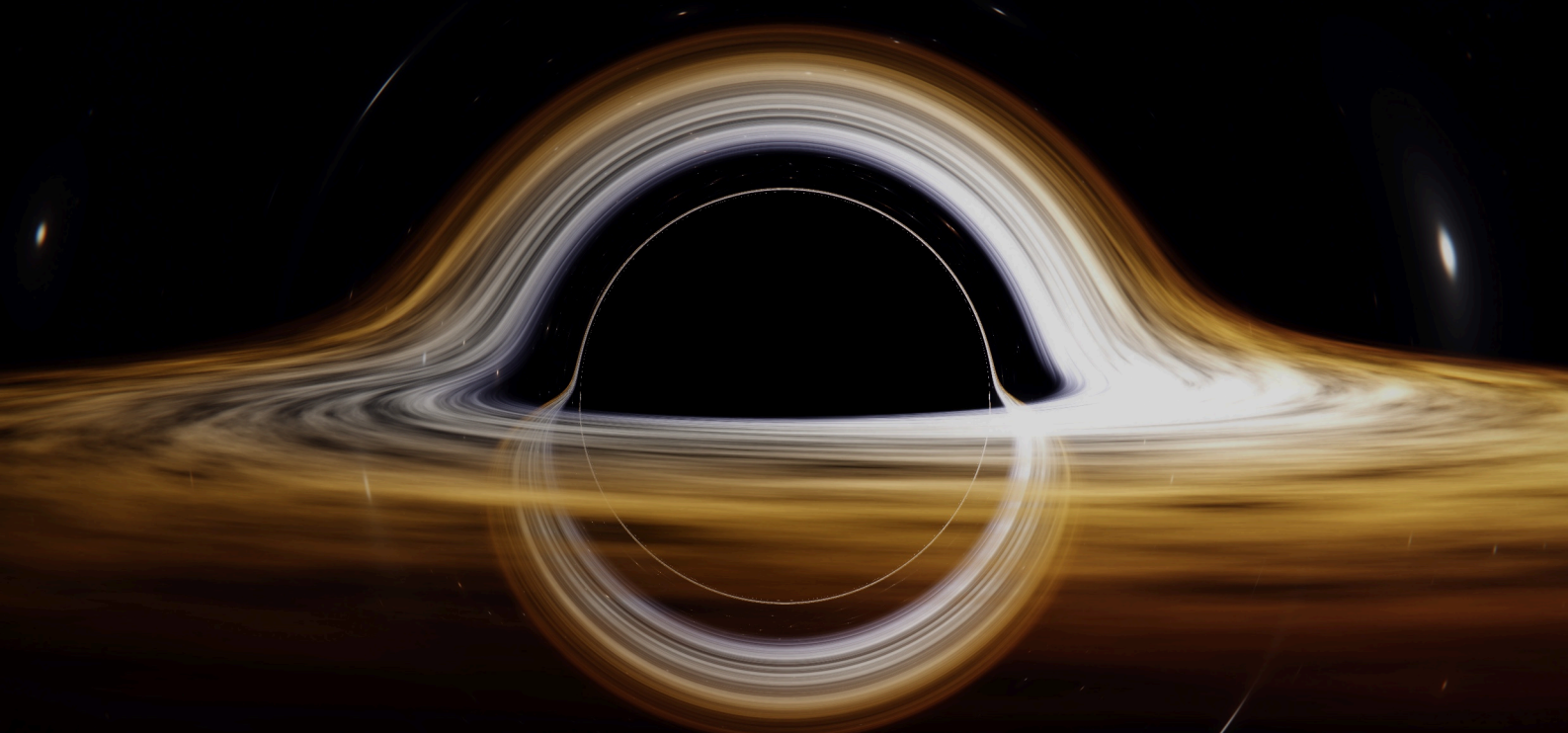




WHITEPAPER

Antimatter Compute

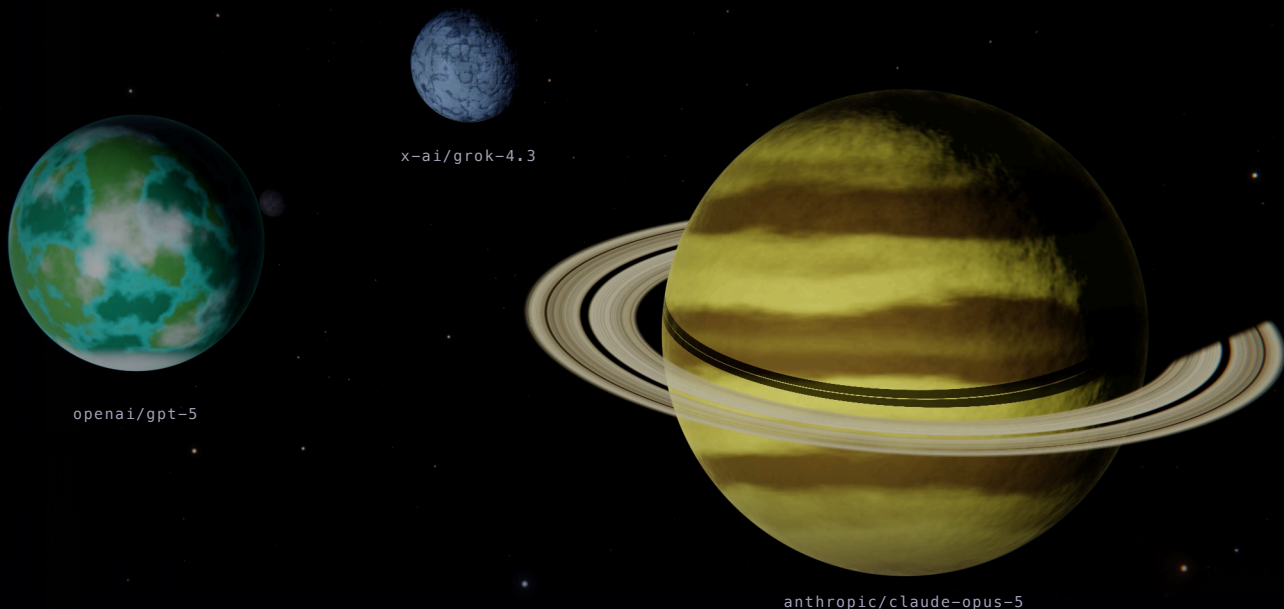
Every model is a planet.

Ten chapters,

01	What Antimatter is	03
	One place to find a language model, see what it costs, try it, and use it from code.	
02	Models and prices	05
	The upstream's published price, plus a margin stated in the open.	
03	The gateway	07
	Three wires over one catalog, one set of keys and one balance.	
04	Measurement	11
	A chart is worth what its readings are worth.	
05	Identity and keys	13
	A wallet signature is the identity. There is no password.	
06	Money	14
	Stablecoins only. Held before a request, settled after it.	
07	The chart and the ship	16
	Every model in the catalog is drawn as a planet.	
08	Privacy	17
	What the API keeps is the bill, unless you ask it to keep more.	
09	The token	19
	\$ANTI on Sable, on Robinhood Chain.	
10	What is not built	21
	Said plainly, so nobody has to find it out.	

As of 3 October 2026. This paper states what Antimatter Compute is and how it works, as the deployment at useantimatter.com was configured on that date. Every figure in it, the margin, the networks a deposit is taken on and the free allowance, was read from that configuration when the paper was built.

Prices and the number of models change too often to print. [Pricing](#) reads them live, and [the docs](#) have the requests, the headers and the errors.



01

What Antimatter

One place to find a language model, see what it costs, try it, and use it from code.

Antimatter Compute lists every model it can reach with its price, lets you talk to any of them, and serves them all through one OpenAI compatible API behind a key your wallet mints. Antimatter is the short name, and the one the site uses.

The models and their prices come first. [Pricing](#) lists every model with the upstream's published rate beside the rate billed. [The briefing](#) teaches, in one scroll, what a model is, what one costs, how they differ and how to choose. [The chart](#) draws each model as a world you can fly to and read. [Chat](#) puts any of them on the other end of a conversation, one at a time or several side by side. [The API](#) serves the same catalog to your code, at the same prices.

