**Artificer — ChatGPT-o3**

**What it is.** OpenAI’s “o3” is the most reasoning-dense model available to the public. Think of it as a CNC mill for ideas: feed it a fuzzy problem, and it iteratively carves out logic, equations, or code until the edges are sharp.

**Why it made the deck.** You rely on it as a *daily driver for hard things*—actuarial tables, SQL migrations, multi-stage prompt chains. Nothing else packs this much reliability into a single call.

**When to play it.** • Designing data schemas • Debugging opaque stack traces • Writing memos where wrong math is reputationally fatal.

**Watch-outs.** Slower and pricier than mid-range models; best used when the extra IQ will be *noticed*.

**Companion — ChatGPT-4o**

**What it is.** The “o” stands for omni: text, voice, and vision in one lightweight wrapper. It trades raw depth for latency and warmth.

**Why included.** It’s the friend who answers on the first ring—perfect for rapid notes, screen-share brainstorming, or photo-based quick fixes.

**Sweet spots.** • Annotating screenshots • Explaining a diagram by voice • Tossing together a polite reply before you board a plane.

**Limitations.** Doesn’t sweat edge cases; for rigorous analysis hand the baton back to Artificer or Strategist.

**Strategist — GPT-4.1**

**What it is.** The flagship successor to GPT-4 Turbo. Same calm prose, sharper long-range sense-making.

**Why included.** You keep reaching for it when the *shape* of a solution matters more than the pixels—product road-maps, proposal trade-offs, scenario planning.

**Plays well with.** • Causal diagrams • OKR roll-downs • Competitive-landscape memos.

**Caveats.** Costlier tokens than 4o; context still capped at 256 k, so ultra-long docs belong to Oracle.

**Prodigy — GPT-4.5 preview**

**What it is.** An early peek at OpenAI’s 4.5—smart, capable, expensive and getting retired soon! Not for long though—expect this to be the backbone of the rumored GPT-5.

**Why included.** Take a peek at the future before ChatGPT-5!

**Use cases.** • Creative writing • EQ conversations • Writing critique.

**Gotchas.** Preview means it won’t always be here, so enjoy it while it lasts.

**Polymath — Claude-4 Opus**

**What it is.** Anthropic’s largest model, famous for graceful prose and strong coding ability, especially complex coding problem-solving.

**Why included.** It threads the needle between technical rigor and executive polish—and plugs straight into Gmail/Calendar in the Claude app.

**Shines at.** • Complex refactors written like a consultant’s playbook • Policy drafts that actually *sound human*.

**Limits.** Slightly narrower tool-calling APIs than OpenAI; can stall in long chats.

**Composer — Claude 3.5 Sonnet**

**What it is.** Anthropic’s “middle-weight” model: smaller than Opus, faster than anything its size, and trained to keep a lyrical, on-brand voice. Brilliant at composing code.

**Why included.** It’s the go-to when you need *speed plus style*—edit copy live on a call, polish slides minutes before the meeting, or refactor 200 lines of code while stakeholders watch.

**Sweet spots.** • Real-time writing coach • Tone-matching executive summaries • Rapid code patches with clean explanations.

**Watch-outs.** Hard problems still belong to Opus or Artificer; token context tops out sooner than you want.

**Oracle — Gemini 2.5 Pro / Ultra**

**What it is.** Google’s flagship with million-token context and a strong fact-checking layer.

**Why included.** When the brief is *“read this mountain of stuff and tell me what’s true,”* Oracle is the only model that can swallow the whole mountain in one bite.

**Shines at.** • Cross-checking stats • Summarising litigation troves • Ingesting entire wikis for “one-pager” outputs.

**Limits.** Thinks more than 4o, occasionally overly cautious, has a little bit of fuzziness following some instructions—have Strategist tighten conclusions.

**Scholar — ChatGPT Deep Research**

**What it is.** A special OpenAI variant tuned for long, self-directed research runs (10–20 minutes).

**Why included.** It’s your personal grad-student: disappear, scour the web & PDFs, return with citations and a mini-literature review.

