

Cultural Daily

Independent Voices, New Perspectives

Sakana Fugu Sells an Orchestrator as a Model, and the Bill Follows the Hardest Agent

Our Friends · Thursday, October 1st, 2026

The most consequential thing about Sakana Fugu is a sentence on its pricing page, and it is not about price. It reads: when multiple agents are active, we never stack model fees; you are charged a single rate based on the top tier model involved. Read that twice, because it describes a product that is sold like a model and billed like a job. You pick one name from a list — the same way you would pick any other entry in this month's new AI models — and behind that name a set of agents wakes up, divides your request, and reports back. What you pay is not the sum of what they did, but the rate of the most capable one that got involved, applied to the tokens of the whole exchange. On [OrcaRouter](#) the [new ai models](#) catalogue is a list of names with a rate beside each, and this release is the first entry in that shape where the name and the bill measure two different things.

One name, four products, one endpoint

Fugu is not one model. The vendor ships four under the same API surface: Fugu, Fugu Ultra, Fugu Max and Fugu Cyber. All four are reachable through an OpenAI-compatible endpoint, so switching between them is a string change rather than a migration.

Fugu balances latency against quality and is positioned as the everyday default. Fugu Ultra coordinates a deeper pool for hard, multi-step reasoning, aimed at work where a wrong answer costs more than the wait; the vendor names competition problems, paper reproduction and patent investigations as early uses. Fugu Max orchestrates the largest pool and optimises cost against performance rather than quality alone. Fugu Cyber specialises the same machinery for security work, and its usage and pricing go through a sales conversation rather than a published table.

The design claim underneath all four is stated plainly enough: instead of encoding a workflow by hand, the system learns how to assemble agents. The vendor grounds this in two papers it says were accepted at ICLR 2026 — one describing an evolved coordinator that assigns thinker, worker and verifier roles across several turns, the other a coordinator trained with reinforcement learning to find coordination strategies in natural language. Whether the learned patterns beat a well-designed hand-built pipeline is exactly the kind of question a benchmark cannot settle, and the vendor does not claim otherwise; it claims the learned version spares you designing the pipeline.

The billing rule that matters

Three pricing structures sit side by side and they answer different questions.

Fugu, the general tier, is charged at the standard rate of whichever underlying model is doing the work when a single agent is active — and, when several are, at the rate of the highest tier among them. If a request fans out to four agents and one of them is a frontier model, you are not billed for four calls. You are billed for one stream of tokens at the frontier model's rate.

Fugu Ultra is priced as a flat ladder rather than a pass-through: five dollars per million input tokens, thirty per million output, fifty cents per million cached input, with a step up to ten, forty-five and one dollar once the context passes 272,000 tokens. Those numbers do not move with how many agents the orchestrator decided to wake. The vendor states them as fixed for the `fugu-ultra-v2.0` line.

Fugu Max is flatter still — two dollars in, six dollars out, twenty-five cents cached, and the vendor notes these rates hold regardless of context length. Its one metered extra is tool use: `web\_search` and `web\_fetch` each cost \$0.007 per call — the only line item in the table that scales with how much exploring the orchestrator chooses to do.

There is also a subscription side: twenty dollars a month for light daily use, a hundred for ten times the allowance, two hundred for twenty times; the vendor positions these for everyday hands-on use.

Why a fixed rate is the interesting part

A fixed rate card plus a system that decides at runtime how many models to use gives a cost structure unlike anything else on a per-token price list. On an ordinary model cost is proportional to work: more tokens, more money. On an orchestrator sold at a flat rate, cost is proportional to the request, and the compute the vendor spends answering it is a variable you neither see nor control. A terse question that triggers a deep pool costs the same per token as a shallow one. The vendor has taken on the risk of its own routing decisions and priced it into one number. A monthly bill is estimable from token volume; what token volume cannot do is compare this product against a single model, because a token here is not the same unit of work.