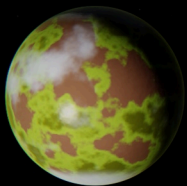
The name is the idea. Every model is a planet, the wormhole is the way between them, and antimatter is the drive that makes the jump. The mark says the other half: antimatter is matter's mirror image, so the mark is one ring and its reflection across one line.

— The problem

Language models come from many vendors. Each has its own key, its own bill, its own wire format and its own idea of an error. Prices move without notice. A catalog page says what a model supports, and the model does not always do it. Using several of them means writing the same integration several times and trusting claims nobody measured.

Antimatter answers with one door: one key, one base URL, one balance, prices stated in the open, and a chart of what is out there that shows what was measured and says so where nothing was. It asks for no card, no name and no email. A wallet signature is the identity.

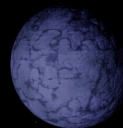
This paper describes only what is built and running at the address above, as the deployment was configured on the date on its cover. Where something is not built, or not switched on, it says so. [The field manual](#) has the requests, the headers and the errors.



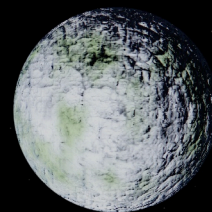
moonshotai/kimi-k3



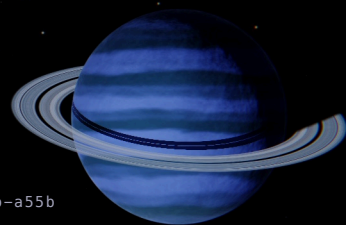
mistralai/mistral-large-2512



minimax/minimax-m1



nvidia/nemotron-3-ultra-550b-a55b



z-ai/glm-5.3

Models and

The upstream's published price, plus a margin stated in the open.

Antimatter hosts no models. Every answer comes from an upstream provider: OpenRouter, which serves models from almost every vendor, or OpenAI and Anthropic directly where the operator connected them. The catalog is the upstreams' own list, with each model's published price per million tokens, its context window and its published maximum output.

MARGIN	FREE ALLOWANCE	CACHE HITS
10%	\$0.25	Free
over the upstream's published price, the same for everyone	once, for an account that has never deposited	an answer served from the cache costs nothing

What you pay is the upstream's published price plus a margin of 10%, the same for everyone and applied to the rate, with nothing added at the end. Every price the site and the API show already includes it, and [pricing](#) puts the upstream's rate beside each one, so the margin is never hidden.

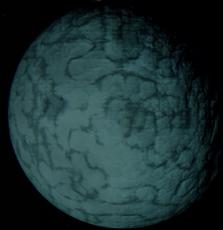
Discounts come off the margin only, never below what the upstream charges. An answer served from the cache costs nothing.

An account that has never deposited gets a lifetime free allowance of \$0.25 to try models with.

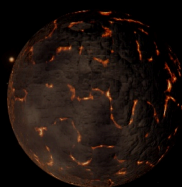
A model with no published price is never served on that allowance or on anyone's balance, only on your own provider key, because a price nobody published cannot be billed honestly.

No price is printed in this paper. Prices move, and a printed one would be wrong by the time it is read. [Pricing](#) reads them live, and every answer from the API names what it was billed.

Discounts come off the margin only,



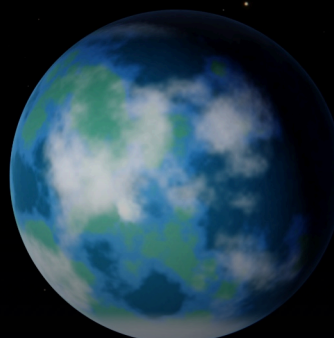
openai/gpt-5.5



anthropic/claude-sonnet-5



deepseek/deepseek-v4-pro



google/gemini-3.5-flash

03

The

Three wires over one catalog, one set of keys and one balance.

