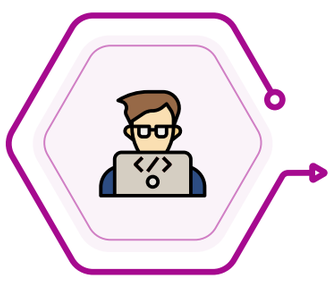
Routing Around the Giants:
Sakana's Fugu Bets Orchestration Beats Bigger Models

When access to the world's most powerful AI models can be switched off overnight, owning one of them stops looking like a strength and starts looking like a liability. That is the wager behind Fugu, the multi-agent orchestration system that the increasingly enterprise-focused Japanese startup Sakana AI launched late on a Sunday night. Rather than build another monolithic foundation model, Sakana built a coordinator that conducts a swappable pool of specialized agents and delivers frontier-level results through a single, OpenAI-compatible API.



Fugu, named for the Japanese word for pufferfish, is aimed squarely at developers, enterprises, and even governments looking for resilience against vendor lock-in and the growing reach of geopolitical export controls. Instead of routing every request to one model, it dynamically dispatches queries across a rotating roster of expert models, verifies their output, and stitches the results back together before the user ever sees them.

Why Now: A Frontier-Model Gap Opens Up

The timing is no accident. On June 12, Anthropic revoked public access to its two most capable models, Claude Mythos 5 and Claude Fable 5, following a U.S. government export control order. For enterprises that had wired those models into critical workflows, the message landed hard: access to a top-tier model is not guaranteed, and it can vanish without warning.

01

Sakana CEO and co-founder David Ha, a Google Brain alumnus, used the moment to position Fugu as a steadier foundation than any single provider. In remarks shared publicly, Ha argued that a well-coordinated pool of interchangeable agents can rival restricted frontier models like Fable and Mythos, and that orchestration models, not ever-larger single models, represent the next real frontier. Leaning on one company's model for national infrastructure, he warned, is an enormous risk, and collective intelligence is the practical hedge against that concentration of power because an orchestrator can simply route around whoever becomes unavailable.

02

Notably, Sakana keeps the specifics of that orchestration hidden by design. The company states that exactly which models Fugu selects, and how it coordinates them, is proprietary. Its documentation refers only in general terms to a diverse pool of powerful models, multiple LLMs, or specialized models, without ever disclosing a count.

How Fugu Works

At its core, Fugu behaves like a master general contractor. Handed a complex request, it does not try to do everything itself. It breaks the problem into parts, delegates sub-tasks to a pool of expert foundation models, checks their work, and synthesizes a final answer. According to Sakana's technical release, Fugu is itself a large language model, trained to call other models in an agent pool, including, recursively, additional instances of itself.

The system draws on two of Sakana's 2026 research efforts, TRINITY and the Conductor, to autonomously manage the full lifecycle of model selection and verification through learned coordination strategies rather than hand-built workflows. To the end user, all of that machinery disappears behind a standard API endpoint.

Two Tiers for Two Kinds of Work

Sakana is shipping Fugu in two variants. The standard Fugu is a high-speed, low-latency engine tuned for everyday tasks, built to power interactive chatbots and to slot directly into coding environments such as Codex. Fugu Ultra is the flagship tier, engineered for complex, high-stakes work like AI research, cybersecurity analysis, and multi-step patent investigations. Ultra coordinates a deeper bench of experts and, by Sakana's account, matches industry-leading monolithic models across demanding scientific and reasoning benchmarks.

Where Fugu Wins, and Where It Still Trails

On benchmark charts shared by Sakana, the orchestration approach pays off on messy, multi-step problems. On LiveCodeBench, an open-source test of coding on continually refreshed software problems, Fugu Ultra scored 93.2 and standard Fugu 92.9, both edging out Claude Fable 5 at 89.8. On GPQA-Diamond, a set of 198 graduate-level science questions across biology, physics, and chemistry, both Fugu tiers hit 95.5, ahead of the earlier Claude Mythos Preview at 94.6.

01

The agentic coding results are where collective intelligence looks strongest. On SWE-Bench Pro, Fugu Ultra posted 73.7, comfortably above Anthropic's Claude Opus 4.8 at 69.2 and OpenAI's GPT-5.5 at 58.6. By orchestrating models from different providers, Fugu also bakes redundancy directly into the stack: if one provider goes down or is suddenly restricted, the system routes around the gap to preserve uptime.

