

PickAIModel Brief

Issue 009 · July 12, 2026 · pickaimodel.com

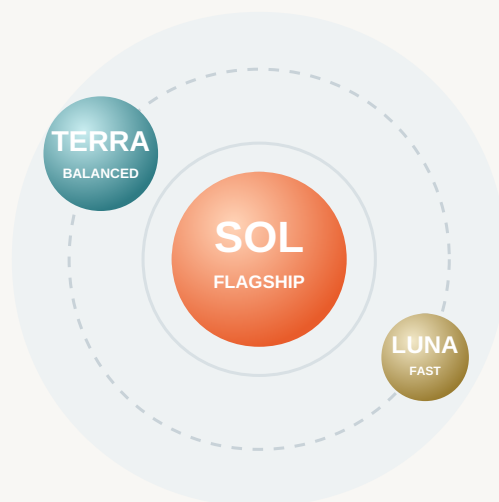
Independent guide to the AI industry, useful tools, and practical model choices.

THE STORY THAT SETS THE ISSUE

OpenAI turns one frontier release into three distinct products

GPT-5.6 Sol, Terra and Luna are not just sizes. They formalise a new buying decision: maximum capability, balanced daily work, or low-cost speed.

A new publication format. Previous PickAIModel newsletters were delivered in Markdown. Starting with this issue, the Brief is published in Adobe PDF format, allowing richer typography, visual summaries and embedded graphics while preserving clickable source links.



One generation · three workload tiers

OpenAI GPT-5.6

What Sol, Terra and Luna change for ChatGPT, Codex and API buyers.

Anthropic pricing

Sonnet 5's tokenizer tax and Fable 5's move beyond flat subscriptions.

Rules for machine traffic

China narrows companion AI; Cloudflare separates Search, Agent and Training bots.

GPT-5.6 makes the model picker a workload decision

OpenAI made **GPT-5.6** generally available on July 9 across ChatGPT, Codex and the API. The important change is not simply a new flagship. OpenAI has split the generation into durable tiers: Sol for the hardest work, Terra for balanced professional use, and Luna for speed and cost control.

In standard ChatGPT conversations, eligible paid users get [Sol through Medium, High and Extra High reasoning settings](#); Sol Pro is reserved for the most demanding work on supported plans. Terra and Luna are available in ChatGPT Work, Codex and the API rather than the normal ChatGPT model picker.

For developers, the family also introduces [Programmatic Tool Calling](#), where the model can write lightweight in-memory programs to coordinate tools and filter intermediate results. The new max setting gives Sol more time to reason; `ultra` coordinates parallel agents for work where time-to-result matters more than token restraint.

Why it matters: Model choice is becoming an operations decision. Pick the tier around task value, latency and failure cost instead of routing every request to the most capable model.

Primary sources: [OpenAI launch announcement](#) · [ChatGPT availability guide](#) · [GPT-5.6 system card](#)

A PRACTICAL MODEL MAP

Use the cheapest tier that can reliably finish the job.

- **Sol:** complex coding, research, computer use, design and long-running multi-step work.
- **Terra:** everyday professional workflows where capability and cost need to stay balanced.
- **Luna:** high-volume, latency-sensitive tasks and lower-cost automation.

Sol

FLAGSHIP

\$5 input
\$30 output
per 1M API
tokens

Terra

BALANCED

\$2.50 input
\$15 output
per 1M API
tokens

Luna

FAST

\$1 input
\$6 output
per 1M
API tokens

Availability nuance. Terra and Luna are not selectable in ordinary ChatGPT chats. They are positioned for Work, Codex and API use, while Sol powers the paid reasoning modes in ChatGPT.

30 min

minimum prompt-cache life, with explicit cache breakpoints

4

parallel agents coordinated by default in the `ultra` setting

The real product change is not “a smarter model.” It is a clearer way to buy intelligence at different levels of urgency and cost.

Anthropic changes both the default model and the subscription bargain

Claude

API migration

Sonnet 5 is the default now. The tokenizer is the invoice.

[Anthropic released Claude Sonnet 5 on June 30](#) and made it the default for Free and Pro users. Introductory API pricing is \$2 per million input tokens and \$10 per million output tokens through August 31, after which it becomes \$3 and \$15.

The migration detail matters more than the list price. [Anthropic says the new tokenizer produces about 30% more tokens for the same text](#). Existing prompt budgets, context assumptions and max_tokens limits therefore need to be measured again. Non-default temperature, top_p and top_k values now return errors, and manual thinking budgets have been replaced by adaptive thinking and an effort setting.