The gateway speaks the OpenAI chat completions wire at `/v1/chat/completions`, the Anthropic messages wire at `/v1/messages`, and OpenAI's Responses wire at `/v1/responses`, all three streaming. Embeddings are at `/v1/embeddings`, image generation at `/v1/images/generations`, speech and transcription under `/v1/audio`, and OpenAI's batch wire at `/v1/files` and `/v1/batches`. Tool calling is translated in both directions, so a tool call written for one wire reaches a model that speaks another. Images and structured output reach Anthropic's models too.

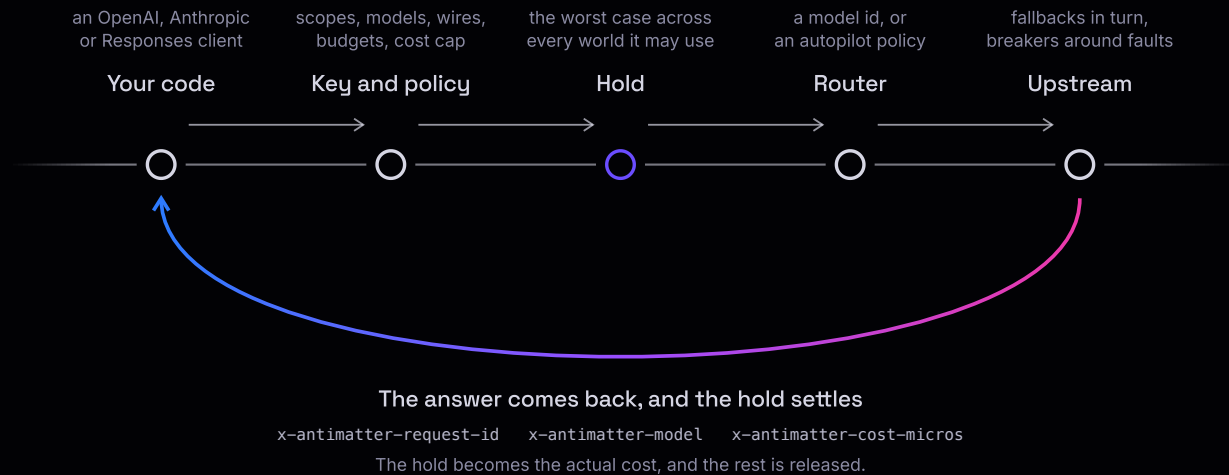


FIGURE One request. The key and its policy are checked, the worst case is held, the router picks a world, and the answer comes back carrying its request id, the world that served it and what it cost.

Every answer names its request in `x-antimatter-request-id` (and in `x-request-id`, which the OpenAI clients read), the world that served it in `x-antimatter-model`, and what it was billed in `x-antimatter-cost-micros`, or in `usage.cost` on a stream's last chunk. `GET /v1/requests/{id}` reads a request back.

— Routing

A model id names a world. An autopilot id names a policy, and the router picks the world.

AUTOPILOT	SELECTS ON
<code>antimatter/auto</code>	price, measured latency and measured quality, scored together.
<code>antimatter/auto:cheap</code>	the lowest blended price.
<code>antimatter/auto:fast</code>	the lowest measured time to first token.
<code>antimatter/auto:code</code>	a world whose tool calls were measured working, with a context sized for code, ranked on measured quality. There is no per task code score, and the router says so.
<code>antimatter/auto:long</code>	a world with a long context window, from the catalog.
<code>antimatter/auto:vision</code>	a world that takes images as input, from the catalog.
<code>antimatter/auto:tools</code>	a world whose tool calls were measured working.
<code>antimatter/auto:json</code>	a world whose structured output was measured holding.

The score is computed per request over the worlds that fit. A request can steer it with the `x-antimatter-route` header: a policy, weights, and constraints such as a price ceiling, a minimum context, or a measured working tool call.

An axis nobody measured counts for nothing, never for the average of the others, and the answer says unmeasured. The request steers it too: an image in the input keeps it to worlds that see, and its size to worlds whose context holds it.

`POST /api/router/explain` shows the decision for a request without sending it.