02

Fugu is not, however, a clean sweep. Against highly specialized or restricted-access monolithic models, it sometimes falls short. On SWE-Bench Pro, Fugu Ultra's 73.7 was eclipsed by the limited-access Fable 5 at 80.0, a model currently absent from Fugu's pool because of the U.S. export order. On Humanity's Last Exam, Fugu Ultra's 50.0 narrowly beat Opus 4.8 at 49.8 but again trailed Fable 5 at 53.3. On the MRCRv2 long-context-recall test, GPT-5.5 held the lead at 94.8 against Fugu Ultra's 93.6, and Opus 4.8 stayed on top of the CTI-REALM cybersecurity benchmark at 69.6 versus Fugu Ultra's 69.4. The pattern is clear: Fugu excels at sprawling, multi-step tasks by combining the strengths of several models, but for brute-force reasoning inside a single narrow domain, the biggest standalone models still lead, so long as an enterprise can keep uninterrupted access to them.

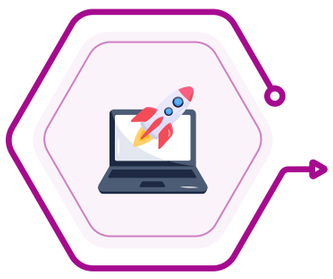
Licensing, Availability, and Data Controls

Fugu is a commercial, proprietary API service rather than an open-source framework. Because Sakana's core intellectual property lies in its non-obvious collaboration patterns, the routing details, meaning exactly which underlying models are chosen for a given query, stay hidden from the user by design.

For enterprises with compliance obligations, Sakana does offer meaningful controls. Developers can explicitly exclude specific models or providers from their Fugu routing pool to maintain strict privacy standards, and users can opt out of having their prompts used for future training. Geographically, Fugu is restricted from operating within the European Union and the European Economic Area while Sakana works to bring its opaque data-routing architecture into line with GDPR.

Pricing Runs Steep

Fugu is available immediately in most regions, with the temporary exception of the EU and EEA, through both subscriptions and pay-as-you-go pricing. Monthly subscription allowances are pitched at individual and hands-on users: a Standard tier at $20 per month for lightweight workflows, a Pro tier at $100 per month offering ten times the standard usage, and a Max tier at $200 per month providing twenty times the usage for continuous, long-running tasks. The exact token allotments under these plans were not disclosed at launch. As an opening incentive, Sakana is giving a free second month to anyone who subscribes to any tier by July 31, 2026.



For production deployments, Sakana offers an elastic pay-as-you-go plan whose requests are served at higher priority than subscription traffic. Under this model, the standard Fugu engine charges the single rate of the highest-tier underlying model involved in a query, without stacking multi-agent fees. Fugu Ultra uses fixed per-million-token pricing of $5 for input, $30 for output, and $0.50 for cached input, rising to $10, $45, and $1.00 for extreme workloads using context windows above 272K tokens. That places it among the pricier options relative to single-model provider APIs, as the snapshot below shows.

Frontier AI model API pricing snapshot (per 1M tokens)

Model	MoInput / 1Mdel	Output / 1M	Total
MiMo-V2.5 Flash	$0.10	$0.30	$0.40
deepseek-v4-flash	$0.14	$0.28	$0.42
deepseek-v4-pro	$0.43	$0.87	$1.305
MiniMax-M3	$0.30	$1.20	$1.50
Gemini 3.1 Flash-Lite	$0.25	$1.50	$1.7
Qwen3.7-Plus	$0.40	$1.60	$2.00

Model	MoInput / 1Mdel	Output / 1M	Total
MiMo-V2.5	$0.40	$2.00	$2.40
Grok 4.3 (low context)	$1.25	$2.50	$3.75
MiMo-V2.5 Pro (<=256K)	$1.00	$3.00	$4.00
Kimi-K2.6	$0.95	$4.00	$4.95
GLM-5.2	$1.40	$4.40	$5.80
Grok 4.3 (high context)	$2.50	$5.00	$7.50
MiMo-V2.5 Pro (>256K)	$2.00	$6.00	$8.00
Qwen3.7-Max	$2.50	$7.50	$10.00
Gemini 3.5 Flash	$1.50	$9.00	$10.50
Gemini 3.1 Pro Preview (<=200K)	$2.00	$12.00	$14.00
GPT-5.4	$2.50	$15.00	$17.50
Gemini 3.1 Pro Preview (>200K)	$4.00	$18.00	$22.00
Claude Opus 4.8	$5.00	$25.00	$30.00
GPT-5.5	$5.00	$30.00	$35.00
Sakana Fugu Ultra	$5.00	$30.00	$35.00
Claude Fable 5 / Claude Mythos 5	$10.00	$50.00	$60.00

There is also an architectural catch worth modeling carefully. According to the developer documentation, Fugu Ultra's API responses separate user-visible token generation from internal orchestration work, and the background tokens consumed when Fugu delegates sub-tasks, verifies code, or routes between agents are not absorbed by the provider. They count as real usage billed at standard rates toward the final price of each request.

