IN-DEPTH ANALYSIS · 4 MODELS · 26 SCENARIOS

OpenAI, Anthropic, Perplexity & Kimi K3 for finance & ops.

Moonshot AI shipped Kimi K3 on July 16, 2026 — a 2.8-trillion-parameter, open-weights-pending model that claims frontier-class benchmarks at a fraction of Western pricing.[1] This brief compares it against OpenAI's GPT-5.6 family, Anthropic's Claude Fable 5 / Opus 4.8, and Perplexity's Sonar / Comet Enterprise stack across the finance and operations workloads that actually matter to a mid-market SME: close, FP&A;, KYC, procurement, supply chain, and board reporting. No single vendor wins every task — and one of the four still carries a compliance flag that rules it out for regulated data. See Perplexity's enterprise positioning for context on where research-grade AI fits into this stack.

MODELS COMPARED	SCENARIOS MAPPED	CLEAR SINGLE WINNER	TIME TO FIRST ROI
4	26	0/4	3–6 mo
Frontier tier, July 2026	Finance + operations	No — task-dependent	Pilot to measurable gain

KEY TAKEAWAYS

1. Claude leads regulated finance work — close, KYC, pitchbooks — on accuracy and retention.
2. OpenAI's Frontier and Codex fit cross-system operations orchestration best.
3. Perplexity is the research layer, not a modeling engine — pair it, don't swap it in.
4. Kimi K3 is cheap and real, but its China-hosted API isn't fit for regulated data yet.

RECOMMENDED BY SME SIZE

MICRO <$1M revenue
Claude Pro/Team + Perplexity Pro for research.

SMALL $1–10M revenue
Claude for finance + OpenAI API for ops + Perplexity Pro.

MEDIUM $10M+ rev, 10+ staff
Claude Enterprise + OpenAI Frontier + Perplexity Ent. Pro.

1. VentureBeat, Kimi K3 launch

SECTION 01 · PROVIDER SNAPSHOT

Four labs, four very different bets.

Flagship model, release cadence and one-line positioning for each vendor in this brief.

01

OpenAI — GPT-5.6 (Sol / Terra / Luna) + Frontier

OPENAI

Three separate models, not modes: Sol is the flagship for analysis and agents, Terra is the balanced enterprise tier, Luna handles high-volume classification. Frontier is the orchestration layer for wiring agents into CRMs and data warehouses.[1] TTMS

02

Anthropic — Claude Fable 5 & Opus 4.8

ANTHROPIC

Fable 5 is the top-of-line reasoning model with long-horizon autonomy; Opus 4.8 is the generally-available workhorse with a zero-data-retention option Fable 5 lacks.[2] Anthropic

03

Perplexity — Sonar & Comet Enterprise Pro

PERPLEXITY

Not a modeling engine — a research and real-time-data layer with citations built in. Comet for Enterprise Pro (launched July 8, 2026) adds agentic browsing on top.[3] Perplexity

04

Kimi K3 — Moonshot AI (Beijing)

KIMI K3

Released July 16, 2026. 2.8T-parameter sparse MoE model matching frontier benchmarks at a fraction of the price — but hosted in China, which matters for any regulated financial data.[4] CNBC

1. TTMS, GPT-5.6 overview 2. Anthropic, Opus 4.8 3. Perplexity blog 4. CNBC, Kimi K3

SECTION 01 · KIMI K3 SPOTLIGHT

The model that triggered this brief, unpacked.

What actually changed on July 16, 2026 — and what it means before you route real budgets through it.

05

What's new

SPEC

2.8T parameters, 16-of-896 expert sparse MoE, 1.05M-token context, native vision, an always-on 'thinking mode' that can't be switched off. Kimi Delta Attention claims ~2.5x scaling efficiency and up to 6.3x faster decoding vs K2.[1] The AI Rankings

06

Why it's not just hype

BENCHMARKS

Intelligence Index 57.1 — 4th place, ahead of Claude Opus 4.8's 55.7. #1 on Frontend Code Arena. Cost per task ~$0.94, about half Opus 4.8's $1.80 and in line with GPT-5.6 Sol.[2] Toms Hardware

07

The catch, for an SME

CATCH

Open weights promised by July 27, 2026 under a Modified MIT license — not yet shipped as of this writing. The bundled web-search tool is flagged 'not recommended for production,' and hallucination rate has reportedly risen vs K2.7.[3] Gamine AI review

08

Verdict for finance & ops

VERDICT

Useful today only for non-sensitive, high-volume, non-regulated tasks (draft generation, internal search, coding glue) where the China-hosted API's data terms aren't a blocker. Revisit once self-hosted open weights actually ship.[4] Layer3Labs

1. The AI Rankings, Kimi K3 2. Tom's Hardware 3. Gamine AI review 4. Layer3Labs compliance review

SECTION 02 · CAPABILITIES MATRIX

Where each model actually wins.

