Product preview · In development

OFFPEAK docs

OFFPEAK is an OpenAI-compatible inference endpoint for open-weight models served from bonded GPU hosts, priced by the hour and by live load, and settled in USDG on Robinhood Chain. These docs describe the service as designed. The public endpoint opens in phase P1. See the roadmap for what is live today.

Getting started

Today

1. **Read the whitepaper** at /whitepaper/ and the **roadmap** at /roadmap/.
2. **Check $OFPK status and official links** on the holder one-pager.
3. **Follow @UseOffPeak** for launch news.

At launch (P1)

1. **Point an OpenAI-style client at OFFPEAK.** Swap the base URL and use an OFFPEAK API key (`Authorization: Bearer op_live_…`).
2. **Fund a prepaid balance.** Deposit USDG on Robinhood Chain (chain id 4663) into the OFFPEAK vault from the dashboard. During the beta, deposits are capped at 500 USDG per buyer.
3. **Set a cap.** Send `X-Offpeak-Max-Price: in=…,out=…` (USD per 1M tokens, including the fee). Requests over your cap fail with `429` instead of charging more.
4. **Keep the record.** Every response ends with a signed usage envelope (`event: offpeak.usage` when streaming). Verify it with the SDK, and fetch its on-chain proof once the hour settles.

```
POST /v1/chat/completions
Authorization: Bearer op_live_…
X-Offpeak-Max-Price: in=…,out=…
X-Offpeak-Flex: 2026-11-02T06:00:00Z   # optional: run as a night job before this deadline

GET /v1/models    GET /v1/prices    GET /v1/usage/:id    GET /v1/status
```

Agents can skip the API key. An unpaid request gets `402 Payment Required` listing a payment channel first, then x402 `upto`.

Core features

FEATURE	WHAT IT DOES	PHASE
OpenAI-compatible API	Chat (streaming, tools, JSON mode), embeddings, batches, models	P1
Off-peak pricing	168-hour host price sheets × router-measured load (0.4×–1.2×), floor ≤ 0.6× lowest base	P1
Price caps	`X-Offpeak-Max-Price`. Over-cap interactive calls return `429 price_above_cap`	P1
Night jobs	`X-Offpeak-Flex` or `/v1/batches`. Price locked up front, runs in a cheap hour before your deadline	P2
Prepaid vault	USDG balance you can withdraw (balance minus reserve). Withdraw-all within 2 h	P1
Agent payments	Payment channels (2 h challenge period), x402 `upto` and `exact`	P2
Signed usage records	Router and host signatures, prompt and output hashes, hourly Merkle roots	P1
Disputes	Dispute any request within 7 days. Token recount and signature checks	P1
Host bonds and audits	≥ 2,000 USDG bond, 72 h probation, statistical swap audits, human panel for 100% slash	P1
Windows	ERC-1155 capacity reservations: cancel before start for 95% back, unused tokens 50% back after the end, resale fee 1% + 10% of markup	P3
Sealed tier	TDX and GPU attestation, client-side HPKE encryption	P3
MCP tools	`offpeak_complete`, `offpeak_prices`, `offpeak_schedule_job`, `offpeak_usage`	P2

Honest limits

- Open-weight models only. Open models trail frontier models on the hardest tasks.
- Prices move. The live price at the start of a request is what you pay.
- Discounts depend on host supply and idle time. With few hosts, discounts are smaller.

- Open hosts and OFFPEAK's router can read prompts unless you use the sealed tier.
- Audit sampling stores request bodies for up to 72 hours, and only if you opt in. Opt-in buyers get 1% off.
- Audits are statistical. They catch sustained model swaps, not every response.
- Withdrawing the reserved part of a balance can take up to 2 hours.
- USDG can be frozen by its issuer. Permit2 is a shared dependency.
- Holding $OFPK alone earns nothing. Token hooks switch on only after legal review.

Pricing and billing

- **Live price** = the host's posted base price for this hour of the week × a router-measured load factor between 0.4× and 1.2×. A host's floor can't exceed 0.6× its lowest hourly base price.
- **Buyer price** = live price + a 4% OFFPEAK fee. Opting in to audit sampling takes 1% off the total.
- **Hold, then capture.** Before serving, the router holds the request's maximum cost against your on-chain reserve. After serving, it captures only the measured usage.
- **Balance states.** *Balance* is your USDG in the vault. *Reserve* is the amount set aside for usage that is still settling. *Withdrawable now* is balance minus reserve.
- **Caps.** Set hourly and daily spending caps. A lower cap applies after the next settled hour.
- **Money is integers.** Amounts are integer micro-USDG. Floating-point money math is never used.

FAQ

Is OFFPEAK live? The website is live. The endpoint and contracts launch in P1, after an external contract audit.

Which models? Open-weight models only. The launch models have not been chosen. Model names describe compatibility only.

What do I pay? The host's live price plus a 4% OFFPEAK fee, in USDG. You set a maximum, and the price at request start is final.

Can I get my money out? Yes, by design. You can withdraw any balance that isn't reserved, and withdraw-all completes within 2 hours.

Do I need $OFPK? No. Inference is paid in USDG. $OFPK adds fee tiers, early Window access and host bond weight once its hooks switch on after legal review.

What is the $OFPK contract address? To be announced. Trust only the official links below.

What does a usage record prove? That the router and host signed the counts and price, and, once settled, that the record is in an on-chain hourly root. It doesn't prove that the answer is correct.

Is it private? Prompts aren't logged or stored, except opt-in audit samples. Hosts and the router can read prompts unless you use the sealed tier (P3). On-chain transfers are public.

How to get help

- **Website:** offpeakai.app (planned public address)
- **X:** @UseOffPeak
- **Telegram:** t.me/offpeakrh
- **Email:** Offpeaky99@atomicmail.io

There is no public status page or support portal yet. OFFPEAK will never ask for your private key or seed phrase.

Roadmap

Live now: the preview website. **Next (P1):** the endpoint, off-peak pricing, withdrawable balances, usage roots, bonded hosts and mainnet contracts under a beta cap. **Later:** agent payments, night jobs and MCP tools (P2), then Windows, sealed hosts and audits at scale (P3), with $OFPK hooks after legal review. See the full roadmap.

Pilot acceptance gates (v1.0 build brief)

The October 12, 2026 build brief defines twelve acceptance gates. All are planned.

GATE	NAME	STATE
PH-00	Operating contract	Planned
PH-01	Scaffold and staging baseline	Planned

GATE	NAME	STATE
PH-02	Wallet identity and signing	Planned
PH-03	Ledger and reservations	Planned
PH-04	Real GPU and quote engine	Planned
PH-05	Finalized deposits	Planned
PH-06	Paid inference loop	Planned
PH-07	Withdrawals	Planned
PH-08	Receipt registry and verifier	Planned
PH-09	Customer console and agent client	Planned
PH-10	Operations and hardening	Planned
PH-11	Capped deployment and pilot	Planned

Third-party names describe compatibility only; no affiliation.

OFFPEAK · Product preview · In development. Third-party names describe compatibility only; no affiliation.