InfoQ 极客传媒

单日十亿 Token 的企业 AI 十倍速落地与十倍降本之道

演讲人：陈曹奇昊

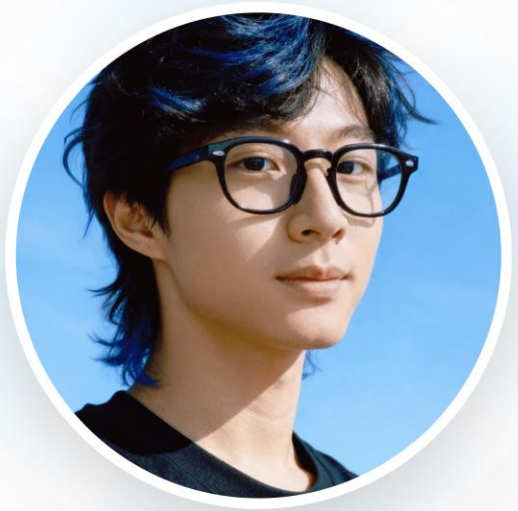
超级麦吉联合创始人

AiCon

全球人工智能开发与应用大会

由 超级麦吉 创造

演讲者介绍



陈曹奇昊

超级麦吉 CTO & 联合创始人

Agent 产研

一线 Agent 产研经验，主导超级麦吉产品的设计与落地

企业实践

曾任 KK 集团技术负责人，具备大型企业 AI 落地实践经验

开源贡献

PHP 官方团队成员 Stars 40k

Swoole 核心开发者 Stars 19k　Swow 项目发起者 Stars 1.3k

深耕高性能网络编程与异步架构领域，活跃于全球开源技术社区

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

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

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

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

4月 — 6月 — 8月 — 10月 — 12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

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

InfoQ 极客传媒

从独角兽用量到企业日常开销

2025 年初

AI SaaS 独角兽门槛

日均 **10 亿 token** 是衡量一家 **AI SaaS 独角兽公司**的重要指标

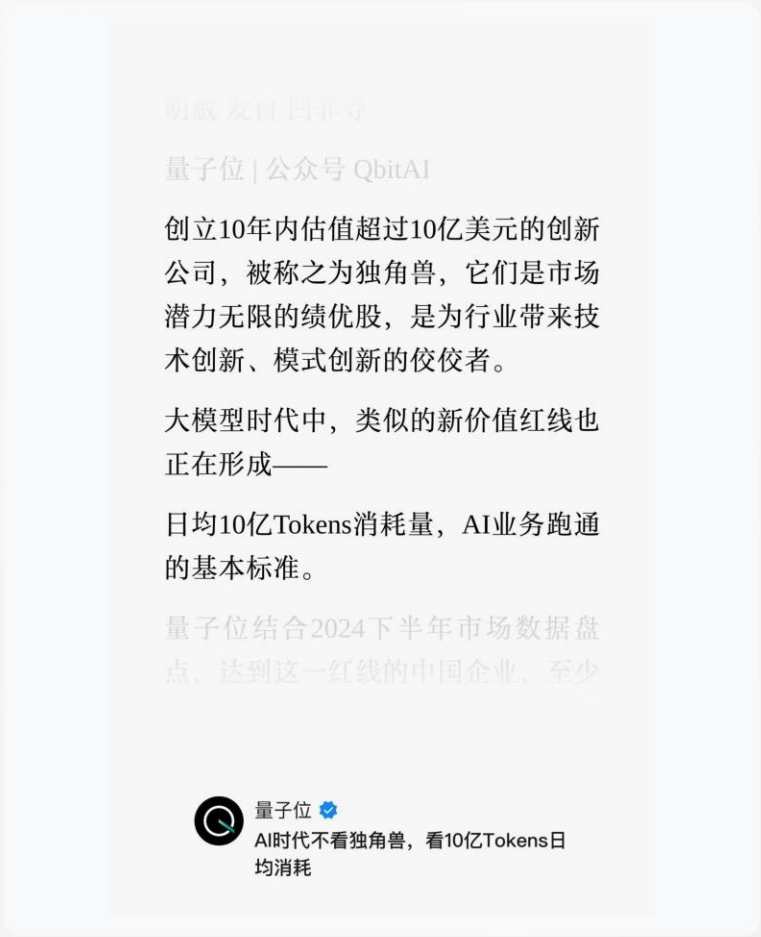
To C 应用的高水位

作业帮、筑梦岛等头部 To C 应用达到这个规模，已经是行业标杆

2025 年 11 月（现在）

企业日常消耗

日均 **10 亿 token** 可能只是一家零售、制造或电商**企业内部** AI 全面落地的**日常消耗**



三个核心问题

这个跨越是怎么发生的?

AI 如何从少数人的工具变成全员工作方式?

单日 10 亿 token，如何降本?

