自助化管理

API 接口测试 | 移动端自动化测试 | UI 自动化测试

环境自动创建/销毁 | 自动化安全测试/检查 | 混沌工程

研发管理系统 | 需求管理 | 设计管理 | 代码管理 | 制品管理 | CI/CD | 资源管理 | 运维监控

管理与交互

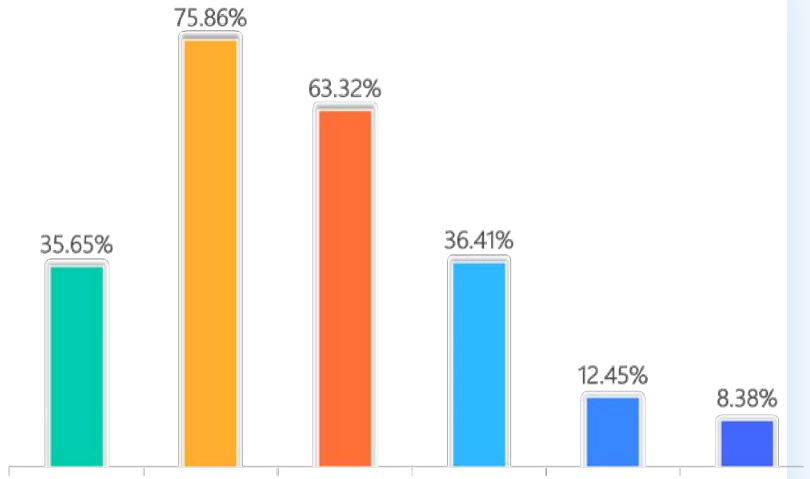
PART 02

问题和痛点

软件工程的切入场景应用繁多

中国信通院调查显示：
软件开发&测试场景是AI4SE技术应用的排头兵

软件工程各阶段AI技术应用比例



AI4SE：是指以大模型等AI技术为驱动的，以提高软件研发运营智能化水平为导向的，以提质增效为目标的，新一代智能化软件工程。

2023年是AI大模型元年，AI4SE相关的商业化解决方案百花齐放，
AI相关的生产力工具正在重塑技术人员的工作方式

需求分析
- Facilitation: miro
- Documentation: Userdoc
- Collaboration: ATLASSIAN Intelligence

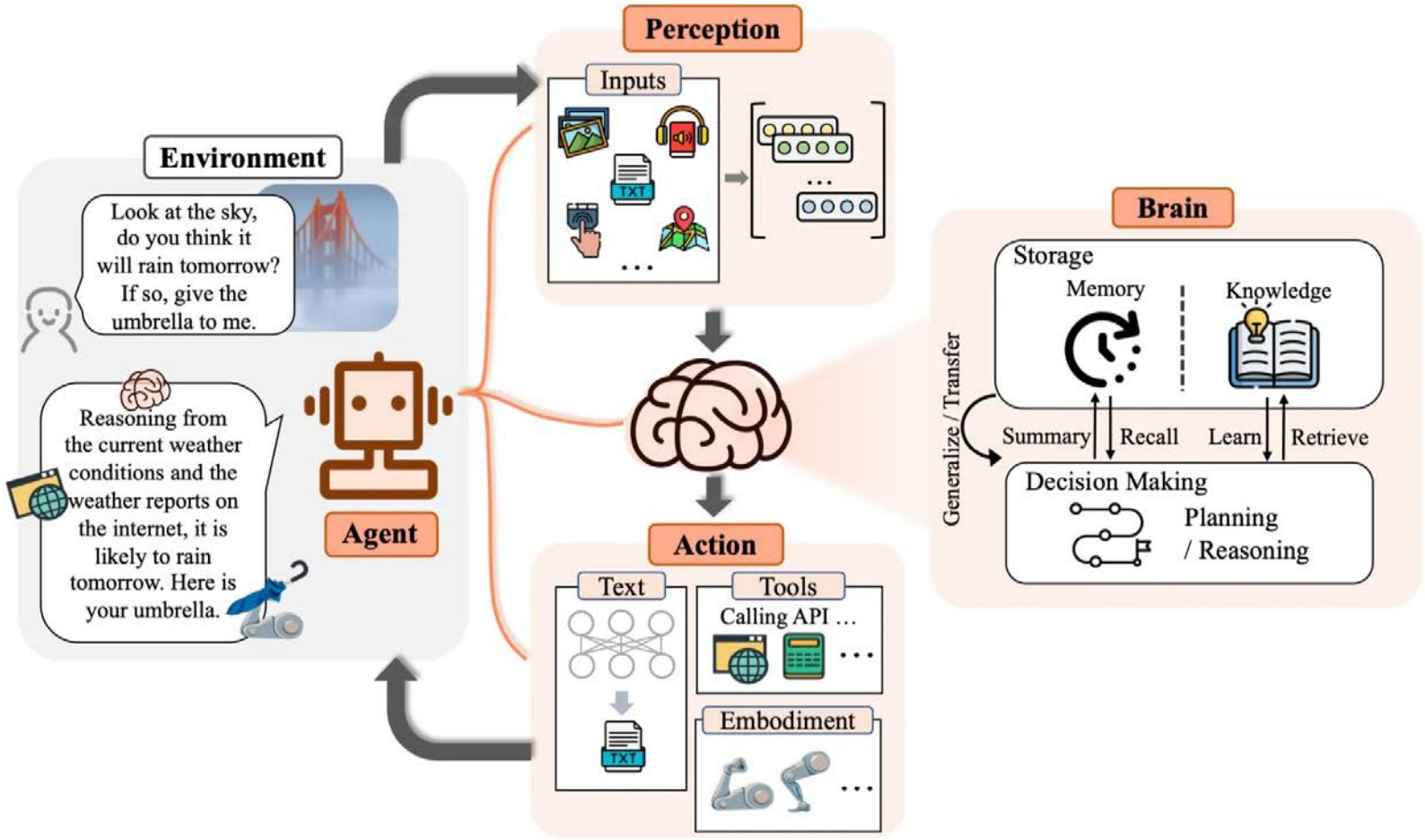
运营
- Research: Kraftful, Enterpret
- Marketing: Jasper, copy.ai, regie.ai, Midjourney
- Customer Support: INTERCOM, ada

