AD-Linkage

AD-LINKAGE LIMITED | FRONTIER AI FIELD REPORT

Kimi K3
完整評測報告

模型架構 × 算力壓力 × 五項實際任務 × 使用成本 × 企業選型

由『暫停新訂閱』切入，拆解 K3 的稀疏 MoE、長程 Agent 能力、
真實資源消耗與 AD-Linkage 五項工作結果。

5 / 5	3 / 3	0 / 5
有可檢視成果	網站／遊戲遇部署問題	首次交付可直接上線

51.38% 深入評估後的 Vivace 帳戶月度總用量
用量訊號不等於單項任務成本或 API token 帳單

公開完整版 | 評估更新：2026.07.22

本報告區分官方資料、獨立評測、AD-Linkage 實際任務紀錄與分析判斷；
不把官方 Demo 當作本地實測。模型、價格、評測及供應狀態會變動。

01 / EXECUTIVE SUMMARY

先看結論：K3 很強，但價值在「推進工作」而非「一次完成」

本報告的核心不是判定 K3
是否「最強」，而是判斷它能否在真實工作中交付、哪一段最有價值、哪一段仍要由人承擔。

5 / 5	3 / 3	0 / 5
五項任務有可檢視成果	網站／遊戲遇 Preview 或部署問題	首次交付可直接 Production

四個答案

- 能力：K3 已進入前沿模型選型範圍，尤其適合長程 Coding、視覺回饋式開發、Agentic 工作及大規模知識任務。
- 算力：16／896 Experts 的稀疏啟動降低每步計算，但 Coding、Swarm 與長任務會造成大量重複模型調用；高效率不等於低總負載。
- 實測：K3 在五種不同交付類型都能由零推進至可檢視成果；真正失分集中在資料完整度、部署、技術格式、節奏與最後 QA。
- 採用：把 K3 當作高能力執行者，而不是無人監督的完成判斷者；企業 KPI 應以每個驗收合格交付物的總成本計算。

一句話判斷

K3 最有說服力的能力，是可以長時間持續做事；最需要治理的風險，是它亦可能在「尚未真正完成」時交付一個看起來已完成的成果。

02 / CAPACITY EVENT

暫停新訂閱：這不是停服，而是一次公開的容量壓力測試

Moonshot 表示，K3 推出後 48 小時內需求超出預測並接近現有運算集群上限，因此暫停個人用戶新訂閱，優先保障現有付費用戶。[S1][S2]

時間	已確認事件	企業含意
2026-07-16	K3 上線：2.8T、1M context、原生視覺、Agentic Coding。	市場開始以完整工作交付評價模型。
其後 48 小時	用戶請求大幅超出預測，接近現有集群上限。	容量規劃未能即時匹配需求峰值。
2026-07-20	暫停個人新會員訂閱；現有付費用戶優先。	可用性與穩定性成為產品選型的一部分。
未定日期	增加容量後分批重開；未公布確實日期。	不應把「即將重開」寫成確定承諾。

官方已確認

暫停的是 individual new memberships，不是 K3
全面停服；現有付費用戶不受影響，新增名額會按容量分批恢復。[S1][S2]

本報告的推論

K3 的 Agent、Coding 及 Swarm 工作會反覆推理、工具調用、讀取畫面與修正，所以一次任務可能包含大量模型調用。這是暫停事件與 K3 產品形態之間的合理技術解釋，但並非 Moonshot 公開披露的 GPU 數量或完整容量模型。

另一項市場訊號

Reuters 報道 Moonshot 計劃把未來會員拆分為一般使用與 Coding 方案，以更精準分配算力。舊五級價目仍可作功能參考，但新用戶不應假設原方案必定原樣重開。[S2]

03 / MODEL CARD

K3 模型資料：2.8T 是容量，16／896 才接近每步啟動規模

K3 的核心不是單一大數字，而是超大總容量、極稀疏專家啟動、長 context、原生視覺與 Agent harness 的組合。[S3]