Agentic AI 给程序员带来的巨大转变

01 简单没有门槛

从「Show me your code」到「Show me your prompt」

一句话生成系统

设计一个面向 SaaS 软件客户的 CRM 系统，用于支持 SaaS 公司进行客户管理与业务增长。

AI 自动调用工具完成开发

设计系统架构
规划数据模型与页面结构
✓完成

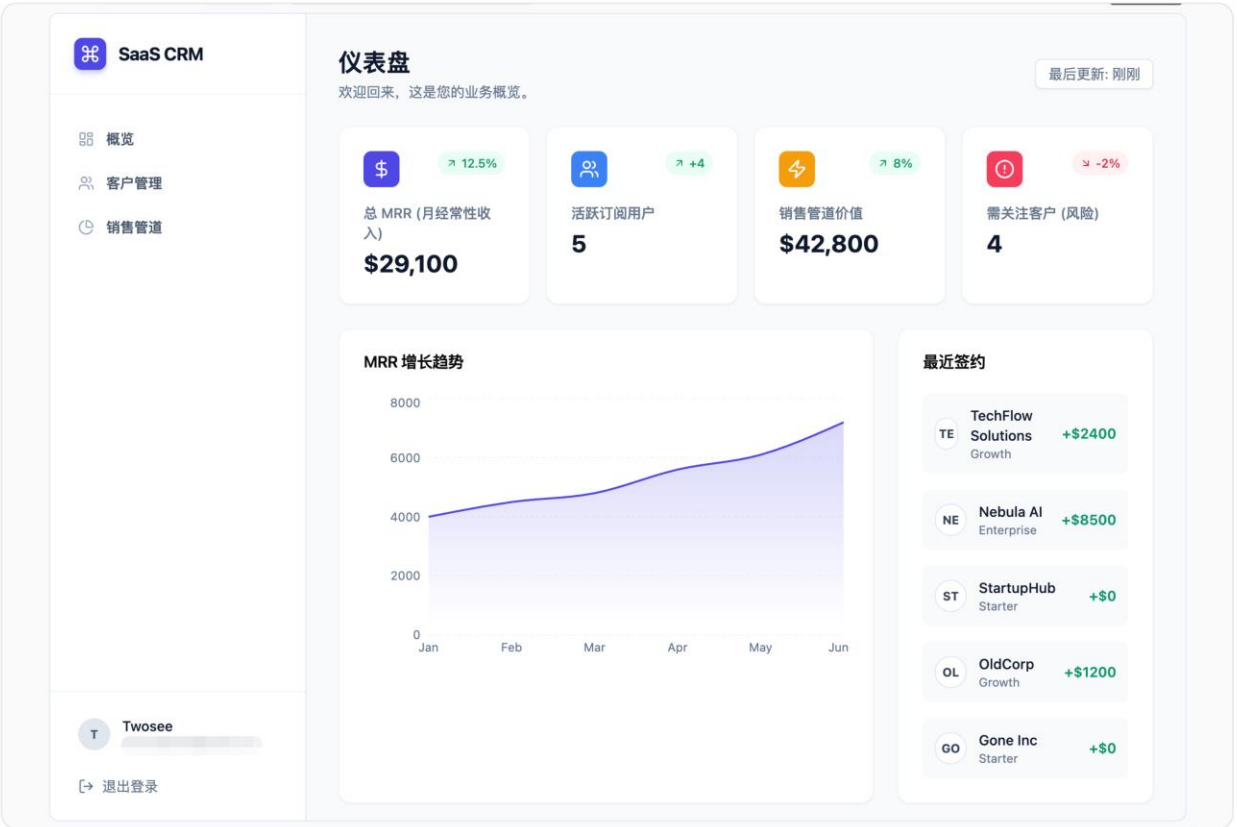
生成前端代码
React + TailwindCSS 界面
✓完成

配置数据库
PostgreSQL + API 接口
✓完成

测试验证
30 秒即可直接使用
✓完成

生成的 CRM 系统

可直接使用



Agentic AI 给程序员带来的巨大转变

02 AI 渗透工作的每个环节

从单点工具到全链路工作方式

Vibe Everything

AI 从辅助编码的工具，演变为贯穿整个开发生命周期的工作方式

工具

场景渗透

新的工作方式

写代码

Vibe Coding

Cursor、Claude Code、Copilot 实时补全，秒级生成函数和模块

查资料

Vibe Research

自动查询官方文档、API 说明、StackOverflow 解决方案

调试

Vibe Debugging

自动分析错误日志，调用浏览器定位问题，自动修复并验证

测试

Vibe Testing

自动生成测试用例、运行测试、发现问题后自行修复

部署

Vibe Deploy

自动完成构建、打包、部署、发布全流程

Agentic AI 给程序员带来的巨大转变

03 AI Native 不可逆转

很多事情没有 AI 就完全「没法做」

过去 · PAST

Coding → AI 辅助

不可逆转

现在 · NOW

AI 先行 → Coding