Why it matters: Treat Sonnet 5 as a migration, not a model-ID swap. Recount tokens and rerun cost tests before the introductory price ends.

Sources: [Anthropic launch post](#) · [Anthropic migration guide](#)

Consumer AI

Usage billing

Fable 5 is back, but the flat-rate subscription no longer covers it.

[Anthropic restored Fable 5 globally on July 1](#) after the US government lifted export restrictions. The free-in-plan period was temporary: [from July 12 at 11:59 PM PT, subscribers must buy usage credits](#) at the API rate of \$10 per million input tokens and \$50 per million output tokens.

That makes Fable 5 a consumer product with two bills: the monthly Claude subscription and metered model usage. Anthropic told WIRED it wants to return Fable to subscriptions when capacity allows, but gave no date. The company is testing whether users will pay separately for frontier-level reasoning rather than expecting every model inside one monthly plan.

Why it matters: Premium subscriptions may become access passes rather than all-inclusive plans. Heavy users should estimate task-level spend before making Fable their default.

Sources: [Anthropic redeployment note](#) · [WIRED reporting](#)

MIGRATION CHECKLIST

- | | |
|---|--|
| 01 Recount representative prompts with the Sonnet 5 tokenizer. | 02 Remove non-default sampling parameters. |
| 03 Retune output limits near truncation boundaries. | 04 Compare cost per completed task after August 31 pricing. |

THE PRICING SIGNAL

Subscriptions are being separated from compute.

Agentic models can consume far more inference than ordinary chat. Anthropic's Fable credits are the clearest consumer test yet of whether premium intelligence will be metered on top of a monthly plan.

China's companion-AI rules were not a nine-day surprise

China's rules for AI services that simulate human personality and sustained emotional interaction were [published on April 10](#) and take effect on July 15. The regulation excludes ordinary customer service, knowledge Q&A, workplace assistants, education and scientific research when they do not involve sustained emotional interaction.

The immediate product impact is now visible. [ByteDance's Doubao said its custom agent feature would go offline on July 15](#); Alibaba's Qwen disabled humanlike and user-created agents on July 10 and broader agent services on July 15. Some settings and prior conversations will no longer be accessible after shutdown.

The regulation does not simply prohibit memory or personalisation. It creates a dedicated compliance regime for systems designed to maintain emotionally humanlike relationships. Builders should distinguish between a useful assistant with a stable profile and a product whose core value depends on simulated attachment.

Why it matters: Companion design choices are becoming regulated product features. Teams should document age controls, emotional-dependency safeguards, data retention and exit paths before entering this category.

Sources: [Official measures](#) · [CAC explanatory Q&A](#) · [South China Morning Post](#)

WHAT THE RULES COVER

REGULATED CATEGORY

Sustained emotional interaction that simulates human personality, thinking and communication



GENERALLY EXCLUDED

Customer service · Q&A · workplace assistants
education · scientific research, without emotional bonding

- Apr 10** | Rules published by five Chinese departments.
- Jul 10** | Qwen disables humanlike and user-created agents.
- Jul 15** | Rules take effect; broader Doubao and Qwen changes complete.

TECHNOLOGY SHIFT OF THE ISSUE

Cloudflare now asks what a crawler is doing, not just who operates it

[Cloudflare has split AI crawler controls into three behaviours](#): Search, Agent and Training. Every customer, including Free-plan users, can allow or block each category independently. From September 15, new domains will block Agent and Training bots on ad-supported pages by default while allowing Search.

The distinction is operationally useful. A search crawler may return visitors; an agent may browse on a user's behalf; a training crawler may extract content without creating a direct audience relationship. Cloudflare also says multi-purpose crawlers will inherit the most restrictive applicable rule, so site owners need to review mixed-use bots rather than assume a search label overrides everything.

Cloudflare AI bot controls

A practical way to separate discoverability, user-directed browsing and model training instead of treating every automated visitor as the same thing.



Search

Crawlers indexing your pages so content can appear in search results and answer systems.



Agent

Automated visits made in real time on behalf of a person, such as browser-use agents.



Training

Crawlers collecting content to train or fine-tune models.

Best for

Publishers, blogs, e-commerce sites and any business whose content, ad inventory or server capacity has measurable value.

Reason to try it now

The controls are already live for all Cloudflare customers. The September 15 defaults for new domains make this a policy decision worth settling deliberately.

Caution

Blocking Training may also affect mixed-purpose crawlers. Cloudflare says Googlebot, Applebot and BingBot can be blocked where a customer applies the restrictive Training rule. Test the impact on indexing and referrals.