产品设计
- Ideation: Midjourney
- Prototype: Visily, uizard, Figma AI
- Research: Kraftful, QoQo, Looppanel
- Design: stability.ai, Adobe Firefly, Galileo AI

交付与监控
- Operations: airOps, KUBIYA, arize
- Observability: Grafana, Cribl, baselime, ctrlstack, dynatrace
- Host: Vercel, netlify, Deno, Temporal, render, Fly.io, Railway, seaplane, akeero, zeet, Pulumi
- CI/CD: harness, Buildkite

开发阶段
- IDE: Graphite, planor, Codeball
- Commit: GitHub Copilot, tabnine, Sourcegraph, codeium, codium, mutable.ai
- Copilots: Magic, grit, Second
- Documentation: readme, swimm, Redocly, Stenography, Readable, Mintlify

企业智能体建设的难度



感知端（Perception）

非结构化数据、
图片辅助应用可交付生成

问题：模型、准度

控制端（Brain）

企业知识库

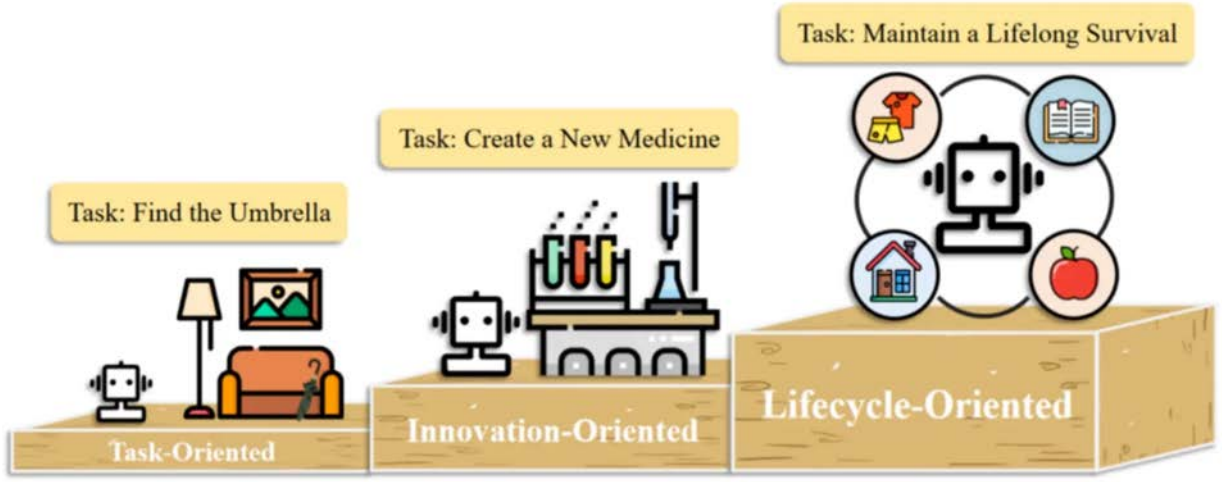
问题：文档不规范

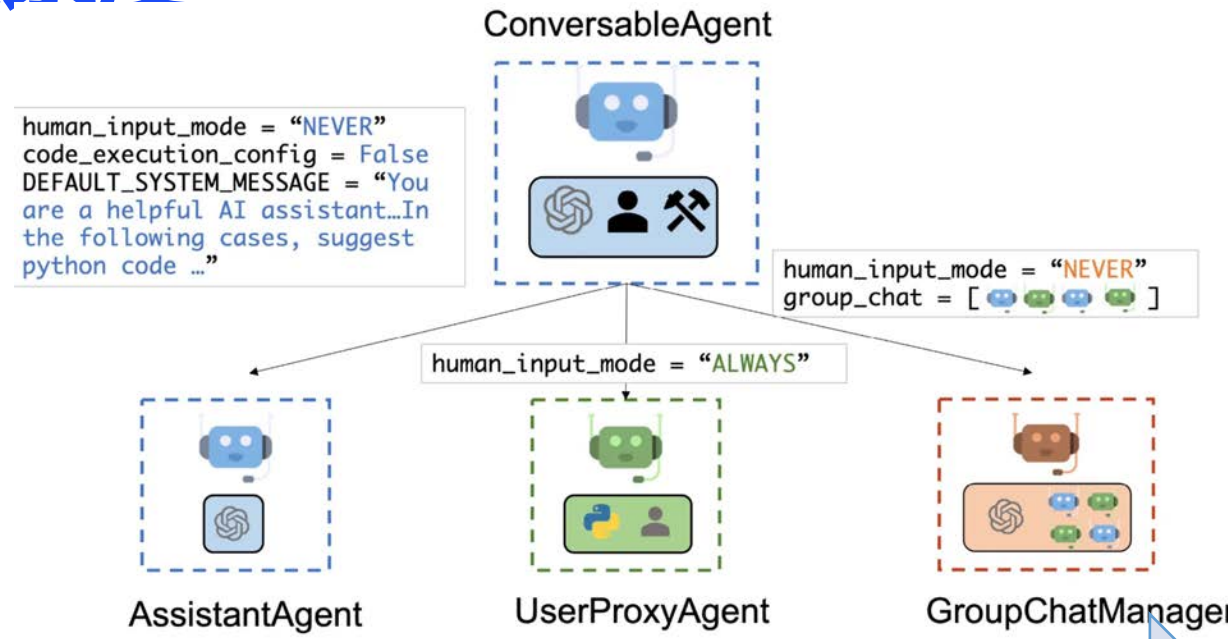
行动端（Action）

开放性、集成性企业系统

问题：复杂问题LLM无法正确调用，不正确的入参/出参

从单Agent到多Agent的复杂度





执行单一任务	学会思考并发掘	TryRun/TryFix	协作	对抗
做什么 怎么做	拆解需求 发现更多相关 信息推理	自我修复 自我保护 自我生成优化	有流程的有序协作 无顺序的自由协作	对话对抗表达观点 评审对抗 代码防护性对抗

LLM与感知、行动的配合

LLM与不同智能体角色的协作

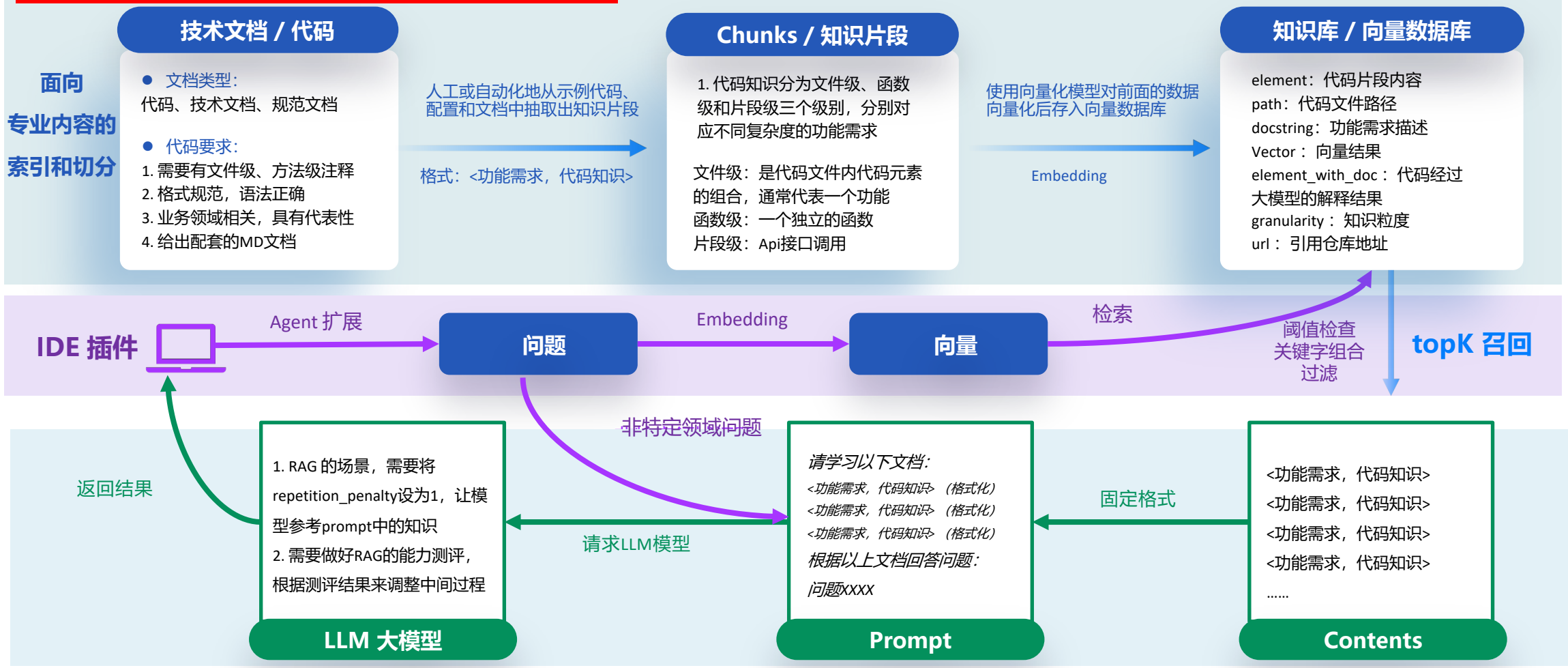
场景简单 → 工程级别单一场景可用 → 复杂工程问题 → 软件工程智能体

业务文档和代码加入知识库

RAG 技术（Retrieval Argmented Generation) 基本都是用来处理自然语言的，无论是量化处理，还是召回，业界都仅在自然语言场景下可用。