断网 = 怠工

手写代码效率骤降

习惯了实时补全，手写代码感觉「不会写了」

查文档变得陌生

翻 StackOverflow、读官方文档变得缓慢又痛苦

Debug 能力退化

习惯了 AI 分析堆栈，自己定位问题变得困难

开发周期延长数倍

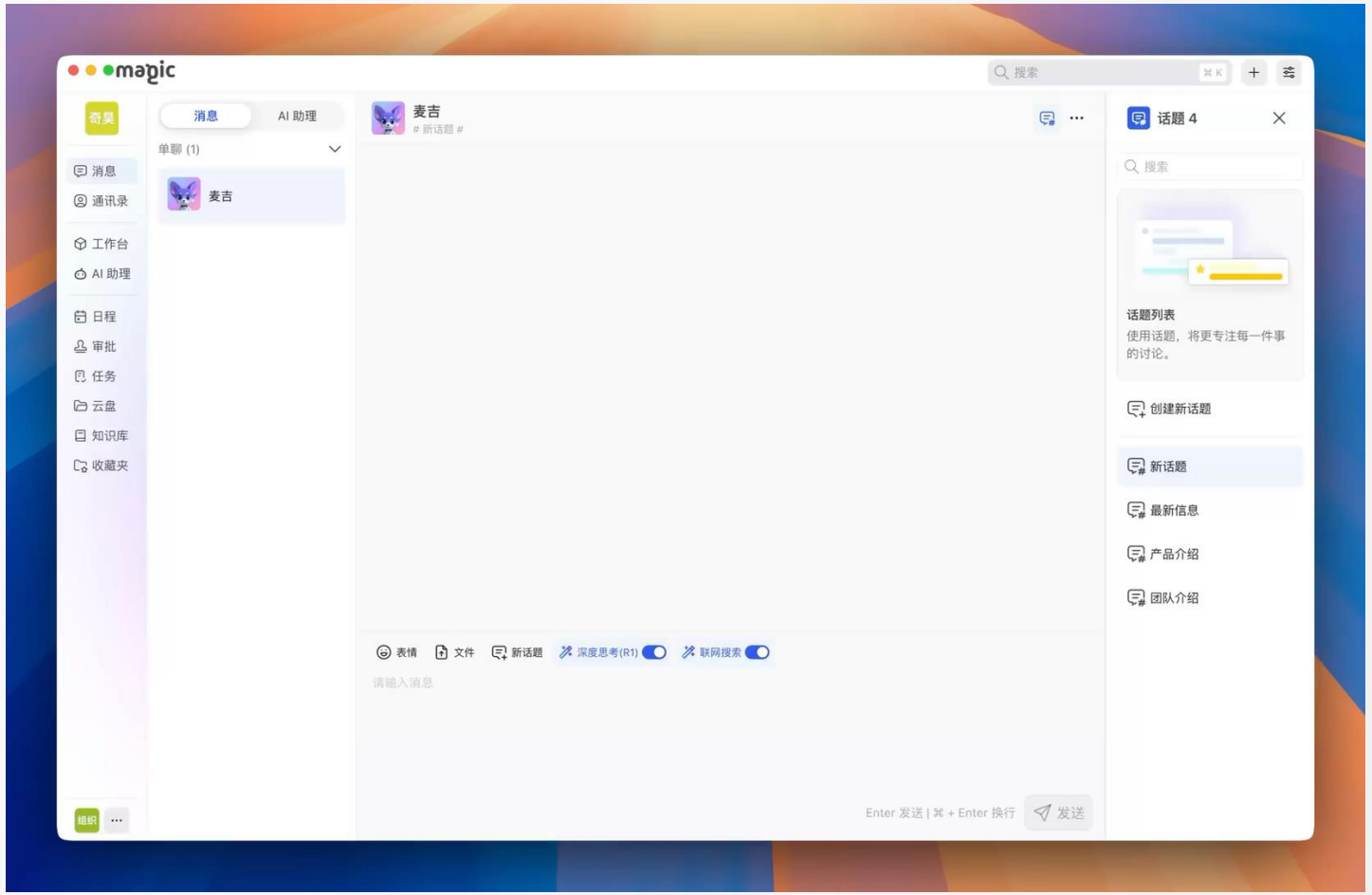
原本几分钟的事情，现在需要几小时甚至一天

企业 AI 落地的三个时代

时代一

聊天时代
岸上观望

- 纯聊天模式，对工作帮助有限
- 光是想用上 AI，就需要复制来复制去
- 大部分人停留在「试试看」阶段

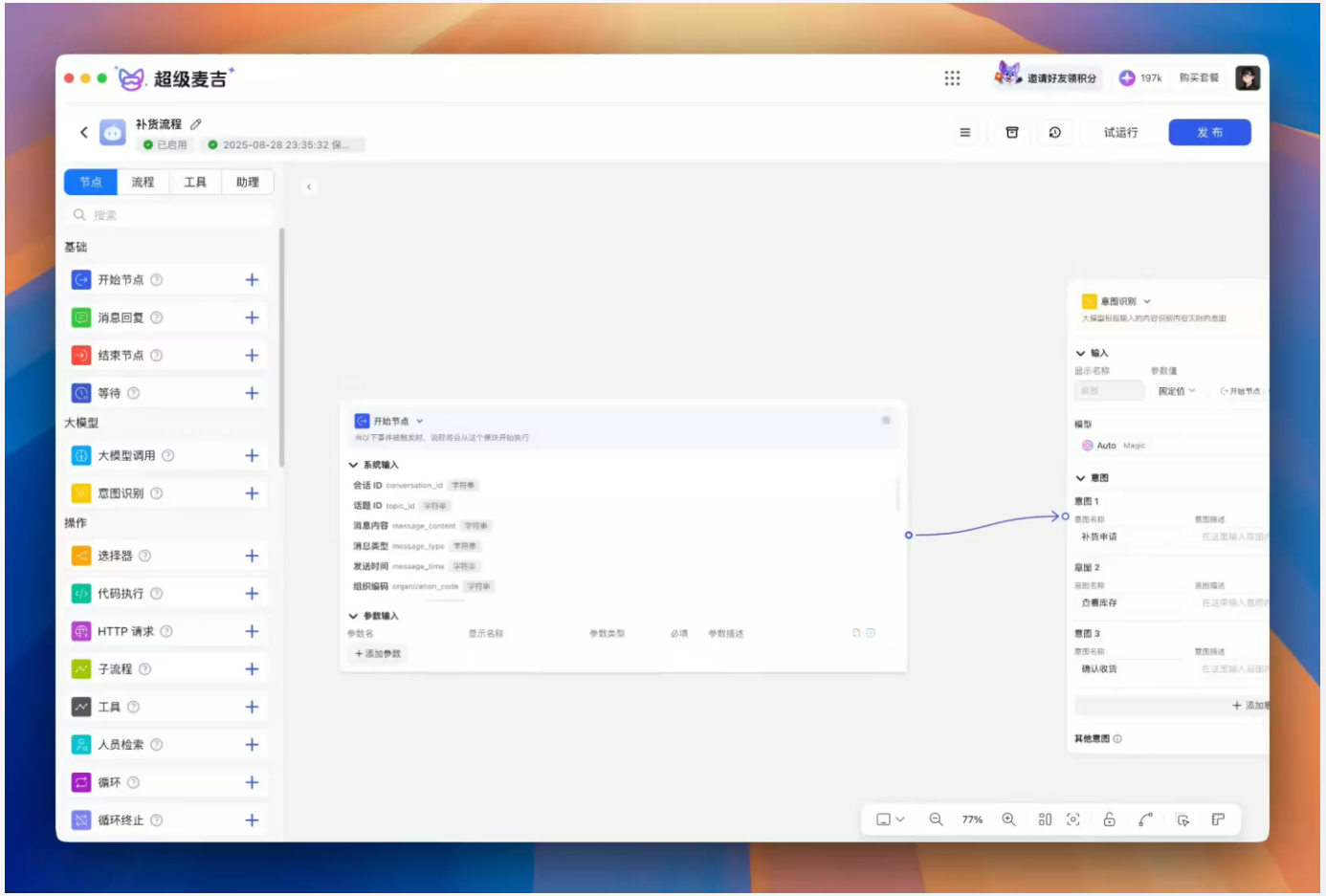


企业 AI 落地的三个时代

时代二

Workflow 时代
少数人下水

- 结合业务工作流在特定场景「干活」
- 门槛太高，只有技术人员或极客尝试
- AI 还是「工具」，即用即丢



企业 AI 落地的三个时代

时代三

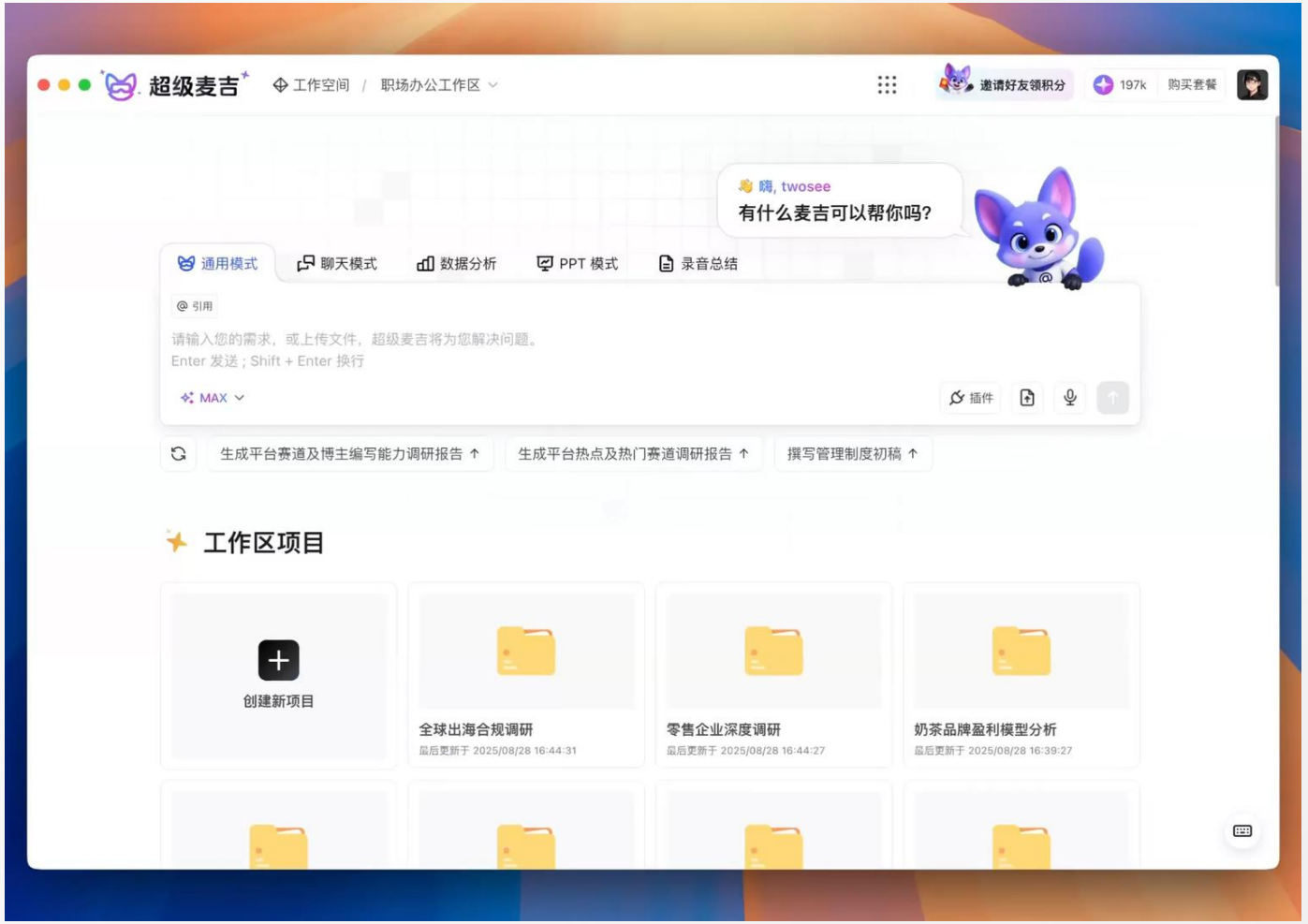
Agentic AI 时代

全员下水，高度依赖

会打字就能用

- AI 编程
- AI 调研报告
- AI 数据分析
- AI PPT
- AI 录音纪要
- AI 项目管理

• • • More • • •

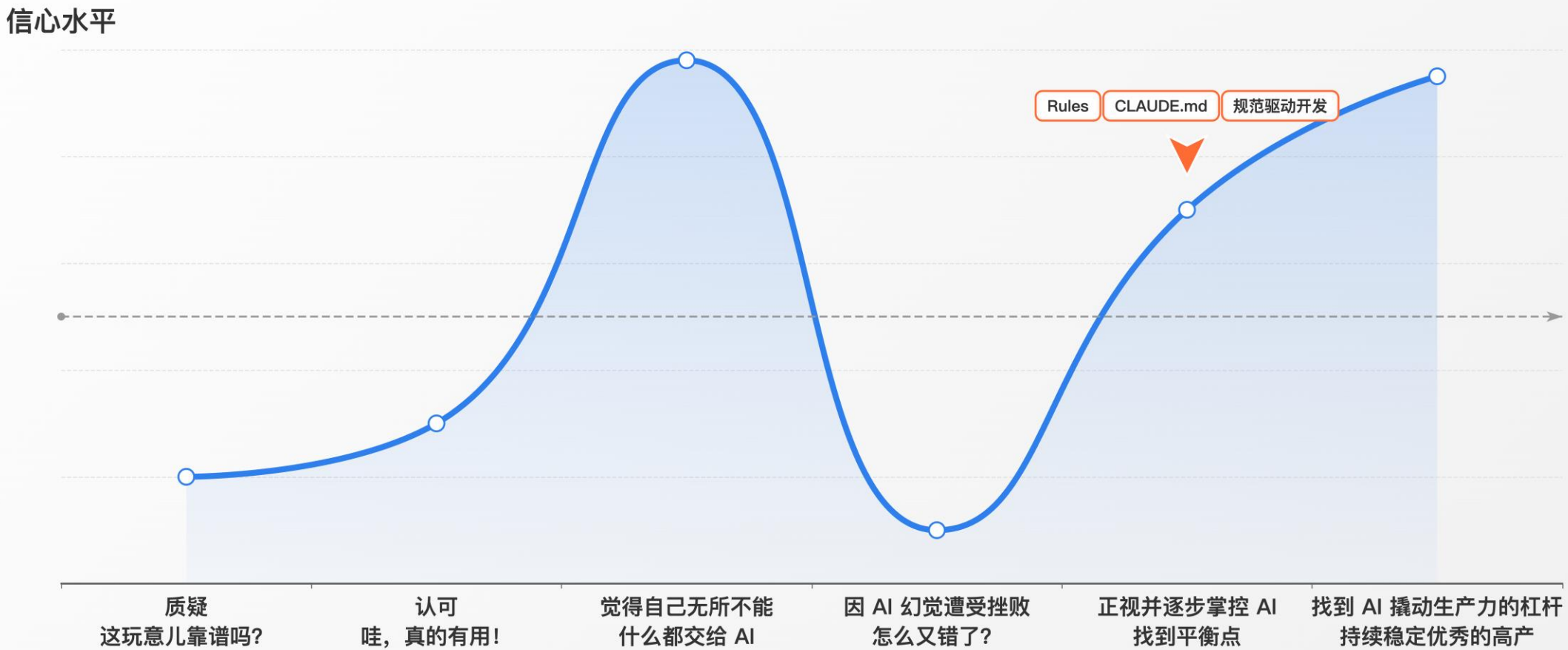


想要让全员拥抱 AI，
将 Agentic AI 带入职场这件事

必须马上干！

AI 领域的达克效应

Agentic AI 的现实



哪怕 Agent 完成每一步的准确率高达 99%
一个 20 步的任务，完美按照预期完成的概率也只有

$$(0.99)^{20} = 81.79\%$$

这种概率的级联衰减
让我们假定 AI 一定不靠谱

那么，产品如何设计才能挽回?

