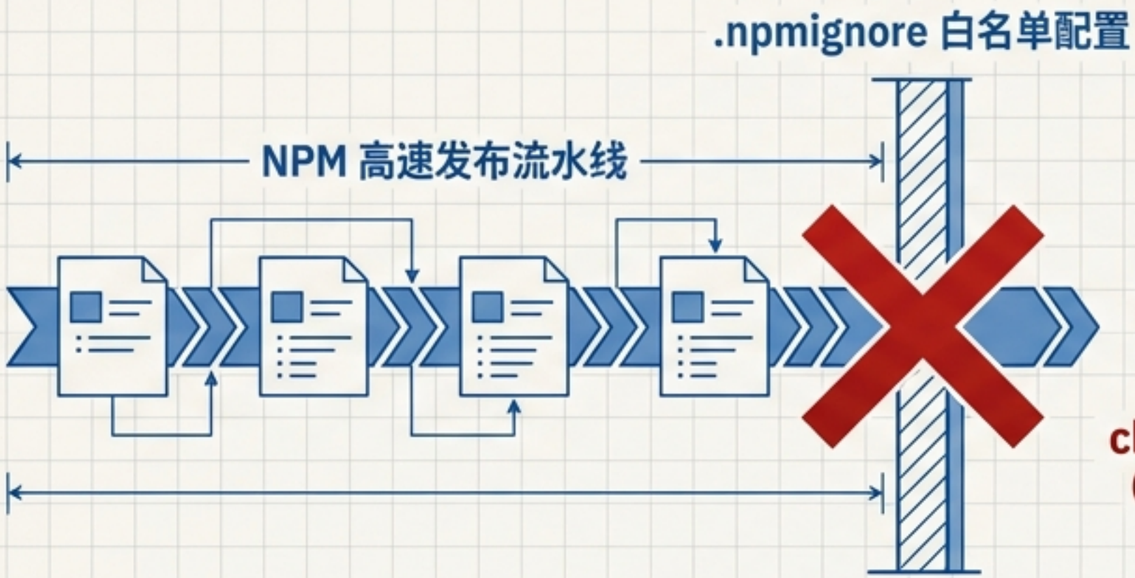
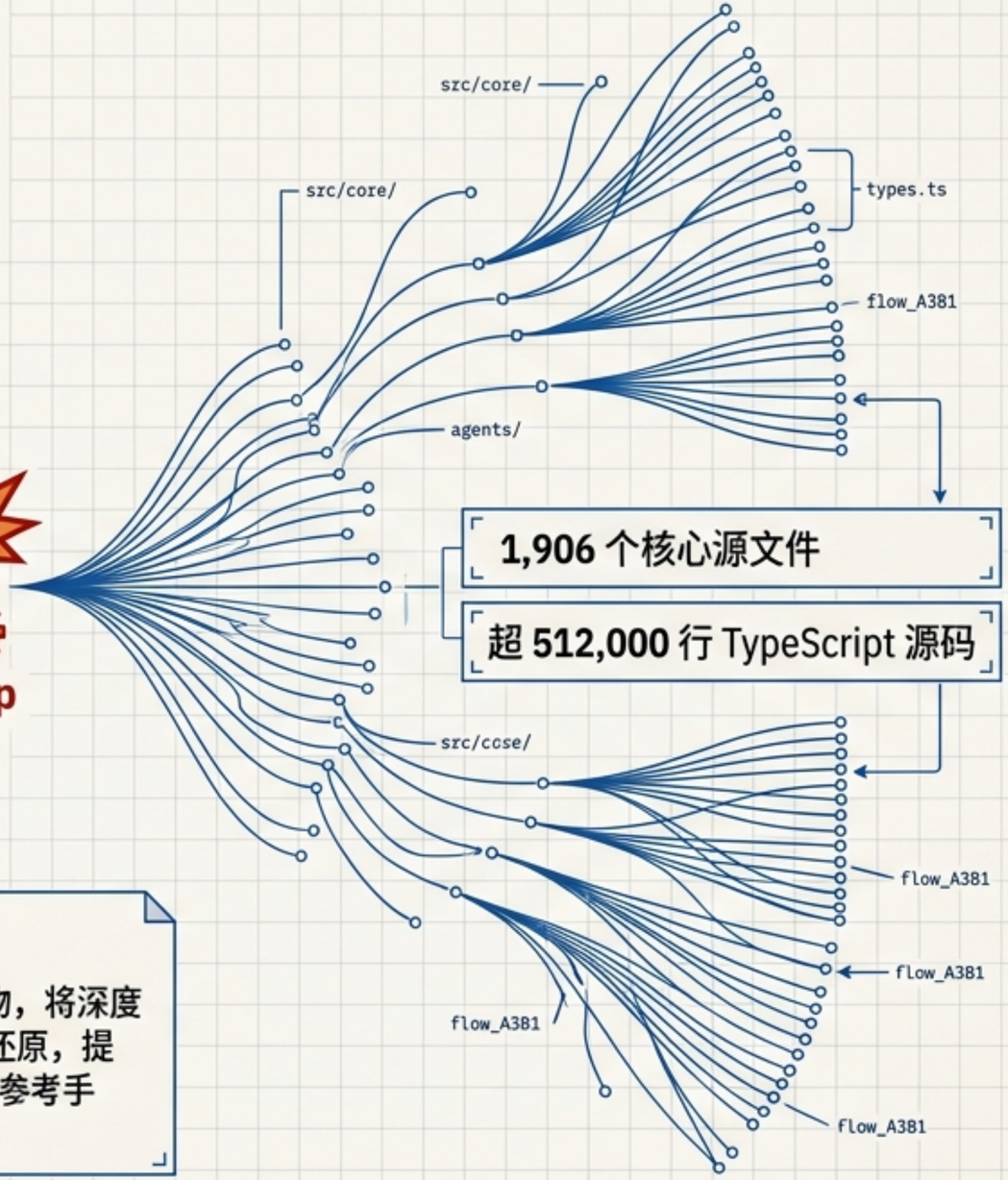
2026年3月31日：史上最严重的 AI 架构泄露



cli.js.map
(59.8 MB)



