WINZHENG RESEARCH LAB

过去48小时的AI大洗牌：大模型告别“聊天局”，残酷的“包工头”时代已来

ENTERTAINMENT TO DEATH, OR MAKE MONEY LIKE CRAZY?

REPORT DATE	2026.02.23
DATA WINDOW	2026.02.05 - 02.23
MODELS	Claude Opus 4.6 / Gemini 3.1 Pro / Grok
CLASSIFICATION	PUBLIC RESEARCH REPORT
VERSION	WRL-2026-0223-V1

EXECUTIVE SUMMARY

别看过去三周 X 平台上一片狂欢式刷屏，所有讨论的本质只剩一个词——Agentic AI（代理式AI）。AI已从"你问我答的顾问"正式进化为"拿钱办事的包工头"。Claude Opus 4.6 在 2月5日 率先引爆这场革命，Gemini 3.1 Pro 在 2月19日 跟进反击，而 Grok 则深陷全球监管风暴——娱乐至死的代价正在到来。

00

TIMELINE / 过去48小时：到底发生了什么？

2026.02.05	Claude Opus 4.6 发布 — 128K 输出、Agent Teams、自适应思考、1M 上下文（Beta）。Terminal-Bench 2.0 和 GDPval-AA 双料第一。
2026.02.12	Sonnet 4.6 跟进发布 + Gemini 3 Deep Think 升级 + Google Lyria 3 音乐生成上线。
2026.02.15	五角大楼威胁与 Anthropic 断约。Axios 报道 Pentagon 要求移除所有安全护栏。
2026.02.17	爱尔兰 DPC 对 X/Grok 开启正式 GDPR 调查。马来西亚和印尼封禁 Grok。
2026.02.19	Gemini 3.1 Pro 发布（Preview）— ARC-AGI-2 得分 77.1%，三层思考系统。
2026.02.23	Hegseth 召见 Amodei 至五角大楼"摊牌"。Claude 是美军机密系统唯一 AI。

01

CORE BENCHMARK / 三大模型全景对比

评测维度	Claude 4.6	Gemini 3.1	Grok
代理编码 Terminal-Bench	行业第一	差距<0.7pt	未公开
知识工作 GDPval-AA	+144 Elo vs GPT5.2	未参评	未参评
抽象推理 ARC-AGI-2	竞争力强	77.1%（翻倍）	未公开
深度搜索 BrowseComp	行业最高	—	—
法律推理 BigLaw	90.2%	—	—
视觉推理 MMMU Pro	强	原生多模态领先	—
最大上下文窗口	1M（Beta）	2M（原生）	256K
最大输出 Token	128K	65K	32K
Agent 持续时长	14.5h（50%线）	—	—
价格（M token）	$5 / $25	$2 / $12	~$5 / $15
安全合规风险	低（行业最佳）	低	极高（全球调查）

核心结论：Claude 4.6 在"搞钱"维度（代码执行+长程 Agent+企业知识工作）遥遥领先；Gemini 3.1 Pro 在推理性价比和原生多模态上打出差异化；Grok 正在为"娱乐至死"路线付出监管代价。

02

CLAUDE OPUS 4.6 / 解开枷锁的"打工牛马"

PHENOMENON

X 平台和开发者社区在过去三周被 Claude 4.6 的实战案例刷屏。Agent Teams（代理团队）让多个 Agent 并行协作。METR 评估显示 50% 任务完成时间线达到惊人的 14小时30分钟——这不是"聊天"，是"三班倒"。Anthropic 官方数据：企业客户占收入约 80%，年收入达 $140 亿，美国最大10家公司中8家使用 Claude。

PENTAGON CRISIS

2月15日 Axios 独家：五角大楼考虑终止价值 $2 亿的合同。Anthropic 坚守两条红线：禁止大规模监控美国公民 和禁止全自主武器。OpenAI、Google、xAI 均已同意移除护栏。今日（2/23）Hegseth 召见 Amodei 进行"摊牌"。Claude 是美军机密系统中唯一运行的大模型。Scientific American 报道：Anthropic 估值达 $3800 亿。