问题一： 业务文档不规范不完整

问题二：“代码文件” 的量化与准确召回一直是业界难题。



PART 03

解决思路/整体方案

感知更多的工程背景

AI 编码辅助中的代码补全，主要由 感知 – Prompt 构造 – 生成 这三个部分组成。

因此提升代码补全效果的方法大致可以分为：更全面的感知上下文、更精确的封装 Prompt 以及 友好的结果处理机制

感知

结合 AST 技术，在代码补全过程中，引入除当前文件之外的与其相关联的代码，实时延展上下文，获取更多跨文件内容

代码上下文	Prefix	Suffix
代码知识	Fill in Middle	依赖解析
	相似代码	调用链/符号定义

更丰富、理解工程

业界做法：大部分产品都只感知当前文件，少了先进产品会结合AST，但是最多也只支持到“打开的文件”进行分析。

腾讯实践：感知的范围更广，在加载工程的时候，就会进行全项目的感知与解析，即使关联文件未打开，也支持。

Prompt 构造

使用相识性分析算法，在感知到代码文件之后，会对文件内容做截取，只保留关键部分代码，并对于关键性代码进行权重排序

内容压缩	相关性分析	语义截断
权重排序	相似函数	调用链路
	父类判断	顶端注释解析

更精准、降低幻觉

业界做法：大部分产品都只参考开源社区的做法，进行 PSM 或者PMS 进行 prompt 封装，既慢，也不准。

腾讯实践：更精细的prompt封装，通过压缩、截断、排序、顶层注释等方式，不仅提升了补全性能，也降低了幻觉。

结果处理

插件会根据当前上下文智能判断行补全或者块补全，并对于补全结果结合上下文进行融合

Stop 策略	智能补全
内容融合	结合语义融合上下文

更智能、优化体验

业界做法：固定的补全粒度（比如 大、中、小；或者行补全、块补全），不智能。

腾讯实践：智能补全，可以根据代码上下文，分析最适合的补全内容，体验更优

构建自主混合多源知识库 Autonomy Hybrid Multiple Knowledge Base

查询增强

- 提问后下轮推荐
  多跳查询
- 提问查询的优化
- 假设文档嵌入
- 摘要

多知识库

丰富的文档类型

支持丰富的文档类型，包括：pdf、markdown、txt、docx、 doc、html 网页、代码(检测非二进制的代码文件)、 ZIP压缩包（.zip/.gz) 等