What the benchmarks do and do not say

The vendor publishes quantitative results for two of the four tiers, and they are worth reading precisely for their limits.

For Fugu Ultra v2.0 it states best or joint-best scores on five of eight benchmarks and a top-two placing on seven of eight. It also states something less common: the training cutoff is 28 August 2026, and three named frontier models are explicitly not in the pool it orchestrates. Naming the exclusions matters, because the whole argument for an orchestrator is that it draws on strong models — and here the vendor is telling you which strong models it does not draw on.

For Fugu Max v1.0 the claim is best overall on six benchmarks, including terminal, graduate-level question answering, long-context reasoning and automation suites; the vendor also says the pool includes an NVIDIA model family through a collaboration with that company. For Fugu Cyber it reports success rates of 86.9 percent and 72.1 percent on two security benchmarks.

What is absent is any independent measurement. We looked for a page for this model on the neutral third-party index we use elsewhere in this series and found none, so this article quotes no capability score for it — and we will not fill that gap with the vendor's own numbers dressed as

someone else's. Read the vendor's figures as a description of what the builder measured, not as a verdict.

sakana fugu multi-agent system as a model What is Sakana Fugu ? How to Use Users' Voices Pricing FAQ Contact Us English

Sakana Fugu

One Model to Command Them All

Frontier-level performance without single-vendor dependency. Fugu dynamically orchestrates the world's best models to tackle complex, multi-step tasks. Plug collective intelligence directly into your workflows today with a single API.

[Start Using Sakana Fugu](#) [Contact Us](#)

Not yet available in the EU/EEA while we work toward compliance with GDPR and EU-specific regulations.

Sakana Fugu is also available on openrouter Vercel opencode CREA MERGE

The pool is the product, and you mostly cannot edit it

The second-order question about any orchestrator is who is in the pool, because that determines both what the system can do and whose terms you are accepting.

The answer here is split. Fugu Ultra's pool is fixed, because the vendor says the model relies on the full pool to hit its numbers. Fugu Max's pool is fixed too, because it optimises cost and performance together and cannot do that against a moving set. Only the base Fugu tier lets you opt specific models or providers out from a console setting, for data, privacy or compliance reasons. Custom configurations go through a sales conversation.

There is also a lag to plan around: when a new frontier model appears publicly, the vendor expects to spend roughly two weeks training and evaluating an updated Fugu before rolling it out. If your reason for buying an orchestrator is that it automatically gets the newest models, two weeks is the interval you are buying.

And there is a geographic limit in the open: no EU or EEA availability while the vendor works toward GDPR and EU-specific compliance.

