

The Executive's Frontier AI Model Playbook

What every frontier model is for, what the names and tiers mean, which one to use for which job, and how to read the next launch announcement.

OCTOBER 2026 EDITION

EXECUTIVE SUMMARY

Pick the model for the job, not the vendor for the year.

Earlier this year, the question was which AI app to buy. In October 2026, the better question is which model to run, because the same frontier models now sit inside ChatGPT, Claude, Gemini, Copilot and dozens of other tools. This playbook is a field guide to those models for non-technical corporate professionals.

Every major lab now ships a family of models in tiers: a top model for the hardest thinking, a balanced everyday model, and a small, fast, cheap model for volume. Anthropic calls them Opus, Sonnet and Haiku (plus Fable above them). OpenAI calls them Astra, Sol and Luna. Google calls them Pro, Flash and Flash-Lite, with Gemini 4 Argon arriving above them. The same models also show up inside other products: Claude Opus 5.5 and GPT-6 Sol began rolling out inside Microsoft 365 Copilot on October 8, 2026.

So the practical skill for a non-technical leader is now twofold: know the platform that fits the workflow, and know which tier of model to choose once you are in it. This edition gives you both.

3

tiers in almost every model family: top, balanced, fast. Learn the pattern once and every vendor's lineup reads the same way.

100×

price gap between OpenAI's top and small API models (\$10 vs \$0.10 per million input tokens). Tier choice is a cost decision.

1M

token context windows are now standard at the frontier across the Claude 5.5 family and Muse Spark 1.3.

What changed in 2026

- **Anthropic** moved from Claude 4.6 to the Claude 5.5 family: Opus 5.5 (September 22), Sonnet 5.5 (September 28) and the new Claude Haiku 5.5 (October 7), with Fable 5.1 above them for the most demanding work.
- **OpenAI** moved from GPT-5.4 through GPT-5.6 (July) to the GPT-6 generation: Astra, Sol and Luna, plus GPT-6.1 Sol (September 29).
- **Google** moved from Gemini 3.1 Pro to a fast-cycling Flash line (3.8 Flash is newest) and announced Gemini 4 Argon, currently limited to vetted cyber defenders.
- **xAI** (now operating as SpaceXAI) shipped Grok 4.5, 4.6 and, this month, Grok 4.7.
- **Meta** replaced Llama as its flagship with the proprietary Muse Spark family (1.3 released September 2). Meta's Manus acquisition is being unwound after Chinese regulators objected, so this edition no longer lists Manus as a Meta product.
- **Microsoft Copilot** is now multi-model. Its remaining limit is workflow scope, not model quality.

All model names, dates and prices in this edition were checked against vendor pages on October 8, 2026. See the source list at the end. This field moves weekly; check before you buy.

01 HOW TO READ A MODEL NAME

Decode the names in sixty seconds.

Model names look like noise, but nearly every lineup encodes the same three things: the family, the generation, and the tier.

PART OF THE NAME	WHAT IT TELLS YOU	EXAMPLE
Family / brand	Who made it and which product line.	Claude, GPT, Gemini, Grok, Muse
Generation number	Higher is newer. A “.5” or “.1” is a meaningful upgrade, not a patch. Do not compare numbers across vendors: Claude 5.5 is not “behind” GPT-6.	Claude Opus 5 → 5.5; GPT-6 Sol → GPT-6.1 Sol
Tier word	How big, smart, slow and expensive the model is within its family.	Opus / Sonnet / Haiku; Astra / Sol / Luna; Pro / Flash / Flash-Lite
Mode suffix	A setting or variant of the same model: more thinking time, faster responses, voice, or a restricted specialist edition.	Deep Think, Extended Thinking, Fast mode, Live, Cyber