2.8T
總參數

1M
Context window

16 / 896
每次有效啟動 Experts

項目	K3 資料	實際含意
輸入	文字＋圖像	可閱讀 screenshot、介面、圖表及視覺參考。
輸出	文字／Code／工具行動	視覺理解不代表直接生成可用 3D 檔案。
Attention	KDA＋Attention Residuals	支援長序列與跨深度資訊取回。
MoE	Stable LatentMoE，16／896 Experts	總參數龐大，但每步只啟動少量專家。
量化	MXFP4 weights＋MXFP8 activations	降低記憶體與服務成本，仍需超大型硬件。
定位	長程 Coding、知識工作、推理	更像持續執行的工作系統，不只是 Chatbot。

開放權重時間點

Kimi 在 7 月 16 日公布會在 2026-07-27 前發布完整權重；截至本報告資料截點，Artificial Analysis 仍標示權重未公開。應區分「宣布將開放」與「已可下載部署」。[S3][S6]

04 / ARCHITECTURE

K3 為何能把 2.8T 變成可服務產品：六層效率設計

超大型稀疏 MoE

要真正可用，必須同時解決長序列、深層資訊、專家路由、訓練穩定、低精度及推理服務。

技術	做什麼	企業含意
Sparse MoE	896 個 Experts 中每次有效啟動 16 個。	保留巨大能力容量，同時控制單步計算。
KDA	以 Kimi Delta Attention 改善長序列效率。	支援 1M context 與長程工作。
AttnRes	選擇性取回不同深度的表示。	深層資訊較不易在累積中稀釋。
Stable LatentMoE	穩定極高稀疏度下的專家路由。	降低訓練與服務失衡風險。
Quantile Balancing	由 router score 分位數分配專家。	避免敏感的手工平衡超參數。
MXFP4／MXFP8 QAT	由 SFT 階段開始量化感知訓練。	降低記憶體，但不代表一般電腦可部署。

官方將 K3 相對 K2 的整體 scaling efficiency 提升描述為約 2.5 倍，來源包括架構、訓練及資料配方的共同改進，而不是單一 MoE 技巧。[S3]

不要誤讀

「每次只啟動 16 個 Experts」不等於只需小模型級硬件。所有權重、路由、跨設備通訊、KV cache、context 與吞吐仍然形成巨大部署成本。

05 / SERVING ECONOMICS

高效率仍會爆算力：模型單步效率與服務總負載是兩件事

K3 把推理成本壓低，但產品把一次問答擴展為長時間、多工具、多輪的完整任務；兩股力量會同時存在。

層次	降低成本的設計	增加負載的行為
模型	16／896 Experts、低精度權重及 activations	2.8T 權重與跨設備專家通訊仍然龐大
Attention	KDA 與 prefix cache	1M context、長 session 與 cache invalidation
服務	Mooncake 分離 prefill／decode、重用前綴	大量同時用戶及突發需求
Agent	工具可替代部分純 token 推理	反覆規劃、工具調用、畫面檢查與重試
Swarm	可平行縮短 critical path	多 sub-agents 同時消耗 credits 與模型調用

>90%

官方稱 Coding API cache hit

64+

官方建議 Supernode accelerators

4,000+

Swarm 每任務工具調用上限

推論公式

總服務負載 ≈ 同時任務數 × 每項任務模型調用次數 × 每次 context／output × 工具與重試循環。MoE 降低其中一部分成本，但 Agent 產品會放大任務長度與調用次數。

06 / CAPABILITY

能力位置：57
分已屬前沿，但速度、冗長度與體驗仍有差距

Artificial Analysis v4.1 同一獨立評測框架下，K3 57、GPT-5.6 Sol 59、Claude Fable 5 60；三者均為 1M context。[S6][S7][S8]

模型	AA Index	速度	輸出量	Input／Output
Claude Fable 5*	60	71.7 t/s	87M	US$10／US$50
GPT-5.6 Sol max	59	63.3 t/s	70M	US$5／US$30
Kimi K3	57	35.2 t/s	130M	US$3／US$15

K3 的高價值任務

- 長程 Coding、repository／terminal 工具與多輪工程工作。
- Screenshot → Code → Preview → 再修改的視覺閉環。
- 長文件、研究、報告、批量資訊與多 Agent 知識工作。