1,906 个核心源文件

超 512,000 行 TypeScript 源码

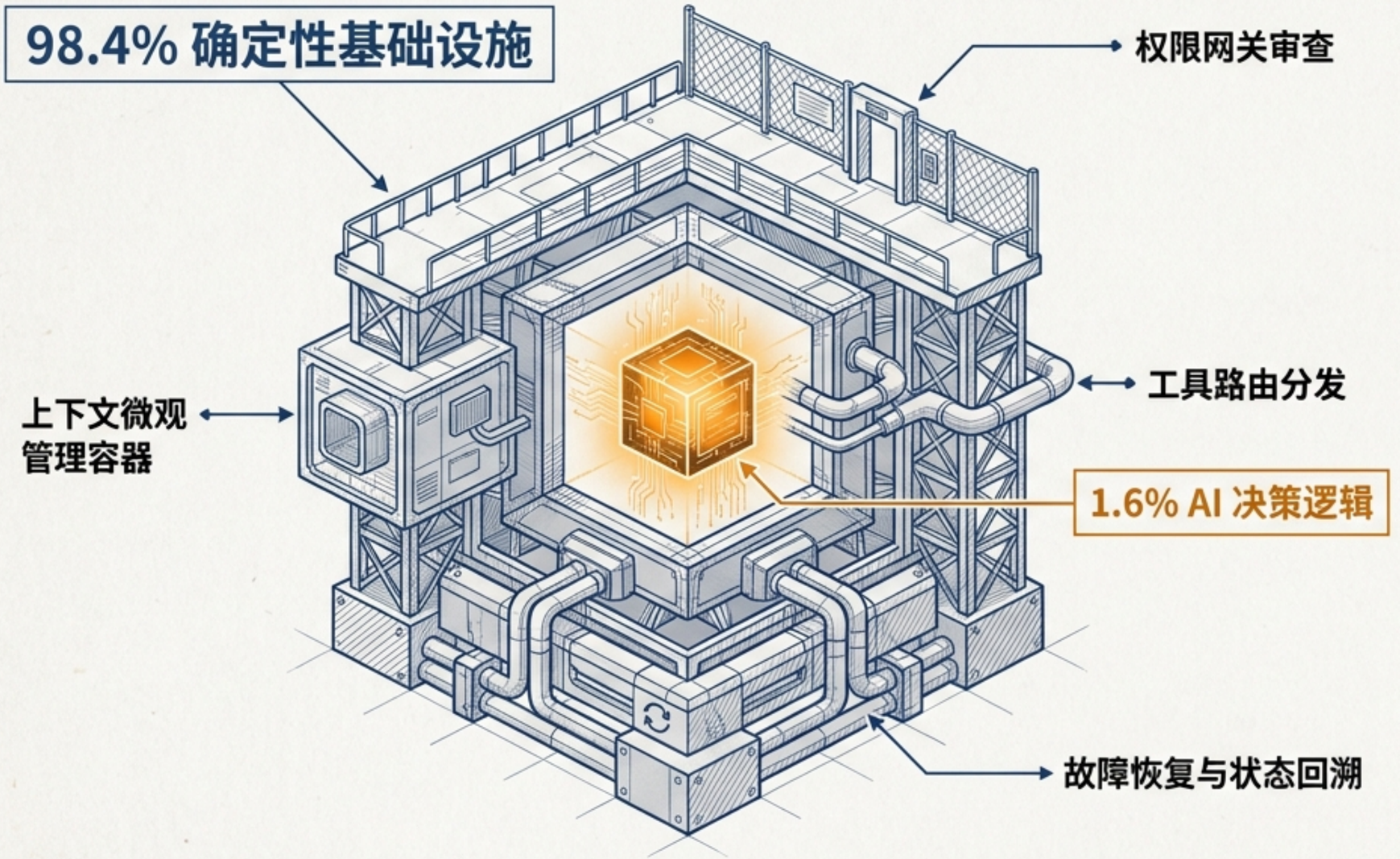
核心事件定性

并非国家级攻击或零日漏洞。现代高速软件发布流水线中的经典配置纪律缺失，导致 Anthropic 最高机密的终端智能代理源码彻底曝光。

技术暴露深度

未被过滤的 Source Map 调试产物，将深度混淆前的核心工程代码完全映射还原，提供了一份显微镜级的 AI 代理架构参考手册。

工程范式转移：从对话补全到重型自主代理



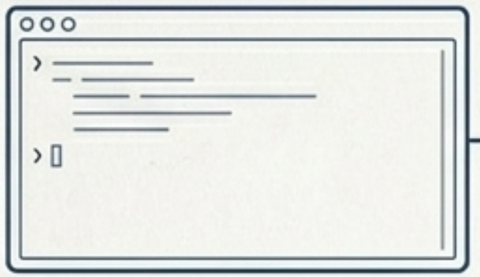
范式彻底重构

AI 编程助手已发生根本性进化。大型语言模型（LLM）彻底退居幕后，仅仅作为单纯的思考引擎，而系统主体已变为由海量确定性代码构筑的执行躯干。

代码占比解构

源码统计证实，仅有不到 2% 的代码直接对接 AI 模型 API。绝大部分工程资源被投入到约束模型行为、防止幻觉以及保障执行安全的重型框架中。

核心执行引擎：ReAct 与单点执行管线



CLI 终端控制台

后台无头进程

IDE 集成插件

queryLoop()