The three-tier pattern, by vendor

VENDOR	TOP TIER: HARDEST THINKING	BALANCED: EVERYDAY WORK	FAST AND CHEAP: VOLUME
Anthropic	Claude Fable 5.1 (beyond-frontier); Claude Opus 5.5	Claude Sonnet 5.5	Claude Haiku 5.5
OpenAI	GPT-6 Astra (Astra Pro in ChatGPT Pro, Business, Enterprise)	GPT-6.1 Sol	GPT-6 Luna
Google	Gemini 4 Argon (limited preview); Gemini 3.1 Pro / 3 Deep Think in the app	Gemini 3.8 Flash	Gemini 3.5 / 3.1 Flash-Lite
xAI / SpaceXAI	Grok 4.7: one flagship with configurable reasoning effort instead of separate tiers		
Meta	Muse Spark 1.3 (proprietary, via Meta AI and its developer API)		Muse Glimmer (open-weight)
Open-weight others	Mistral Large 4 (preview); Qwen 3.8-Max; DeepSeek V4 Pro	DeepSeek V4.1 Flash; Mistral Small and Minstral lines	

The rule of thumb

Start on the balanced tier. Move up when the work is high-stakes, long, or ambiguous (a board brief, a contract review, a strategy you will defend). Move down when the work is repetitive and checkable (tagging 2,000 survey comments, routing tickets, extracting fields from invoices). Most professionals over-use the top tier for simple tasks and under-use it for the few that matter.

Anthropic · Claude

LONG DOCUMENTS · STRUCTURED OUTPUT · AGENTS

Claude’s signature strength, holding a large document set in mind and returning a clean, structured deliverable, is now available at every tier. All four current Claude models offer a 1M-token context window and up to 128K tokens of output, according to Anthropic’s model documentation.

MODEL	RELEASED	WHAT IT IS FOR	API PRICE / 1M TOKENS (IN / OUT)
Claude Fable 5.1	Sep 1, 2026	Anthropic’s most capable generally available model, for ambitious, long-running, asynchronous work. Slower and premium. A restricted sibling, Mythos 5.1, is open only to verified security organizations.	\$10 / \$50
Claude Opus 5.5	Sep 22, 2026	The default for complex work requiring careful judgment: long agentic tasks, analysis and knowledge work. Anthropic says it performs at Fable 5.1’s level on most work at lower cost.	\$4 / \$20
Claude Sonnet 5.5	Sep 28, 2026	The speed–intelligence balance. Strongest at well-scoped everyday tasks and polished documents, slides and spreadsheets. 30%+ faster than Sonnet 5.	\$2 / \$10
Claude Haiku 5.5 New	Oct 7, 2026	High-volume, cost-sensitive and real-time work: classification, extraction, routing, quick drafts, sub-tasks inside agents. Anthropic calls it its cheapest, fastest and most capable small model yet, and the first Haiku with effort controls.	\$0.10 / \$0.50 (prompts ≤100K); \$0.50 / \$2.50 above

Where it excels

- Board-ready synthesis across many long documents in a single session
- First drafts that read like third drafts: proposals, strategy memos, policies
- Careful, hedged analysis where wrong answers are costly (regulatory, clinical, legal)

Where it falls short

- Fewer native SaaS connectors than ChatGPT or Gemini
- Fable usage is metered tightly on consumer plans
- Live web and social signals are not its home ground

Where to get it: Claude.ai (Free includes Sonnet and Haiku; Pro \$20/month or \$17 annually adds Opus and Fable credits; Max from \$100/month), Team and Enterprise plans, Claude Code, Microsoft 365 Copilot, and the AWS, Google Cloud and Microsoft cloud marketplaces.

What’s new: Claude Haiku 5.5

For non-technical teams, Haiku 5.5 changes the economics of bulk work. Every Claude.ai user, including Free, can select it. At \$0.10 per million input tokens for normal-length prompts, processing a 300-page document set costs well under a cent through the API. Use it to triage inboxes, tag CRM notes, pull fields from contracts, or summarize call transcripts at volume, then hand the few that matter to Opus 5.5.

03 OPENAI

OpenAI • GPT and ChatGPT

GENERALIST • APP ECOSYSTEM • COMPUTER USE

OpenAI has replaced the GPT-5.x line with the GPT-6 generation and adopted named tiers. OpenAI describes GPT-6 Astra as its best model across the board. Sol and Luna were trained with similar methods to bring that quality to faster, cheaper models. GPT-6 rolled out to all ChatGPT users on October 7, 2026, with “Intelligent UI”: answers that include visuals and interactive elements.