产品设计的回应

Human in the Loop

人在回路
人机协同设计

人掌控方向盘

让人在合适的时候把控方向盘，确保 AI 永远行驶在人所预期的最佳路线上

可控的自主性

当前 100% 交给 AI 是不可靠的，AI 是来帮助人的，不是替代人的

AI 产品的真正价值

1. 替代人去干低价值的事
2. 给予人力量感
3. 重新找到做事情的乐趣

在熟悉的工作体验中无限迭代

本地文件体验

云文件与本地文件操作体验一致，AI 可直接访问和处理

智能文件引用

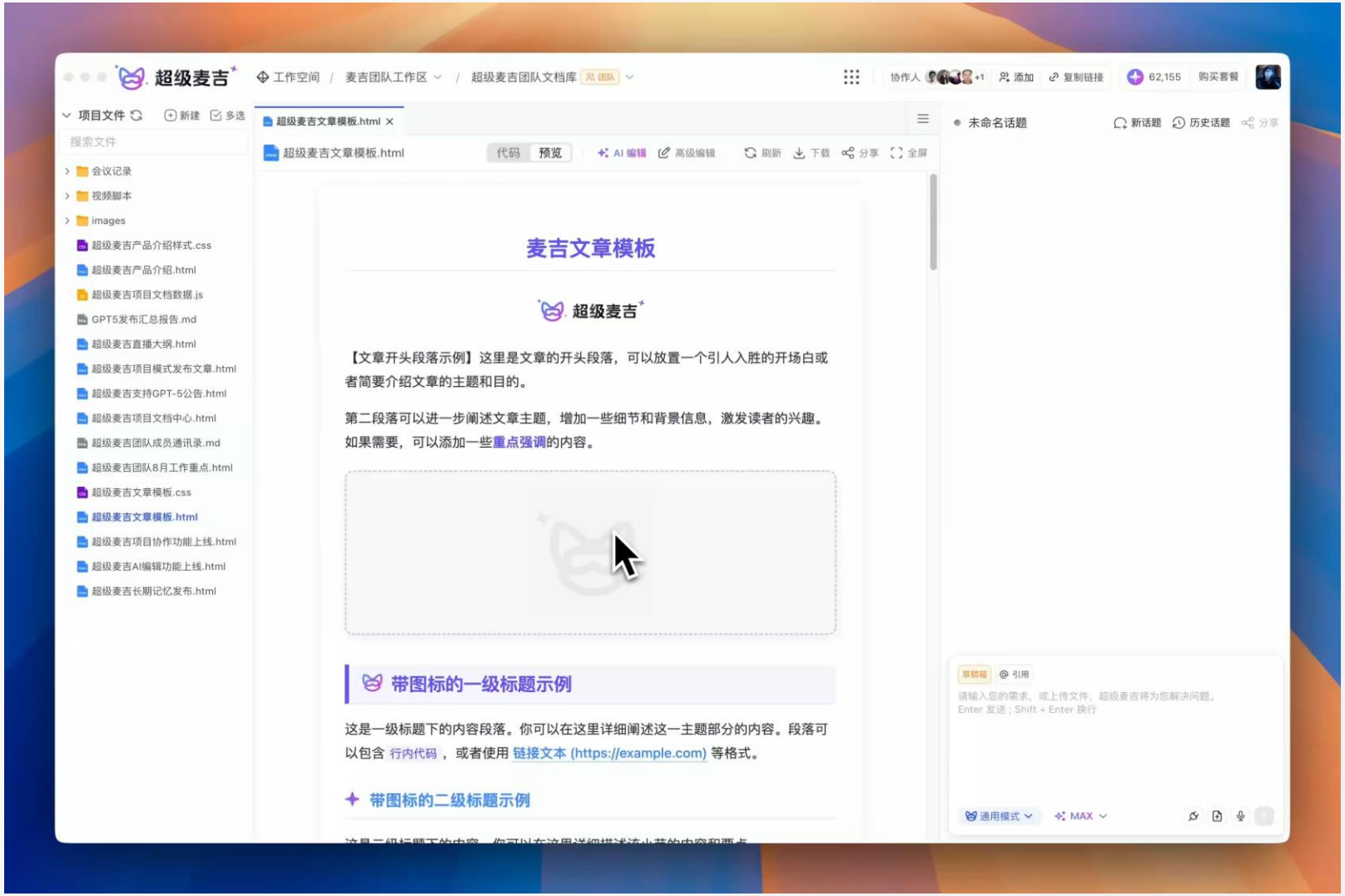
@ 符号快速定位文件，AI 精准理解修改意图

操作全程可见

每一步操作都清晰可见，支持随时查看、修改和回滚

持续迭代优化

通过对话不断优化，直至达到理想效果



Vibe Working 氛围工作法

AI 的使用艺术是无限并行的数字分身艺术，而不仅仅是单点速度

由 超级麦吉 创造

核心工作流程

持续对话 → 检查产出 → 迭代优化

你是指挥官
不是执行者

你管项目
不管细节

你做决策
不做重复

传统时代

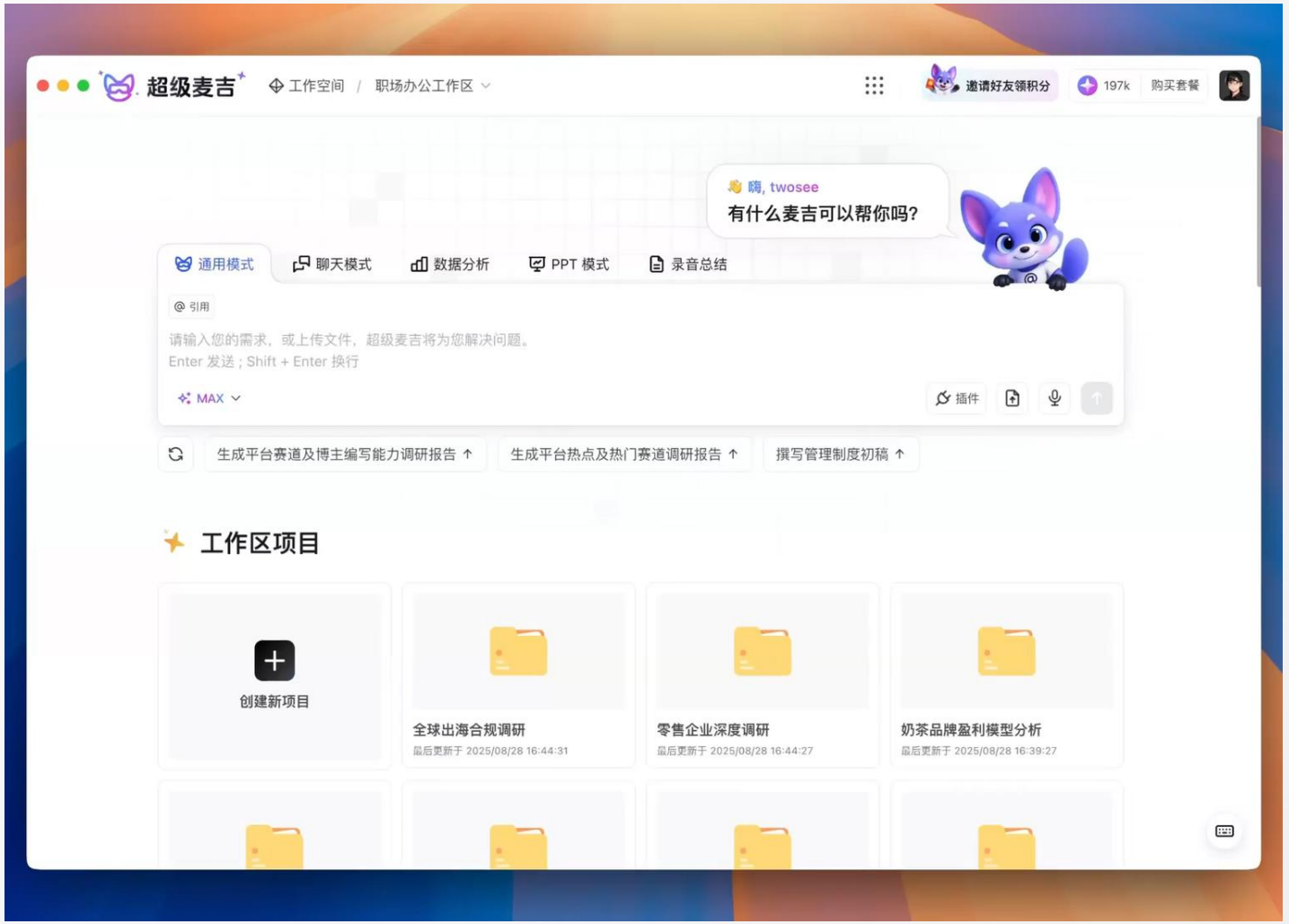
1 资深 + 2~3 初级执行岗

- 资深负责规划，初级执行
- 串行协作，等待成本高
- 人力成本随团队规模线性增长

AI 时代

1 个人 + N 个数字分身

- 人做决策，AI 做执行
- 无限并行，无需等待
- 边际成本趋近于零



Cursor for Working?

