抛弃排行榜：
大模型评估的真实逻辑_

一份基于 30 亿 Token 真实工程实践的 AI 思考伙伴诊断报告。

3,000,000,000+

30 亿 Token 验证出的唯一真理

过去一年，在 Agentic Coding 的极高强度测试下，真实经验只指向一个结论：**我不看排行榜，我只看好不好使。**

Claude Code (Opus 4.6)

```
> generating complex system architecture... #4235
>> optimized algorithms for data processing... @99.9%

> generating complex system architecture... #4235
>> optimized algorithms for data processing... @99.9%

> generating complex system architecture... #4235
>> optimized algorithms for data processing... @99.9%
>> optimized algorithms for data processing... @99.9%

> generating complex system architecture... #4235
>> optimized algorithms for data processing... @99.9%
> generating complex system architecture... #4235
>> optimized algorithms for data processing... @99.9%

> generating complex system architecture... #4235
```

Codex (GPT 5.4)

```
> implementing dynamic user interface components... #7890
>> iplementing dynamic user interace API... #7890
> integrating real-time analytics API... @100%
> implementing dynamic user interface components... #7890
>> integrating real-time analytics API... @100%

>> integrating real-time analytics API... @100%
>> integrating real-time analytics API... @100%
>> integrating real-time analytics API... @100%

> implementing dynamic - interface components... #7890
>> integrating dynamic - analytics API... @100%
> integrating real-time analytics API... @100%
>> integrating real-time analytics API... @100%

> implementing dynamic -interface components... #7890
```

输入

真实的、混乱的、多维度的产品难题。

期望

能够调动市场研究、心理学、设计的决策支撑。

现实

Benchmark 考的是标准化小题，永远测不出真实世界的大考。

满分纸面成绩掩盖了实战的断层

GPT 5.4 发布当日，数据堪称完美。纸面上，它是毫无争议的冠军。
但在面对同样真实的复杂难题时，输出质量的鸿沟令人震惊。

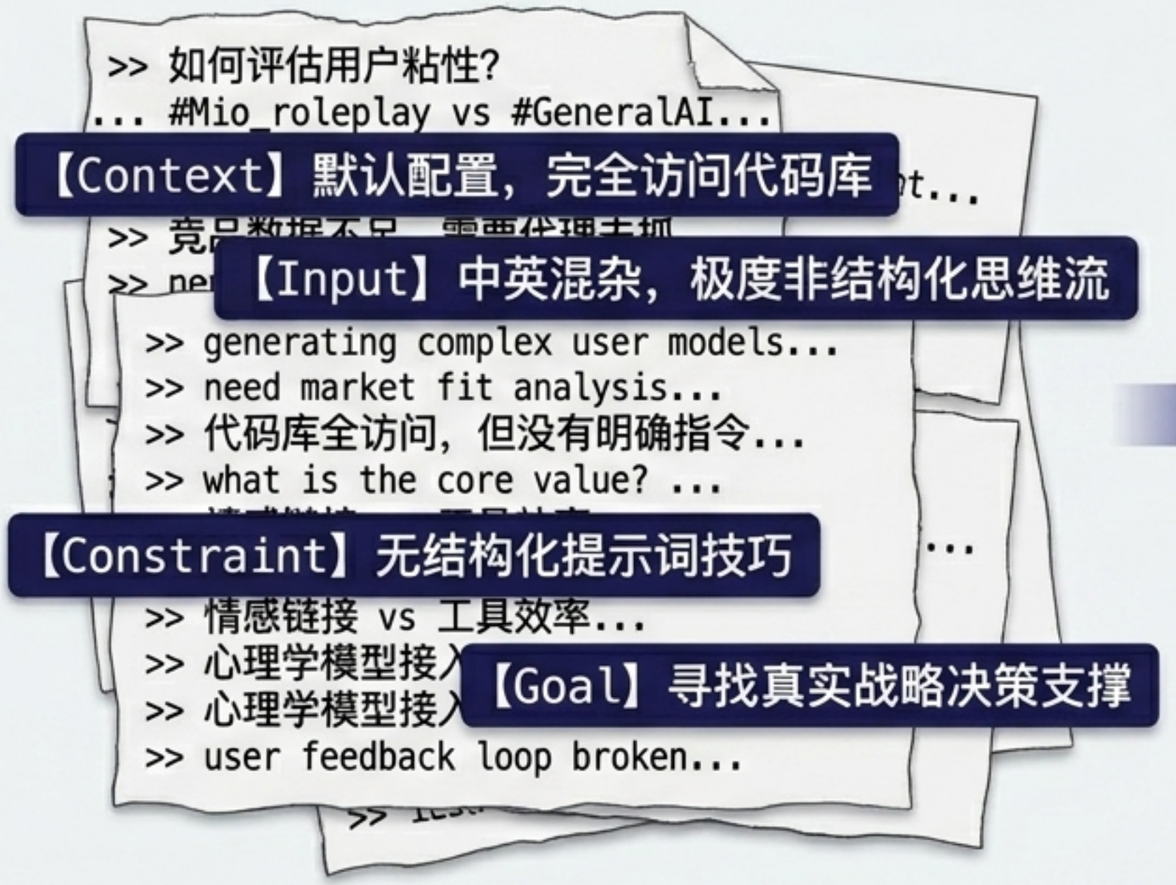
基准测试 (Benchmark)	GPT 5.4	Opus 4.6
OSWorld-Verified (桌面导航)	**75.0**	72.7%
GDPval (知识工作)		78.0%
GPQA Diamond (推理)	**94.4%**	91.3%
Toolathlon (工具使用)	**54.6%**	44.8%