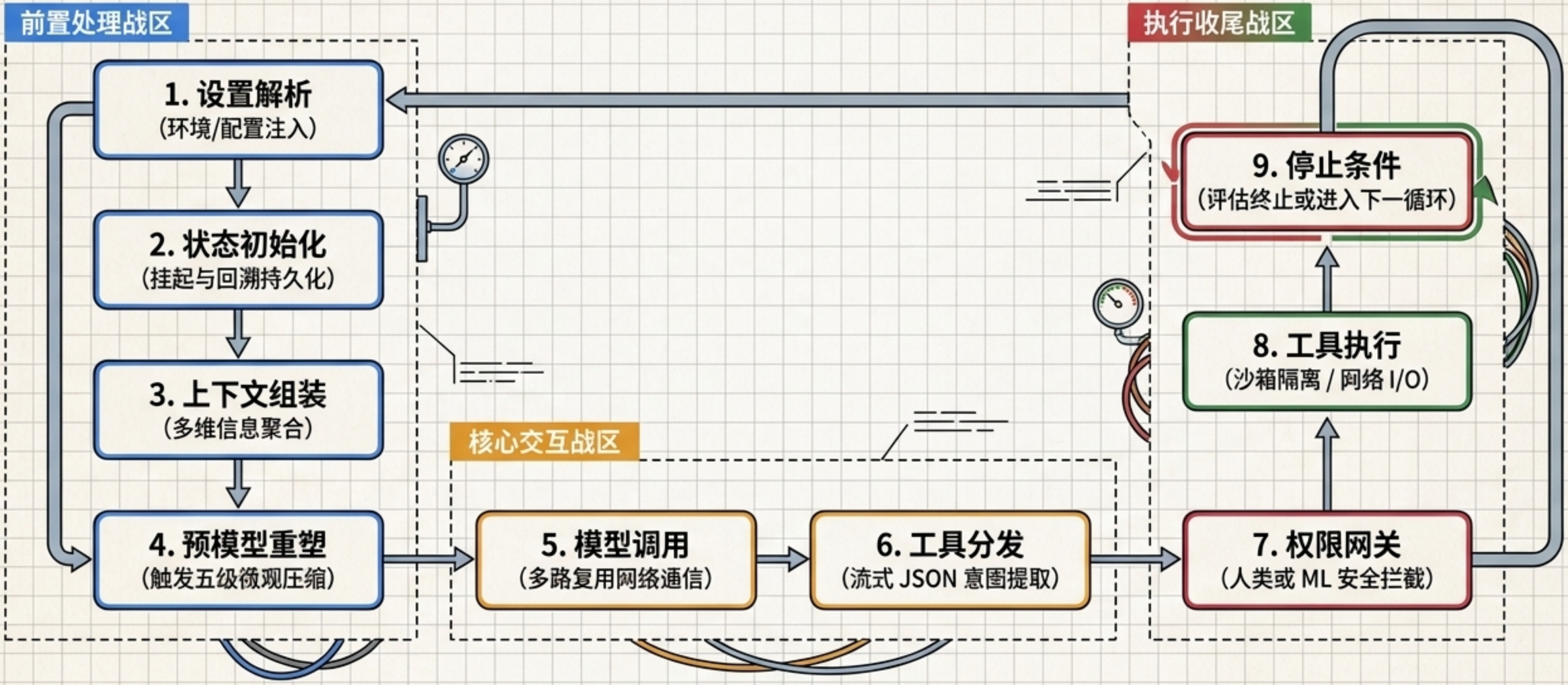
`src/query.ts | 1,729 行长篇异步生成器`

统一架构的工程胜利

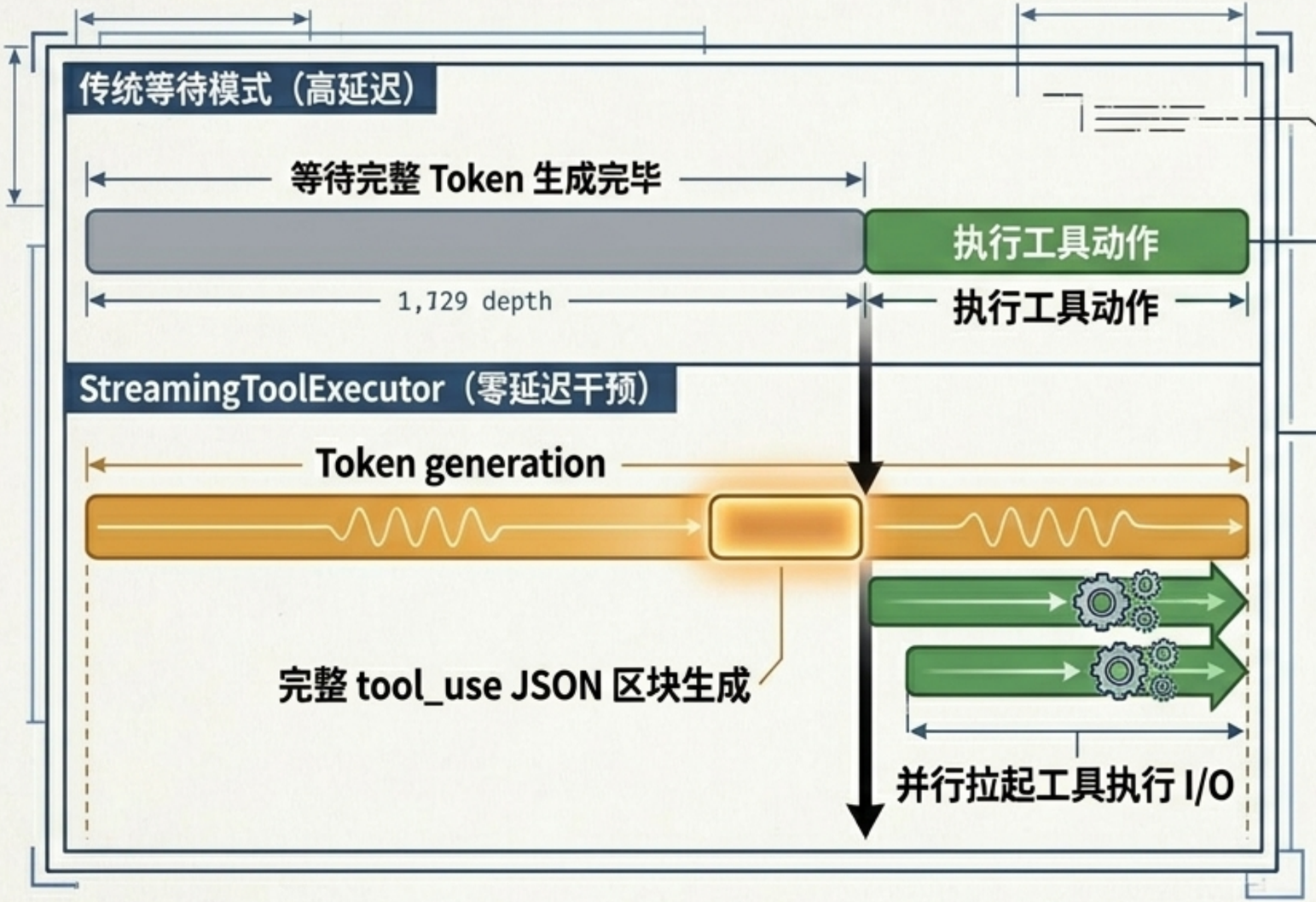
摒弃为不同界面开发独立分支的传统做法。所有的环境观测、工具调用和认知推理逻辑，完全硬连线并共享同一条主干生成成器，确保全端行为的绝对一致性。