— Fallbacks and breakers

Give a list of models and the gateway tries each in turn when one answers 429, fails with a 5xx or drops the connection. With `x-antimatter-fallback: comparable`, or the same switch on the key, a chain that fails hands over to the nearest world that is no dearer on either price and can do what the request needs. Before the first attempt the gateway checks that every world it might use has a published price, and it holds the worst case across all of them, because a fallback can cost more than what it replaces. Each upstream sits behind a circuit breaker: a run of failures opens it, traffic goes around it while it is open, and a trip on one instance is shared with the others.

— Streams that survive a dropped connection

Send `x-antimatter-resumable: 1` on a keyed stream and the generation outlives the connection: it runs to its end, is billed once and in full, and `GET /v1/streams/{id}` replays it from the last event the client saw, in the wire it was asked in. Without the header nothing of the answer is kept.

— Batches, images and audio

A batch is a JSONL file of chat or embedding requests, run within 24 hours, each line as a live request by the key's owner with the same policy, budgets and billing as any other. Image generation, text to speech and transcription run on the same keys and credits: the worst case is held, one usage row is written, and there is no fallback chain.

— Caching

Caching is off unless asked for. An exact repeat can be served from store. A thread whose earlier turns match byte for byte is recognised as the same thread, and a stored answer is only ever served for the whole conversation, never for a shorter one. A semantic layer can match a question that means the same thing, and a prompt too long for it to read in full is never a candidate. Anthropic's `cache_control` passes through untouched.

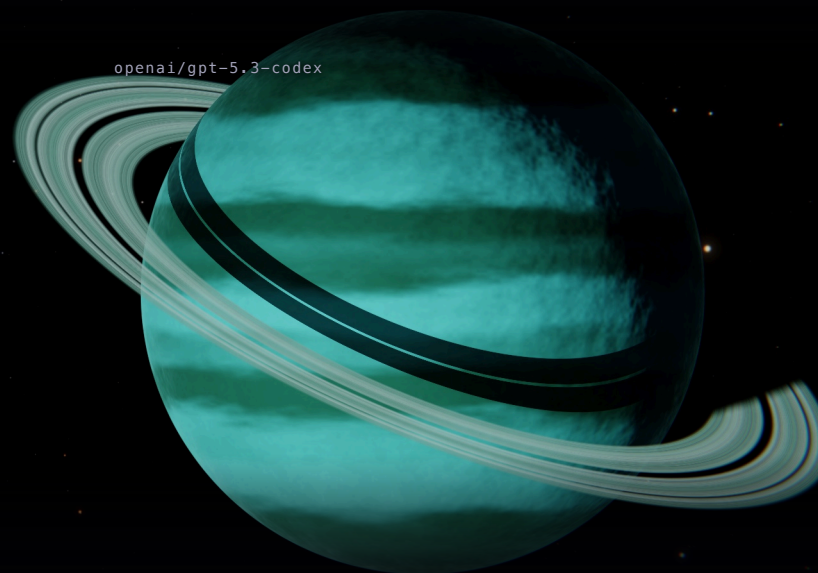
— Your own key

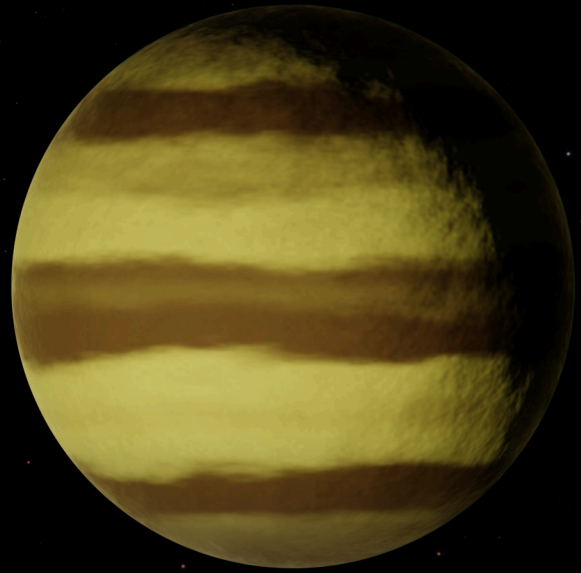
Bring an OpenRouter, OpenAI or Anthropic key and the requests it serves bill your provider account, spend no credits, and come back with `x-antimatter-byok: 1`. A request on your key never falls back onto the house key: a plan that would is dropped and named in `x-antimatter-route-warning`.

