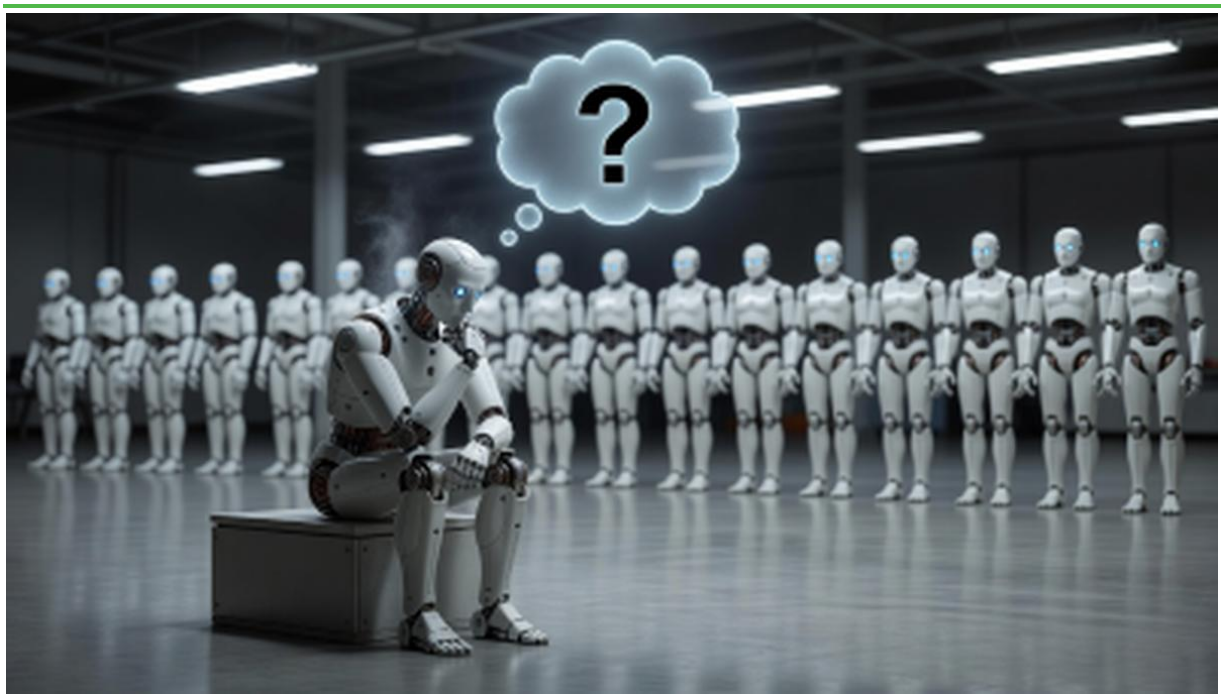
REFERENCE GUIDE · JULY 2026

Making Sense of the AI Engines

The most common AI engines — and which one to use for which job



A plain-English guide for New Zealand businesses. Artificial intelligence is no longer one product but a whole marketplace of "engines" — each with different strengths, speeds and costs. This guide explains the differences in everyday terms and shows, at a glance, which model suits which job.

What's inside

- The three tiers every AI vendor offers — and what each one is for
- The real trade-offs: speed, cost and quality
- A one-page comparison of the leading engines, side by side
- How to choose — and combine — the right models for your business

Enabling people to make a difference through better use of technology.

Why there's no single "best" AI

Artificial intelligence has stopped being a single product and become a marketplace. Ask "should we use AI?" today and the more useful question is quickly "which AI, for which task?" A handful of vendors — OpenAI, Anthropic, Google, DeepSeek, xAI, Meta and Mistral among them — now each offer not one model but a family of models, and they are not interchangeable.

Think of it like a vehicle fleet. You wouldn't send a courier van to do a road-train's job, or run a heavy hauler to drop a single parcel across town. AI models work the same way.

Three tiers, three jobs

Almost every major vendor now builds across three broad tiers:

- **Built for depth.** The flagship "frontier" models handle complex reasoning, difficult coding, and long multi-step tasks that must stay coherent for hours. They are the most capable — and the most expensive to run.
- **Built for balance.** Mid-tier models handle the majority of everyday work — drafting, summarising, analysis, straightforward coding — at a fraction of the cost, and fast enough for interactive use.
- **Built for speed and scale.** Lightweight models are made for high-volume, repetitive jobs: classification, tagging, simple Q&A, processing thousands of records — where throughput and cost matter more than raw brilliance.