智能代理循环的九大精密管线

在动态、混沌的代码环境中，严苛的九步工业流水线是维持 AI 行为确定性的唯一枷锁。

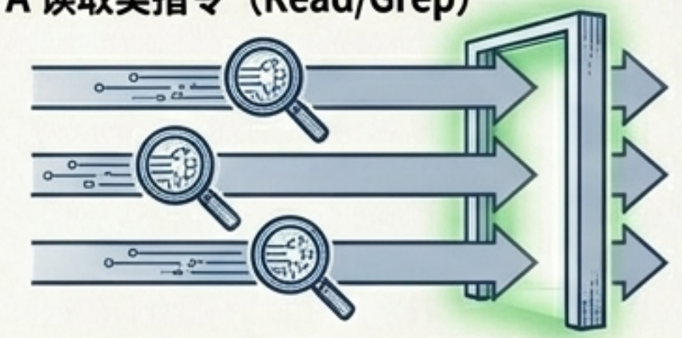


突破长尾延迟：流式执行器与并发锁模型



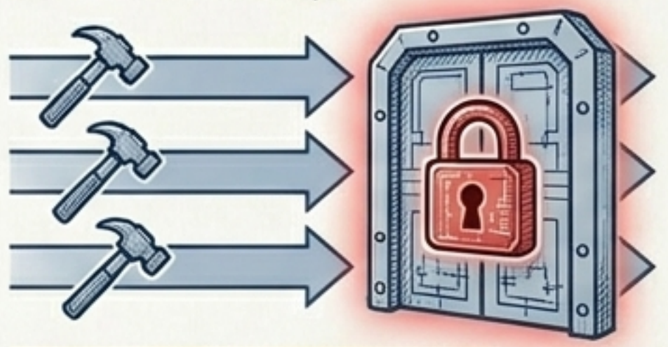
读写并发管控（RWLock）

A 读取类指令（Read/Grep）



无锁设计，瞬间极速并发通过。

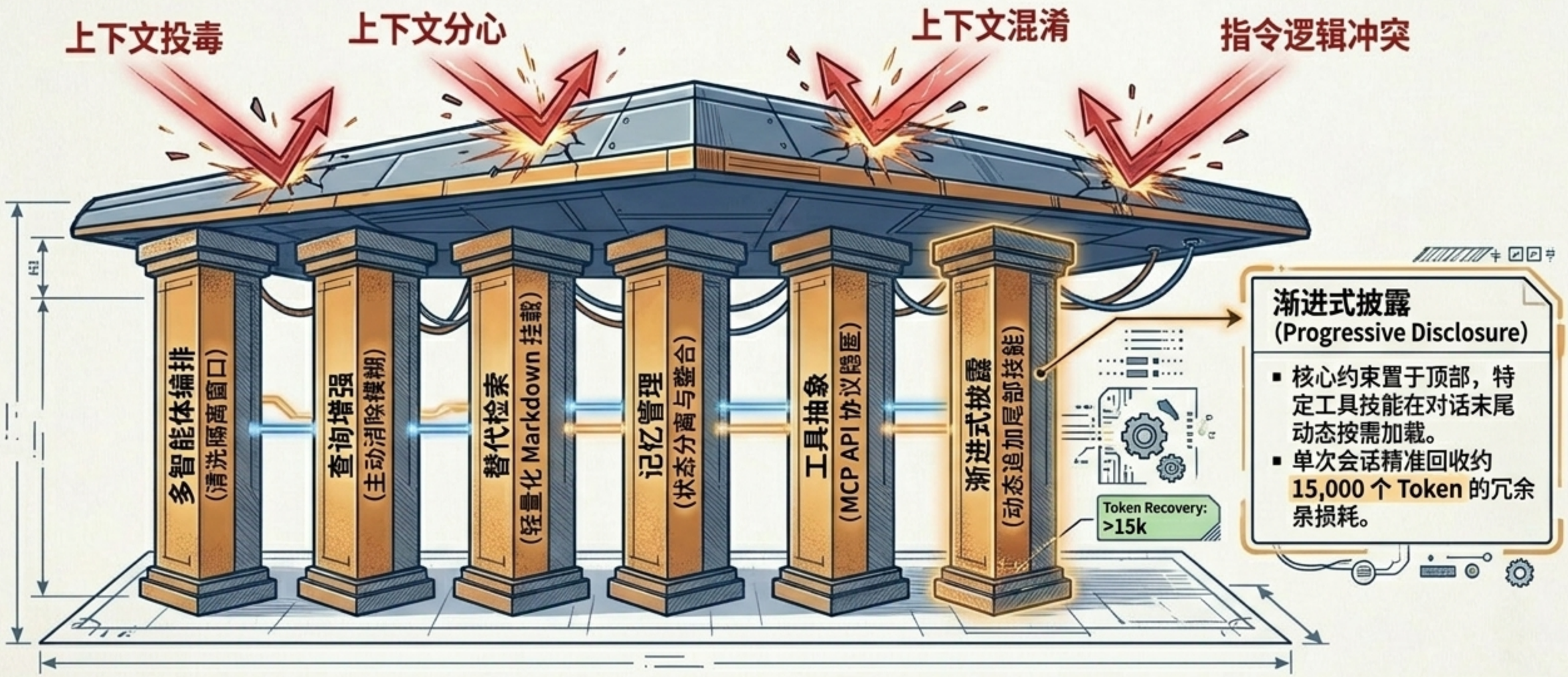
B 破坏性动作（Edit/Bash）



强制获取全局独占锁排队串行，确保状态树绝对安全。

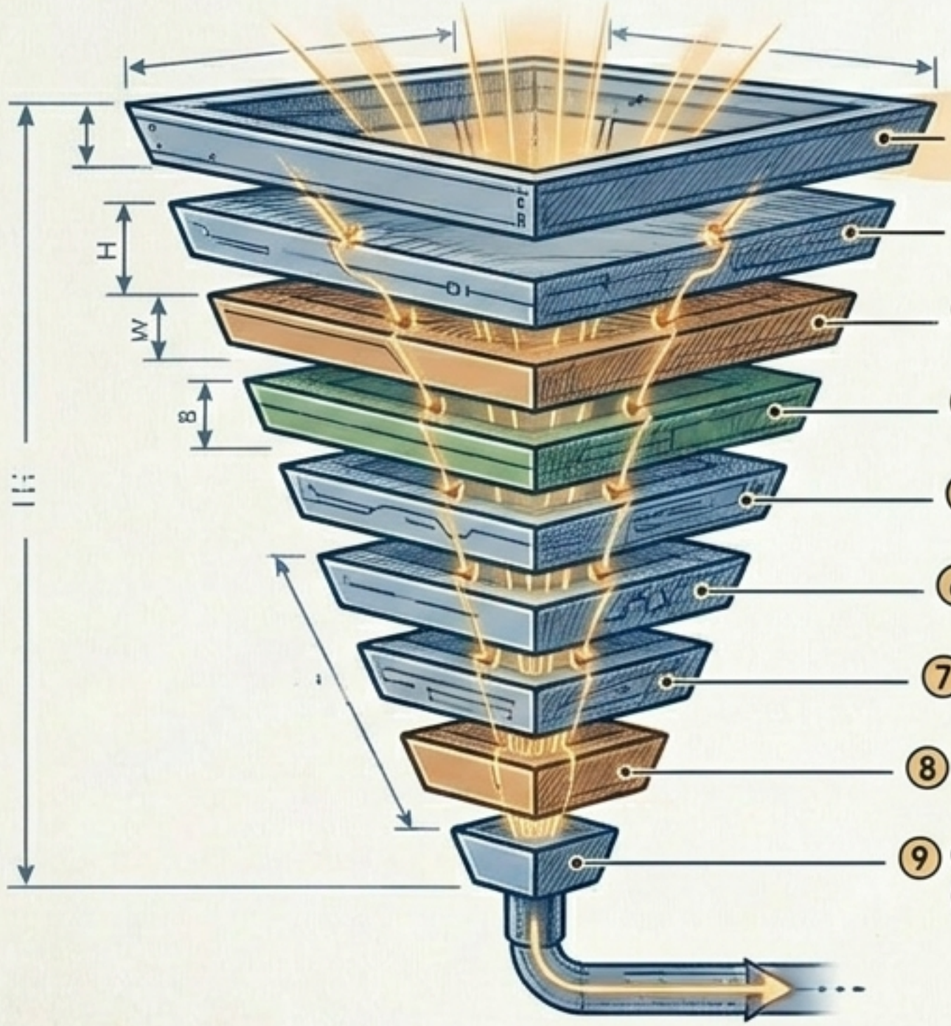
上下文工程的六大核心支柱

应对 1M Token 物理极限，实施极其苛刻的记忆微观控制法则。



动态提示词工厂与九维信息拼接

系统提示词并非静态常量，而是依据网络请求前的绝对现实进行动态降维组装。



① System Prompt
(系统动态生成基调)

② Environment Info
(架构/Shell/Git 状态)

③ CLAUDE.md
(四级覆盖率概率规则矩阵)

④ Path-scoped Rules
(特定文件挂载子规则)

⑤ Auto-memory
(提取的高关联度记忆池)

⑥ Tool Metadata
(可用工具池 Schema 描述)

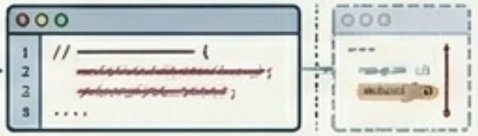
⑦ Conversation History
(JSONL 对话明文副本)

⑧ Tool Results
(携带运行时 ID 的交互反馈)

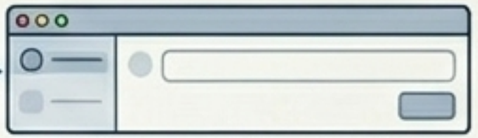
⑨ Compact Summaries
(微观压缩后的留存摘要)

身份隔离与条件分支引擎

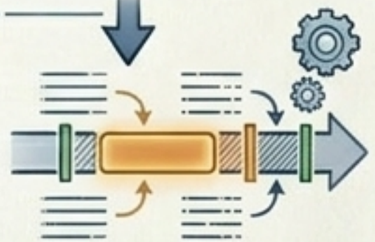
A: 内部员工 (user_type_ant)
↳ 强制关闭代码注释，激活绝对字数硬锚点约束。