MODEL	RELEASED	WHAT IT IS FOR	API PRICE / 1M TOKENS (IN / OUT)
GPT-6 Astra	Sep 3, 2026	Hardest work: computer use, coding, science, high-stakes professional analysis. In ChatGPT Plus, Pro, Business and Enterprise; Astra Pro on Pro, Business and Enterprise.	\$10 / \$50
GPT-6.1 Sol	Sep 29, 2026	OpenAI’s recommendation for repeated, long-running work across code, apps and documents: near-Astra quality at one-fifth of Astra’s API price. Supersedes GPT-6 Sol (Sep 22).	\$2 / \$10
GPT-6 Luna	Sep 22, 2026	Everyday and high-volume tasks at the lowest cost.	\$0.10 / \$0.50
GPT-Live	Jul 8, 2026	Natural voice conversation in ChatGPT Voice; an API version adds telephony.	Per-use

Where it excels

- The broadest generalist: one tool for drafting, data analysis, brainstorming and summarizing
- Deepest app and connector ecosystem; an expanding Atlassian partnership (Oct 6)
- Agentic computer use and coding (Codex)

Where it falls short

- Rarely the single best tool for a specialist job
- Plan and model availability is complex; Enterprise admins must switch Astra on
- Still verify facts and citations on research work

Plans: Free; Go \$8; Plus \$20; Pro tiers including a \$500/month Pro 500 with Astra Ultrafast; Business and Enterprise per seat.

Google · Gemini

WORKSPACE · MULTIMODAL · ENTERPRISE SEARCH

Google ships fast and in many variants. The pattern to remember: **Pro** is the deep reasoner, **Flash** is the fast workhorse (and now the main line of improvement), **Flash-Lite** is the cheapest, and **Deep Think** is a slow, careful reasoning mode. **Live** models handle voice; **Nano Banana** and **Veo** handle images and video.

MODEL	STATUS	WHAT IT IS FOR
Gemini 4 Argon	Sep 2026, limited	Google's newest frontier model, with advanced reasoning and a 1M-token output limit. Rolling out only to vetted cyber defenders through the Fairwind Program; broader availability to follow.
Gemini 3.1 Pro / 3 Deep Think	Feb 2026	Complex problem-solving in the Gemini app today; Deep Think for science, research and engineering.
Gemini 3.8 Flash	Sep 2026	Google's best reasoning and coding yet, at Flash speed and cost; built for long agentic workflows.
Gemini 3.6 Flash	Jul 2026	Everyday app model for document analysis and prototyping (free tier).
Flash-Lite (3.5 / 3.1)	Stable	High-throughput, lowest-cost work.
Gemini 3.8 Live	Sep 2026	Natural voice in Gemini Live, Search Live and Workspace.

Where it excels

- Native in Gmail, Docs, Sheets, Meet and Drive
- Multimodal: images, video, audio and long PDFs in one model
- Voice (Live) and media generation (Nano Banana, Veo)
- Strong price-performance on the Flash tier

Where it falls short

- The newest frontier model (Argon) is not yet generally available
- Many overlapping versions; app and API names differ
- Most valuable to organizations already on Workspace

Plans: Free; Google AI Plus \$4.99; AI Pro \$19.99; AI Ultra from \$99.99/month. Gemini Enterprise from \$21 per seat (Business) and \$30 (Standard/Plus). In September, Gems were replaced by "Skills" for task automation.

05 XAI / SPACEXAI

xAI / SpaceXAI • Grok

REAL-TIME • SPEED • KNOWLEDGE WORK

xAI now operates as SpaceXAI. Instead of separate tiers, it offers one flagship, Grok 4.7 (October 2026), with configurable reasoning effort and a 500K-token context, at \$2 / \$6 per million tokens. SpaceXAI positions it as its most capable model for coding and knowledge work, better at checking its own work and at creating documents and presentations. Separate Voice and Imagine APIs handle speech and image/video generation.

Strengths: one simple lineup, low API price, a large context window, and live public signals through X.

Weaknesses: thinner enterprise governance and fewer workplace integrations.