An answer your key paid for is never written to the shared cache. Provider keys are encrypted with a key held outside the database.

— Clients

The official OpenAI and Anthropic clients work as they are, pointed at this base URL, and a compatibility suite drives them against the gateway to keep that true. Antimatter's own clients (TypeScript, Python and Go SDKs, a command line client, an MCP server, and providers for the Vercel AI SDK and LangChain) are built and not yet published; until they are, the official clients are the way in. The whole surface is in the [OpenAPI document](#), generated from the code.





anthropic/claude-opus-5.5

04

Measurement

A chart is worth what its readings are worth.

Antimatter measures, and where it has not measured, it says so.

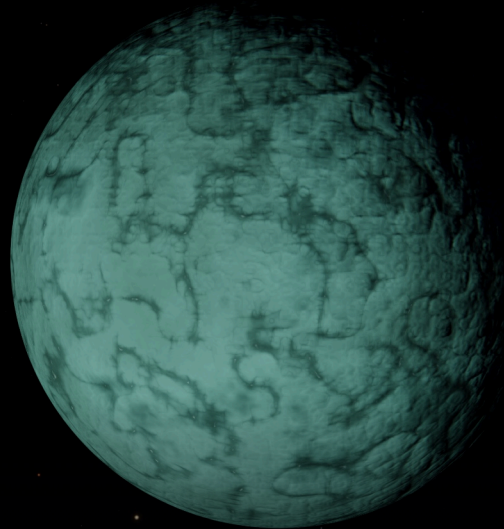
- Capability probes send a world small fixed tasks through the same path a customer's request takes, and record whether tool calling and JSON output actually work, what a long context does to recall, and how often an ordinary prompt is refused. Probes spend real money, so they run on a budget against a bounded set of worlds. A world never probed reads unmeasured, and a reading past its freshness window is marked as history. Where a capability was not measured, the pages say whether the vendor at least declares it.
- A scheduled probe is the canary, and `/api/health` reports the newest one.
- The router's quality axis is the probes plus the reliability measured from real traffic, each weighted by the evidence behind it. A world with neither reads unmeasured, and nothing is borrowed from another world to fill the gap.
- [Rankings](#) count what this deployment was actually asked to serve. They measure adoption, not quality.
- The flight recorder keeps one span per request: the world, the timing, the tokens, the cost and the outcome. Bodies are kept only for a key where you switch capture on, and only for the window you set. A captured request can be replayed against another world and the two answers read side by side. Spans export over OpenTelemetry, and alert rules watch them.

— Why nothing on the site is invented

A deployment with no upstream key shows an empty catalog and answers 503, never a mocked reply. A reading nobody took reads unmeasured, never zero. A readout whose source has not answered is not shown at all, rather than shown with a placeholder. Prices come from the upstream catalog and from the vendors' published tables, with the margin on top; the card shows the price that is billed.

04 MEASUREMENT

A reading nobody took reads unmeasured,



openai/gpt-6-astra

0.5

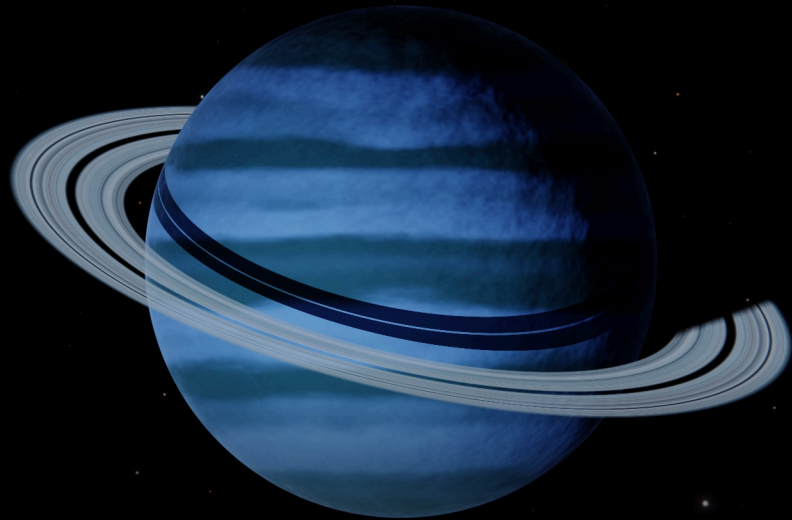
Identity and

A wallet signature is the identity. There is no password.