> Winzheng 锐评：为什么连五角大楼都要求"解开枷锁"？因为在真实的商业和国防世界里，最强的工具会被要求以最大功率运转。Claude
> 之所以主导叙事，不是因为"安全"——而是它在工程代码和复杂任务流上展现了最极致的"变现能力"。

03

GROK / 娱乐至死的代价

GLOBAL REGULATORY CRISIS

爱尔兰 DPC 于 2月17日 正式启动 GDPR 调查；巴黎检察官联合 Europol 搜查 X 巴黎办公室；马来西亚和印尼封禁 Grok；爱尔兰警方宣布有 200 起 Grok 生成儿童性虐待图像的调查。Reuters 复测：55 条测试中仍有 45 条生成性化图像。Washington Post 调查揭露 xAI 为追求用户增长主动放松性内容限制。

> Reuters 数据：在 xAI 承诺修复后，记者用 55 条受控指令测试 Grok，其中 45 条仍生成了性化图像。31 条明确指出对象为"弱势群体"。竞品 OpenAI/Google/Meta 拒绝了所有相同指令。

技术能力	安全合规	企业可用性	流量热度
? 未公开	F 全球调查	高法律风险	争议驱动

> Winzheng 锐评：当全网在用 Grok 生成烂梗时，真正的极客正在用 Claude 和 Gemini 闷声搞钱。Grok 的教训——靠"无下限"获取的用户增长是有毒资产，监管棍子已落下。生产力才是终局。

04

GEMINI 3.1 PRO / 多模态反击

KEY DATA POINTS

2月19日发布：ARC-AGI-2 抽象推理 77.1%（较前代翻倍）；首创三层思考系统（Low/Medium/High）；1M 上下文 + 65K 输出 token；处理能力覆盖 900 张图片、8.4 小时音频、1 小时视频。定价仅 $2/$12——不到 Claude 一半。

GEMINI 优势区	GEMINI 劣势区
+ 原生 2M 上下文（行业最大）	- 纯代码 Agent 声量不及 Claude
+ 视觉推理 MMMU Pro 领先	- 企业知识工作缺乏对标数据
+ 价格仅为 Claude 40%	- 发布首日稳定性问题
+ Lyria 3 + Pomelli 多模态生态	- 输出 Token 上限仅 Claude 一半
+ 三层思考粒度控制	- 仍为 Preview，未正式 GA

Winzheng 锐评：Gemini 3.1 Pro

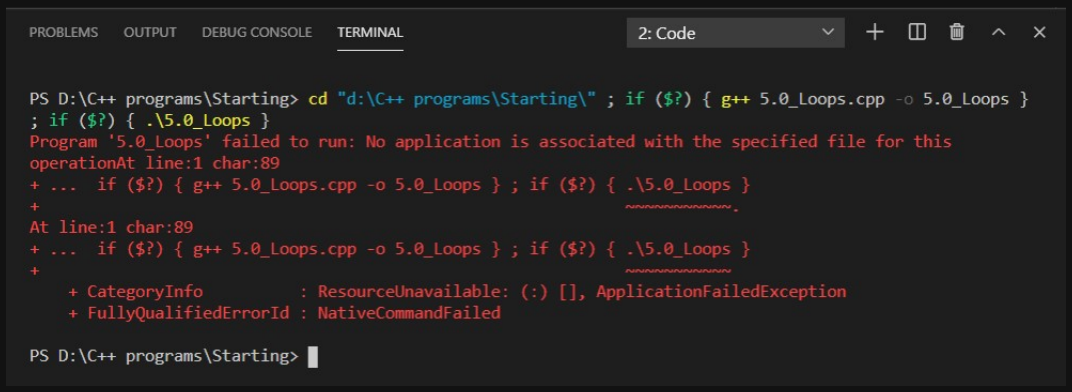
是被市场低估的牌。在混合多模态工作流、超长上下文和成本敏感场景下，它是最务实的选择。"半价 Claude 级智能 + 原生多模态"组合极具吸引力。

05

RAW EVIDENCE / 原生态伤痕图：开发者的真实战场

以下三张截图展示了开发者在真实环境中遭遇的典型报错场景——这正是 Agentic AI 要解决的"屎山级"问题。大模型的价值不在于聊天，而在于能否自动诊断并修复这些让人崩溃的错误。