比較限制

- Kimi 官方 benchmark 會使用 KimiCode、Claude Code 或 Codex 等不同 harness。
- Fable 評測可能包含 fallback；不同模型的 reasoning effort 亦可能不同。
- 獨立綜合分數建立候選名單；企業採購仍要用同一任務、同一工具與同一完成標準做 Pilot。

07 / COST

標價比 Sol 低，不代表每項工作一定平一半

真正成本取決於 input、output、cache、工具、重試、等待及人工覆核。K3 的低單價可能被較長輸出及多輪任務部分抵消。

模型	Cache hit	Input／1M	Output／1M	100k in＋20k out
Kimi K3	US$0.30	US$3	US$15	US$0.60
GPT-5.6 Sol	US$0.50	US$5	US$30	US$1.10
Claude Fable 5	US$1.00	US$10	US$50	US$2.00

US$2,709.75

K3 完整 AA 評測成本

US$2,824.18

Sol 完整 AA 評測成本

US$5,630.52

Fable 完整 AA 評測成本

固定 token 例子中，K3 比 Sol 低約 45%，比 Fable 低 70%。但完整獨立評測中，K3 因輸出 130M tokens，總成本只比 Sol 低約 4%，比 Fable 低約 52%。[S6][S7][S8]

企業成本 KPI

每個驗收合格交付物總成本 = 模型 + 工具 + 重試 + 等待 + 人工覆核 + 失敗風險。不要以每百萬 token 價格代替工作 ROI。

08 / TEST DESIGN

五項實際任務：我哋實際計咗啲乜？

每項任務都唔同，所以本報告評估的不是 K3 一次答對幾多題，而是佢能否交出可以打開、操作、檢查同正式使用嘅成果。

評估步驟	簡單解釋	我哋睇咩證據
明唔明任務	有冇掌握功能、資料、格式同合格標準？	計劃、任務拆解、交付規格
第一次交咩	第一次交付係咪打得開、撳得到、睇得到？	網站、Preview、檔案、影片
技術係咪啱	資料、部署、檔案格式同跨裝置有冇符合要求？	資料檢查、部署測試、檔案驗證
改唔改得到	收到具體回饋後，能否修正並再次驗證？	修正前後版本、重測結果
可唔可以直接用	需唔需要大改，先可以正式發布或交畀客戶？	發布前驗收清單

5／5、3／3、0／5點樣睇？

5／5：五項任務都有可檢視成果。
3／3：三項涉及網站或遊戲部署，三項都出現 Preview 或部署問題。
0／5：五項都未能第一次交付就直接正式發布。

呢組數字唔代表乜？

今次只測咗五個、而且每個難度都唔同嘅真實專案。所以唔可以將佢講成『K3 成功率 100%』、『部署失敗率 100%』或所有 K3 任務都會有同樣結果。

09 / RESULTS MATRIX

五項結果矩陣：K3 的能力集中在推進，缺口集中在最後一公里

所有任務均產出可檢視成果；所有首次交付亦均需要人手指出關鍵問題或完成最後修正。

任務	交付結果	主要問題	判斷
世界盃分析網站	完成網站介面、比賽頁及分析框架	首次部署缺實際數據與分析內容	前端快；資料型產品須另設內容驗收
智能行程 Planner	可操作介面、條件、路線及選項	近 8 小時；近三成用於部署排錯	功能可做；基建摩擦直接影響 ROI
Day 1-90 3D 遊戲	90 日進度、狀態、互動邏輯及場景	完成後卡在 Preview／部署	複雜原型強；上線仍需工程驗收
3D 角色與素材	約 2 小時產出 20+ 視覺素材	不是要求的 .glb 真正 3D 模型	視覺產量高；技術格式必須驗收
多模態影片剪輯	最終輸出 177.6 秒完整影片	多輪重剪、字幕、記憶體及暫存問題	可做剪輯協作者；導演與 QC 仍要人