Use it for: fast reads of live public conversation and market reaction, quick-turn analysis, and low-cost API work. **Be cautious:** enterprise governance remains thinner than at Microsoft, Google, OpenAI and Anthropic. Confirm certifications directly before using it with regulated data.

06 META

Meta • Muse (and Llama)

CONSUMER REACH • LOW-COST API

In April 2026, Meta Superintelligence Labs introduced Muse Spark as Llama's successor powering Meta AI. Muse Spark 1.3 (September 2, 2026) is a natively multimodal reasoning model tuned for agentic work, with a 1M-token context. It is proprietary and available in meta.ai, the Meta AI app and Meta's developer API. That API comes in two price tiers: \$1.25 / \$4.25 per million tokens, or \$0.10 / \$0.20 if you let Meta use your data to improve its products. A smaller Muse Glimmer model is open-weight. Llama 4 (April 2025) remains available as open-weight but is no longer Meta's frontier.

Strengths: huge consumer reach through WhatsApp, Instagram and Facebook, and a very low API price.

Weaknesses: a young enterprise offering, and the cheapest tier trades away your data.

For corporate users: read the data terms before choosing the cheaper tier. That tier trades your data for a lower price, so it is wrong for confidential work.

07 OTHERS TO KNOW

Others to know

OPEN-WEIGHT · SOVEREIGNTY · COST

MODEL	MAKER	WHY A BUSINESS LEADER SHOULD KNOW IT
Mistral Large 4	Mistral (France)	Public preview October 8, 2026; open weights promised by end of October. Mistral's largest model, with an EU-operated deployment and on-premises options. The pick when European data sovereignty is the requirement.
Qwen 3.8-Max	Alibaba	August 2026. Alibaba's most capable model for coding and professional work; its first Max-class model to get open weights.
DeepSeek V4 Pro / V4.1 Flash	DeepSeek	V4 Pro generally available August 13; V4.1 Flash September 10. Very low cost. Most corporate policies restrict sending sensitive data to the hosted service; self-hosting is the usual route.

08 PLATFORMS THAT SIT ON TOP OF MODELS

Platforms that sit on top of models

WHERE THE MODELS SHOW UP

Most professionals reach these models through platforms, and those platforms increasingly let you choose the model underneath.

PLATFORM	ROLE IN YOUR STACK (OCTOBER 2026)
Microsoft 365 Copilot	The ambient layer in Outlook, Teams, Word, Excel and PowerPoint. Now multi-model: GPT-5.6 became the preferred model in July, and Claude Opus 5.5 and GPT-6 Sol began rolling out on October 8 across Word, Excel, PowerPoint, Chat, Cowork and Copilot Studio.
Perplexity	Cited research and always-on monitoring. It routes queries across top models from several labs. Enterprise Max unlocks audit logs and SCIM.
Genspark	Turning one prompt into finished assets (slides, sheets, docs, sites, phone calls) by orchestrating several frontier models.
Manus	Autonomous background execution on its own virtual machine. Meta's \$2B acquisition is being unwound under Chinese regulatory pressure. Treat it as an independent vendor and re-check its ownership in procurement.

09 FIELD GUIDE

How to read a model launch announcement.

A major model launches nearly every week. Most announcements follow the same template. Here is how to pull out what matters for your team in five minutes.

#	LOOK FOR	WHAT IT TELLS YOU	EXAMPLE FROM THIS AUTUMN
1	Where it sits in the family	Is it the top, balanced or fast tier, and what does it replace? This tells you whether it is for your hard work or your volume work.	"Sonnet 5.5 is a faster, lower-cost complement to Opus 5.5"
2	Availability: who, where, when	"Rolling out", "preview", "trusted testers" and "API only" all mean you probably cannot use it today. Check your plan and whether your admin must switch it on.	Gemini 4 Argon is limited to vetted cyber defenders; GPT-6 Astra is off by default for Enterprise
3	Price, and the fine print	Look at per-token prices and also surcharges for long prompts, fast modes and discounts for caching. A model that needs fewer tokens can cost less even at the same price.	Haiku 5.5 costs 5× more above 100K-token prompts; Astra Fast mode costs 2×
4	Benchmarks: read the names	Vendors choose the benchmarks they win. Look for ones that resemble your work (knowledge work such as GDPval, computer use such as OSWorld) and ignore the rest. Treat a lead of a few points as a tie.	Opus 5.5 at 1846 vs Sonnet 5.5 at 1844 on GDPval-AA: effectively equal
5	Customer quotes	Named companies with specific numbers are a better signal than benchmarks for real workflows.	HubSpot's CRM task score for Haiku 5.5
6	Context and output limits	How much it can read at once and how long its answers can be. This matters for document-heavy roles.	1M-token context across the Claude 5.5 family
7	Safety and data terms	Look for the system card, safeguards and whether your data is used for training on that tier.	Meta's Contributor tier: lower price in exchange for your data
8	What it retires	Deprecation dates can break your team's saved workflows and custom assistants.	DeepSeek rerouted V4 Pro traffic to V4.1 Flash on September 14

