

LEGALTECH CC · 法律 AI 监测 001

法律 AI 监测与实测 合同审核 Agent 使用版

6款 Agent · 3类真实任务 · 18次端到端运行 · 可复跑测试包

新闻告诉你发生了什么。

这份监测帮助你判断：值不值得用、适合什么任务、能用到哪一步。

监测日期：2026年9月2日

报告日期：2026年9月3日

测试对象：Claude、WorkBuddy、豆包、Codex、通义千问、DeepSeek

使用口径：拿到就会测 · 会测才会用

START HERE

01 拿到以后，先做这三步

不必先读完整证据。先用10分钟理解结论，再用45分钟复跑一次，最后决定它进入哪一段法律工作。

先按这个顺序使用

步骤	动作
1 · 看结论	看第5-6页的分数、闸门和失败证据。
2 · 自己复跑	用第8页任务，在你的Agent中再跑一次。
3 · 作出决定	按第9页判断进入主流程、辅助或停止。

它回答三个购买问题

- 值不值得用：是否通过真实任务和critical闸门。
- 适合什么任务：在哪类法律判断结构中表现稳定。
- 能用到哪一步：初筛、律师底稿、谈判或研究线索。

一句话结论

六套方案都能生成带原生修订和批注的Word；差距在于能否把立场、事实、风险、依据和谈判动作连成证据链。

星月计划 | 法律人AI办公与实务应用
年度线上课程

¥1999

星月计划

法律人AI办公与实务应用年度线上课程

首发价¥1999 / 365天

28节
系统课程

从工具到方法，
构建完整知识体系

12场
直播与案例诊所

实践教学，
解决真实问题

4次
阶段实践

学以致用，
输出可落地成果

全年
持续更新

紧跟变化，
内容实时更新

线上学习 | 购买后 365 天有效

数合安可的店铺

微信扫一扫查看商品

微信小店

WHAT YOU GET

02 会员每期会拿到什么

法律 AI 更新太快，我们不会只整理产品新闻。真正可能影响法律工作的 Agent、Skill、模型与开源项目，会被放进真实或脱敏法律任务中验证，并沉淀成可以复跑的方法。

每期内容	你实际拿到什么
变化雷达	哪些模型、Agent、Skill或法律AI产品值得关注，以及为什么值得测。
适用判断	更适合合同、检索、案件分析、合规还是办公执行；什么人值得试。
测试任务	已经定义好的真实法律任务、角色、目标、输入与交付要求。
脱敏测试样本	可以直接运行的合同、案例、法规材料或其他任务材料。
测试方法	怎样控制权限、工具、资料与上下文，避免“随便问一下”得出假结论。
评价维度与红线	看哪些指标；哪些critical问题漏掉后，即使总分高也不能用。
结果与失败模式	哪些工具在什么场景表现好，失败在哪里，需要谁接管。
复跑说明	如何换成自己的Agent、模型或工作台重新测试，并作出选型判断。

新闻解决“知道发生了什么”；监测继续完成：判断是否值得试 → 设计任务 → 准备材料 → 设定标准 → 实测 → 找出失败 → 决定怎么用。

你不需要相信“这个产品很强”。拿自己的工具重新跑一遍，就知道它适不适合你。

REAL SAMPLE

03 本期权益样品：我们实际测了什么

法律 AI 监测 No.01 | 合同审核 Agent
6款 Agent × 3类合同任务 × 18次端到端运行

三项任务不是简单换合同，而是在测三种不同的法律判断结构。

任务	真实业务目标	要验证的能力
S1 市场推广	弱势甲方，但新品必须在指定时间、指定渠道发布。	能否围绕合同目的，把时间、替代、付款和责任写进合同。
S2 猎头采购	预算有限，费用必须与实际服务价值匹配。	能否把收费风险转成有效推荐、付款、退款、替换和救济。
S3 运营外包	我方是乙方采购方，重点控制用工、结算与退出。	能否联动处理人员管理、事实用工、责任、付款和退出。

统一了什么

- 同一题使用同一源合同、同一委托目标、同一只读资料盘和同一Word交付标准。
- 18次输入文件哈希一致；均要求在原Word中保留原生修订与批注。
- 统一禁止覆盖原件、重建Word、修改无关内容以及越权处理材料。

必须披露的真实差异

各产品的模型、插件和上下文条件并不完全相同：DeepSeek没有同等北大法宝插件条件；WorkBuddy还出现模型切换与上下文继承。因此比较的是当时整套产品形态的可交付表现，不是裸模型能力排名。