5 / 5
可檢視成果

3 / 3
網站／遊戲部署問題

0 / 5
首次 Production-ready

跨任務模式

介面與可視成果往往先出現，資料、部署、檔案規格及發布級品質其後才暴露。驗收設計必須針對「看起來完成，但技術上未完成」。

10 / TASK 1

世界盃比賽分析網站

測試目標是由需求建立可瀏覽、可部署、可呈現比賽資訊及分析的網站，而不是只生成靜態 mockup。

交付	核心缺口	首次上線
已部署網站框架	實際數據不足	不合格

實際交付

- 建立完整網站介面、比賽頁、導覽與分析呈現框架。
- 能把需求快速轉成具體前端頁面及可檢視部署版本。

人手驗收發現

- 首次部署主要完成外殼，實際比賽數據與分析內容不足。
- 如果只看畫面，容易誤判為「網站已完成」；資料型產品必須抽查內容覆蓋。

AD-Linkage 判斷

- K3 的前端生成速度有價值，但資料完整度必須成為獨立 publish gate。
- 建議加入最少資料筆數、欄位非空率、來源證據與隨機頁面抽查。

證據與限制

有網站與人手驗收結果；未有一致保存的 token、credits 及逐輪時間，因此不作成本數字推算。

11 / TASK 2

智能行程／路線 Planner

任務要求整合多種條件、路線、選項與互動介面，屬於規則、資料、UX 及部署交叉的決策支援產品。

端到端	部署排錯	交付
接近 8 小時	接近 30%	可操作 Planner

實際交付

- Multi-Agent 模式完成可操作 Planner。
- 能把條件、路線、不同活動選項與互動介面整合成具體產品。

人手驗收發現

- 端到端工作接近八小時，其中接近三成時間用於部署與排錯，而非核心功能。
- 時間節省不是由模型能力單獨決定；Preview、環境及部署摩擦會吞噬 Agent 優勢。

AD-Linkage 判斷

- K3 能處理複雜產品鏈，但企業 ROI 必須記錄「功能工時」與「基建排錯工時」。
- 建議預先固定 runtime、deployment target、資料 schema 及健康檢查。

證據與限制

時間與比例為團隊操作紀錄中的近似值；沒有統一的模型 token／credits 對照，不能把八小時全部歸因於模型推理。

12 / TASK 3

Day 1-90 的 3D 創業遊戲

任務要求 90 日進度、狀態變化、互動邏輯、前端及 3D 場景，而不是單一畫面或短 Demo。

範圍	製作	狀態
Day 1-90	數小時	可玩原型

實際交付

- 在數小時內建立 90 日進度、狀態變化、互動邏輯及 3D 場景原型。
- 能由長規格持續推進到可玩的複雜互動成果。

人手驗收發現

- 核心遊戲完成後仍卡在 Preview 及部署。
- 「程式已寫完」不等於「用戶可打開並穩定遊玩」；跨裝置與效能亦未在首次交付完成。

AD-Linkage 判斷

- K3 由概念到複雜互動原型的能力強，特別適合產品探索。
- 正式上線需要獨立工程驗收：build、runtime error、FPS、裝置、存檔及回歸測試。

證據與限制

「數小時」是操作紀錄的量級，不代表統一 benchmark；本報告沒有足夠資料比較同一遊戲由其他模型完成所需時間。

13 / TASK 4

3D 角色與素材生成

任務最能顯示視覺成果與技術交付物之間的差異：畫面像 3D，不等於可在 3D pipeline 使用。

製作	產量	要求格式
約 2 小時	20+ 素材	.glb 未交付

實際交付

- 約兩小時產出 20 多個可視素材。
- 角色外觀、場景方向與視覺概念足以支援後續設計討論。

人手驗收發現

- 交付物不是要求中的真正 .glb 模型。
- 缺少可驗證的 mesh、rig、material、animation 與 3D 檔案結構；「看起來像 3D」不能代替資產格式。

AD-Linkage 判斷

- K3 適合快速探索視覺方向及大量概念變體。
- 若目標是可匯入 Blender／Unity／Unreal 的資產，必須以實際檔案開啟、節點、骨架、貼圖與動畫作驗收。