The three-question test

1. Can my team use it this week, on our plan? 2. Is it better on a job we actually do, according to someone other than the vendor? 3. Does it change the cost or risk of that job? If any answer is no, note the launch and move on.

10 USE-CASE MATRICES

Which model for which job.

These matrices are written for non-technical corporate professionals using the chat apps their company provides. “Primary” is the best fit. “Alternative” is a capable backup if that model is not licensed. Tier guidance applies inside any vendor.

Matrix 1 • Everyday work, by function

FUNCTION	JOB	PRIMARY	ALTERNATIVE	TIER
Executive / strategy	Board brief from 5+ long documents	Claude Opus 5.5	GPT-6 Astra	Top
	Stress-test a strategy; find the counter-argument	GPT-6 Astra	Claude Fable 5.1	Top
Sales	Account research before a call	Perplexity	Gemini 3.8 Flash	Balanced
	Proposal or SOW first draft	Claude Sonnet 5.5	GPT-6.1 Sol	Balanced
	Tag and summarize 500 call notes	Claude Haiku 5.5	GPT-6 Luna	Fast
Marketing	Campaign concepts and copy variants	GPT-6.1 Sol	Claude Sonnet 5.5	Balanced
	Live audience and market reaction	Grok 4.7	Perplexity	Balanced
	Images and short video	Gemini (Nano Banana, Veo)	Grok Imagine	—
Finance	Spreadsheet analysis and variance commentary	Claude Sonnet 5.5	GPT-6.1 Sol	Balanced
	Extract fields from 1,000 invoices	Claude Haiku 5.5	Gemini Flash-Lite	Fast
Legal / compliance	Contract or regulation review	Claude Opus 5.5	GPT-6 Astra	Top
HR / people	Policy drafts and job descriptions	Claude Sonnet 5.5	Copilot (Word)	Balanced
	Survey open-text themes at scale	Claude Haiku 5.5	GPT-6 Luna	Fast
Operations	Email triage and meeting follow-ups	Microsoft 365 Copilot	Gemini in Workspace	Balanced
	Overnight multi-step research or report	Claude Fable 5.1	GPT-6 Astra; Manus	Top

Everyone	Quick question, rewrite, translation	Whatever is open	—	Fast / balanced
----------	--------------------------------------	------------------	---	-----------------

Matrix 2 • Choose by what the task demands

IF THE TASK NEEDS...	REACH FOR	WHY
Many long documents at once	Claude 5.5 family; Muse Spark 1.3	1M-token context windows
Cited, verifiable facts	Perplexity; Gemini Deep Research	Built around sources
Today’s news and social sentiment	Grok 4.7; Perplexity	Live data
Polished slides, sheets or documents	Claude Sonnet 5.5; Genspark	Vendor-stated design strengths
Working across apps on a computer	GPT-6 Astra; Claude Opus 5.5	Leading computer-use models
Voice conversation	GPT-Live; Gemini 3.8 Live	Purpose-built voice models
Thousands of repetitive items	Claude Haiku 5.5; GPT-6 Luna; Flash-Lite	Cents per million tokens
Data that cannot leave the EU or your servers	Mistral Large 4; open-weight Qwen	Self-hosting and EU deployment
Staying inside Microsoft 365	Copilot (pick Claude or GPT inside it)	No workflow change