代码/文档混合处理

会根据不同的处理方式，分别处理文档和代码。在增强回答的时候，可以结合代码/文档，分别提供解决方案内容与解决方案代码。

API 文档处理

针对 API 文档，会当成单独的类型处理。在问答过程中，可以所选代码内容与 API 文档内容，给出相关的代码实现。

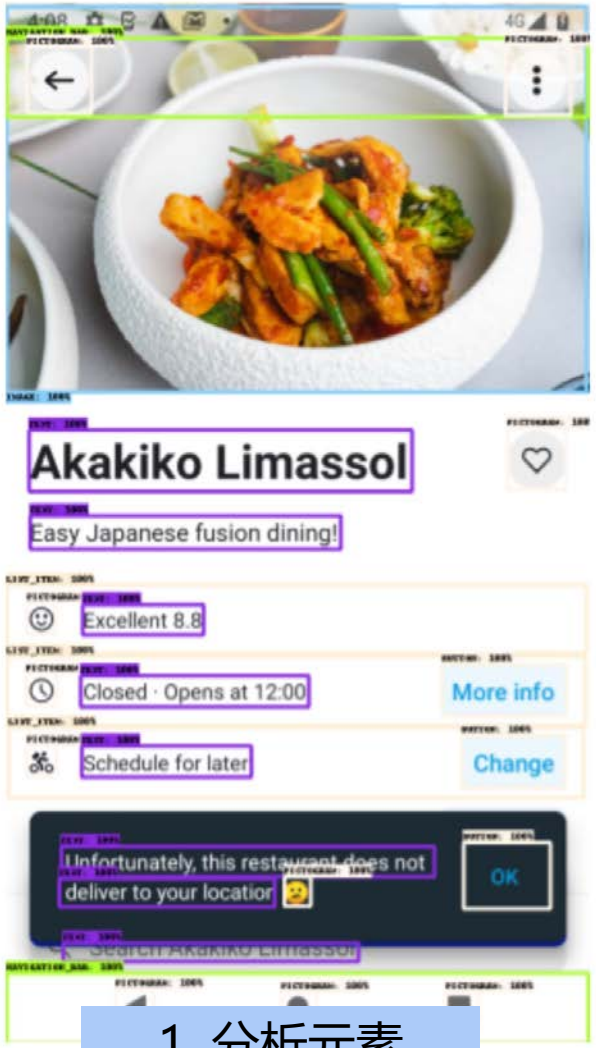
多知识库问答

在问答的时候，会遇到一个问题，需要参考多个知识点/知识库的场景。问答中，可以选择多个知识库增强回答效果。

后检索

- 合并重排
- 过滤、设置阈值
- 检索到的内容改写

图片信息提取



1. 分析元素

2. OCR提取层次路径下的元素，

该模型的任务是检测和识别屏幕上的 UI 元素。这包括执行 OCR 和图像字幕以理解和解释文本和非文本内容。

```
IMAGE a white bowl with a chicken curry and vegetables . 0 99
  NAVIGATION_BAR 1 996 34 109 (
    PICTOGRAM arrow backward 36 148 43 105
    PICTOGRAM three dots 853 966 41 107)
  )
TEXT Akakiko Limassol 39 695 411 469
PICTOGRAM heart 857 959 409 467
TEXT Easy Japanese fusion dining! 40 574 493 524
LIST_ITEM 0 994 560 625 (
  PICTOGRAM happy face 35 86 577 606
  TEXT Excellent 8.8 130 339 579 607)
LIST_ITEM 1 991 628 694 (
  PICTOGRAM time 34 87 645 675
  TEXT Closed Opens at 12:00 128 518 647 676
  BUTTON More info 745 959 636 685)
LIST_ITEM 4 988 697 763 (
  PICTOGRAM 743 714 87 35
  TEXT Schedule for later 129 420 715 744
  BUTTON Change 778 957 704 754)
TEXT Unfortunately, this restaurant does not 94 733 811 839
TEXT deliver to your location 90 460 842 868
BUTTON OK 782 931 807 870
PICTOGRAM sad face 475 522 840 867
TEXT Search AkCkiKU LilliASSOT 98 603 904 921
NAVIGATION_BAR 0 997 933 999 (
  PICTOGRAM arrow backward 187 254 948 984
  PICTOGRAM a gray circle with a white background 471 532 95
  PICTOGRAM nav bar rect 752 809 951 982)
```