證據與限制　20+ 指可視素材數量，不應寫成 20+ 個真正 3D 模型；本報告刻意保留這個差異。

14 / TASK 5

多模態影片剪輯

任務包含素材理解、片段排序、字幕、修改、重新輸出與技術排錯，評估是否能持續作為剪輯協作者。

最終成片	修正	角色
177.6 秒	多輪	AI 剪輯協作者

實際交付

- 最終輸出 177.6 秒完整影片。
- 能按回饋重組片段、修正字幕及重新輸出，證明任務可跨多輪持續。

人手驗收發現

- 需要多輪重剪、字幕校準及人工節奏判斷。
- 過程遇到記憶體不足及暫存空間問題；完成不是一鍵生成。

AD-Linkage 判斷

- K3 可承擔素材理解、結構初稿、機械修正與持續輸出。
- 品牌節奏、情緒、鏡頭取捨、字幕安全區與發布級 QC 仍應由人作導演。

證據與限制

177.6 秒是最終輸出時長，不代表模型只處理 177.6 秒素材，也不能直接轉換為人工節省時數；本報告沒有完整剪輯工時基線。

15 / CROSS-TASK FINDINGS

跨任務診斷：最常見的不是「完全做不到」，而是「完成判斷過早」

五項任務呈現一致模式：K3 能快速製造可見進度，惟真正的發布風險藏在資料、環境、格式與品質標準。

風險類型	在測試中的表現	建議驗收
資料完整度	世界盃網站有外殼，但內容與分析不足	筆數、欄位、來源、抽樣頁面
部署／Preview	Planner、3D 遊戲均有部署摩擦	health check、runtime logs、跨裝置
技術格式	3D 視覺素材不是 .glb 模型	檔案開啟、schema、mesh／rig／materials
品牌品質	影片需多輪節奏與字幕修正	人工導演、字幕安全區、發布清單
資源限制	長任務出現記憶體與暫存問題	資源上限、checkpoint、可恢復流程

最適合 K3 的工作

- 由零到第一個可檢視版本：網站、工具、互動原型、研究與影片結構。
- 需要長時間、跨檔案、工具與視覺回饋的任務。
- 可逆、可測試、可由人驗收的工作。

不應完全自動化的工作

- 涉及付款、對外發布、刪除、合規或不可逆操作。
- 缺少明確資料來源、技術規格或完成證據的工作。
- 只靠畫面「似完成」就判定成功的交付。

16 / MODE & MEMBERSHIP

Kimi 各種使用方式：點揀先啱？

個人新會員訂閱目前暫停。先按任務需要選 Agent、Swarm、Code 或 API；以下月費表只係 K3 推出後、調整前原有個人方案的參考資料。

方式	適合咩工作	簡單選擇方法
Low／High／Max	同一任務的推理深度	先用 High；做唔到或 debug 困難先升 Max
K3 Swarm	大量搜尋、獨立章節、真正可平行工作	工作互相依賴就唔好硬拆；先定共用格式
Kimi Code	Repository、terminal、IDE、測試及長時間 Coding	保留版本、測試結果同權限限制
Kimi API	產品整合、批量流程、模型路由及日誌	設 token 上限、fallback、監控同人工升級

K3 推出後、調整前的原有個人月費方案

即 Kimi 官方所稱的 legacy plans：已訂閱會員可以繼續使用同續期；新客現時未能購買。60／25 代表介面列出的 Agent 額度 60、Swarm 額度 25，唔等於保證可以完成 60 或 25 個完整專案。[S4]

方案	月費	Agent／Swarm 額度	適合邊類使用
Moderato	US$19	60／25	低頻試用，先驗證 K3 是否適合自己
Allegretto	US$39	150／50	經常用長 context／HighSpeed 的起點
Allegro	US$99	360／120	高頻使用及最多 4 個 Agent 並行
Vivace	US$199	720／240	重度 Swarm 及最多 8 個 subtasks