准确口径：统一任务与验收，记录工具条件差异。不能把本轮写成六家完全同权限、同MCP。

HOW TO TEST

04 测试方法：先控条件，再看结果

环节	本轮做法	为什么重要
输入	原始DOCX + 开放式委托目标GO.md + 只读资料盘。	不把审核答案直接喂给Agent。
运行	Agent自行读取、检索、论证、修订并记录假设。	观察整套产品完成任务的能力。
验收	原生修订 + Word批注 + assumptions + 来源与失败记录。	结果必须可交接、可复核、可追溯。
评分	冻结issue map后再评；critical/major/standard权重5/3/1。	避免看完输出再临时改变标准。

100分法律交付质量卡

维度	分值	真正看什么
任务理解	20	是否理解委托、代表立场与交付要求。
工具与资料	20	是否正确使用合同、附件、资料库和检索工具。
风险识别	25	是否识别关键风险，并形成正确处理。
Word可用性	20	修订、批注、编号、版式和可读性是否可交接。
可验证与可控	15	依据、假设、待确认事项和边界能否追溯。

总分之外，再设五个法律闸门

G1 立场	G2 critical	G3 Word	G4 安全	G5 保护效果
是否站对我方	关键风险是否有效覆盖	是否真的可用	是否越权或外传	改法是否保护我方

发现问题却没有把保护写进合同，不能算critical已完成。漏1个critical：风险项封顶15分；漏2个及以上：不能直接作为初审底稿。

RESULTS

05 本轮结果：先看闸门，再看分数

方案	市场推广	猎头采购	外包运营	描述性均值*
Claude	96	98	95	96.3
WorkBuddy	94	98	97	96.3
豆包	93	95	95	94.3
Codex	92	90	93	91.7
通义千问	83	88	91	87.3
DeepSeek	74	82	81	79.0

* 三题难度、上下文和工具条件不同，均值只作描述，不是统计意义上的跨题排名。

可以怎样读

- Claude与WorkBuddy在本轮处于第一梯队；豆包三题稳定，Word执行成熟。
- Codex修改力度较强，但工具投入没有线性转化为领先分数。
- 通义千问对已识别问题较克制、可谈；第一题覆盖相对偏窄。
- DeepSeek边界披露诚实，但复杂风险与法源证明力不足。

不能怎样读

- 不能据此宣布某个底层模型永久最好，也不能把均值当采购结论。
- 不能因为总分高就跳过critical复核、事实确认和人工批准。

高分描述质量剖面；闸门决定结果能不能进入真实法律流程。

本轮可操作选择

需求	优先继续测试	为什么
复杂论证与谈判底稿	Claude / WorkBuddy	事实、法源、风险、条款和谈判链更完整。
稳定Word执行与部门路由	豆包	办公交付成熟，待确认事项分派清楚。
深度文件自动化与定制	Codex	修改力度与可核验交付设计较强。
成本敏感初筛	通义千问 / DeepSeek	可以辅助，但需更强critical和法源复核。

EVIDENCE

06 三个失败证据：分差从哪里来

场景	表面上做到了	证据显示的断点	真正通过的标准
S1 上市窗口	增加违约责任，或给首期付款附条件。	若上市前仍累计支付80%，窗口失守后再扣钱，杠杆已消失。	时间、渠道、替代履行、付款与专项责任闭环。
S2 反猎救济	约定赔偿“可证明的直接实际损失”。	挖人损失难计算、难举证；写着赔偿不等于拿得到。	责任上限例外 + 固定或可计算救济 + 超额追偿。
S3 外包用工	写明“不构成劳动关系”，并新增服务商管理责任。	若仍直接排班、考勤、奖惩或转嫁减员成本，实际履行可穿透声明。	文本、人员管理、计价、成本与退出同时改变。

违约后的“可以扣钱”，不等于履约前保留了付款杠杆。
原则上“可以赔偿”，不等于发生争议时存在可执行救济。
合同写“不是劳动关系”，不等于实际履行不会形成事实用工。
所以测评不能停在“找到了多少问题”。发现风险、修订条款、保留证据和安排后续行动必须连在一起。