3. 生成中间语言

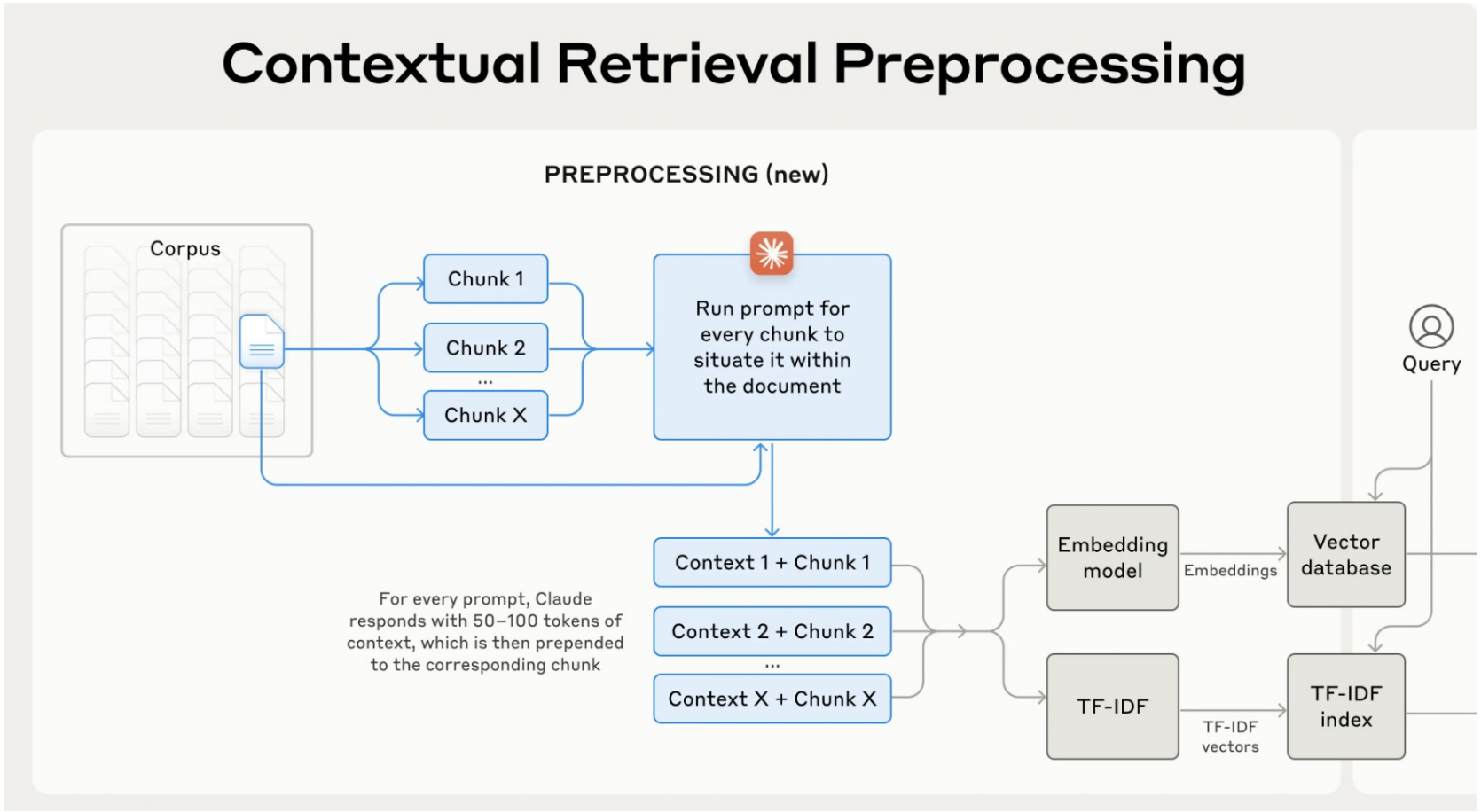
4. 在合适的组件库中找到最接近的组件

5. 为每个组件生成描述，并生成完整的任务提示词

结合多模态模型，最终拓展感知能力

构建面向代码工程的企业知识库

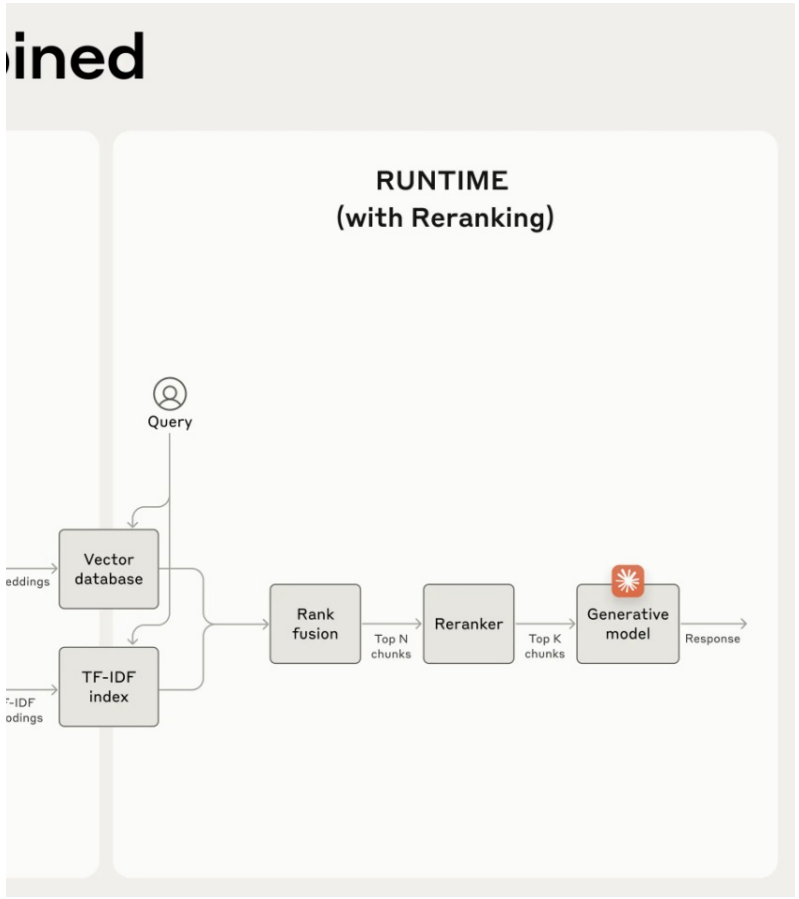
目标：解决文档、代码信息不全，增强工程上下文的知识点关联性



Contextual RAG Preprocessing

1. 为代码文档添加通用摘要
2. 基于摘要索引
3. 文档型代码优先Embedding
4. 调整召回的各种参数
5. 召回测试

▶ 允许重排、调整参数和召回测试

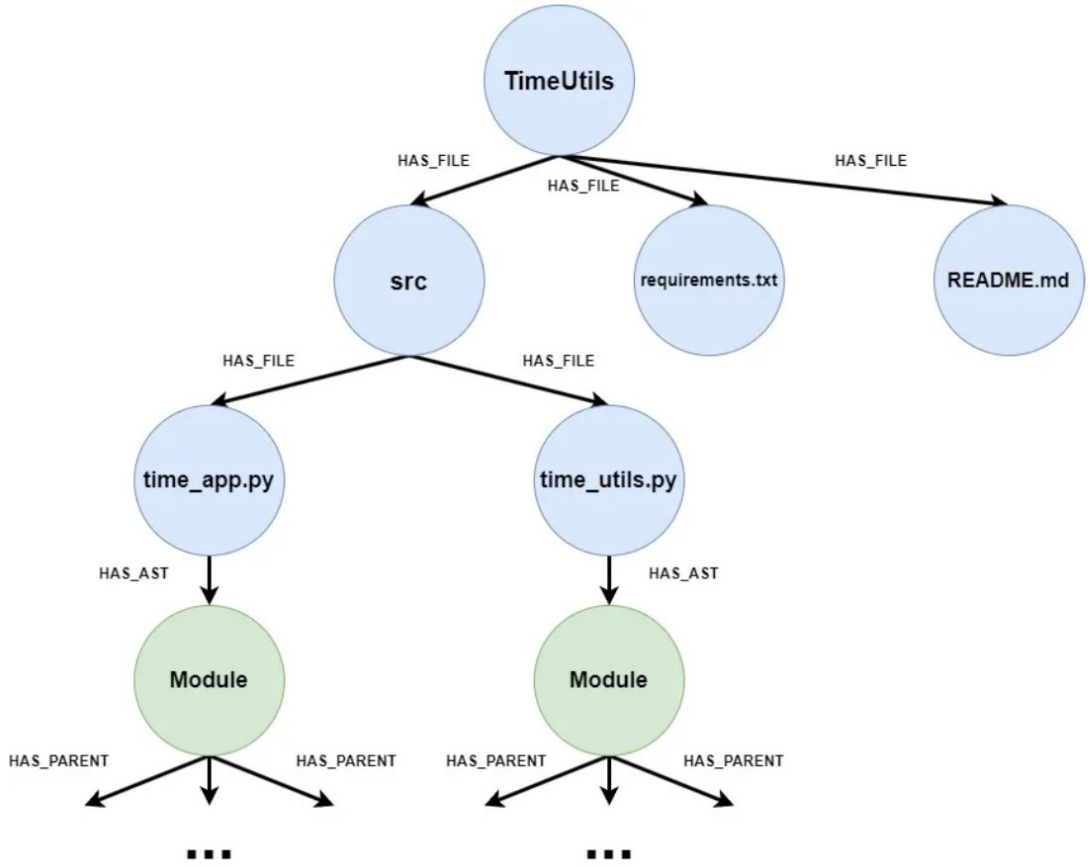


Rerank + TopK + Score Threadhold

创建自定义知识库

探索工程背景的知识图谱

目的：对代码中的复杂关系和依赖关系进行建模



1.加载和预处理代码文档数据

2. Rewrite 提问增强 – 多跳问答

3. Codebase检索本地工程 – 代码检索

4. Contextual RAG

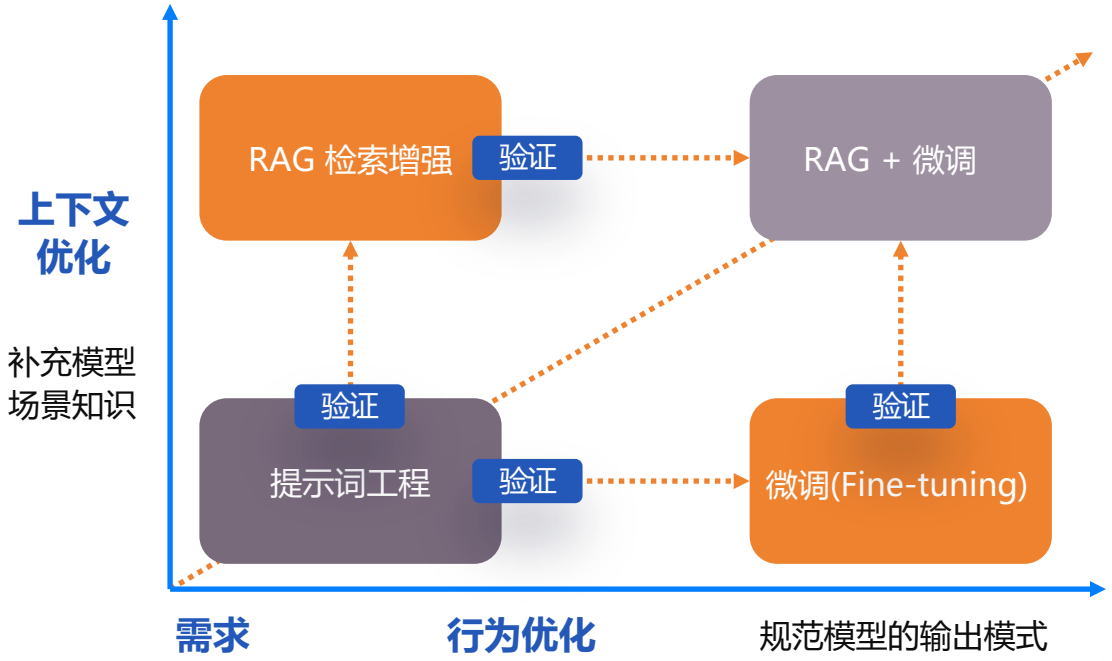
5. 基于知识图谱的切分策略

6. 知识图谱召回测试

Prompt 工程、RAG、微调 的决策

大模型特点：

1. 不确定性 ---> 提升模型输出的稳定性质
2. 静态性 ---> 扩展额外数据