Speed and cost: the real trade-offs

The differences that matter most day to day are speed, cost and quality — and you rarely get all three at once. The gap is dramatic: a flagship model can cost 20 to 50 times more per unit of work than a lightweight one from the same vendor. For a one-off strategic analysis that's money well spent; for a chatbot answering thousands of routine questions a day, it would be a budget blowout for no real benefit.

What else sets the vendors apart

Beyond raw capability, four factors increasingly separate the options:

- **Context window —** how much information the model can weigh at once. Some now read a million-plus words (an entire codebase, or years of contracts) in a single pass.
- **Multimodality —** whether it handles images, audio and video, not just text.
- **Openness —** some models can be downloaded and run on your own infrastructure for full data control; others are cloud-only.
- **Ecosystem fit —** how neatly the model plugs into the tools you already run (Microsoft 365, Google Workspace, and so on).

The smart approach for most organisations isn't picking one winner — it's matching the right model to each task. Many teams now route work across two or three vendors depending on the job.

The engines at a glance

The table below summarises the leading AI engines and where each one earns its keep. Model names and versions are current as of July 2026.

Vendor	Model	Version	When to use it
OpenAI	**GPT-5.5**	5.5 (5.5 Pro for max precision)	Highest-stakes reasoning, coding and multi-tool "agentic" workflows where quality justifies the premium.
	GPT-5.4 mini / nano	5.4	Fast, low-cost, high-volume work — chatbots, classification, summarising and routing.
	o-series	o3	Deliberate, step-by-step reasoning: hard maths, logic and analytical problems.
Anthropic (Claude)	**Claude Opus**	4.8	Deepest reasoning and the strongest coding / agent work; long-running, quality-critical tasks.
	Claude Sonnet	4.6	The everyday workhorse — writing, analysis and coding at near-flagship quality for less. Sensible default.
	Claude Haiku	4.5	Fastest and cheapest Claude; high-volume, latency-sensitive jobs.
Google (Gemini)	**Gemini Pro**	3.1 Pro	Flagship reasoning with huge context and full multimodality (text, image, audio, video); deep Google Workspace integration.
	Gemini Flash	3.5 Flash	Fast, cost-effective default for production apps and high-volume tasks.
	Gemini Flash-Lite	3.1 Flash-Lite	Cheapest Gemini; translation, classification and latency-critical work.
DeepSeek	**DeepSeek V4-Pro**	V4-Pro	Open-weight flagship with strong coding and maths at a fraction of Western pricing; can be self-hosted.
	DeepSeek V4-Flash	V4-Flash	Ultra-low-cost, high-volume workloads; open weights for on-premise use.
xAI	**Grok**	4.3	Cost-efficient frontier chat with real-time X / social data and native video input.
Meta	**Llama 4 (Scout / Maverick)**	4.5 refresh	Open-source and self-hostable — when data privacy, on-premise control or avoiding per-use API costs matter most.
Mistral	**Mistral Large / Medium**	Large 3 / Medium 3.5	Europe's open-weight family; strong multilingual performance, data-sovereignty options and cost-efficient enterprise deployment.

Speed and cost signals are folded into the "when to use it" notes; per-token pricing is left out because it changes almost weekly.

The bottom line

There is no single "best" AI engine — only the best fit for a given task, budget and risk profile. The frontier models from OpenAI, Anthropic and Google trade top honours back and forth month to month; the open models from DeepSeek, Meta and Mistral compete hard on cost and control; and xAI carves out a niche on value and real-time data. For most businesses, the win comes from understanding the trade-offs and deploying the right tool for each job — not betting the house on one.

A NOTE ON PACE

This space moves extraordinarily fast — new versions and price changes land almost weekly. The details in this guide are current as of July 2026. Always check the vendor's own model and pricing pages before committing to a production workload.

How Kinetics can help

Choosing and combining AI engines — safely, cost-effectively and in line with your obligations — is exactly the kind of decision Kinetics helps New Zealand businesses make every day. If you'd like to talk through where AI fits in your organisation, we'd love to hear from you.

kinetics.co.nz · Auckland 09 379 8200 · Christchurch 03 974 3139