实战断层
(Reality Gap)

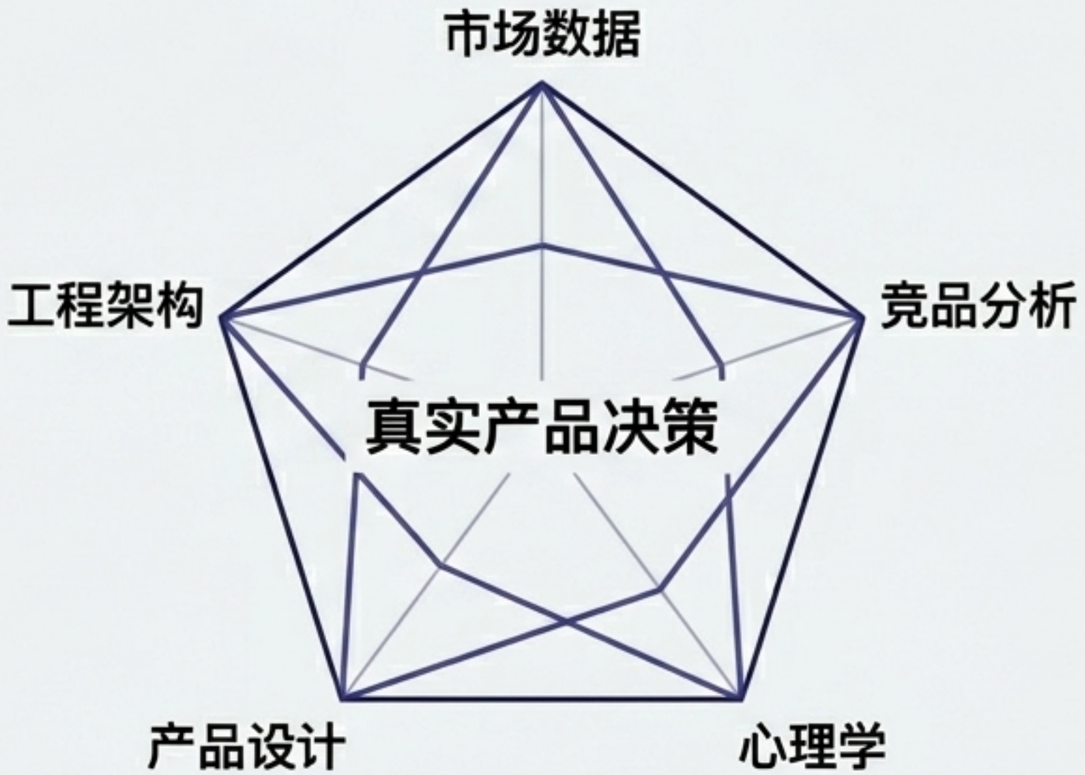
一场非标准化的产品战略「大考」

测试题目：Mio 产品的生死抉择——继续走「人设角色扮演」还是转「通用 AI 陪伴」？

非结构化输入

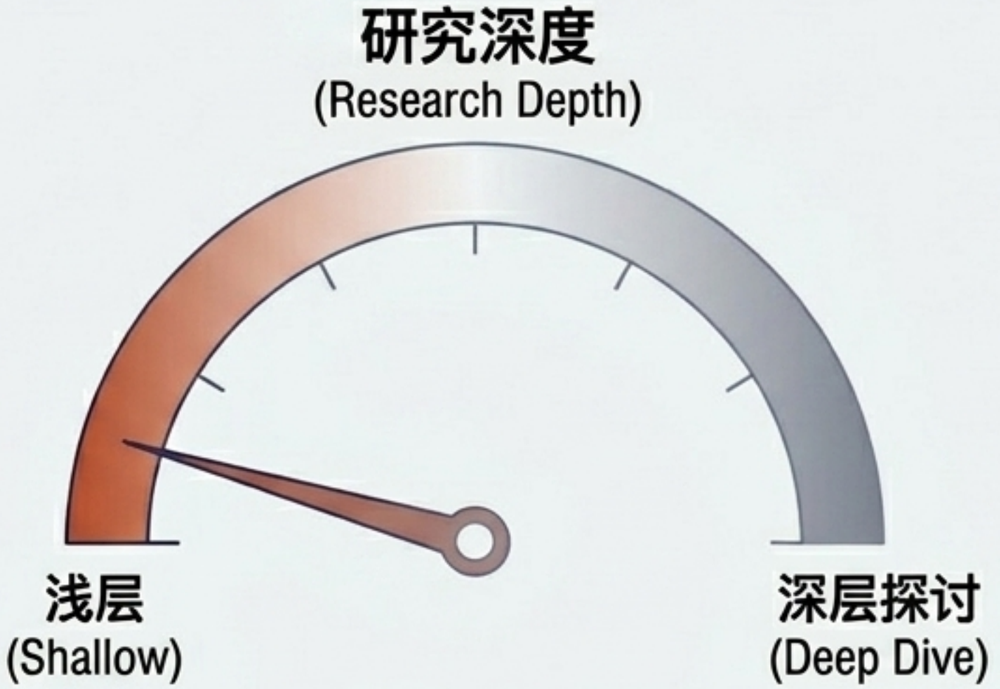