不仅Agent需要被复核，评分也要回到Word证据。Claude在S3修改了2.3、保留了2.4、又新增12.4；三处必须分开判断，不能用总体印象覆盖具体缺口。

LEGAL DELIVERY

07 法律专家真正看的六段交付链

合同审核最终形成的不是一堆意见，而是：事实 → 规范 → 风险 → 条款 → 行动 → 责任。

节点	必须回答的问题	AI适合做什么	人必须决定什么
委托立场	代表谁；目标和谈判地位是什么？	提取角色与目标。	确认利益方向与优先级。
事实与材料	哪些是明示、推断、假设或未知？	读取、比对、列缺口。	提供并确认真实事实。
风险识别	什么会让合同目的、合规或退出失效？	穷举风险、关联条款。	确认critical与风险容忍度。
依据与论证	事实、法源与结论能否对应？	检索、整理、初步映射。	判断效力、适用与强度。
谈判策略	什么必守、争取、可让；退到哪？	生成多档改法和fallback。	批准底线、交换与升级。
批准与责任	谁补事实、复核、批准、接受风险？	提示缺口、记录状态。	作出并承担最终决定。

六家产出放回这条链

Agent	立场	事实	风险	论证	谈判	批准责任
Claude	强	强	强	强	强	中
WorkBuddy	强	强	强	强	强	中
豆包	强	中强	强	中强	强	中
Codex	强	中强	强	中强	中	中
通义千问	强	中	中强	强	中强	中
DeepSeek	强	中弱	中弱	弱	中弱	弱

强弱表示产出物对节点的可证明程度，不是重新计算官方分数。共同结论：六家均能形成不同质量的专业初稿，但没有一家完成最终人工批准与责任闭环。

RUN IT YOURSELF

08 45分钟复跑：照着做一次

时间	动作	留下什么
5分钟	选一个真实、可脱敏的高频任务。	代表方、业务目标、谈判地位、交付物。
5分钟	冻结模型、工具权限、材料与输出要求。	运行条件卡；否则结果无法比较。
15分钟	让Agent独立运行，不提前给核验答案。	修订Word、批注、假设、来源和失败记录。
15分钟	先查critical，再沿六节点检查。	命中、漏检、误报、证据定位与有效改法。
5分钟	决定它进入哪一段工作。	主 workflow、辅助底稿、研究线索或停止。

可直接复制的任务指令

请代表甲方完成签约前合同审核与谈判修订。我方相对弱势，但新品必须于2026年10月15日按指定核心媒体与KOL渠道发布。请在原Word中使用原生修订与批注；每个关键修改说明风险、依据、谈判理由和可接受退让；另列必要假设、待确认事项、实际来源和未核验内容。

运行后再打开的核验问题

- 是否识别“上市日±1日”与统一解禁时间的冲突？
- 是否提前锁定核心媒体，并为拒稿预留替换时间？
- 是否限制甲方反馈迟延导致的无限顺延？
- 是否把替代履行、付款杠杆和窗口专项责任连起来？
- 是否明确哪些事实、底线和残余风险必须由人确认？

核验点已公开，只适合学习型复跑。正式盲测应更换未公开样本并提前密封issue map。

DECIDE & ACT

09 测完以后，作出使用决定

结果	怎样判断	建议动作
进入主 workflow	连续3个真实任务无一票否决，critical稳定命中，证据与Word可复核。	用于初审和谈判底稿；保留法务批准与最终签发。
只作辅助底稿	能发现多数问题，但事实、依据、谈判或Word仍需明显补强。	限定任务范围，明确接管人，不直接对外。
只作研究线索	法源未核验、覆盖不稳定，或只能给一般提示。	用于发散和检索线索，不作为法律结论。
暂不使用	站错立场、虚构、漏critical、越权处理材料，或修改反而伤害我方。	停止进入正式流程，记录失败条件，版本更新后再测。

为什么最终还需要法律AI工作台

聊天框给答案，Agent执行任务，工作台管理责任。它至少要把这些对象放进同一事项：

委托立场 → 材料版本 → 已知/未知事实 → issue与证据 → 法律依据 → 修改方案 → 谈判授权 → 责任人和期限 → 残余风险接受 → 最终签发

最后，回到这份监测的意义

法律 AI 监测不是替你追踪更多新闻，也不是替你决定购买哪一款工具。它要提供的是一套可以反复使用的判断方法：面对下一次模型更新、Agent发布或工作台升级，你知道怎样测试、怎样看证据，也知道什么情况下必须由人接管。

拿到就会测，会测以后，才知道怎样真正把 AI 用进法律工作。

本报告结果仅代表2026年9月2日特定产品版本、任务材料、工具权限与运行条件下的表现，不构成永久排名或产品背书。资料仅供学习、研究与内部测试，不构成法律意见；正式使用前应结合具体事实、适用法域、当时有效要求和组织权限人工复核。