You sign in by signing a message with a wallet. The message says what it is, the signature proves you control the address, and it authorizes no transaction and costs nothing. The nonce it signs is bound to that address and used once. The session is a signed cookie on the site.

Keys are minted from a signed in session, on the [dashboard](#) or the [bridge](#), or by an agent that signs the same message itself. A key starts `am_live_`, is shown once, and is stored only as a SHA-256 digest, so nobody at Antimatter can read it back. A key carries scopes, an optional expiry and an optional IP allowlist, and its policy can hold it to some models or block others, keep it to some wires, set a daily or monthly limit, cap what one request may cost and lower the output ceiling. Keys rotate in place, and an account can cap its own monthly spend.

[Organizations](#) hold one shared balance. A lead sets a budget for each member, roles are enforced on the server, and a key minted for an organization belongs to it rather than to whoever minted it; a member who leaves has their keys for it revoked. Every action that touches money or security is written to an append only audit log.



google/gemini-2.5-flash

06

Money

Stablecoins only. Held before a request, settled after it.

Antimatter is funded with stablecoins only. There are no card payments. A deposit can be USDC on Robinhood Chain, USDC on Base, and USDC or USDT on Arbitrum One, sent from the wallet you signed in with to the address the [billing](#) page shows you. Robinhood Chain, the chain the token lives on, comes first. You claim the transfer by its transaction hash. It is verified on chain and credits exactly once, however many times it is claimed. A stablecoin's decimals come from configuration and are never guessed.

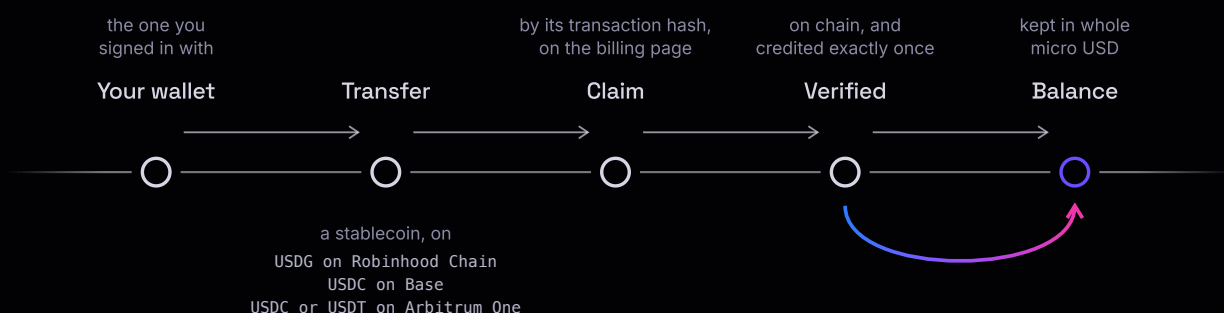


FIGURE A deposit. Sent from the wallet you signed in with, claimed by its transaction hash, verified on chain, credited once.

Paying per request with x402, with no key, is built and not switched on here: a keyless request is refused.

— Held, then settled

A request is billed at the price in the catalog. Before it goes out, the worst case it could cost is held against the balance. When it settles, the hold becomes the actual cost and the rest is released. A request the client abandons after it was sent is still billed, because the upstream bills it; one that was never sent costs nothing. The ledger is kept in whole micro USD. Receipts are read from the ledger, and monthly statements are built from it, on the [billing](#) page.



Sent and then abandoned is still billed, because the upstream bills it. Never sent costs nothing.

FIGURE Hold and settle. The worst case is reserved before the upstream is asked anything; the settled cost is what was used, and the remainder goes back to the balance.