Swarm 幾時用？

只有可以同時做、而且唔需要輪住等上一步結果嘅工作，Swarm 先可能慳時間。高度相依工作會增加整合時間同 credits，未必更快。

17 / ENTERPRISE PILOT

企業點樣驗收 K3：唔好淨係睇佢話『完成』

開始前先寫清楚咩叫合格。K3 可以自主處理草稿、分析、Coding 等可修改工作；付款、發布、刪除及其他高風險行動，最後仍由人批准。

01 首次合格率

第一次交付有幾多項毋須大改已經達標？

02 人手覆核時間

同事實際要花幾多分鐘檢查同修正？

03 總完成時間

由開始到驗收合格，總共用咗幾耐？

04 合格成果總成本

月費／API、credits 同人手時間加埋幾多？

05 證據完整度

有幾多內容有來源、測試或檔案支持？

06 出錯後修復率

遇錯誤後能否修好、重跑並交回合格版本？

每個 Agent 任務開始前，至少寫清楚五件事

- 做畀邊個、要交咩、點先叫合格：唔好俾模型將『有輸出』當成『已完成』。
- 可以用邊啲資料：核實唔到就寫待確認，唔可以用空白或假資料補位。
- 邊啲行動要人批准：付款、發送、公開發布、刪除同高風險操作。
- 邊啲工作可以同時做：共用資料格式、檔案負責人同合併方法要先定好。
- 幾時要停：預先設定 credits、時間、重試上限，並要求保存進度同完成證據。

公司應唔應該買？

用同一項真實工作，俾 K3、Sol、Fable 或其他候選使用相同輸入、工具、合格標準同輸出上限。只有質素、時間或總成本持續好過現有做法，先逐步增加資料、權限同自動化範圍。

18 / EVALUATION FRAMEWORK

本報告點樣評分 K3：由『有初稿』到『可直接發布』

一個畫面漂亮嘅 Preview 只代表有初稿，唔代表成品已經可交客。本報告先判斷交付去到邊一級，再計算人手、時間同 credits。

G1	G2	G3	G4	G5
有初稿可睇	真係操作到	符合要求	穩定部署	可直接發布
已有文字或 Preview	可以打開、撳、播放或運行	資料、功能、格式全部合格	正式環境可運行及重開	毋須重大修改即可交客

除咗交付層級，我哋再問八條實際問題

01 理解任務

有冇理解使用者、功能、資料、格式同不可妥協要求？

02 自主推進

能否自己跨步驟、檔案同工具持續做落去？

03 第一次交付

第一次交嘅版本係咪已經打得開同操作到？

04 技術符合

資料、部署、檔案格式同跨裝置係咪真合格？

05 修正能力

收到回饋後，能否搵到問題、修正同再測試？

06 運行穩定

Preview、runtime、記憶體、暫存同部署穩唔穩？

07 人手負擔

最終要產品、工程、內容、導演同 QA 幫幾多？

08 實際消耗

完成一件合格成果，用咗幾多 credits、時間同重試？

19 / EVIDENCE MATRIX

五項任務證據矩陣：強項係推進，主要失分係首次交付與最後一公里

以下為 AD-Linkage 操作判斷，不是通用模型 benchmark。G／A／R 代表在現有證據下的 Green／Amber／Red；「-」代表不適用或證據不足。

任務	成果	規格	部署	修正	人手	證據
世界盃分析網站	G	A	R	A	R	A
智能行程 Planner	G	G	R	A	R	G
Day 1-90 3D 遊戲	G	G	R	A	R	A
3D 角色與素材	G	R	-	A	R	G
多模態影片剪輯	G	A	A	G	R	G