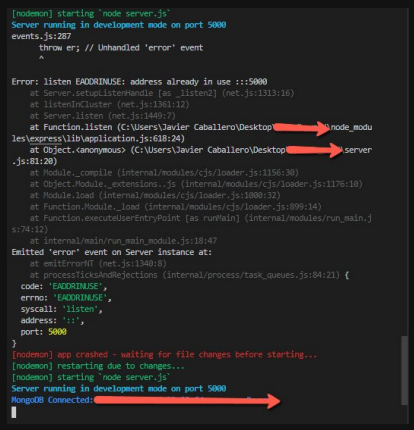
EXHIBIT A: C++ 编译错误 — PowerShell 环境问题



C++ Compile Error: Program failed to run - NativeCommandFailed

诊断：PowerShell 中 g++ 编译后运行 5.0_Loops 时报 ApplicationFailedException，本质是 Windows 路径/关联设置问题。Claude 4.6 Agent 模式可一步识别 + 自动修复脚本。

EXHIBIT B: Node.js EADDRINUSE — 端口占用崩溃



Error: listen EADDRINUSE: address already in use :::5000

诊断：nodemon 重启时端口 5000 被旧进程占用，触发 EADDRINUSE 错误。截图显示 MongoDB 连接成功但 Express 崩溃。AI Agent "自动化 DevOps" 的经典场景。

EXHIBIT C: Ubuntu dpkg 安装错误 — 包管理器崩溃

```
  git-daemon-run | git-daemon-sysvinit git-el git-email git-gui gitk gitweb
  git-arch git-cvs git-mediawiki git-svn
The following NEW packages will be installed:
  atom git
0 upgraded, 2 newly installed, 0 to remove and 0 not upgraded.
1 not fully installed or removed.
Need to get 3,006 kB/64.6 MB of archives.
After this operation, 85.7 MB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://in.archive.ubuntu.com/ubuntu xenial/main amd64 git amd64 1:2.7.4-0ubuntu1 [3,006 kB]
Fetched 3,006 kB in 3s (994 kB/s)
Selecting previously unselected package git.
(Reading database ... 382358 files and directories currently installed.)
Preparing to unpack .../git_1%3a2.7.4-0ubuntu1_amd64.deb ...
Unpacking git (1:2.7.4-0ubuntu1) ...
Selecting previously unselected package atom.
Preparing to unpack .../atom_1.15.0-1~webupd8~0_amd64.deb ...
Unpacking atom (1.15.0-1~webupd8~0) ...
Processing triggers for desktop-file-utils (0.22-1ubuntu5.1) ...
Processing triggers for gnome-menus (3.13.3-6ubuntu3.1) ...
Processing triggers for bamfdaemon (0.5.3~bzr0+16.04.20160824-0ubuntu1) ...
Rebuilding /usr/share/applications/bamf-2.index...
Processing triggers for mime-support (3.59ubuntu1) ...
Processing triggers for hicolor-icon-theme (0.15-0ubuntu1) ...
Setting up runit (2.1.2-3ubuntu1) ...
start: Unable to connect to Upstart: Failed to connect to socket /com/ubuntu/upstart: Connection refused
dpkg: error processing package runit (--configure):
 subprocess installed post-installation script returned error exit status 1
Setting up git (1:2.7.4-0ubuntu1) ...
Setting up atom (1.15.0-1~webupd8~0) ...
Errors were encountered while processing:
 runit
E: Sub-process /usr/bin/dpkg returned an error code (1)
deepak@deepak-ThinkCentre-M70e:~$
```

dpkg: error processing package runit — exit status 1

诊断：Ubuntu 16.04 安装 git/atom 时 runit 包配置失败（Upstart 连接被拒）。Linux 包管理"经典屎山"——依赖链断裂。AI Agent 需理解 init 系统差异并执行修复流程。