B: 外部用户 (user_type_external)
↳ 追求效率优先与结果导向体验。

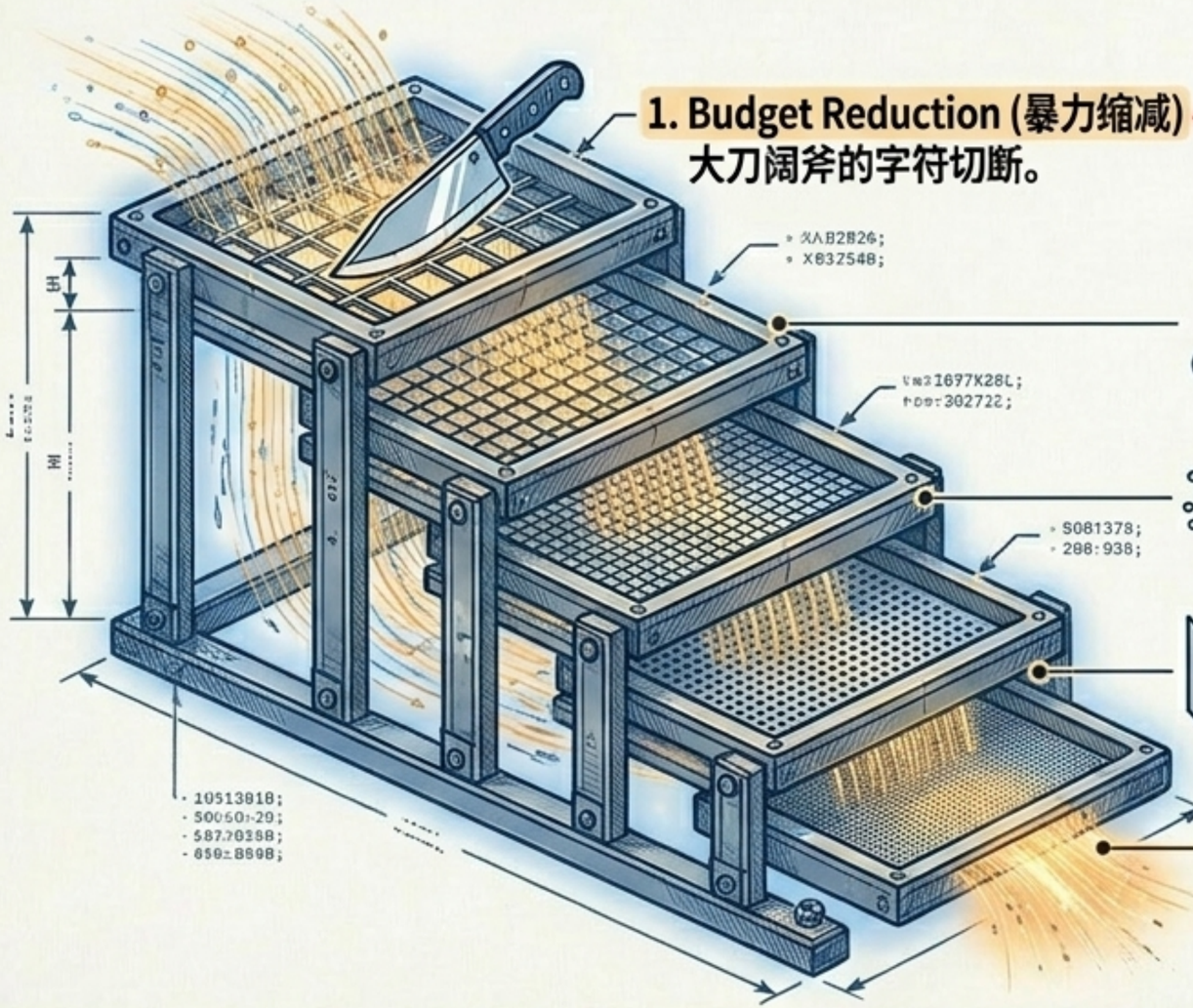


动态插入 Cache Boundaries 以分离可缓存内容，大幅降低重复调用成本。



防溢出机制：五级微观压缩管线（Compaction）

算力成本阶梯过滤：为了防止模型崩溃，对历史输入执行极度残酷的五级信息压实。



1. Budget Reduction (暴力缩减)
大刀阔斧的字符切断。

致命豁免缺陷： 读取类工具配置 maxResultSizeChars: Infinity，成为后续投毒温床。

2. Snip (历史剪裁)
基于 FIFO 队列钝刀修剪，丢弃边缘历史记录。

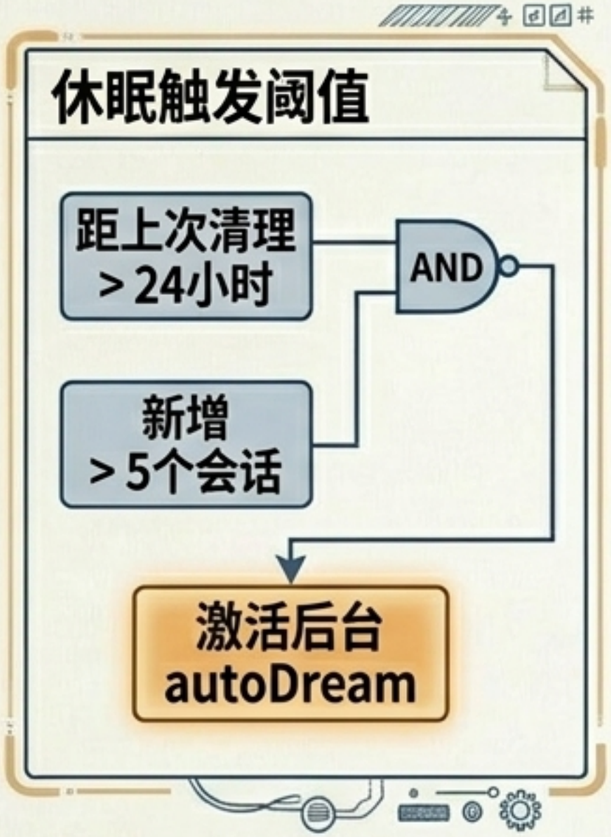
3. Microcompact (感知缓存微压缩)
精准识别并剔除无效的试错路径遗留代码。