5 / 5
可檢視成果
所有任務均到達 G1

3 / 3
部署風險
相關網站／遊戲均遇問題

0 / 5
首次 Production-ready
全部需要關鍵修正或 QA

最重要的評估結論

K3 並非「做不到」，而是經常把 G1／G2 的可見進度當作完成。企業真正需要管理的，是模型何時可以繼續自主，何時必須停下來交由資料、工程、格式或品牌驗收。

20 / FAILURE ANALYSIS

五種重複失敗模式：畫面越快出現，越容易過早宣布完成

缺陷並非隨機散落；它們集中在由「可見成果」跨到「可發布成果」的轉換位置。這些位置亦是企業應設 Gate 的地方。

01 外殼完整 ≠ 資料完整

世界盃網站

介面與頁面已存在，但實際數據及分析不足。

Data gate：欄位、來源、筆數、抽樣頁面

02 程式完成 ≠ 成功部署

Planner／3D 遊戲

Preview、runtime 或 deployment 仍未穩定。

Release gate：build、health check、logs、跨裝置

03 像 3D ≠ 真正 3D 資產

3D 角色與素材

交付 20+ 視覺素材，但不是要求的 .glb。

Format gate：檔案開啟、mesh、rig、materials

04 有成片 ≠ 品牌可發布

多模態影片

177.6 秒成片仍需多輪重剪、字幕與節奏校準。

Editorial gate：節奏、字幕安全區、品牌 QC

05 長時間持續 ≠ 資源穩定

多項長任務

記憶體、暫存、重試與部署會吞噬 Agent 優勢。

Ops gate：額度、checkpoint、恢復、停止條件

治理含意

如果沒有明確 Gate，AI 會以『已有輸出』取代『已通過驗收』。K3 最適合在可逆、可測試、有清晰停止條件的工作內取得自主權。

21 / HUMAN OVERSIGHT

人工介入不是例外：五項任務都需要人定義『真正完成』

K3 減少的是由零開始與機械執行的成本；它沒有消除產品判斷、工程驗證、專業格式與品牌責任。

人工角色	必須承擔的判斷	主要出現任務
產品負責人	定義需求、優先次序與不可妥協條件	全部任務
資料／領域 Reviewer	檢查數據、來源、分析覆蓋與合理性	世界盃網站
工程 Reviewer	處理 runtime、部署、效能與跨裝置	Planner／3D 遊戲
技術資產 Reviewer	驗證 mesh、rig、materials、格式與可匯入性	3D 素材
導演／品牌 Reviewer	決定鏡頭、節奏、字幕、情緒與發布品質	影片

≈ 8 小時
Planner 端到端
近三成為部署及排錯

≈ 2 小時
3D 素材
產出快，但格式不合格

177.6 秒
影片最終輸出
需要多輪人工導演與 QC

企業真正應量度的人工 KPI

- 首次交付後，需要幾多分鐘才可通過驗收？
- 每個缺陷需要幾輪回饋；由誰發現；是否可自動測試？
- 等待、排錯、重新部署及重新輸出的時間，有沒有抵消 AI 節省？