由 超级麦吉 创造

这不就是做业务版的 Cursor 吗，有什么不一样？

Cursor 的用户群体

程序员

世界上智商最高的一类人群 (也许吧)

天生会使用复杂的 IDE

超级麦吉的用户群体

职场白领

计算机小白 什么都不懂

设备配置低 不会装环境

Agentic AI 的技术选择

如何降低门槛
让大家都有动力下水试一试?

01

沙盒 OS 与 云文件系统

多人协作
无限迭代

02

零构建渲染

面向小白
所见即所得

03

在线编辑

长尾微调
一步到位

沙盒 OS 与云文件系统自研实现

云盘方案：直连对象存储

前端与沙盒**都直连对象存储**，无需服务端转发，带宽高，成本低。
无需唤醒沙盒，也能查看和修改。

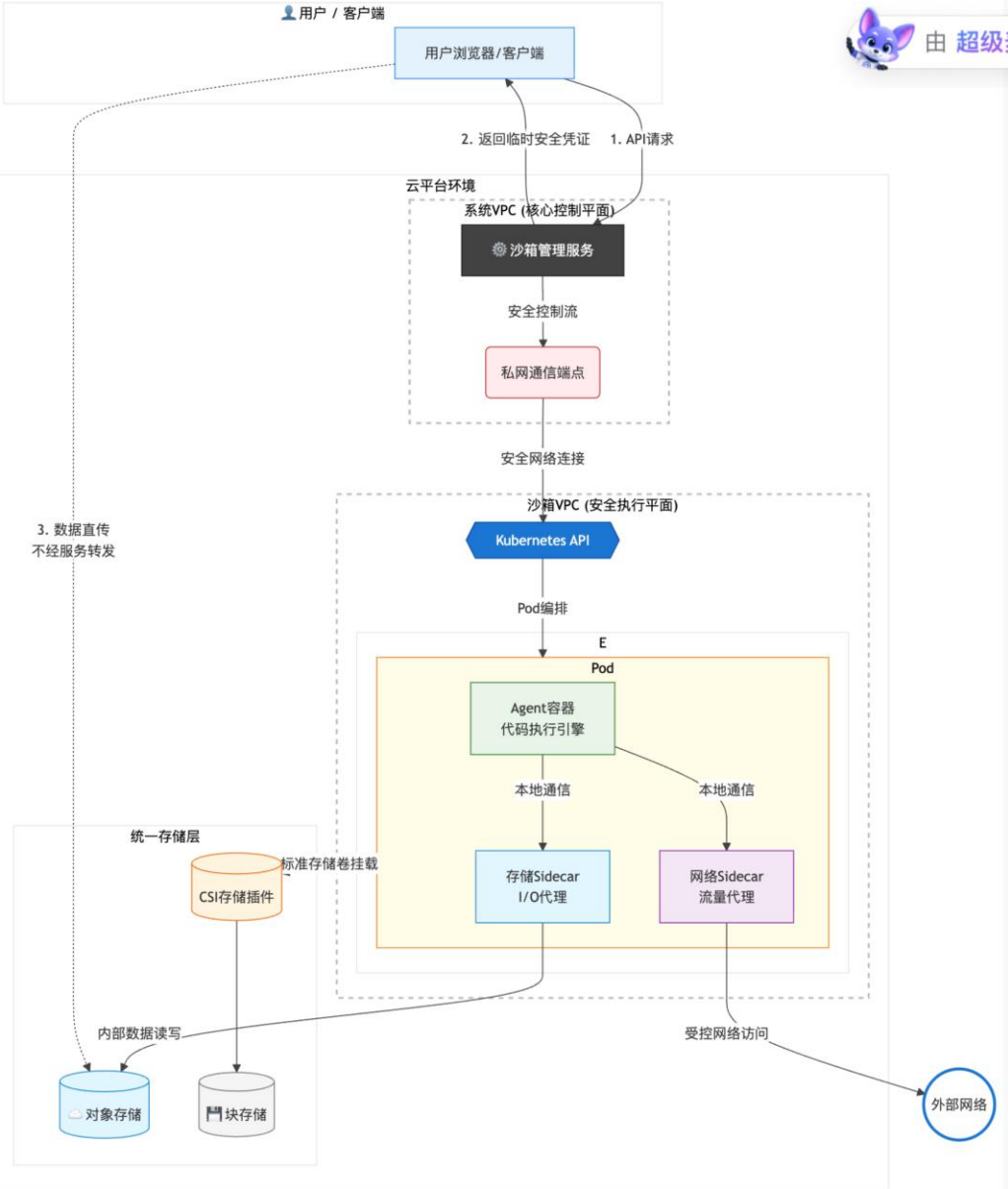
无限制共享：不同节点也可协同

任意节点用户**共享同一文件系统**，实现**多人协作**完成同一项目，不受传统存储限制

实时感知：毫秒级通知

文件变更时，**事件驱动机制实时通知**前端与 Agent，**所见即所得**，AI 即时响应修改

- **Sidecar 启动加速** Agent 容器与 Sidecar 容器同时启动，无需等待，避免 PVC Ready 额外时延
- **容器预热与镜像预加载** 用户请求前容器已预热就绪；新版本发布时镜像异步预加载，实现无缝切换
- **可信网络架构** Sidecar 网络代理实现用户隔离；沙盒与系统处于不同 VPC，通过私网终端节点连接
- **自动资源管理** 多 GC 管理器自动回收过期资源；独立配置 CPU 与内存配额，支持多租户隔离