Six capability dimensions that decide finance & operations outcomes, mapped to today's leader.

09 — Research & real-time market data — PERPLEXITY

Citation-backed answers, live quotes, SEC table extraction. No other vendor here treats sourcing as a first-class feature.

10 — Financial modeling & FP&A; — CLAUDE

Top score on the Hebbia Finance Benchmark; Fable 5 handles multi-step forecast logic with fewer silent errors.[1] Anthropic

11 — Document & compliance review — CLAUDE

Native vision for diagrams, charts and tables inside PDFs — built for contract, KYC and audit-trail review.

12 — Agentic process automation — OPENAI

Codex reads screens and drives Excel, Outlook and Salesforce directly; Frontier schedules recurring cross-system agents.[2] OpenAI

13 — Coding & data pipelines — OPENAI

GPT-5.6 Sol plus the Agent SDK is the default choice for ETL, ERP integration glue and supply-chain copilots.

14 — Cost efficiency — KIMI K3

Lowest cost-per-task of the four at frontier-adjacent quality — if the China-hosting risk is acceptable for the workload.[3] The AI Rankings

1. Anthropic, Claude Fable 5 2. OpenAI, Frontier 3. The AI Rankings, cost per task

SECTION 03 · FINANCE SCENARIOS

Six finance jobs, six honest picks.

Recommended engine per workflow — not the vendor with the loudest press release.

15 — **Month-end close & consolidation** — CLAUDE

Anthropic's ready-made close template plus sub-agent delegation for reconciliation steps.[1]

Anthropic finance agents

16 — **AP/AR automation & 3-way match** — OPENAI

Codex-driven document capture plus Frontier orchestration across ERP and email; Operator-tier flows report large drops in manual reconciliation.

17 — **FP&A; & rolling forecasts** — CLAUDE

Fable 5's long-horizon reasoning holds up best on multi-scenario forecast logic and variance narratives.

18 — **KYC/AML & compliance screening** — CLAUDE

Purpose-built KYC agent template; SOC 2, ISO 27001/42001 and HIPAA BAA already in place.[2]

Anthropic

19 — **Fraud & anomaly detection** — HYBRID

Claude flags the anomaly and explains it; Perplexity corroborates against external filings and news in the same session.

20 — **Board reporting & pitchbooks** — HYBRID

Claude drafts and formats the deck; Perplexity supplies sourced market comparables and live figures to slot in.

1. Anthropic, Finance Agents 2. Fortune, Anthropic on Wall Street

SECTION 04 · OPERATIONS SCENARIOS

Six ops jobs, where orchestration wins.

Operations is mostly about wiring systems together — the category OpenAI has invested in hardest.

21 — **Supply chain visibility & demand forecasting** — OPENAI

Agent SDK + Databricks MCP cookbook covers inventory, demand and disruption queries out of the box.[1] Beam.ai

22 — **Procurement-to-pay automation** — OPENAI

Operator-tier flows automate end-to-end procurement; Stripe reported over 60% less manual reconciliation after adoption.

23 — **Inventory & ERP copilot** — OPENAI

Codex operates Excel and native ERP screens directly via clicks and keystrokes — no API integration required to start.

24 — **Customer ops & support automation** — HYBRID

GPT-5.6 Luna handles high-volume triage and routing cheaply; escalate anything financially sensitive to Claude.

25 — **Vendor management & contract review** — CLAUDE

Document vision on Claude reads redlines, tables and clause changes across contract versions reliably.

26 — **Cross-system exception handling** — OPENAI

Frontier's 90+ launch plugins (Jira, MS365, Salesforce, HubSpot) make it the practical default for routing exceptions across tools.

1. Beam.ai, enterprise agents 2026

SECTION 05 · PRICING SNAPSHOT

What it actually costs, per token.

List API pricing as of July 2026. Real spend depends on caching, thinking-mode use and context length — treat this as a floor, not a forecast.