Matrix 3 • Risk and stakes

STAKES	EXAMPLES	MODEL TIER	HUMAN CHECK
Low, reversible	Internal notes, brainstorm, rewrites	Fast or balanced	Skim
Medium, external	Client emails, proposals, marketing copy	Balanced	Edit and fact-check figures
High, decision-grade	Board materials, pricing, M&A, legal positions	Top (Opus, Astra, Fable)	Expert review and source check
Regulated data	Patient, financial or personal data	Only enterprise-licensed tools under contract (BAA, DPA)	Compliance sign-off

11 ARCHITECTURE

The 5-layer stack, updated.

Treat AI as a layered stack, not a single tool. Today the same frontier models now power several layers. That makes it easier to standardize on two or three model families while keeping the best workflow tool for each layer.

LAYER	PURPOSE	PLATFORM	MODEL TO SET BY DEFAULT
5	Specialized functions	Gemini, Grok, Genspark	Gemini Live / Veo for media; Grok 4.7 for live signal
4	Autonomous execution	Claude Fable, ChatGPT agents, Perplexity Computer, Manus	Fable 5.1 or GPT-6 Astra
3	Strategic production	Claude, ChatGPT	Opus 5.5 / Astra for high stakes; Sonnet 5.5 / Sol for drafts
2	Research and intelligence	Perplexity, Genspark, Gemini Deep Research	Multi-model routing
1	Workflow acceleration	Microsoft 365 Copilot, Gemini in Workspace	GPT-6 Sol or Claude Opus 5.5 inside Copilot

The Copilot correction, revisited

A Copilot-only strategy used to mean settling for weaker models. Copilot now runs Claude and GPT-6 models, so the model-quality gap has narrowed. The scope gap has not: Copilot is still strongest inside the Microsoft 365 boundary. Keep it as Layer 1, choose the model inside it deliberately, and keep specialist tools for research, autonomy and real-time work.

12 PRICING

What it costs, October 2026.

Published list prices as checked on October 8, 2026. Enterprise pricing is negotiated and varies by volume and contract.

Seats in the apps people use

PRODUCT	INDIVIDUAL	TEAM / BUSINESS	NOTES
ChatGPT	Free; Go \$8; Plus \$20; Pro tiers up to \$500/mo	Business and Enterprise per seat	Astra Pro on Pro, Business and Enterprise
Claude	Free; Pro \$20 (\$17 annual); Max from \$100	Team and Enterprise per seat	Free tier includes Sonnet and Haiku 5.5
Google Gemini	Free; AI Plus \$4.99; AI Pro \$19.99; Ultra from \$99.99	Gemini Enterprise from \$21–30/seat	Also bundled in Workspace
Microsoft 365 Copilot	Copilot Chat included with eligible M365	Business from \$21/user (promotional); Enterprise \$30 list	Plus the required M365 base license

API price ladder (per 1M tokens, input / output)

TIER	ANTHROPIC	OPENAI	XAI	META
Top	Fable 5.1 \$10 / \$50 Opus 5.5 \$4 / \$20	GPT-6 Astra \$10 / \$50	Grok 4.7 \$2 / \$6	Muse Spark 1.3 \$1.25 / \$4.25
Balanced	Sonnet 5.5 \$2 / \$10	GPT-6.1 Sol \$2 / \$10		
Fast	Haiku 5.5 \$0.10 / \$0.50	GPT-6 Luna \$0.10 / \$0.50	—	Contributor tier \$0.10 / \$0.20

Caching and batch discounts can cut these sharply (Anthropic: up to 90% with caching and 50% with batch on Haiku 5.5). Long-prompt surcharges apply on Haiku 5.5 (above 100K tokens) and on OpenAI's long-context tiers.

13 COMPLIANCE

Procurement first filter.