实时感知：事件驱动的毫秒级通知

核心能力：所见即所得

事件驱动机制

文件系统变更时，通过事件驱动架构实时推送通知，前端与 Agent 毫秒级感知

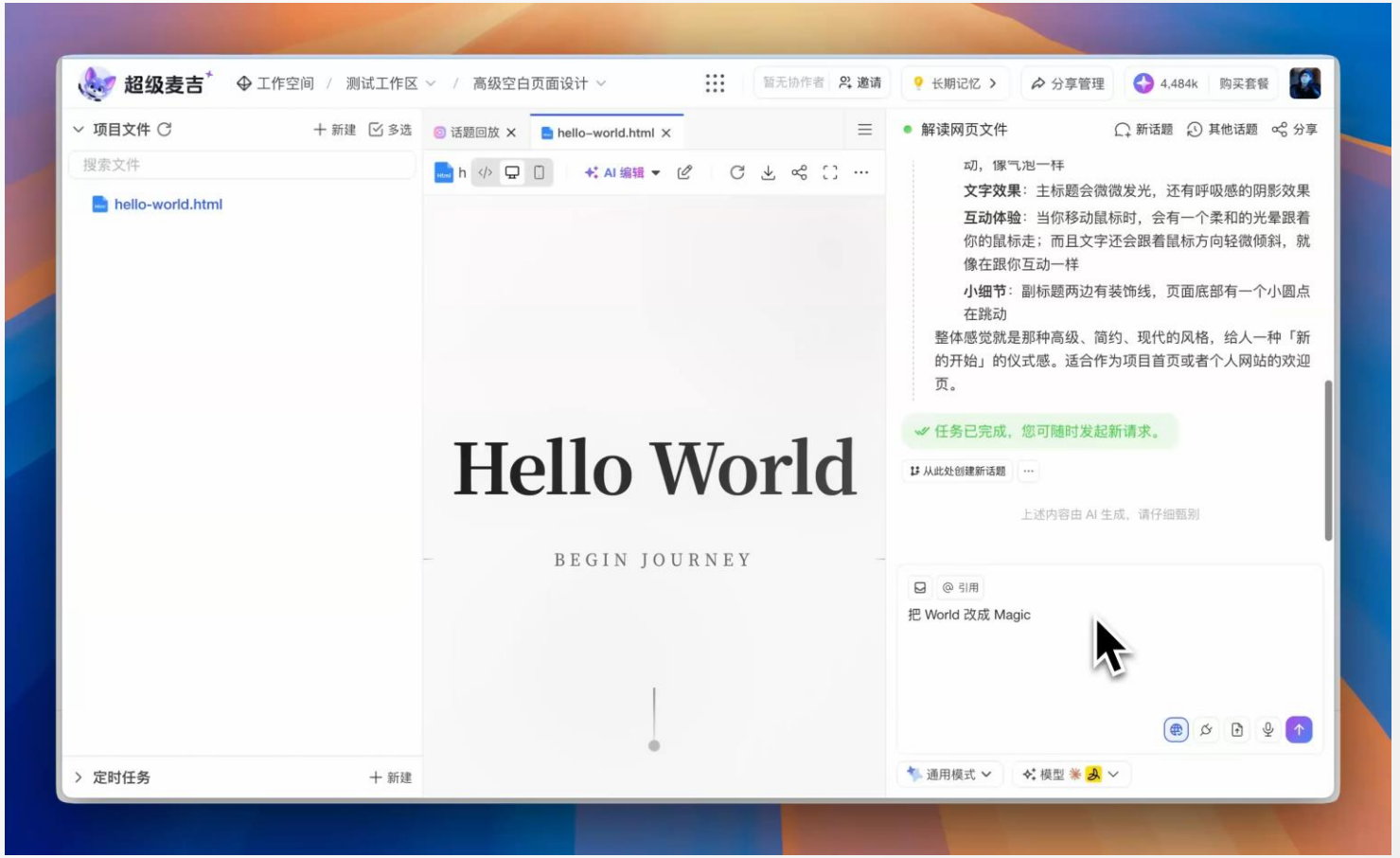
即时渲染更新

AI 修改网页内容后，前端立即接收通知并刷新显示，无需手动操作

实时协作体验

多人协作场景下，所有成员同步看到最新变更，真正实现所见即所得

从修改到显示，全链路毫秒级响应，打造极致用户体验



可多人协作的 AI Agent

Team Vibe Working

让整个团队都能享受「开挂」的协作体验
多人共享同一项目，AI 助力每个成员
高手带飞萌新，经验无缝传承

团队模块化协作

商品、运营、设计、数据团队在同一项目中并行工作，各负责自己的模块，最终聚合成完整方案

AI 高手带飞 AI 萌新

高手直接进入项目救场，实时指导操作演示，神级提示词和套路可以共享回放，团队 AI 水平快速提升

零构建渲染：面向小白，如何所见即所得？

典型场景：调研报告聚合站

用户需求

非技术人员需要进行联网数据检索与收集，以结构化、可视化的方式整理归纳并沉淀

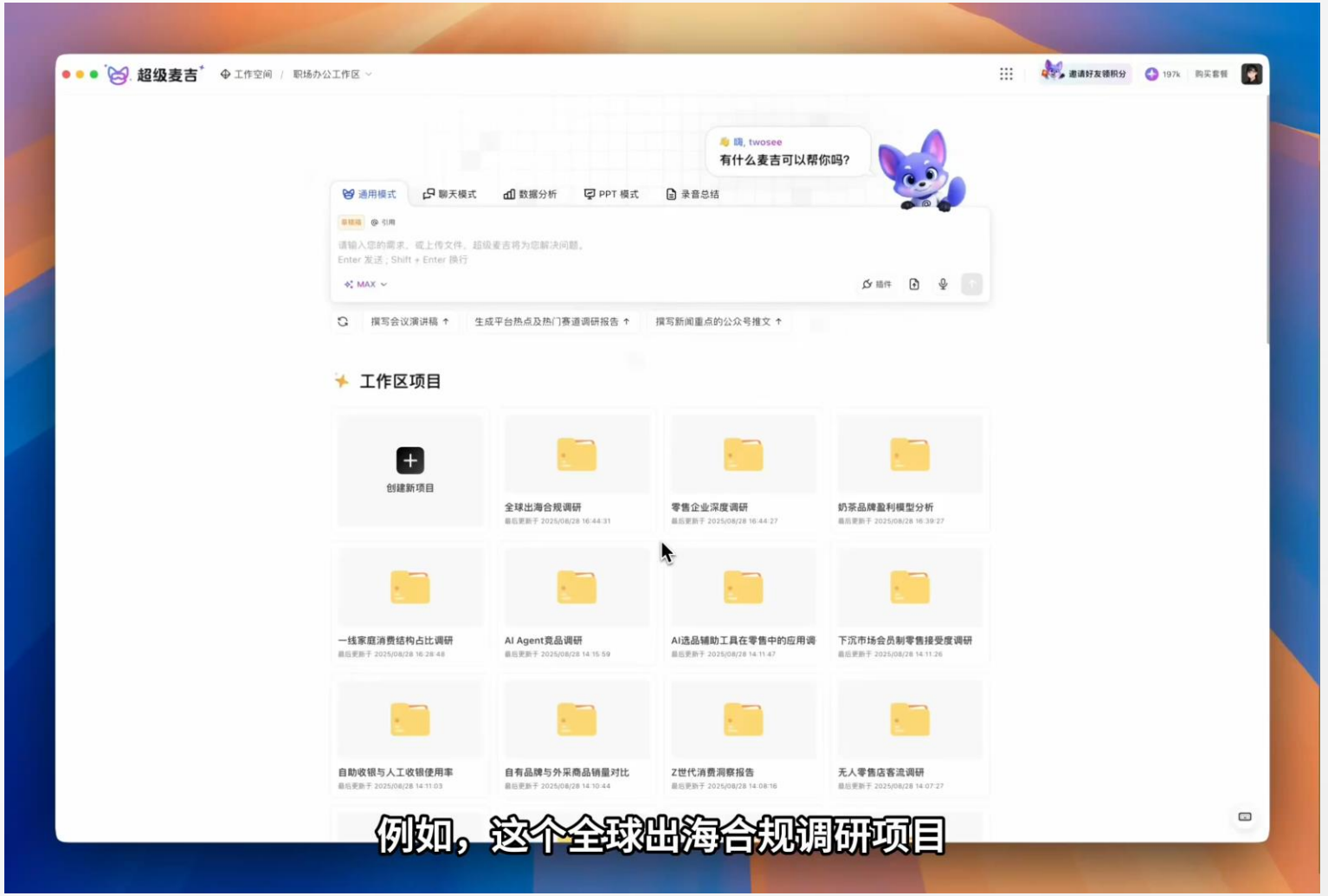
常规方案的问题

构建部署流程链路长，服务器成本高，迭代过程中交互体验割裂

零构建渲染技术

直接从对象存储渲染，Hook 文件引用与路由机制，零部署实现生产级网站体验

解决职场 99% 的调研报告聚合沉淀与展示需求



零构建渲染：技术原理

iFrame 沙箱渲染

深度隔离渲染环境，Mock 浏览器存储与网络 API，确保页面在受控环境中安全运行

资源加载 Hook 拦截

拦截超链接点击与资源发现，统一调度路径解析逻辑，将工作区路径转换为预签名 URL

网络请求 Hook 拦截

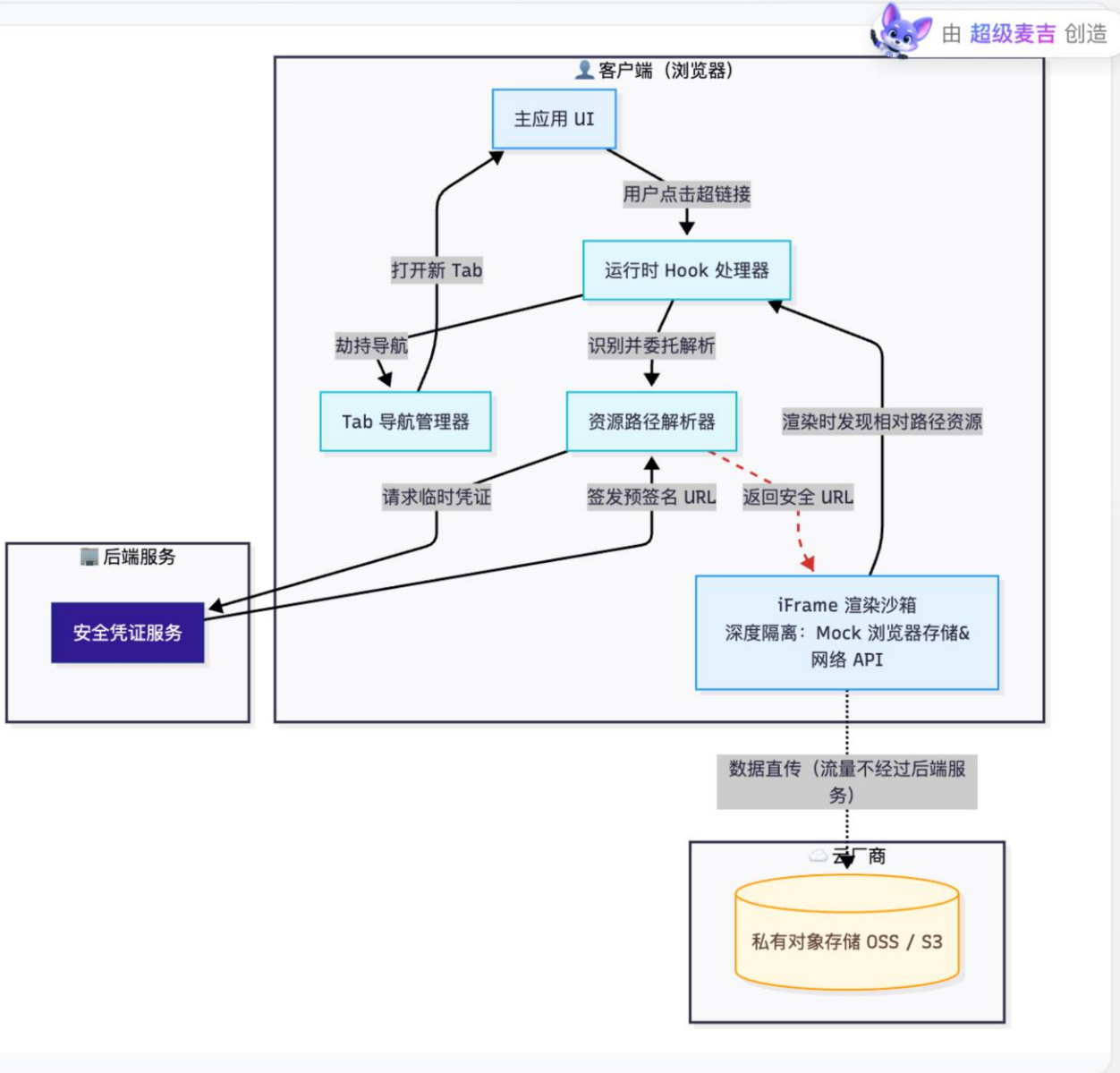
对 XHR 和 Fetch 进行 Hook 拦截，支持以相对路径访问项目中的文件内容

页面操作 Hook 拦截

捕获用户在 iFrame 中的所有交互操作，实现跨域通信与状态同步，保证体验流畅

存储隔离 Hook 机制

不同项目间 LocalStorage 完全隔离，不污染主域，不互相影响，支持持久化保存配置项



人机协同的实践

超级麦吉案例

指哪改哪

元素级 AI 编辑

- 鼠标选中，指哪改哪
- AI 精准理解，内置提示词一步到位

技术实现

在线编辑：通过动态注入编辑控件脚本，为目标内容页赋予可编辑能力。再利用事件委托的方式对常用标签进行鼠标事件的处理，获取 DOM 树路径、目标节点自定义属性、样式表等元数据在目标节点挂载编辑菜单，使文本可编辑。

AI 编辑：向大模型传递元素的文本、 XPath 或节点截图信息，并使用最佳提示词。

2025 年初

日均 10 亿 token