[Open the Cloudflare configuration guide](#)

Non-affiliate link

THE NUMBER THAT CHANGES THE PICTURE

43%

of all global startup funding in the first half of 2026 went to just two companies: OpenAI and Anthropic.

[Crunchbase reported \\$510 billion in global startup investment during H1 2026](#), with OpenAI and Anthropic accounting for \$217 billion. The number matters less as a trophy than as a market-structure warning: capital is concentrating around the labs that can afford frontier training, distribution and data-centre commitments.

For application startups, that creates two opposing effects. Better models and lower inference costs arrive faster, but investors may reserve a larger share of late-stage capital for the infrastructure and model providers beneath those applications.

THE WEEK IN ACCIDENTAL HONESTY

Anthropic says Fable 5 will return to subscription plans when capacity allows. Until then, the subscription buys access to the meter.

A five-minute operating checklist

Using ChatGPT or Codex?

- Use Sol for work where a wrong or incomplete result is expensive.
- Test Terra on recurring professional workflows before paying Sol rates.
- Route high-volume, time-sensitive automation to Luna when quality is sufficient.

Building companion AI?

- Separate stable personalisation from simulated emotional dependency.
- Give users a clear export and deletion path.
- Document age, escalation and mental-health risk controls.

Migrating to Sonnet 5?

- Recount tokens on real prompts.
- Remove unsupported sampling parameters.
- Retest context and output limits.
- Price the workflow again after August 31.

Running a content site?

- Review Cloudflare's Search, Agent and Training settings.
- Check mixed-use crawler behaviour before blocking Training.
- Separate human engagement metrics from automated traffic.

NEXT ISSUE PREVIEW

What we are watching

GPT-5.6 outside the launch window

The first useful evidence will be cost per completed task, reliability across long workflows and whether Terra can replace more expensive defaults in production.

China's July 15 implementation

We will watch whether companies redesign companion features for compliance or remove them, and what users can recover from stored conversations.

Fable 5's metered consumer test

The question is not whether some users will pay. It is whether usage credits become a normal layer on top of premium AI subscriptions.

Cloudflare's crawler taxonomy

The practical issue is how multi-purpose bots behave once publishers start applying separate rules at scale.

Forwarded this issue?

Subscribe at pickaimodel.com/newsletter.

Have a question for a future issue? Reply to the email - we read everything.

Editorial note: This revision was checked against primary product announcements, official regulatory material and original reporting available through July 12, 2026. Product availability and pricing can change after publication.

Editorial and legal notes

Editorial independence

PickAIModel.com produces independent editorial content. Model rankings, quality scores, and value scores are determined by our published methodology and are not influenced by commercial relationships with any AI vendor. No company can pay for ranking position, score changes, inclusion in rankings, or favourable treatment in our methodology outputs.

AI-assisted content disclosure

Portions of the editorial summaries and commentary in this newsletter may be drafted with the assistance of AI language models and reviewed by the PickAIModel editorial team. Benchmark scores, pricing fields stored in PickAIModel's own system, and ranking data are not invented by the model and are not editorially altered to favour advertisers or affiliates. Routine leaderboard comparisons are handled on PickAIModel.com rather than repeated in this newsletter unless a benchmark event is unusually consequential.

Affiliate disclosure

This newsletter may contain affiliate links. If you click a qualifying link and make a purchase, PickAIModel.com may earn a commission at no additional cost to you. Affiliate relationships do not influence our rankings, scores, or methodology outputs.

Not financial or legal advice

Nothing in this newsletter constitutes financial, investment, or legal advice. References to funding rounds, valuations, IPO timelines, pricing, contracts, or company strategy are provided for informational purposes only. Make decisions based on your own judgment and, where appropriate, qualified professional advice.

Accuracy and currency

AI model pricing, capabilities, availability, and company claims change frequently. While we aim to be accurate at the time of publication, information may become outdated or be revised after publication. Verify critical purchasing details directly with the relevant vendor before acting.

Source and methodology note

Product facts were checked against OpenAI, Anthropic and Cloudflare materials. Regulatory claims were checked against China's Cyberspace Administration. Product shutdowns, consumer billing and funding concentration were cross-checked with original reporting.

Publication format

Issue 009 is the first designed PDF edition of PickAIModel Brief. It uses embedded vector graphics and clickable hyperlinks. An accessible text manuscript is available from the publisher.

To unsubscribe, use the unsubscribe link in your email footer. Copyright © 2026 PickAIModel.com. All rights reserved. NeuroSync AI Dynamics Pty Ltd.