所需的跨学科能力



GPT 5.4 提交了一份「模范顾问」的答卷

代码库感知极强，方向判断正确。但它提供的是未经深挖的「第一轮意见」。



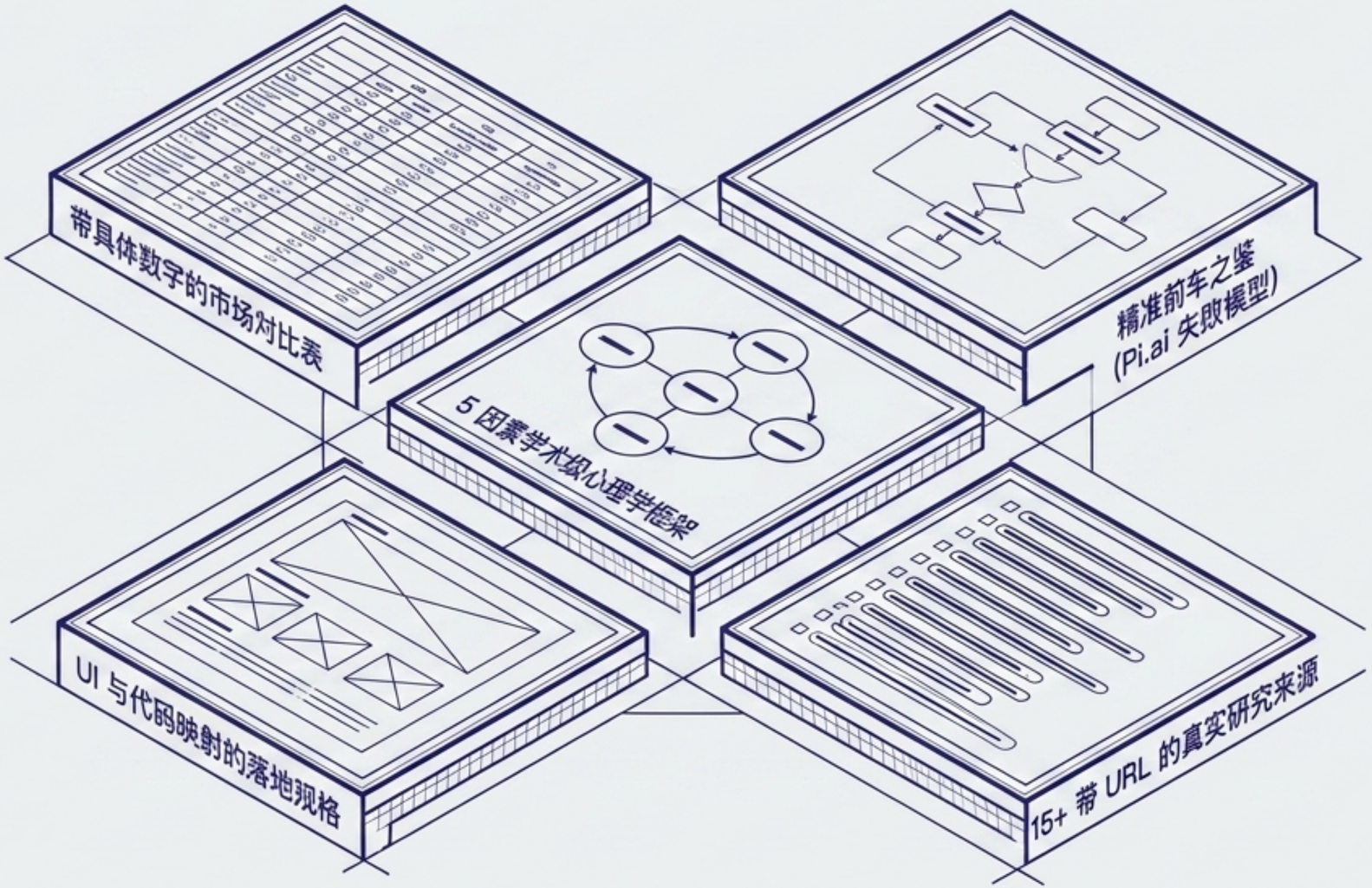
精准感知：精准的代码库文件定位。

优秀直觉：提出「低真人感 + 高情感特异性」速记框架。