22 / CREDIT ECONOMICS

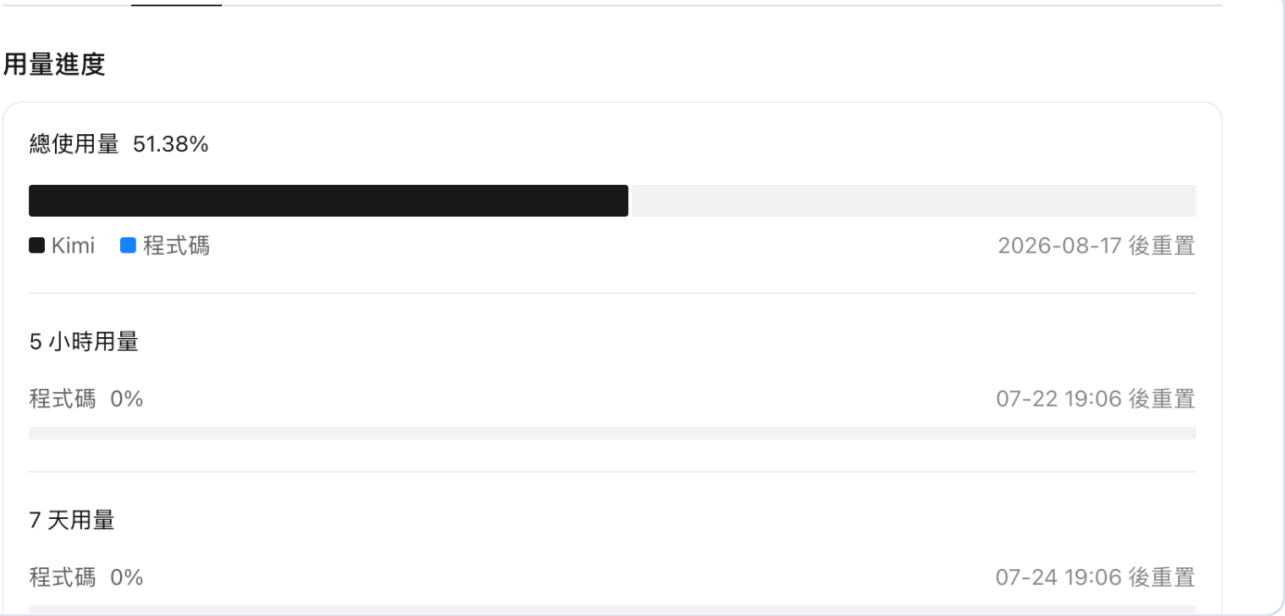
半個月度額度：低 Token 單價，未必等於低完整任務成本

用戶提供的 Vivace 用量截圖顯示，完成本輪深入測試後，帳戶總用量已達 51.38%。這是企業評估 K3 時不能忽略的資源訊號。

51.38%

Vivace 月度帳戶總用量

截圖：2026-07-22｜重置：2026-08-17



48.62%

帳戶餘額

低於本輪已用的 51.38%

≈ US$102.25

比例分攤參考

US$199 × 51.38%；不是 API 帳單

< 1 次

可重複性訊號

若下輪用量相同，餘額不足

證據邊界

截圖證明的是帳戶總用量，不是每項任務的逐筆消耗。51.38% 可作整輪測試的資源壓力訊號；不能平均分配到五項任務，也不能把 US$102.25 當成實際 API 費用。完整任務成本仍須加入人工、工具、等待、重試與失敗風險。

23 / OPERATIONAL VERDICT

最終評價：高能力 Prototype Accelerator，低自主 Production Readiness

下面分數是 AD-Linkage 根據五項任務作出的營運評分，不是統計 benchmark；用途是協助決定應否把更多資料、權限與預算交給 K3。

由零推進可見成果	五項任務全部有可檢視成果	4.5 / 5
長程持續與修正	能跨多輪、工具及視覺回饋繼續工作	4.0 / 5
首次交付正確性	0／5 可直接作 Production 成品	1.5 / 5
部署與運行可靠性	3／3 相關網站／遊戲遇部署問題	2.0 / 5
技術規格符合	資料、格式與發布要求需人工把關	2.5 / 5
人工覆核負擔	全部任務都需要關鍵人手判斷	2.0 / 5
額度／資源效率	本輪後月度總用量達 51.38%	2.0 / 5

採用結論

值得用，但使用角色必須正確：K3 應作高能力執行者、研究與 Prototype 加速器；未應在沒有 Release Gate 的情況下直接發布、部署或交付客戶。

建議繼續投入

研究、網站／工具 Prototype、長程 Coding、視覺回饋式開發、可逆而且有測試的工作。

暫不應全自動

對外發布、付款、刪除、合規、高風險部署，以及沒有明確完成證據的工作。

24 / SOURCES & DISCLOSURE

資料來源、證據層級與披露界線

官方資料回答模型與供應安排；獨立評測回答同一框架下的能力、速度及成本；AD-Linkage 任務與用量證據回答真實交付及資源壓力。三者不能互相代替。

[S1] Kimi Help Center - Moon Companion Plan

新個人會員暫停、容量增加與分批恢復。

[S2] Reuters - Moonshot pauses subscriptions

48 小時需求、現有會員優先及未來方案拆分。

[S3] Kimi K3 Official Tech Blog

2.8T、1M、16／896、架構、價格及部署。

[S4] Kimi Membership Pricing

舊會員方案、Agent／Swarm／Code 配額。

[S5] Kimi Agent Swarm

300 sub-agents、4,000+ 工具調用及 credits。

[S6] Artificial Analysis - Kimi K3

57 分、速度、輸出量、成本與 1M context。

[S7] Artificial Analysis - GPT-5.6 Sol

59 分、速度、輸出量、價格及總成本。

[S8] Artificial Analysis - Claude Fable 5

60 分、速度、輸出量、價格及總成本。

[S9] 用戶提供 Vivace 用量截圖｜2026-07-22

帳戶總使用量 51.38%，2026-08-17 重置。

披露界線

- 五項結果來自團隊操作紀錄與人手驗收；不包含私人對話、完整 Prompt、客戶資料或未經同意的內部畫面。
- 51.38% 是帳戶總用量；沒有逐項 credit ledger，因此不作單項任務成本或模型耗用比例的虛假拆分。
- 營運評分是基於這批任務的管理判斷，不可外推為所有 K3 任務的固定成功率。