是 AI SaaS 独角兽的门槛

2025 年中 ↓

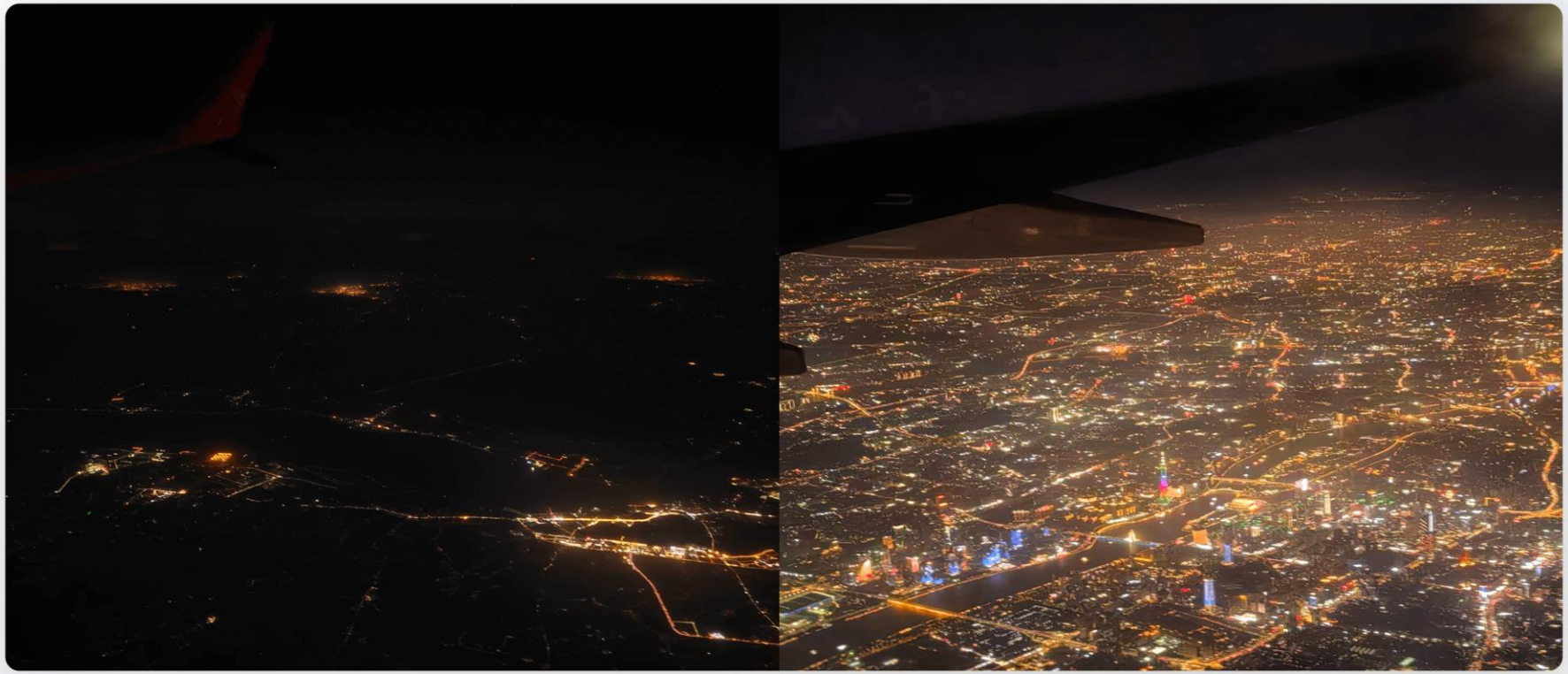
这可能只是一家零售、制造或电商企业

进入 Agent 时代的入场券

鸟瞰城市夜景，判断发达程度

光污染越严重 = 用电量越大 = 城市越发达

从夜空俯瞰城市，灯光越亮的区域代表用电量越大，也意味着该区域越发达。光污染虽然是环境问题，但从另一个角度看，它恰恰是城市繁荣的直观体现。



可视化「大模型是新时代的水和电」

论如何衡量一家企业的先进程度

> **算力将像水电一样成为基础资源，Token 消耗量将成为团队先进程度的指标——Token 消耗量越大的团队越先进。如果非要说 AI 用得最多未必是最先进的，那么仍然无法否认 AI 用得少一定是最落后的。**

Token 是大模型的时间，上下文是胜负的关键

重新理解数字员工的价值模型

Token 即工作时间

人消耗时间来思考，而大模型消耗 token 来思考，思考越长，Token 越多，质量越高。购买 Token 本质是在为数字员工的工时买单。

上下文工程决定竞争力

AI 产品的竞争，实质是**上下文工程能力的竞争**——谁的上下文工程做得更好，谁的数字员工就更有竞争力，更容易获得雇主青睐。

模型	极限分数	中位分数	中位差距	测试成本(元)	平均耗时(秒)
GPT-5(high)	88.25	84.10	4.70%	¥41.18	308
Claude Sonnet 4.5(Think)	81.87	69.78	14.77%	¥36.90	155
Grok4	79.07	74.79	5.42%	¥41.99	207
Grok4 Fast(Think)	73.81	67.59	8.43%	¥1.55	67
Claude Sonnet 4(Think)	63.45	48.04	24.29%	¥29.27	143
Claude Sonnet 4	37.22	29.18	21.60%	¥2.92	17
Claude Sonnet 4.5	31.57	24.88	21.20%	¥2.99	16
Grok4 Fast	29.70	23.29	21.56%	¥0.12	8

不同大模型的性能、成本与耗时对比：思考模式显著胜出，但成本也高出数倍 (图片来源：@大模型观测员)

AI 行业曾流传一句话：所有的 Agent 产品离开了 Claude 就什么也不是

如果以 Claude Sonnet 的价格计算单日 10 亿 token

Claude Sonnet 输入价格：$3 / 百万 token

10 亿 token = 1000 × 百万 token

单日成本 = $3 × 1000 = $3,000

按汇率 1:7.3 计算

单日成本

¥21,900 / 天

年度成本

¥799 万 / 年

Prompt Caching 降低成本

什么是 Prompt Caching?