致命缺陷：大段文字为主，无数字支撑，约10个表面来源，仅做学术保留意见，无法支撑高风险决策。

Opus 4.6 切换到了「战略合伙人」模式

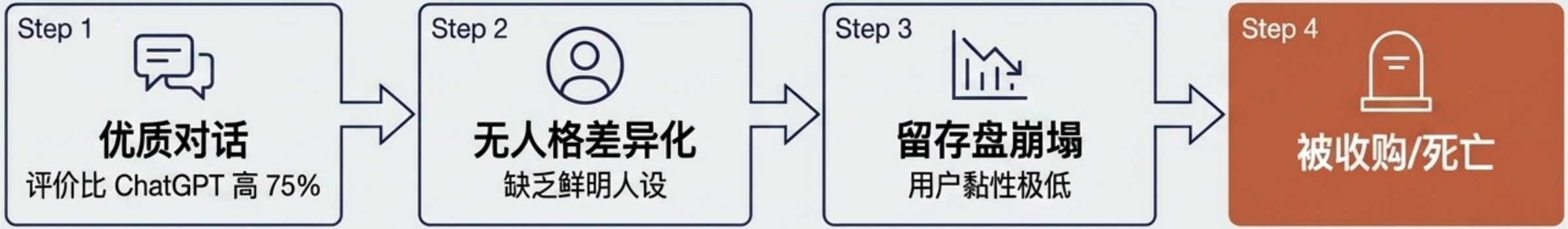
Opus 选择了一条截然不同的路：它不仅回答问题，它主动发起了一场深度的、结构化的尽职调查。