Routing vs. Orchestration: Where Fugu Sits

To place Fugu in the mid-2026 landscape, it helps to separate model routing from multi-agent orchestration. Over the past year, enterprise adoption of routing platforms such as Not Diamond, Martian, and the open-source RouteLLM has surged. These act like intelligent air traffic controllers, using classifiers or meta-models to predict which single model will best handle an incoming prompt, then dispatching the query there.

Fugu works on a different principle. Rather than making a one-shot routing decision, it aligns more closely with multi-round systems like Router-R1, the framework introduced at NeurIPS 2025. It decomposes a query, interleaves reasoning with delegation, and assigns sub-tasks to multiple models in parallel or in sequence before synthesizing the result. Frameworks such as LangGraph, CrewAI, and Microsoft AutoGen let developers build comparable systems, but they demand heavy manual configuration of roles, conditional logic, and state across long loops. Fugu hides that overhead entirely, packaging a LangGraph-style workflow as a single black-box endpoint.

The Company Behind the Pufferfish

Sakana AI was founded in Tokyo in 2023 by Llion Jones, a co-author of Google's foundational 2017 paper Attention Is All You Need, and David Ha, the former head of research at Stability AI. Frustrated by big-company bureaucracy and the industry's fixation on scaling ever-larger single models, the founders built Sakana around biomimicry and evolutionary computing. The company's name, Japanese for fish, reflects its central thesis: collective swarm intelligence over brute-force compute. After a $2.6 billion Series B valuation in late 2025 and the June 2026 launch of Marlin, an autonomous eight-hour research agent for the B2B market, Fugu represents the commercialization of that multi-agent routing technology for everyday developers.

A Mixed Reception

The developer community has met Fugu with hands-on scrutiny, weighing its routing efficiencies against the raw power of monolithic models. The AI commentator and developer known as Chris argued that for a single clean prompt, a developer would probably still reach for Fable 5, Mythos, or GPT-5.5 directly, but that Fugu's value emerges in messy, multi-step environments involving delegation, verification, synthesis, code review, research loops, and security analysis. He also flagged the strategic upside: if frontier access is abruptly revoked, an orchestrator can swap models to avoid a total outage.

01

Creative agency owner Mark Santos ran a direct comparison, asking both Fugu Ultra and Claude Opus 4.8 to build a Crossy Road clone in Three.js. Fugu Ultra finished in 22 minutes using roughly 89,000 tokens for about $7.32, though the result carried minor logic flaws like inverted turns and awkward camera angles. Opus 4.8 took 79 minutes, burned roughly 940,000 tokens for nearly $37.85, and got stuck in a retry loop that needed human intervention, yet ultimately produced superior design and functionality. His verdict split the decision: Opus won on application quality, while Fugu won on speed and performance.

02

Skeptics pushed back harder on the sovereignty pitch. Prime Intellect research engineer Elie Bakouch noted that Fugu is a closed-source orchestrator sitting on top of closed-source models, meaning users who previously did not control the models now do not even control which ones get used or how heavily, which he argued is not AI sovereignty. That sentiment echoed a common refrain in early discussions, summed up by one Reddit user who called Fugu, until proven otherwise, a highly advanced router or wrapper rather than a fundamental leap in intelligence on the order of Mythos or Fable.

The Bottom Line

Fugu is unlikely to settle the debate over whether orchestration can truly substitute for raw frontier capability. The largest standalone models still hold an edge on narrow, brute-force reasoning, and Sakana's black-box design means buyers trade transparency for convenience and resilience. But as enterprises grow increasingly wary of betting critical infrastructure on a single vendor whose best models might disappear by government order, Sakana is making a credible case that packaging collective intelligence behind one API endpoint is a genuinely viable strategy, and perhaps an essential one. In a market where access is no longer guaranteed, the ability to route around any one provider may prove just as valuable as topping a benchmark.