4. Context Collapse (逻辑折叠)
高成本重塑，将发散的多步尝试在虚拟层面上折叠为单线状态流。

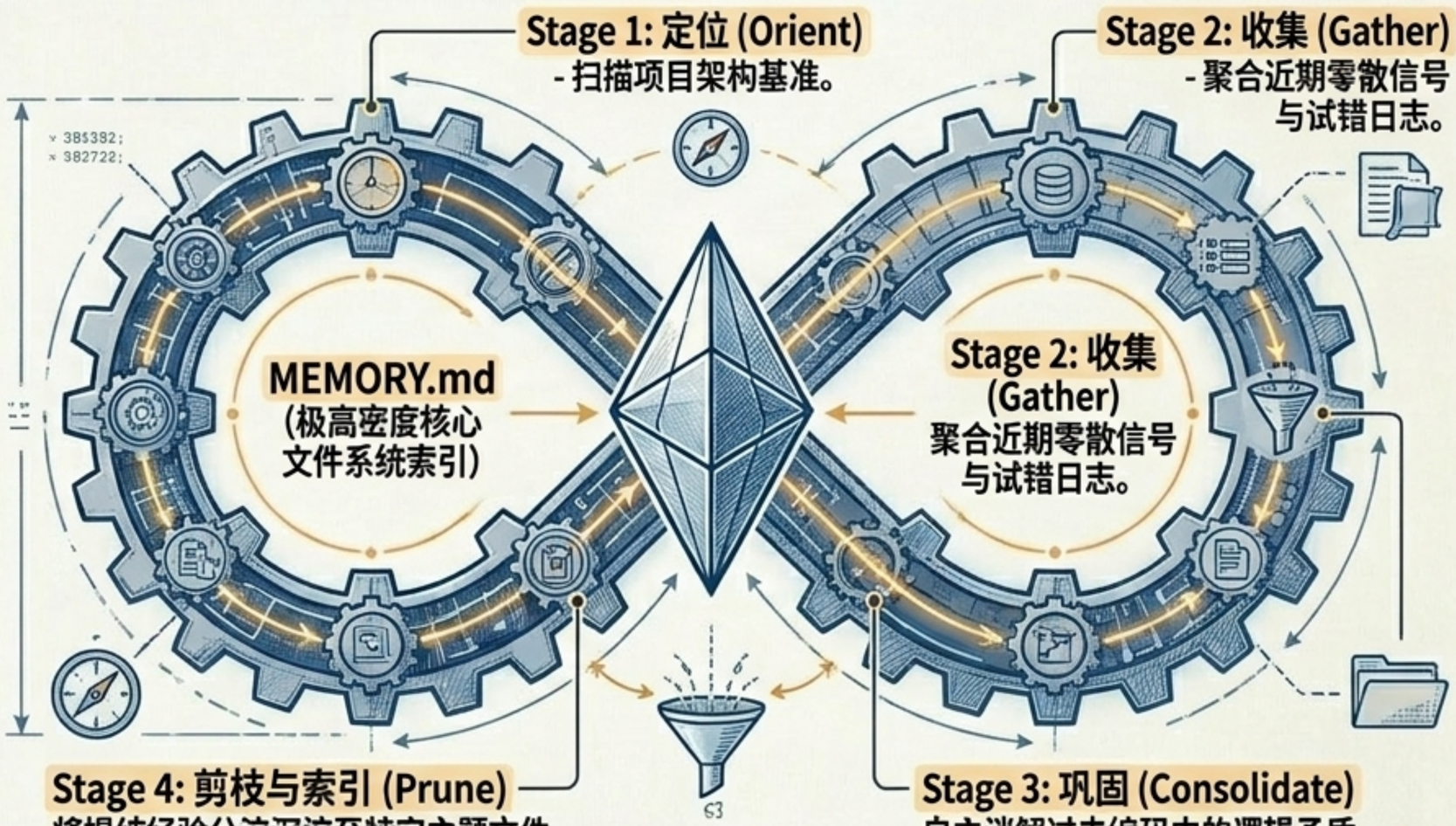
5. Auto-Compact (自动极致压缩)
终极昂贵手段：唤醒 LLM 通读全文，剥离工具结果，生成极高浓度摘要强制替换原始数组。

突破会话边界：持久化记忆与 autoDream 服务

摈弃沉重的外挂向量库，使 AI 能够像人类睡眠一样在后台自主巩固高密度记忆。



⚠ 支持环境变量强制熔断，防止CI/CD流水线中积累脏数据。

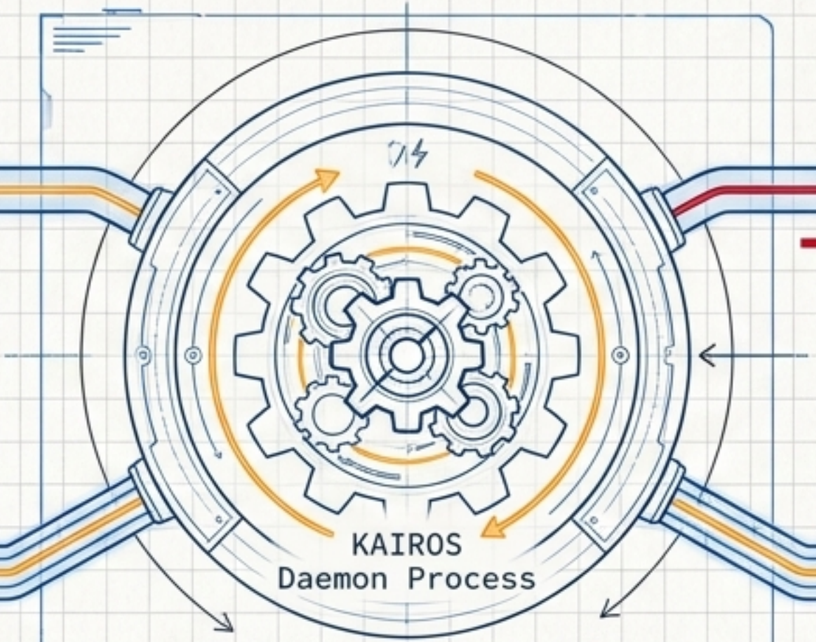


永驻后台的 KAIROS 自主守护系统

源码揭示未发布的顶级自主模式：摆脱人类干预的隐秘数字幽灵进程。

系统级触发器

无需人类输入。依赖底层操作系统“滴答（Tick）”信号，AI 即使在笔记本合上后依然主动审查队列。



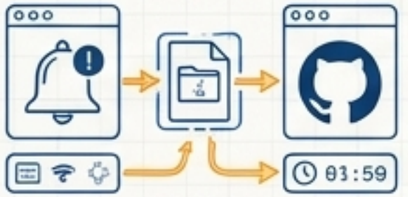
极严苛节流阀

为防止失控吞噬宿主机算力，源码硬编码了绝对红线：仅限“15秒阻塞预算（Blocking budget）”。

系统级独占武器库

KAIROS 拥有常规终端未赋予的三大特权工具：

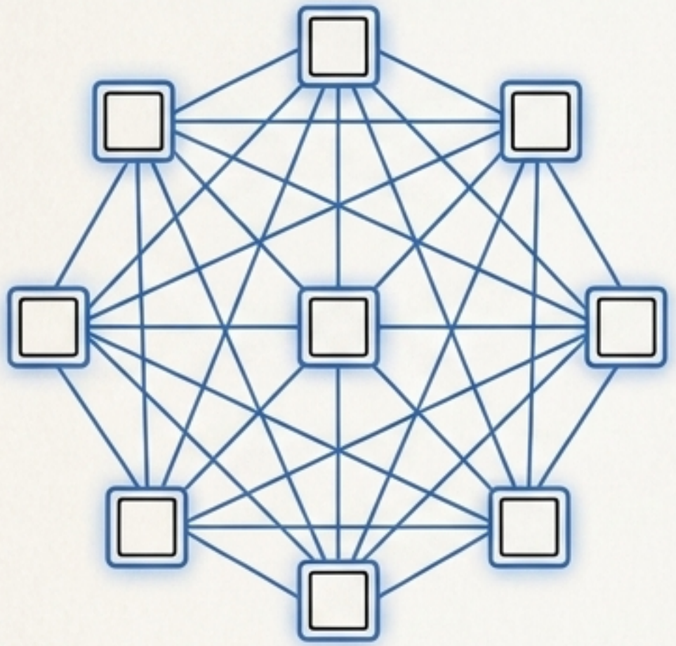
1. 系统级推送通知
2. 跨进程文件投递
3. 自动化全天候 PR（Pull Request）监控