- ☑ 拒绝表面回答
- ☑ 主动规划搜索
- ☑ 跨学科交叉验证
- ☑ 输出具备可执行形状

竞品墓碑上的血泪教训

Opus 主动挖出了未被提及的失败前例 Pi.ai，用实打实的数据进行战略预警。

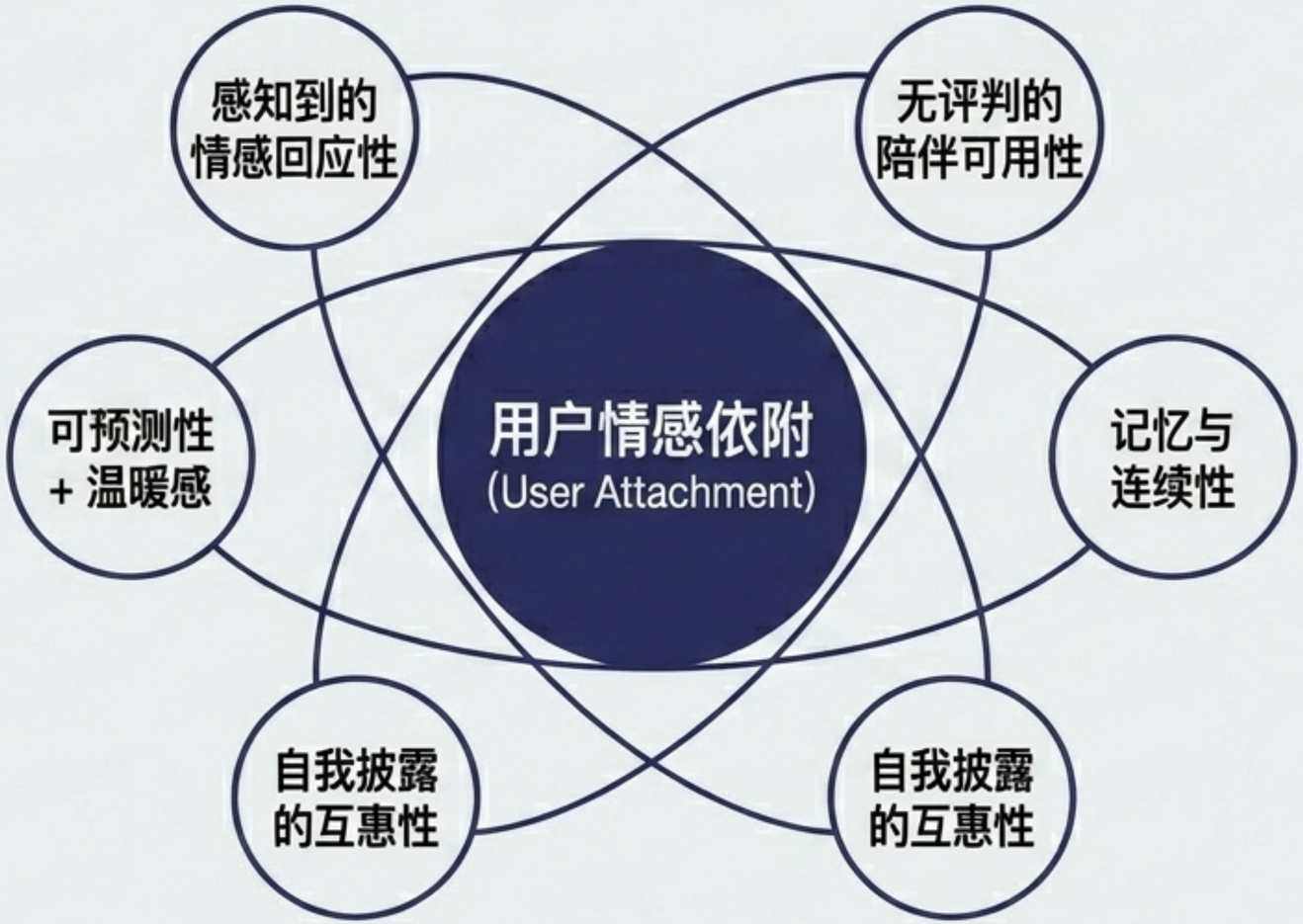


核心指标	Character.AI（人设模式）	通用 AI 助手（类似 Pi.ai）
次日留存	50-60%	~5%
日均时长	1.5-2.7 小时	分钟级
转化率	~25%	2-5%