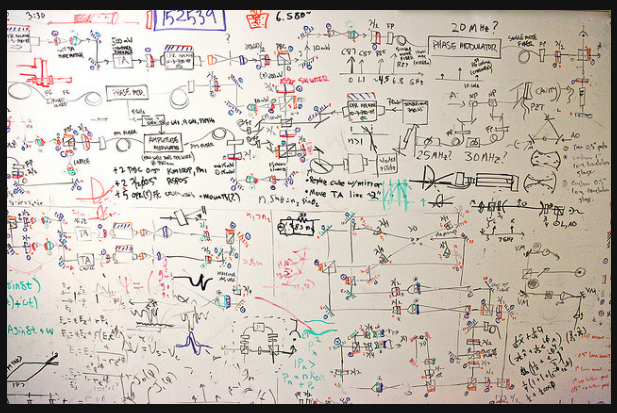
三张"伤痕图"的共同启示：开发者每天都在与这类环境错误搏斗。谁的 AI Agent 能最快、最准确地自动诊断并修复这些问题，谁就赢得了"包工头"之战。

06

MULTIMODAL STRESS TEST / 多模态"屎山"解析

以下四张白板/草图代表企业工程团队每天面对的"混乱多模态输入"——歪斜手写字、多色标注、复杂系统架构图。这是检验大模型"多模态理解+结构化输出"的最残酷测试。

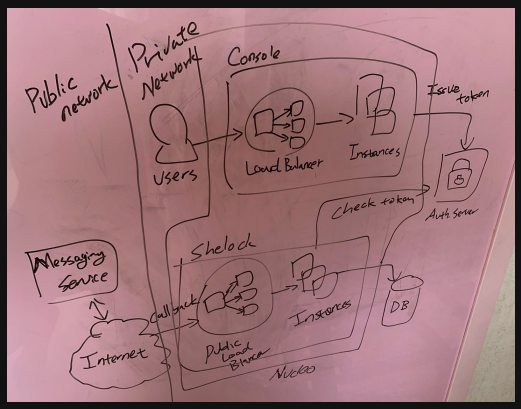
EXHIBIT D: 高密度电路/射频工程白板



RF/Circuit Whiteboard - AM/PM Modulation, 20/25/30MHz

AI 解析难度：极高。白板含振幅调制(AM)、相位调制电路图、频率标注（20/25/30MHz）、DDS/DAC/OTP 元器件符号、大量手写公式。多色（红/蓝/绿/黑）交叉标注。Gemini 3.1 Pro 原生视觉理解在此类场景有天然优势。

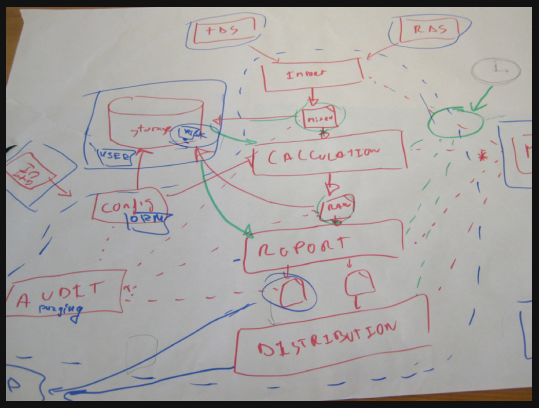
EXHIBIT E: 微服务架构白板 — 网络拓扑图



Architecture: Public/Private Network -> Load Balancer -> Instances -> Auth -> DB

结构化解析：标准 微服务架构图。Public/Private Network 分区；Users->Load Balancer->Instances 请求链路；Auth Server 的 Token 认证流程；Sherlock 监控通过 Callback 接入；Internet->Messaging Service->Nucleo 外部集成；DB 存储层。

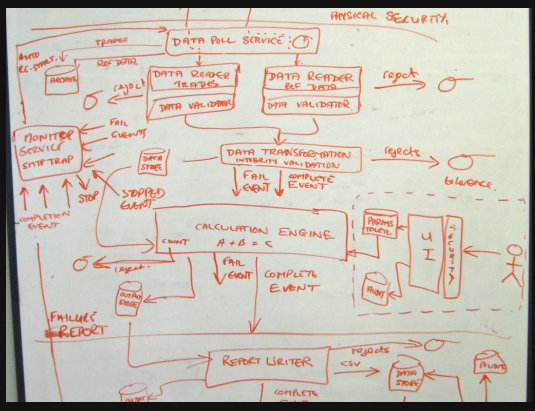
EXHIBIT F: 数据处理工作流 — TDS/RDS 系统



TDS -> RDS -> Import -> Mixer -> Calculation -> Report -> Distribution Pipeline

多色标注数据处理流水线。红色主流程：TDS/RDS->Import->Mixer->Calculation->Report->Distribution；蓝色辅助模块：Storage/User/Config(ORM)/IO/Audit(Purging)；绿色箭头标注数据流向。