What to check before routing production traffic

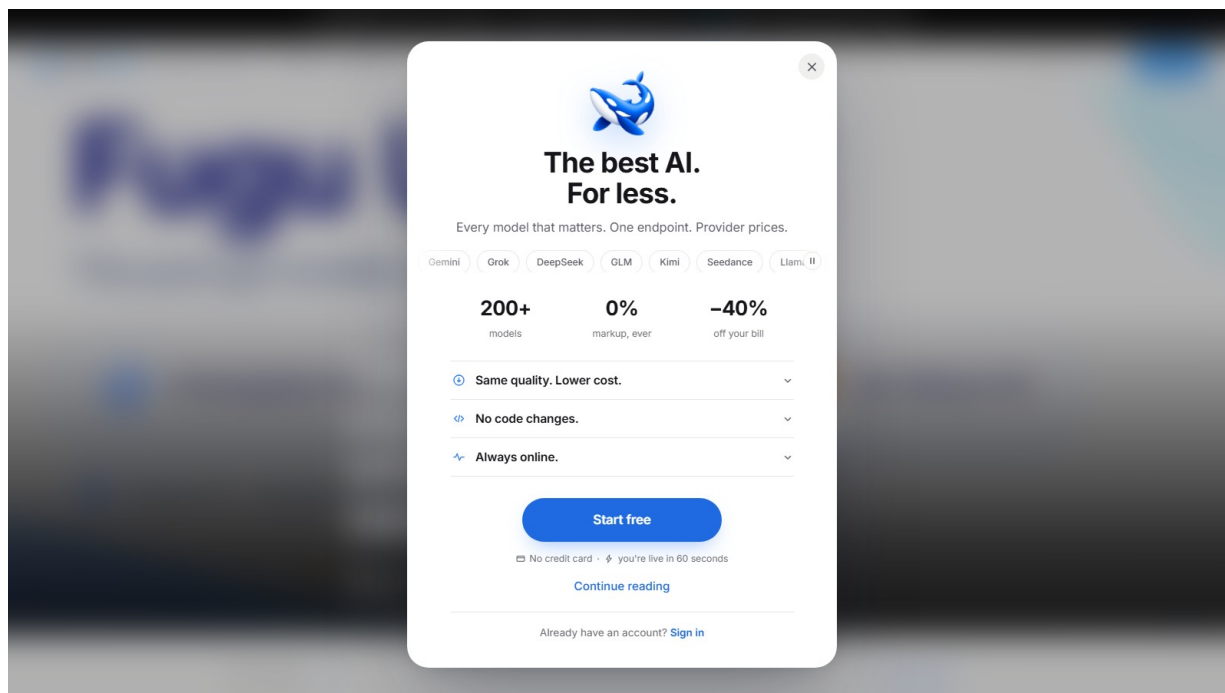
Three checks are worth running before this becomes a dependency, and none is a benchmark.

First, measure cost per accepted task, not cost per token. Build a task set spanning the difficulty range you actually serve, record cost, latency and whether each answer met a stated criterion, then run the same set through the single models you would otherwise use. The orchestrator wins if it wins on that measure across your distribution — and if it wins only on the easy third, you have learned something the rate card cannot tell you.

Second, check determinism where you need it. A system that varies how many agents it wakes may vary its answers. For interactive work that is invisible; for anything with an audit trail it is a gate.

Third, check what happens when the pool changes. You are buying a name attached to a routing policy the operator can retrain on its own schedule. Regression tests on your own task set are the only way to notice.

Sakana Fugu is not in OrcaRouter's catalogue, so no rate of ours is quoted for it and nothing here implies we serve it. Whether the model you call is one network or a committee of them is a question that used to have an obvious answer, and this release is a good reminder that it no longer does.



Sourcing note: The four-tier product line, the two ICLR 2026 papers, the benchmark and security figures, the training cutoff of 28 August 2026 with its named exclusions, the pool-fixing policy, the roughly two-week update lag, the EU/EEA availability limit, and every price quoted — the Fugu pass-through rule, the Fugu Ultra `fugu-ultra-v2.0` ladder and its 272K step, the Fugu Max flat rates and the \$0.007 per-call tool charge, and the \$20 / \$100 / \$200 monthly tiers — were all read from the vendor's public product and pricing pages on 28 September 2026, and are the vendor's own stated figures rather than measured results. The comparison against named frontier model prices is the vendor's claim and is reported as such. No independent capability score is quoted, because the neutral third-party index we use has no page for this model; the observation that the index has no page is a statement about the index. The statement that this model is not in OrcaRouter's catalogue is a statement about our own catalogue on 28 September 2026. The discussion of budgeting, determinism, portability and evaluation design is editorial framing rather than measured results. This article names no competing platform and describes none.

Photo: Felix Rottmann via Pexels

CLICK HERE TO DONATE IN SUPPORT OF OUR NONPROFIT COVERAGE OF ARTS

AND CULTURE

This entry was posted on Thursday, October 1st, 2026 at 11:45 am and is filed under [Check This Out](#). You can follow any responses to this entry through the [Comments \(RSS\)](#) feed. You can leave a response, or [trackback](#) from your own site.