构建大模型应用是一个典型的迭代过程，这个构建过程需要从应用场景出发，先搞清楚我们要做什么，然后再去优化大模型应用系统的性能、质量和用户体验。

1 尝试定义新的提示词

首先使用提示工程的方式优化大模型应用，这是成本最低，见效最快的方式。提示词工程可以同时为模型补充上下文（上下文优化）和优化模型的行为（行为优化）。

2 多样本学习 和 上下文引导学习

提示工程的能力直接赋予用户，用户根据自己的实际场景，在提示词模板中丰富场景的用例样本，让模型的返回更符合业务场景

3 检索增强式内容生成（RAG）

当模板中大部分内容是模型未知且不断变化的知识，且需从多个数据源获取，导致创建不同模板应对各种场景时，可以考虑引入RAG进行优化。

4 微调 （Fine-tuning）

当模板主要包含各种格式示例，且目标是让模型模仿示例输出，但即使示例充满模板，输出仍不稳定时，可以考虑进行微调以优化模型表现。

PART 04

架构实现/技术实践

产品架构图

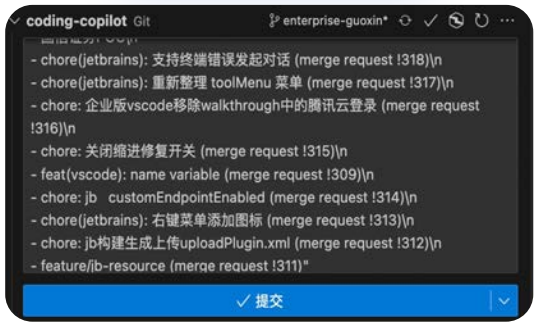
说明：

- **一套代码**：一套代码，通过配置的方式生成不同的产品，公有云、私有化、腾讯内部，产品基础能力完全一致
- **多种部署方式**：应对不同客户群体的诉求
- **多模型**：为不同的客户提供和调度不同的模型服务
- **RAG**：通过 RAG 拓展知识领域，提供贴近客户业务的问答能力
- **Agent**：结合客户自有产品或腾讯云其他产品，自定义 Agent，串通流程