Provider	Flagship model	Input $/M	Output $/M	Context	Notes
OpenAI	GPT-5.6 Sol	$5.00	$30.00	400K	Terra/Luna cheaper tiers available
Anthropic	Claude Fable 5	$10.00	$50.00	1M	Opus 4.8: $5 / $25, ZDR option
Perplexity	Sonar (Enterprise Pro)	$40/seat/mo	—	—	Seat-based, not token-metered
Kimi K3	Kimi K3	$3.00 / $0.30 cache	$15.00	1.05M	~5x pricier than K2.7 predecessor

TYPICAL SME MONTHLY STACK SPEND

Micro (<$1M rev)
$40-80/mo
Claude Pro/Team + Perplexity Pro, per seat.

Small ($1-10M rev)
$400-1,500/mo
Mixed seats + light API usage for ops automation.

Medium ($10M+ rev)
$3,000-15,000+/mo
Enterprise seats + Frontier/Agent SDK API spend at volume.

1. TTMS, GPT-5.6 pricing 2. Caylent, Opus 4.8 3. Perplexity plans 4. Tom's Hardware, Kimi K3 pricing

SECTION 05 · COST & ROI

Cheap tokens aren't the same as ROI.

Per-task cost is a footnote next to integration effort, verification overhead and compliance risk.

27

Cost per completed task

Kimi K3 ~$0.94, GPT-5.6 Sol ~$1.04, Claude Opus 4.8 ~$1.80 per benchmarked agentic task — but these figures ignore integration and review time.[1] The AI Rankings

28

Where ROI actually shows up first

AP/AR automation and month-end close see measurable time savings within one quarter; FP&A; and board reporting gains compound over 2-3 quarters as templates mature.

29

The hidden cost line: verification

Every model still hallucinates. Budget a human-review step for anything that reaches a regulator, auditor or board — this is the real cost driver, not token price.

1. The AI Rankings, cost per task

SECTION 06 · RISK & COMPLIANCE

The part everyone skips until the audit.

Five compliance dimensions that matter more than any benchmark score for regulated SME data.

30 — **Data residency** — KIMI K3

Beijing-hosted, Alibaba-backed. Flagged as unsafe for regulated financial data under Chinese-law jurisdiction until self-hosted weights ship.[1] IAPS

31 — **Retention policy** — CLAUDE

Opus 4.8 offers zero data retention; Fable 5 requires a 30-day retention window — a hard blocker for some regulated contexts.

32 — **Certifications** — ALL FOUR

OpenAI and Anthropic both hold SOC 2, ISO 27001 and sector add-ons (HIPAA BAA). Perplexity Enterprise adds SSO and admin controls. Kimi K3 publishes none of this yet.

33 — **Hallucination & verification** — KIMI K3

Reported hallucination rate rose vs. K2.7; the bundled search tool is explicitly marked not production-ready.[2] Gamine AI

34 — **Open-weights self-hosting** — KIMI K3

Modified MIT license promised by July 27, 2026 — the only realistic path to using K3 on sensitive workloads is self-hosting once weights actually ship.

1. IAPS, Kimi Claw risks 2. Gamine AI, Kimi K3 review

SECTION 07 · RECOMMENDATIONS

Buy for the size you are, not the demo.

Enterprise platform pitches assume enterprise headcount to run them. Most mid-market firms don't have it yet.

35

Micro — under $1M revenue, 1-5 staff

MICRO

Claude Pro/Team (Sonnet 5) for finance drafting + Perplexity Pro for research. Skip enterprise platforms and Kimi K3 entirely — the overhead isn't worth it at this scale.

36

Small — $1-10M revenue

SMALL

Claude (Opus 4.8/Sonnet 5) for finance workflows + Perplexity Pro/Enterprise Pro for research + OpenAI GPT-5.6 API (Terra/Luna) for lightweight ops automation glue.

37

Medium — $10M+ revenue, 10+ staff

MEDIUM

Claude Enterprise with finance templates + OpenAI Frontier/Agent SDK for cross-system ops orchestration + Perplexity Enterprise Pro for continuous market intelligence. Add Kimi K3 (self-hosted, post-July 27) only for internal, non-regulated workloads.

SECTION 07 · BY INDUSTRY VERTICAL

Manufacturing isn't fintech. Stack accordingly.

The verticals DJV Consulting works across most, mapped to the stack that fits their actual workflow.

38

Manufacturing & automotive parts

OpenAI-led (ERP/MES integration, supplier orchestration) + Claude for cost accounting and standard-cost variance analysis.