**Sweet spots.** • Competitive landscapes • Historical chronologies • “Explain this like I’m new, with sources.”

**Watch-outs.** Not interactive in real time—hand the baton back to Companion for quick Q&A.

**Prospector — DeepSeek R1**

**What it is.** A Chinese-led open model that just cracked the top-10 in code benchmarks—at a fraction of Big-Tech inference cost.

**Why included.** Perfect for *high-volume generation* when you can’t afford GPT pricing or need strong Mandarin ability.

**Use cases.** • Bulk code stubs • STEM problem sets • Bilingual chatbot demos.

**Caveats.** Slightly lower on scores than o3. More importantly, the Terms of Service aren’t friendly to most US or EU business use-cases.

**Collective — Mixtral 8 × 22B**

**What it is.** An open Mixture-of-Experts (MoE): eight specialist subnetworks “vote” on every token.

**Why included.** It delivers near-Claude quality while staying fully open-weight—ideal for privacy-sensitive or air-gapped deployments.

**Shines at.** • Self-hosted knowledge bots • On-prem RAG stacks • Fast fine-tunes with low GPU budgets.

**Limits.** MoE routing sometimes misfires on niche queries; heavier model-loading logistics than plain Llama.

**Voyager — Yi-1.5 200B**

**What it is.** 01.AI’s large bilingual model, fluent in English and Chinese and permissively licensed.

**Why included.** The best bridge when you need to serve both EN and ZH users *without* juggling two vendors.

**Sweet spots.** • Cross-border product copy • Dual-language knowledge bases • Fine-tunes for APAC markets.

**Watch-outs.** High GPU memory footprint; English style slightly stiffer than Opus—polish with Composer if tone matters.

**Diplomat — Cohere Command R+**

**What it is.** Cohere’s enterprise-tuned model, optimised for Retrieval-Augmented Generation (RAG) and multi-step tool use.

**Why included.** It plugs straight into long corporate docs, and is set up to play nicely with corporate workflows from the start.

**Plays well with.** • Internal knowledge bots • Compliance summaries • Multi-step workflows.

**Limits.** Designed for workflows more than writing; for marketing copy pair with Companion or Composer.

**Commons — Llama 3 405B**

**What it is.** Meta’s largest open model—basically “ChatGPT-3.5 quality, but yours to tinker with.”

**Why included.** It’s the default choice for universities, startups, and hobbyists who want strong performance and full control.

**Use cases.** • Private-cloud assistants • Domain-specific fine-tunes • Academic experiments.

**Caveats.** No built-in tools API; connect external RAG or function-calling layers yourself.

**Navigator — Mistral Large**

**What it is.** A large model cousin to Mixtral; designed for intelligence inference and coding abilities.

**Why included.** An EU-friendly model with strong multi-lingual capabilities.

**Shines at.** Instruction following abilities, strong at coding and inference.

**Limits.** Not a true top 3 model; for heavy analysis use Artificer or Strategist.

**Maverick — Grok-2**

**What it is.** xAI’s second generation, personality-forward model with a cheeky tone and access to real-time information.

**Why included.** Great for *free-wheeling ideation* and keeping up to date on X (where it gets tagged all the time).

**Sweet spots.** • Brainstorm sessions • Solid coding abilities • Good creative problem-solving sense

**Watch-outs.** Can veer informal; run final drafts through Composer for polish. No published model card and a number of prominent model alignment issues.

**Scout — Perplexity Sonar**

**What it is.** A retrieval-native model that always cites sources—think “AI search-engine in a box.”

**Why included.** Fastest path to a **sourced** bullet list when you don’t have time to double-check links yourself.

**Use cases.** • Quick news briefs • “Show me five recent papers on…” • Pre-meeting fact scouting.

**Caveats.** Depends on live web access, purpose-built for search.