For regulated industries, compliance comes before capability. Certifications change often and differ by plan, so treat this as a checklist to verify with each vendor rather than a scorecard:

- SOC 2 Type II and ISO 27001 reports, plus ISO 42001 for AI management
- HIPAA Business Associate Agreement for the exact plan you are buying
- FedRAMP status for public-sector work

- A written no-training-on-your-data commitment for your tier. Check Meta's Contributor tier and other discounted tiers carefully.
- Data residency (Mistral's EU deployment; regional hosting on AWS, Google Cloud and Azure)
- SSO/SAML, SCIM, audit logs and retention controls
- Model access through your existing cloud contract: Claude on AWS, Google Cloud and Microsoft; GPT on Azure

Our view: the four large vendors (OpenAI, Anthropic, Google, Microsoft) have the most mature enterprise compliance. Newer platforms are powerful for individual productivity but need specific diligence before enterprise-wide use.

14 CLOSE

The bottom line.

The AI tool wars became the AI model wars, and the lesson is the same: clarity beats loyalty. Learn the three-tier pattern, set sensible defaults by job, keep the top models for the work that deserves them, and let the cheap, fast models handle the volume. Organizations that do this will get better decisions at lower cost than those that pick one vendor and stop thinking.

Ready to put the right model on the right job?

Cresera helps leadership teams build AI into how their people actually work, with measurable outcomes in weeks.

[Schedule a strategy session →](#)

creserai.com

15 ABOUT THE AUTHOR



Brett Jansen

CEO, Cresera · Human-led growth for the AI era

Brett Jansen has more than 16 years of experience in commercial growth strategy, primarily in the heavily regulated healthcare technology ecosystem. She has closed \$175M in enterprise deals, advised 75+ companies from pre-seed through IPO, and helped companies she advised raise \$310M+, including two exits. She created the PLT (Proof, Like, Trust) sales methodology for complex B2B enterprise deals.

creserai.com · linkedin.com/in/thebrettjansen

16 GLOSSARY

Key terms, in plain English.

Token

The unit models read and write, roughly ¾ of an English word. Pricing and limits are counted in tokens, usually per million (MTok).

Context window

How many tokens a model can consider at once, including your documents, the conversation and its answer. 1M tokens is about 750,000 words.

Output limit

The maximum length of a single answer (for example, 128K tokens for the Claude 5.5 family).

Model vs. app

The model is the engine (Claude Opus 5.5, GPT-6 Astra). The app is the product you use (Claude.ai, ChatGPT, Copilot). One app can offer several models, and one model can appear in many apps.

API

The programmatic way to use a model, billed per token. Developers and IT teams build on it. Chat apps are billed per seat instead.

Tier

The size and price level within a model family: top (Opus, Astra, Pro), balanced (Sonnet, Sol, Flash) or fast (Haiku, Luna, Flash-Lite).

Frontier model

A model at or near the leading edge of capability at the time of release.

Reasoning / thinking model

A model that works through a problem step by step before answering. It is better on hard problems but slower and costlier. Often controlled by an “effort” or “thinking” setting.

Effort control

A setting that trades speed and cost against depth of thinking on the same model (for example, Claude Haiku 5.5, Grok 4.7, DeepSeek V4 Pro).

Multimodal

Able to take in or produce more than text: images, audio, video, documents.

Agent

An AI system that plans and carries out multi-step tasks using tools such as browsing, files, apps or code, rather than only answering a prompt.

Computer use

An agent capability to operate a desktop or browser the way a person does: clicking, typing, navigating software.

Tool use / function calling

A model's ability to call external tools or systems (search, a calculator, your CRM) during a task.

MCP (Model Context Protocol)

An open standard, introduced by Anthropic, for connecting AI models and agents to tools and data sources through a common interface. Many vendors now support it.

RAG (retrieval-augmented generation)

Grounding a model's answer by first retrieving relevant passages from your own documents or the web, then answering from them. It reduces hallucination and keeps answers current.

Hallucination

A confident answer that is false or unsupported. All models still hallucinate. Verify facts, figures and citations on anything external.

Grounding / citations

Typing answers to specific sources you can check, which is the core design of tools like Perplexity.

Fine-tuning

Further training a model on your own examples to specialize its behavior or style. It is usually unnecessary for knowledge workers; good prompts and RAG cover most needs.

Prompt

The instructions and context you give the model. Clear goals, audience, format and examples produce better output.

System prompt / custom instructions

Standing instructions applied to every conversation, such as Custom GPTs, Claude Projects or Gemini Skills.