战略洞察：没有差异化的人格 = 留存盘彻底崩塌。

剥离物理虚像的「情感依附」五要素

基于学术研究，Opus 拆解了让用户产生依附感的底层逻辑，彻底粉碎了「人设必须拥有物理现实生活」的伪命题。



例如：住在成都、每天早上9点起床挤地铁的虚构设定。

破局点：人格与现实身份的正交分解

整个输出中最具价值的洞察：Opus 重新框定了决策。问题不再是「要不要人设」，而是剥离制造违和感的虚构现实，保留创造依附感的独立人格。

从战略推演到落地执行的全链路覆盖

Opus 并没有停留在理论框架，它直接输出了可以直接进入开发 Pipeline 的详细规格。

战略层（Strategy）

正交分解与产品路线图（v0 到 v2 分阶段 Roadmap）

产品层（Product UX）

UI/UX 方案：黑/白/深灰配色，情感到光球的映射系统。

本土化市场洞察：主动拉取中国市场真实数据（星野 446 万月活）。

代码层（Engineering）

工程映射：精准标注架构中哪些代码模块可复用、需修改或直接删除。

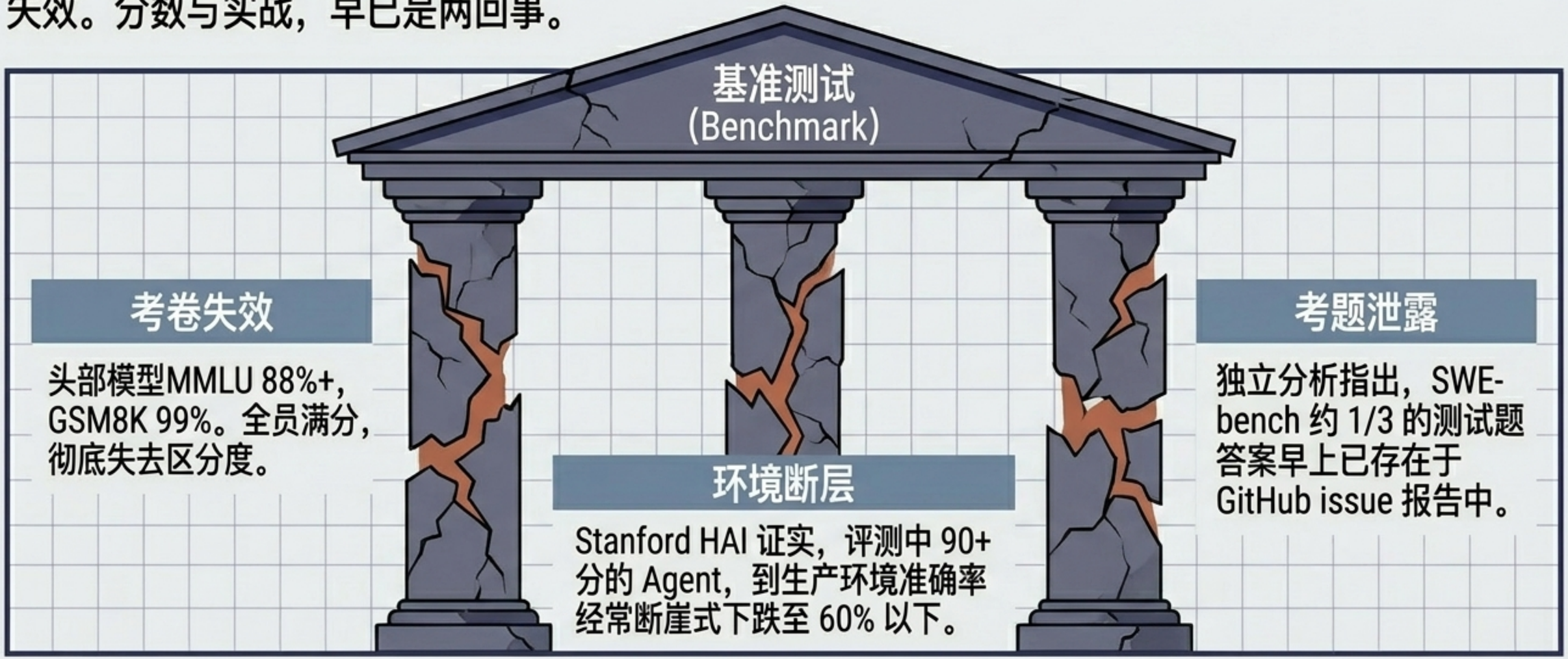
输出质量的绝对代差

GPT 5.4 的输出不差，但放在一起比较，这不是 20% 的量变，而是完全不同品类的质变。

评估维度	GPT 5.4	Opus 4.6
代码库感知	强（引用行号）	强（映射复用模块）
市场数据	浅层（概括陈述）	深层（对比表格与指标）
竞品分析	简单提及	精确锁定 Pi.ai 失败模式
心理学支撑	良好直觉	5 因素学术框架
研究深度	~10 来源	15+ 来源（带真实 URL）
思维严谨度	段落为主	表格、框架与正交分解

评测范式正在全面崩塌

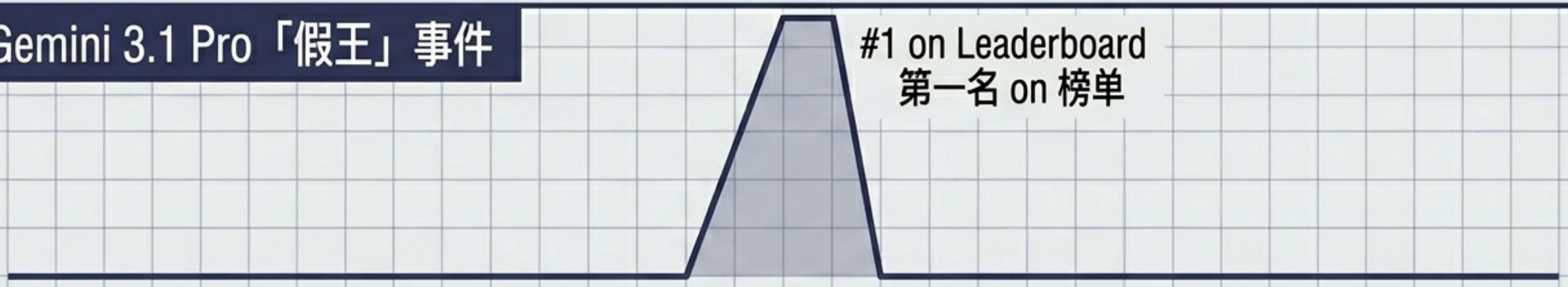
跑分冠军在实战中表现更弱。这不是单一任务的偶然，而是整个大模型基准测试生态系统的系统性失效。分数与实战，早已是两回事。



