

Sakana AI launches Fugu, an API that routes work across several frontier models

Source copy kept by wheresthe.ai for the story "Sakana AI launches Fugu, an API that routes work across several frontier models" (<https://wheresthe.ai/d/jp-sakana-fugu-orchestration-api-2026-06/>).

Source: <https://datanorth.ai/news/sakana-ai-launches-fugu-and-fugu-ultra>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: a8c9e9212f3f6f7bf88b21ab349ec5b1a1ae14ccc3f2ca445f4a00e14503732d

Sakana AI launches Fugu and Fugu Ultra 24-06-2026 Sakana Fugu and Fugu Ultra are a multi-model orchestration API released by Sakana AI with a 1 million token context window and benchmark scores that the company reports as matching Anthropic Fable 5 and Mythos Preview. Written by:

Jorick van Weelie Marketing Lead at DataNorth | AI Enthusiast & Tech Storyteller

Sign up for our Newsletter

Published: June 24, 2026

Sakana AI, the Tokyo-based AI lab, has released Fugu and Fugu Ultra, a pair of models delivered through a single API that route each request across a pool of frontier language models rather than running as one fixed model. Fugu is the lower-latency default and Fugu Ultra is the quality-first variant built for harder multi-step tasks, with Sakana AI reporting that Fugu Ultra matches Anthropic Fable 5 and Mythos Preview on several engineering, science, and reasoning benchmarks. Both models offer a 1 million token context window and are available now through the Sakana AI API and OpenRouter.

What are Sakana Fugu and Fugu Ultra?

Sakana Fugu is a multi-agent orchestration system delivered as a single model API. Instead of answering from one fixed network, Fugu breaks a problem into smaller tasks, assigns each task to one of several frontier language models, and combines the results into a single response. Depending on the difficulty of the request, it routes work to between one and three agents. Sakana AI says the design draws on two peer-reviewed papers accepted at ICLR 2026.

The pool of underlying models is publicly accessible and swappable, which means Fugu can compose its answers from whichever frontier models are available rather than depending on a single provider. Fugu is the lower-latency default for everyday work, while Fugu Ultra is the quality-first variant Sakana AI positions for harder multi-step tasks where answer depth matters more than response time. Both are reached through the same API, so callers select the tier rather than managing individual models.

Fugu Ultra benchmarks and technical specifications

On Sakana AI's own testing, Fugu Ultra scores:

95.1 on GPQA Diamond,

93.2 on LiveCodeBench v6,

54.2 on SWE-bench Pro.

Sakana AI reports that these results place Fugu Ultra above Anthropic Opus 4.6, Gemini 3.1 high, and GPT-5.4 high on all three benchmarks, and shoulder to shoulder with Anthropic Fable 5 and Mythos Preview. These figures are vendor-published rather than independently verified, so they should be read as Sakana AI's own evidence.

Fugu Ultra has a context window of 1 million tokens and a maximum output of 131,000 tokens. Because Fugu Ultra is an orchestration layer rather than a single trained model, its capability comes from how it decomposes a task and routes the parts to specialist models, then merges their output. That design is what Sakana AI credits for the benchmark scores, and it is also why latency is higher than a single-model call, particularly on Fugu Ultra.

How do Fugu and Fugu Ultra compare to Anthropic Fable 5 and Mythos?

Sakana AI positions Fugu Ultra as a direct alternative to Anthropic Fable 5 and Mythos Preview, the models Fugu

Ultra claims to match on benchmarks. The practical difference is access. On June 12, 2026, Anthropic's most capable models became subject to United States export controls citing national security, which made Fable 5 and Mythos inaccessible to organizations in a broad set of countries. Fugu is built to route around that constraint by composing its answers from publicly accessible frontier models, which Sakana AI frames as a way to reduce vendor lock-in.

The benchmark parity comes with caveats. The scores are Sakana AI's own and have not been independently confirmed, and early testers have reported a gap between benchmark results and real-world use. Ethan Mollick noted that his usual coding tests took about 30 minutes to run on Fugu Ultra and that the output did not match Fable 5 in practice. The orchestration approach also means response times are longer than a single-model call, which matters for interactive workloads.

Fugu and Fugu Ultra availability and pricing

Fugu and Fugu Ultra are available now through the Sakana AI API and through OpenRouter, alongside a subscription option. Sakana AI lists the fugu-ultra-20260615 model at 5 US dollars per 1 million input tokens, 30 US dollars per 1 million output tokens, and 0.50 US dollars per 1 million cached input tokens, with higher rates applied to context above 272,000 tokens.

Standard Fugu is the lower-latency, lower-cost default for general tasks, while Fugu Ultra carries the higher rates above and is intended for the hardest multi-step work. Because both tiers share one API and a swappable pool of underlying models, teams can move between them without re-integrating, which is central to Sakana AI's argument that Fugu reduces dependence on any single model provider.

For full details, see Sakana AI's Fugu release announcement .

Add DataNorth AI to your Google favorites Discover more news

Meta's Muse Spark Open runs on two Mac Studios

ChatGPT Voice gets GPT-Live-1.1, with 13 new voices in seven languages and more natural accents

Meta's AI camera glasses face a Dutch court case that asks for changes or a sales ban

GPT-4o and GPT-4.1 leave OpenAI's API on 14 January 2027, fine-tuned versions included

Back to the newspaper