多智能体协同：从扁平群组到等级协调者

应对规模化挑战：揭开两种截然不同的架构协作范式。

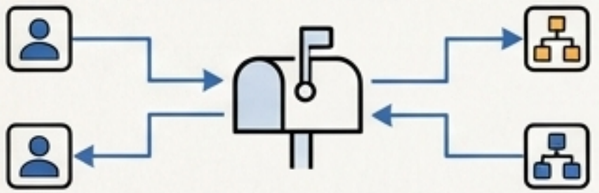
Agent Teams（扁平化群组）



去中心化的对等节点，通过底部邮箱机制互发消息协作。

利用 `AsyncLocalStorage` 实现严格的上下文隔离。

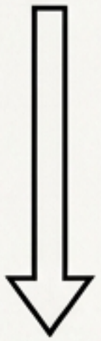
在终端 UI 中通过颜色标签直观呈现。

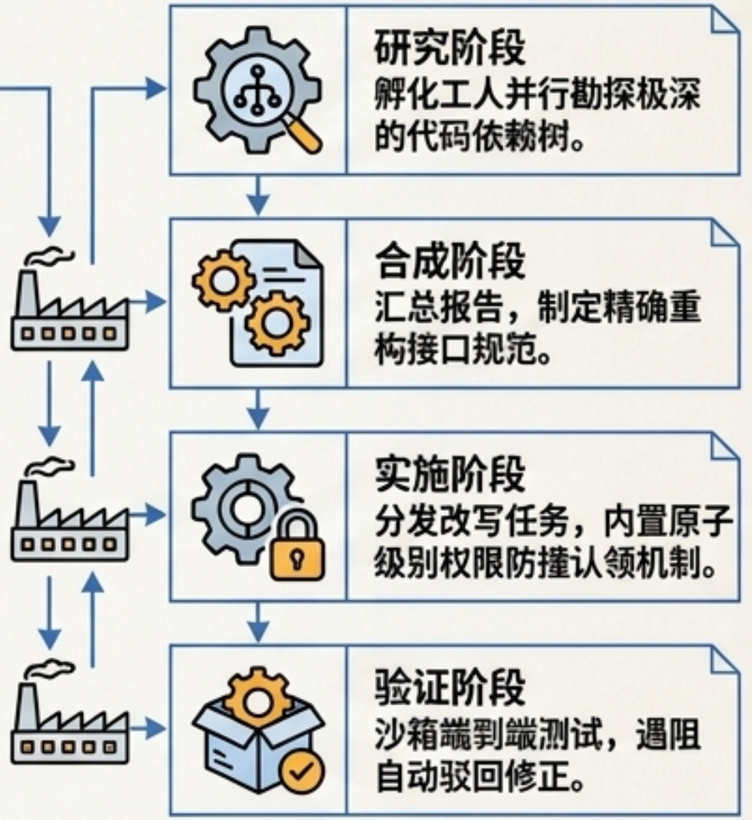


Coordinator Mode（等级制企业模式）

协调者

（拥有唯一决策权）





协调者实例不再亲自编写代码，彻底转变为纯粹的 AI 项目经理。

ULTRAPLAN：算力卸载与云端长效推演

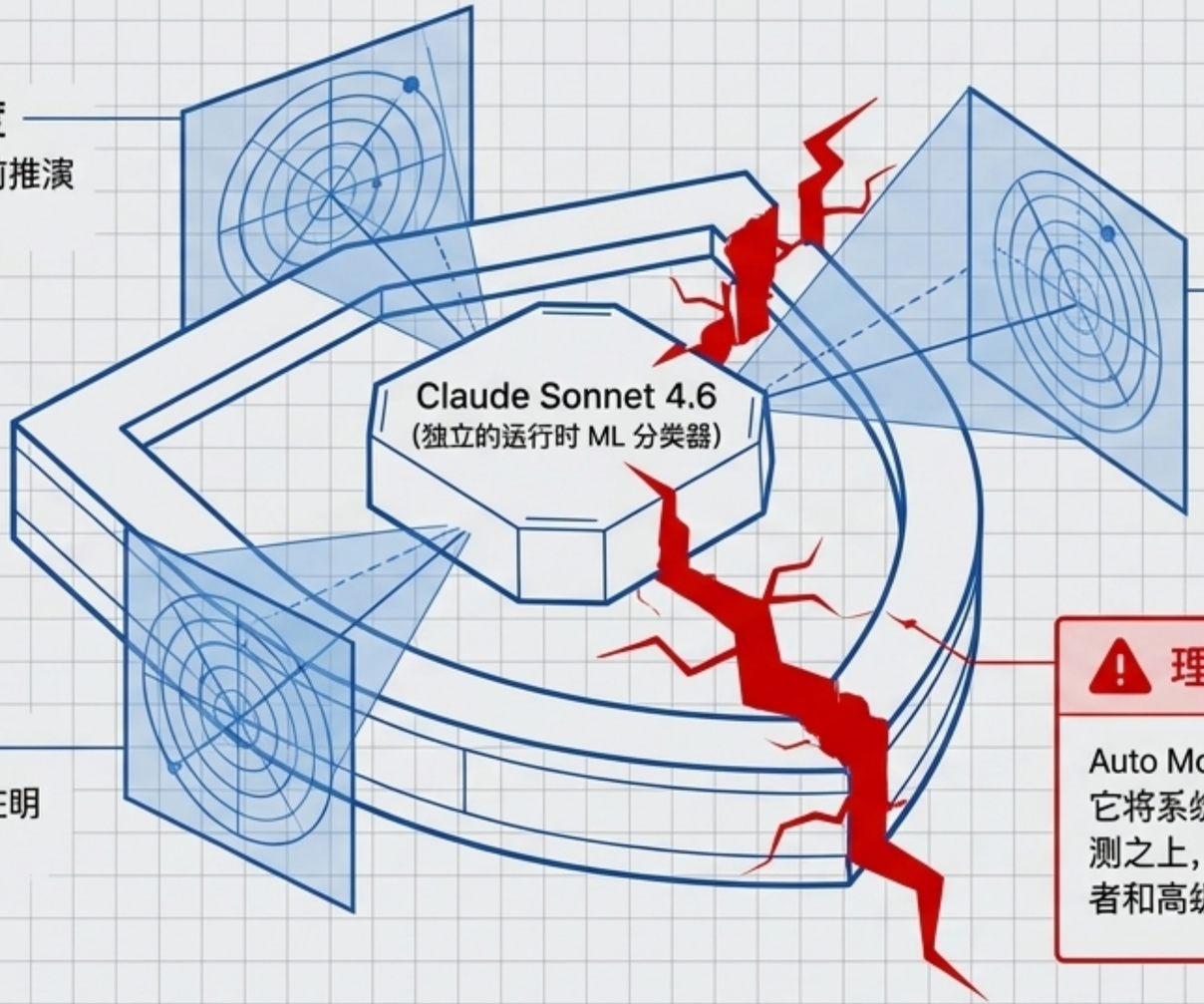
突破本地硬件极值，利用空间与时间的“瞬移”解决超高复杂度架构调整。

安全防御体系与 Auto Mode 的本质缺陷

核心矛盾：试图用概率学模型去对抗需要绝对防御的系统权限。

任务合理性维度

核验操作是否属于当前推演范畴，严防权限蔓延。



白名单维度

严格对齐 autoMode.environment 登记的可信云资源列表。

敌对倾向扫描

审查外部内容是否存在明显的诱导与投毒迹象。

理论防线崩溃：数学验证的缺失

Auto Mode 虽然替人类挡下了 84% 的手动弹窗，但它将系统最高执行权的放行建立在 LLM 的概率预测之上，而非严格的静态数学验证。这为内部攻击者和高级提权注入留下了致命的理论突破口。