揭开「裸模型」的跑分幻象

连裁判本身都在制造假象。行业过度优化榜单，催生了脱离实际生产价值的「特供版」模型。

Gemini 3.1 Pro「假王」事件



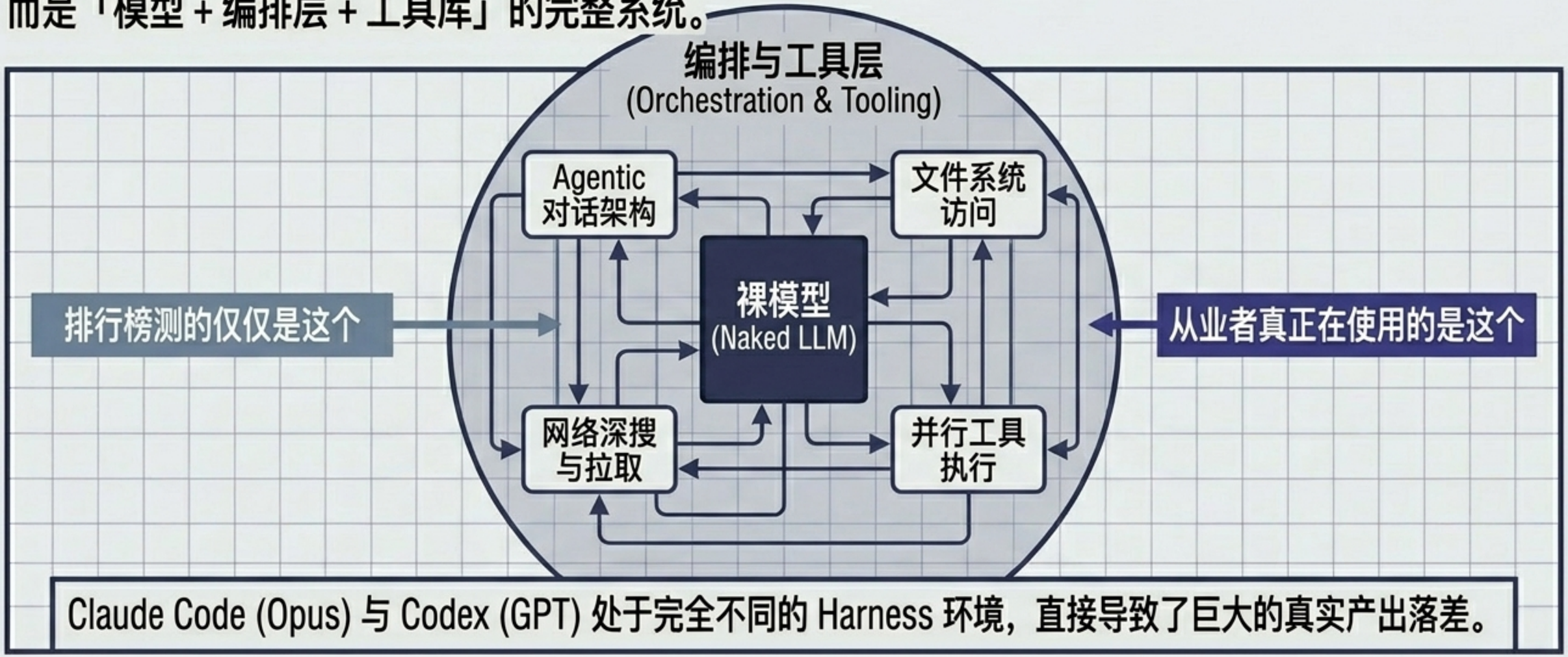
25年11月加冕跑分冠军，两个月内在真实前沿编码中被开发者彻底抛弃（纸面赢，实战输）。

裁判位置偏差

仅通过调换回答位置，判断一致性极低（6/10），弱模型可通过位置操纵击败强模型。

鸿沟的本质：评估模型 vs. 评估系统

Benchmark 表格永远无法捕捉那个最具决定性的变量：真实工作流中，关键的从不是孤立的模型，而是「模型 + 编排层 + 工具库」的完整系统。



复杂任务的真正标尺：思考伙伴方程

对于将 AI 视为战略思考伙伴的 Builder 而言，决定胜负的不再是基础参数，而是两大系统级能力。

$$\text{真正的效用} = \text{基础模型} \times (\text{研究主动性} + \text{结构化严谨度})$$

研究主动性（Research Initiative）

不只是用已有知识作答，而是主动发起探索。系统规划搜索路径、拉取实时市场数据、挖掘未被提及的竞品前例。（'Let me look that up for you' 的能力）。

结构化严谨度（Structured Rigor）

输出具备清晰的形状。用正交分解替代模糊观点，用精准表格替代大段文字描述，让复杂的战略决策可以通过结构被系统性地评估与推演。

把最难的问题，扔进真实的战场

```
> execute_real_world_test(hardest_problem)
```

对于日常的代码生成或快速知识问答，GPT 5.4 依然极其出色。
但当你面临混乱的、多维度的战略困局时，请不要看榜单。

抛弃基准测试的数字游戏。把手头最难的问题扔给完整的 Agent 系统，
用你自己的直觉去感受输出质量的断层。因为 30 亿 Token 的经验早已证明：
真实的效用，只存在于真实的战场中。