39

Logistics & supply chain

OpenAI Agent SDK for demand forecasting and disruption detection, Perplexity for live freight/commodity market context.

40

Retail

OpenAI Luna for high-volume customer ops, Claude for inventory-linked FP&A;, Perplexity for competitive pricing checks.

41

Fintech

Claude-first across the board — KYC/AML, close, and compliance review all sit on Anthropic's ZDR-capable tier. Kimi K3 off the table entirely.

42

Professional services

Claude for client deliverables and document review, Perplexity for research-heavy engagements, OpenAI for internal ops (billing, scheduling).

SECTION 08 · IMPLEMENTATION ROADMAP

A phased rollout, not a big-bang switch.

Four phases from pilot to governed scale — sized for a team that isn't hiring an AI department to run this.

43

Month 1 — pilot one workflow

Pick the single highest-friction task (usually AP/AR or close) and run it on Claude or OpenAI in parallel with the existing process. Measure time saved and error rate, don't automate yet.

44

Months 2-3 — standardize prompts & templates

Turn the pilot into a reusable template (Anthropic finance agent template or a Frontier workflow). Add Perplexity as the standing research layer for anything requiring live data.

45

Months 4-6 — scale to adjacent workflows

Extend the working template to 2-3 more finance or ops tasks. Add role-based access and a review checkpoint for anything customer- or board-facing.

46

Ongoing — governance & vendor review

Revisit this comparison quarterly — model releases in this market move faster than most procurement cycles. Re-test Kimi K3 once open weights ship if cost pressure is high.

SECTION 09 · BEST PRACTICES

Five habits that separate pilots from production.

None of this is exotic. Most failed AI pilots skip one of these five things.

47

Always specify the output format

Ask for a table, a specific currency, a named template — vague prompts produce vague, unreviewable numbers.

48

Never let a model be the sole approver

Every finance output that reaches a regulator, auditor or board needs a named human sign-off, logged separately from the model run.

49

Pin the data retention setting before you start

Check ZDR/retention terms per provider before the first real client file goes in — not after.

50

Fact-check external claims separately

Use Perplexity or another citation-backed tool to verify any market figure a modeling engine produces before it goes into a deliverable.

51

Re-test quarterly, not annually

Pricing, benchmarks and compliance posture in this category have all shifted materially within single quarters this year.

SECTION 09 · PROMPT LIBRARY

Prompts you can paste in this quarter.

Eight starting prompts for finance & ops — adjust the bracketed fields to your data.

52

Month-end close checklist

"Build a month-end close checklist for [entity], flag any account with >10% variance vs. last month, and draft the variance explanation."

53

KYC screening summary

"Summarize this counterparty's KYC file against [jurisdiction] requirements and list any missing document."

54

AP exception triage

"Review these 20 unmatched invoices, group by exception type, and draft a resolution email for each group."

55

Supplier risk scan

"Given this supplier list, flag any with recent news of financial distress or delivery disruption."

56

Market comparable pull

"Find the three closest public comparables to [company] by revenue and margin, with sourced figures."

57

Board deck data check

"Verify every figure in this board deck draft against public filings and flag any discrepancy."

58

Forecast sensitivity

"Re-run this forecast at +/-15% on [key driver] and summarize the impact on EBITDA in one paragraph."

59

Contract redline summary

"Compare these two contract versions and list every substantive change, ignoring formatting."

QUICK REFERENCE

The one page you'll actually reuse.

Section-by-section winners, the overall call, and the rule of thumb.

CAPABILITIES

Research & real-time data
Perplexity

Financial modeling / FP&A;
Claude

Compliance & document review
Claude

FINANCE

Month-end close
Claude

KYC / AML screening
Claude

Board reporting / pitchbooks
Claude + Perplexity

OPERATIONS

Cross-system orchestration
OpenAI

Supply chain forecasting
OpenAI + Perplexity

Procurement-to-pay
OpenAI

COST & RISK

Lowest cost per task
Kimi K3

Best data-residency posture
Claude (ZDR)

Weakest compliance today
Kimi K3

OVERALL RECOMMENDATION

Claude-first for finance, OpenAI-first for ops, Perplexity always-on.

Run Kimi K3 only on non-sensitive, high-volume workloads once open weights ship — never on data that needs to stay off a China-hosted API.

RULE OF THUMB

If it touches regulated financial data, it runs on Claude. If it touches five internal systems at once, it runs on OpenAI. Everything gets fact-checked on Perplexity first.

1. CNBC, Kimi K3 context