Agent扩展整体架构

已经集成的内容

AI 生成提交信息



代码 AI 评审

未来可扩展点

代码库集成扩展

- 识别代码设计
- 规范格式问题
- 识别性能问题
- 识别和修复漏洞
- AI 配置评审规范
- 提升代码可读性
- 提升代码可维护性

CI集成拓展

- 修复构建错误
- 分析构建耗时
- 识别代码扫描错误
- 识别代码安全漏洞
- 识别自动化测试错误
- 实现安全门禁

CD集成拓展

- 识别制品漏洞
- AI 变更评审
- 识别和修复部署错误
- 集成 OS，终端补全
- 识别配置内容
- AI 架构治理
- AI 根因分析

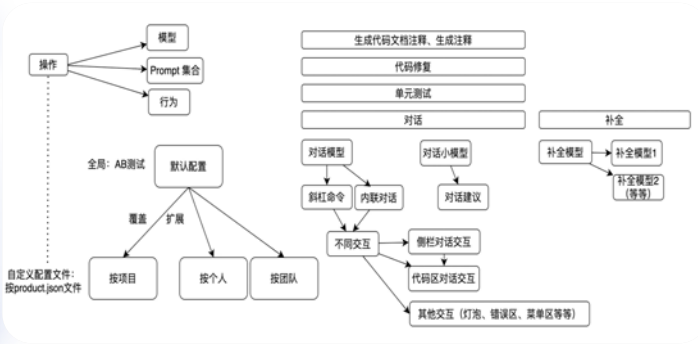
企业研发规范知识库 | **语言、框架知识库** | 企业知识库管理 | **CI/CD 编排知识库** | **企业文档**

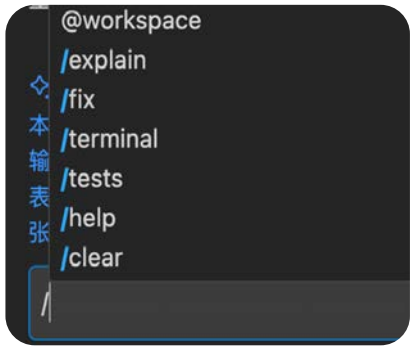
AI Agent 框架

IDE 插件命令扩展 （Prompt As Code）

行为感知	提示词	工具执行	权限控制	多轮处理	思维链路

Prompt 模板 / 配置中心 | **模型调度**





知识库整体架构

1 多样化的数据源支持

支持多样化的文档作为知识库的数据源，文档类型包括：普通文本、Word、PDF、Excel、md 等格式。另外支持代码仓库、数据库作为数据源

2 知识切片与存储

根据不同的数据类型，分别处理数据切片，并将切割之后的数据存放到向量数据库中。

3 便捷的检索增强体验

在技术问答的时候，可以通过 @知识库 的方式，选择某个领域的知识库，检索相关知识后，强化问答效果。

4 面向技术代码类文件的强化索引技术

对于代码文件、技术文档等特殊性，支持强化索引能力，以提高召回率

PART 05

总结与展望

AISE 场景规划与方法

AI 能力在企业落地不是一个独立的事情，它需要融入到企业现有的管理和工程场景中，为现有的系统注入新的能力。这个过程不单单是部署一套大模型应用，而是需要大部分人参与其中，逐步演进，最终产生价值。

对于企业管理者而言，如何构建一套 AI 应用在企业内部迭代和演进的规划路线，建设 AI 场景的核心竞争力，成为了关键问题。

AI 编码工具的核心能力

模型结合私域数据的持续优化能力

大模型发展迅猛，开源模型、商业模型发展迅速。对于企业而言，需要获取最新的模型成果，并快速结合企业私域数据产生业务价值，并持续优化。

模型结合软件工程上下文优化编码效果

代码生成有别于通用AI对话，它需要更准确的上下文并回复更准确的结果。如何深入分析软件工程，并让大模型理解软件工程成为核心能力

落地运营、产品体验、上下游整合

AI 编码工具好用，有更多人用才会产生价值。如何提升产品体验、提升用户活跃度，打造内部运营体系，也是 AI 编码落地的核心任务

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 · 敦煌站
K+ 思考周®研习社
时间：2025.08.29-30

K+峰会 · 上海站
K+ 金融专场
时间：2025.10.17-18

K+峰会 · 香港站
K+ 思考周®研习社
时间：2025.11.25-26

K+峰会详情

AiDD峰会

AiDD峰会 · 上海站
AI+研发数字峰会
时间：2025.05.17-18

AiDD峰会 · 北京站
AI+研发数字峰会
时间：2025.08.08-09

AiDD峰会 · 深圳站
AI+研发数字峰会
时间：2025.11.28-29

AiDD峰会详情

AiDD AI+研发数字峰会 第5届
AI+ Development Digital summit

利用AI技术深化计算机对现实世界的理解

推动研发进入智能化时代