Open-weight vs. closed

Open-weight models (Llama, Mistral Large 4, Qwen, Muse Glimmer) can be downloaded and run on your own infrastructure. Closed models (Claude, GPT, Gemini, Grok, Muse Spark) are used only through the vendor's app or API.

Distillation

Training a smaller, cheaper model to imitate a larger one. It is one way fast tiers inherit capability, and labs guard against competitors distilling their models.

Benchmark

A standardized test of model performance (for example, GPDPval for knowledge work, OSWorld for computer use, Terminal-Bench for coding). Useful directionally, but vendors choose the ones they win.

Latency

How long you wait for a response. Fast tiers and "fast modes" cut latency, usually at some cost in quality or price.

Caching

Reusing previously processed input (such as a long document) at a large discount, often up to 90% off input price.

Batch processing

Submitting many requests to run asynchronously at a discount, often 50%, in exchange for slower turnaround.

Knowledge cutoff

The date after which a model has no built-in knowledge. Web search or RAG fills the gap.

Deprecation

A vendor retiring an older model on a set date, which can break saved workflows and custom assistants.

System card

The vendor's safety and evaluation report for a model. Read it for limitations and safeguards.

BAA / DPA

Contracts covering regulated data: a Business Associate Agreement (HIPAA) or a Data Processing Agreement (GDPR). Required before putting patient or personal data into a tool.

Sources

All checked October 8, 2026.

1. Anthropic, Models overview: platform.claude.com/docs/en/models/overview
2. Anthropic, Claude Haiku 5.5: anthropic.com/claude/haiku
3. Anthropic, Introducing Claude Opus 5.5: anthropic.com/claude-opus-5-5
4. Anthropic, Introducing Claude Sonnet 5.5: anthropic.com/claude-sonnet-5-5
5. Anthropic, Claude Fable 5.1: anthropic.com/claude/fable
6. Anthropic, Claude pricing: claude.com/pricing
7. OpenAI, Release index: openai.com/research/index/release/
8. OpenAI, GPT-6 Astra: openai.com/index/gpt-6-astra/
9. OpenAI, Introducing GPT-6 Sol and Luna: openai.com/index/introducing-gpt-6-sol-and-luna/
10. OpenAI, API pricing: developers.openai.com/api/docs/pricing
11. OpenAI, ChatGPT What's new, Sept 28–Oct 2, 2026: learn.chatgpt.com/docs/whats-new/september-28-october-2-2026
12. OpenAI, ChatGPT pricing and Pro tiers: chatgpt.com/pricing; help.openai.com/en/articles/9793128
13. Google, Gemini API models: ai.google.dev/gemini-api/docs/models
14. Google, AI updates September 2026: blog.google/innovation-and-ai/technology/ai/google-ai-updates-september-2026/
15. Google, Gemini app release notes: gemini.google/release-notes/
16. Google, AI plans: gemini.google/subscriptions/
17. SpaceXAI, Models and pricing: docs.x.ai/developers/models
18. SpaceXAI, Introducing Grok 4.7: x.ai/news/grok-4-7
19. Meta, Introducing Muse Spark: ai.meta.com/blog/introducing-muse-spark-msl/
20. Meta, Muse Spark 1.3 developer page: dev.meta.ai/models/muse-spark
21. Wikipedia, Muse Spark; Llama (language model)
22. Mistral, Introducing Mistral Large 4: mistral.ai/news/mistral-large-4
23. DeepSeek, API change log: api-docs.deepseek.com/updates/
24. Alibaba Qwen, Qwen3.8-Max: qwen.ai/blog?id=qwen3.8
25. Microsoft, More Models, One Copilot: techcommunity.microsoft.com/blog/microsoft-copilot-blog/more-models-one-copilot/4559035
26. Microsoft, Copilot business pricing: microsoft.com/en-us/copilot/pricing/business
27. TechCrunch (Jun 13, 2026) and Reuters (Jun 18, 2026) on Meta unwinding the Manus acquisition

© 2026 Cresera. May be shared with attribution. Vendor names and trademarks belong to their owners. Not affiliated with or endorsed by any vendor listed.