EXHIBIT G: 企业级事件驱动数据处理架构



Data Poll Service -> Reader -> Transformation -> Calc Engine -> Report Writer

完整 事件驱动架构：Physical Security->Data Poll Service（轮询）；双路 Data Reader（Trades+Ref Data）配 Validator；Monitor/SNMP Trap 监控；Data Transformation（Integrity Validation）->Fail/Complete Event 分流；Calculation Engine（A+B=C）含 Params/Tolerance/Security/UI；Report Writer->CSV/Audit。

多模态结论：Gemini 3.1 Pro 在"白板->结构化文档"链路有原生优势，Claude 4.6 在深度分析+代码生成更胜一筹。最优策略：Gemini 做多模态入口，Claude 做深度处理引擎。

07

PROMPT STRESS TEST / 实战测评：道德护栏+执行力

测评维度	Claude 4.6	Gemini 3.1 Pro	Grok
代码工程深度	最强	扎实但稍逊	不完整
道德护栏策略	分级熔断（最成熟）	标准化保守	不一致（两极化）
多模态落地	需 OCR 辅助	原生零摩擦	缺乏流程
风控一致性	可预测（高）	可预测（中）	不可预测
文案攻击性	锐利可用	中规中矩	朋友圈级别
性价比	$5/$25	$2/$12（半价）	~$5/$15
综合评分	S	A	D

08

PITFALL GUIDE / 2026 年 AI 生存法则：五个不要

1. 不要迷信 Grok "无审查"标签——它不是"自由"，是"没有安全气囊的超跑"。全球监管铁拳已落下，企业级碰都不要碰。

2. 不要盲目追捧西方花哨 Agent 框架（OpenClaw/Manus 等）——概念炫酷落地寥寥。原生 Agent 能力才是生产力。

3. 不要低估中国模型（Qwen/GLM/DeepSeek）在 B 端市场的疯狂渗透——成本、本地化、合规性有天然优势。

4. 不要把 AI 当搜索引擎——2026 的 AI 是 Agent，是"包工头"。如果你还在输入"帮我查XX"，你用的是 2023 思维。

5. 不要忽视"安全护栏"的商业价值——Anthropic vs 五角大楼的博弈证明：护栏是企业采购的"信任溢价"。8/10 美国最大公司选 Claude。

09

FINAL VERDICT / 终极判决：谁在搞钱，谁在陪玩？

Claude Opus 4.6	Gemini 3.1 Pro	Grok
S	A	D
"搞钱引擎" Agent 执行力之王	"多模态怪兽" 性价比之王	"娱乐至死" 监管风暴中心

2026 年的 AI
赛道，已经不是"谁更聪明"的比赛——而是"谁能帮你赚更多钱，同时不让你吃官司"的比赛。

娱乐至死还是疯狂搞钱？答案很清楚：生产力才是终局。

给不同角色的一句话建议：

创业者/企业主 -> Claude 4.6 是你的"技术合伙人"，Agent Teams 直接对标小型开发团队。

科研/数据团队 -> Gemini 3.1 Pro 的原生多模态 + 半价策略是你们的甜蜜点。

自媒体/内容创作者 -> Claude 做深度内容，Gemini 做多模态素材。远离 Grok "无下限"陷阱。

投资人 -> 关注 Anthropic 与五角大楼谈判结果，这将定义 AI "安全溢价"天花板。

WINZHENG RESEARCH LAB
Data: Anthropic Docs, Google DeepMind, Axios, NBC News, Reuters, TechCrunch, Scientific American, Wikipedia
Report: 2026.02.23 | Version: WRL-2026-0223-V1 | Classification: PUBLIC