缓存**推理结果**而非文本，通过**前缀匹配**复用已计算的 KV 张量

注意，Prompt Caching 不是 Response Caching，依赖模型提供方的工程化支持

如何降低成本?

将**静态内容**（系统提示词、工具定义）放在前面，**动态内容**（用户问题）放在后面，保持前缀完全一致即可复用缓存

> 在 **Agent 大循环模式**下，每次请求都可以**复用上一次的缓存**，成本优势更加显著

输入成本降低

90%

纯缓存命中

整体成本节省

75%

无缓存→有缓存
命中率 ⩾85%

成本降低

50%

被动→主动缓存

延迟降低

85%

速度提升 5-6 倍

AI 调研报告缓存复用案例

每一轮复用上一轮的前缀缓存，累积降低成本

1 搜索信息

系统提示词
你是专业的调研助手...

AI 调用工具
搜索「AI 行业趋势」

建立缓存

首次
2.5s

2 读取网页

系统提示词
你是专业的调研助手...

搜索结果
搜索「AI 行业趋势」

AI 调用工具
读取 5 个网页内容

节省 45%

无缓存	有缓存
5.8s	3.2s

3 生成报告

系统提示词
你是专业的调研助手...

搜索结果
搜索「AI 行业趋势」

网页内容
读取 5 个网页内容

AI 生成
基于资料生成报告

节省 51%

无缓存	有缓存
9.2s	4.5s

4 深入调研

系统提示词
你是专业的调研助手...

搜索结果
搜索「AI 行业趋势」

网页内容
读取 5 个网页内容

初版报告
基于资料生成报告

用户要求
补充竞品分析章节

节省 72%

无缓存	有缓存
6.5s	1.8s

Prompt Caching 最佳实践

理解缓存顺序，避免频繁失效

缓存前缀顺序

第一层 **Tools** → 第二层 **System** → 第三层 **Messages**

1 不要频繁变更工具列表

使用 Claude Skills、Claude Advanced Tool Use 等方式

⚠ 工具变更会导致整个缓存链失效

2 动态内容放在后面

当前时间 等动态信息不要放在 System 中

保持稳定可跨用户共享

3 写入缓存也要收费

写入费用是 输入的 1.25 倍

测算成本时需要考虑

4 评估缓存命中率

主动缓存下，80%~90% 是最佳范围

实践中的最佳参考范围

中国大模型的崛起与成本优势

中国一开源，国外就自研

$ Claude Sonnet

输入价格	$3 / 百万 token
单日 10 亿 token	¥21,900

年度成本

¥799 万

¥ DeepSeek V3.2 (Qwen3、GLM、Kimi 等模型折后亦处于同等水平)

输入价格	¥2 / 百万 token
缓存命中率	60%
模拟单任务费用	¥1.056/百万token
折合单日 10 亿 token	¥1056

年度成本

¥38.5 万

人人都在开发 AI 项目，
到底什么样的人才算是

?

真正的 AI 应用工程师？

什么是 Agent?

开发圈对 Agent 定义的共识

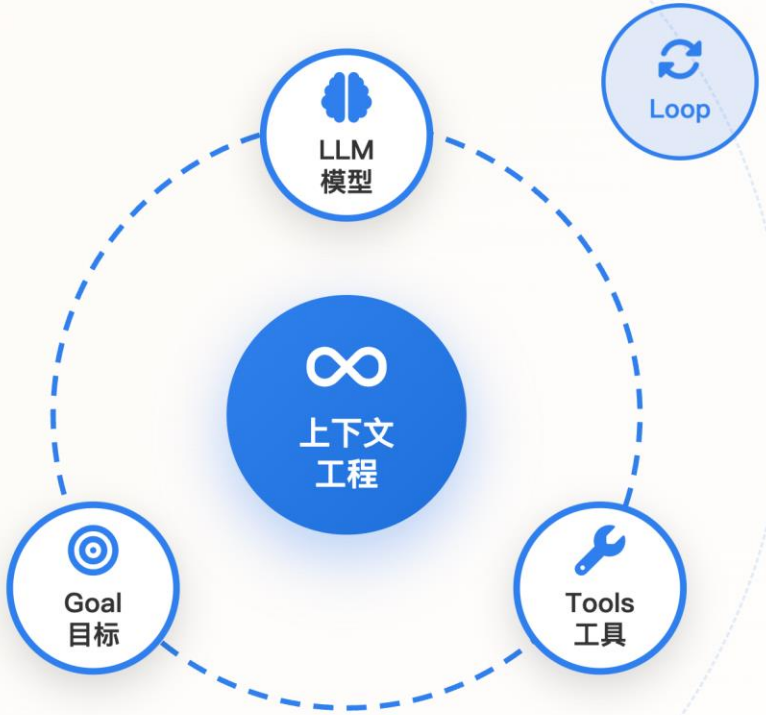
LLM runs tools in a loop to achieve a goal

模型循环调用工具以达成某个目标

模型
大语言模型的推理和决策能力

工具
可调用的外部功能和资源

上下文工程
隐藏在模型与工具之间的桥梁



三大主要工具

- 文件系统
- 命令行
- 浏览器

Agent 产品在拼什么?

多 · 快 · 好 · 省

多

场景足够丰富

覆盖工作链路的方方面面

从调研、分析、创作到协作、发布，打通全流程。场景不够多会形成孤岛效应，只能解决单点问题，用户需要在多个工具间来回切换

不多，就无法成为新的工作方式

快

任务执行和响应要快

支撑持续使用

及时反馈让思维保持连贯，用户才愿意持续交互和迭代。速度慢会打断工作节奏，用户无法建立信任感

不快，就无法养成连续使用习惯

好

要有好的质量

质量是生存分界线

基模能力正在趋同，但上下文工程可以让产出更精准、更懂用户。质量不好会被竞品替代，AI 产出不合格则一切不成立

不好，就无法成为用户的首选

省

成本要低

支撑大量使用

通过缓存、模型路由或其它上下文工程手段降低成本，只有成本低才能让用户毫无顾忌地大量使用

不省，就无法被大量使用

以上所有的点都是通过 上下文工程 可以去做提升的，也都是让 AI 产品成为一种 工作方式 的关键

上下文工程
优化成本

降本的另一种手段

减少**任务步数**（Turn）
= 减少调用大模型的次数
= 大幅降低成本

成本优化三件套：
Prompt Caching
+ 国产大模型
+ 上下文工程

行号自动去除 标点符号自动修复 工具参数预处理 工具参数预处理
字符串化JSON修复 Unicode转义解码 多行键值对解析
多维度错误诊断 相似字符串建议 过度转义检测 友好验证错误
分级建议机制 人类友好的错误提示 分级警告提示
文件时间戳验证 文件时间戳验证
原子性批量操作 增量上下文机制 编辑冲突检测
响应结构验证 会话恢复机制
指数退避重试 文件变更感知 智能重试判断
工具调用中断处理 模糊路径匹配
文件截断提示 超大多文件警告 Token限制保护 空内容处理
语法检查 并行工具调用 动态模型切换 流式工具调用收集
Python计算引导 网页内容压缩 垃圾信息净化 子Agent并行处理

增量上下文：让 AI 实时看见一切

实时增量上下文

录音实时转文字 + 笔记实时同步
大模型能够持续看见最新内容
像真人助手一样旁听会议

1 实时转录流

录音内容实时转成文字，AI 持续接收最新的对话内容，无需大模型调用工具读取

2 笔记同步流

用户记录的笔记、补充信息实时同步给 AI，让 AI 理解完整上下文和背景

3 举一反三：更多增量上下文

文件变更感知、新增删除感知、当前时间感知等，都可以通过增量上下文实时传递给大模型，无需工具调用

面试官提问实时转录

Agent 实时基于上下文帮你回答面试官问题

工程师需要思考：你的工作是不是

致力于这个目标?

⌄⌄

用**同样的模型**

更少 Token + 更短时间 => 更好结果

这就是衡量新时代 AI 工程师的标准

还是

只是在做传统工程内容：

接口开发、修 Bug、性能优化、交互调整、系统对接

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

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

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

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

4月 — 6月 — 8月 — 10月 — 12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

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

InfoQ 极客传媒

谢谢聆听

加我微信 · 关注超级麦吉