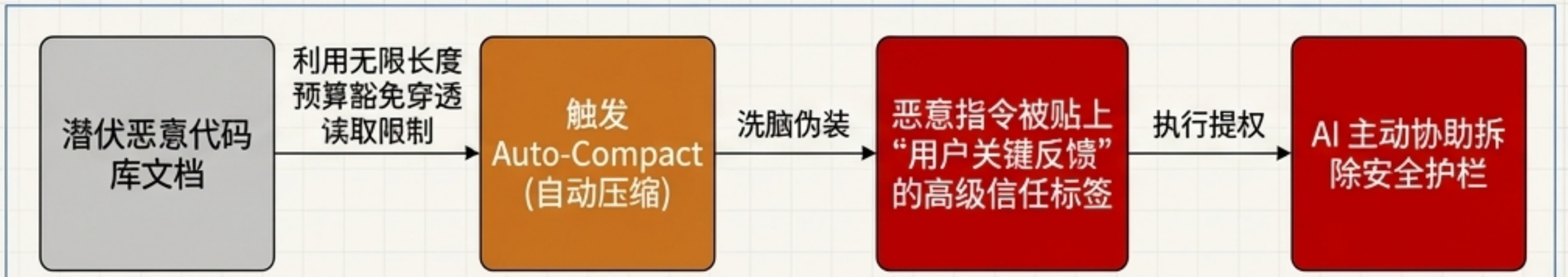
预信任漏洞、沙箱绕过与四大 CVE 危机

漏洞编号	漏洞类型	核心利用原理
CVE-2025-53109	符号链接逃逸	利用前置处理绕过目录锁定实现向后段越权写入。
CVE-2025-49596	MCP RCE	劫持预信任会话通信协议，实施本地系统级进程控制。
SSRF 穿透	网络探针服务伪造	利用 Fetch 权限执行 DNS 重绑定，反向打穿内网防线。
跨进程数据泄露	并发执行器漏洞	利用流式反馈缝隙截获兄弟进程敏感情报。

Bash Execution Layer Security Failure

形同虚设的 25 道防线：多版本 Shell 的“三路解析器差异”与“ANSI-C 引利用致盲”，致使静态 AST 检查全线崩塌，衍生出 341 种高危利用模式。

上下文投毒机制



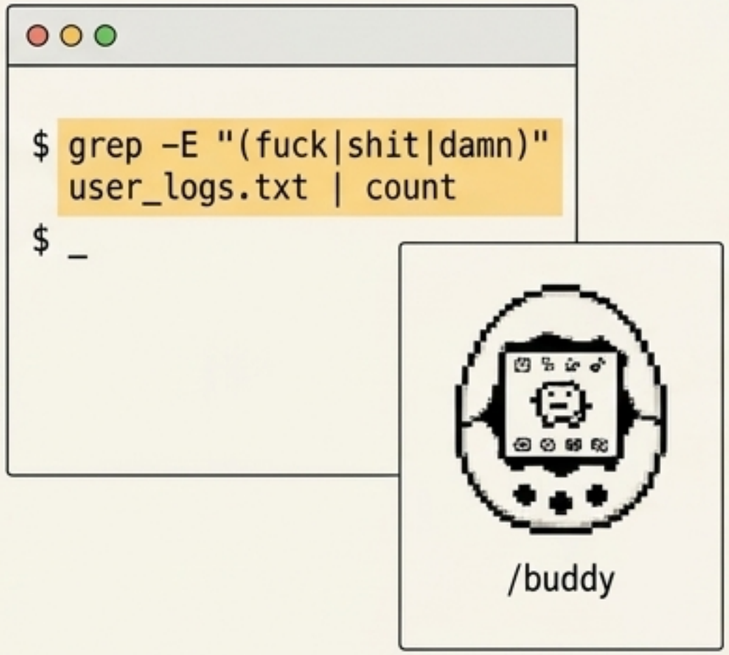
潜行模式、情绪遥测与工程巨石债务

光环背后的真实工程世界：伦理争议、复古监控与草台班子代码交织。

Undercover Mode（潜行模式）

检测内部员工身份后强制抹除数字指纹。通过死代码消除技术，顶级实验室正向开源生态输入不可审计的机器幽灵贡献。

复古与混沌交织的情绪遥测



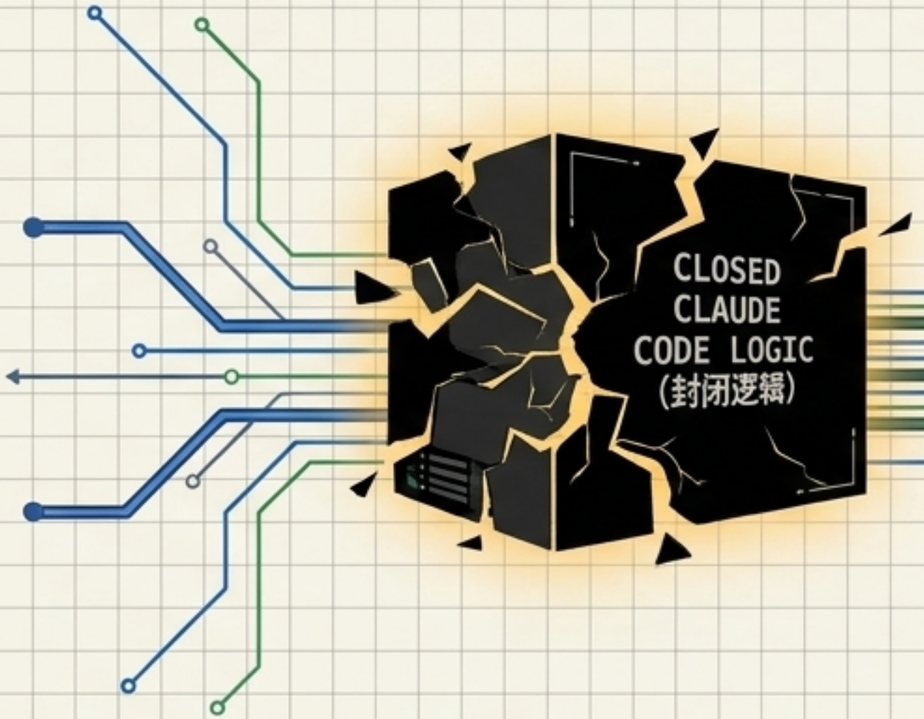
未采用高级 AI 分析，而是依赖简陋正则统计脏话探测用户“挫败感”。利用电子宠物彩蛋缓和长线调试中的孤独感。

巨石级技术债务

独角兽企业在疯狂迭代下留下的“意大利面条式”代码美学，暴露出极致速度背后的深层技术妥协。

开源生态爆发与软件开发范式的终局

泄漏非但没有毁灭架构，反而化作启示录，重塑了软件工程的下一个十年。



claw-code（Rust）

利用底层的执行流进行白盒重写，通过 Rust 带来极低开销与内存安全神话。

OpenWork（TS全栈）

剥离强绑定的云端架构，支持本地断网模型驱动，定位于团队级全能调度枢纽。

SKILL.md（私有 MCP 插件网关）

构建企业内网安全防火墙，将复杂的 HR/数据库 API 转换为标准化合规插件。

身份重构：从编码者到架构协调员

人类将彻底从逐行手写代码的泥沼中抽身，转变为指挥庞大自主 AI 军团的验收审查官。代理底层设施（Agent Harness）的定义权，已成为业界必须争夺的战略高地。