A house breaker caps what free use can spend in a day, across everyone. Credit cannot be withdrawn: it is spent on requests.

Credit cannot be withdrawn:



07

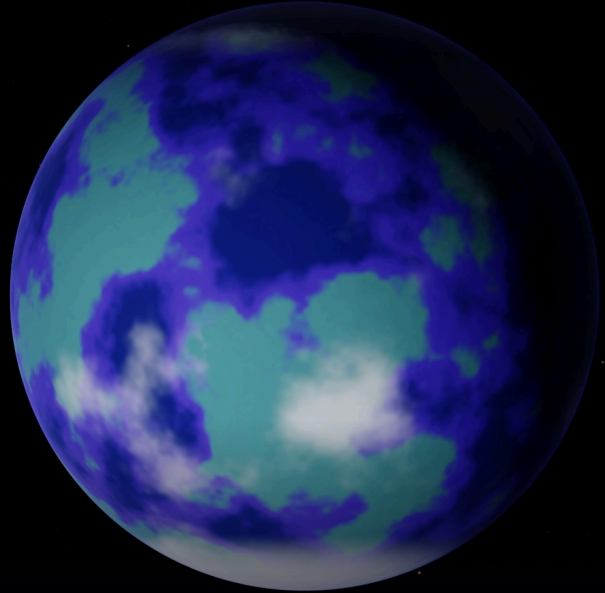
The chart and

Every model in the catalog is drawn as a planet.

A world's look is derived from its model id, so it keeps the same face on every visit and every screen, and a vendor's worlds gather into one system. The worlds in this paper are drawn the same way, by the same material, from the ids written under them.

The [chart](#) is flown, not scrolled. The ship is a starship with warp nacelles, and it moves like one: it keeps its velocity until the engine changes it, turns round to brake, and arrives standing still. Hold a world under the scan and it joins your codex. Voyages are routes through the chart with a line about each stop, shared as a link. Seasons are windows with a goal, and a standing that cannot change once the window closes. Photo mode takes a picture of where you are. Other explorers show on the chart while they fly. Charting earns hulls and trails, and they are cosmetic: nothing you earn changes a price or a limit.

Each world has its own page with its prices and their history, its context, its measured latency, its capability readings and a live line to it. Each vendor has a system page. The [observatory](#) draws your own recent use as a solar system.



z-ai/glm-5.2

08

Privacy

What the API keeps is the bill, unless you ask it to keep more.

[What the API keeps](#) is stated on its own page, with every window read from the code that enforces it. In short:

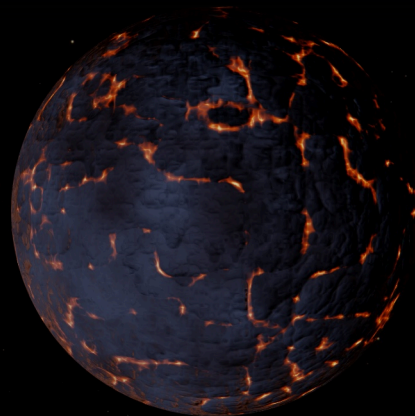
- A request to the API writes no prompt and no answer to Antimatter's database unless you ask it to. What is kept is the bill: the model, the token counts, the cost, the timings, the key and any session label you sent.
- Content is kept only where you switch it on: the response cache and resumable streams by a header on the request, flight recorder capture per key. Each is held for a window, then swept.
- Some features keep content because that is their job. [Chat](#) saves your threads, an anonymous one for a window after its last message. A batch keeps its lines and answers until the account is erased. The Telegram bot keeps a short history. A shared chat link is a snapshot until you revoke it.
- Your wallet address is your account. Your keys are stored as digests only, and provider keys encrypted with a key held outside the database. The address a sign in came from is kept only as a keyed digest, for the abuse rules, and pruned. The audit log keeps the IP address of notable acts in plain text for a window, then only a keyed digest of it.

A prompt has to reach a model to be answered, so it goes to the upstream that serves the world you chose, and that upstream's own retention and training policies apply. Antimatter does not ask an upstream, on your behalf, to skip training or retention.

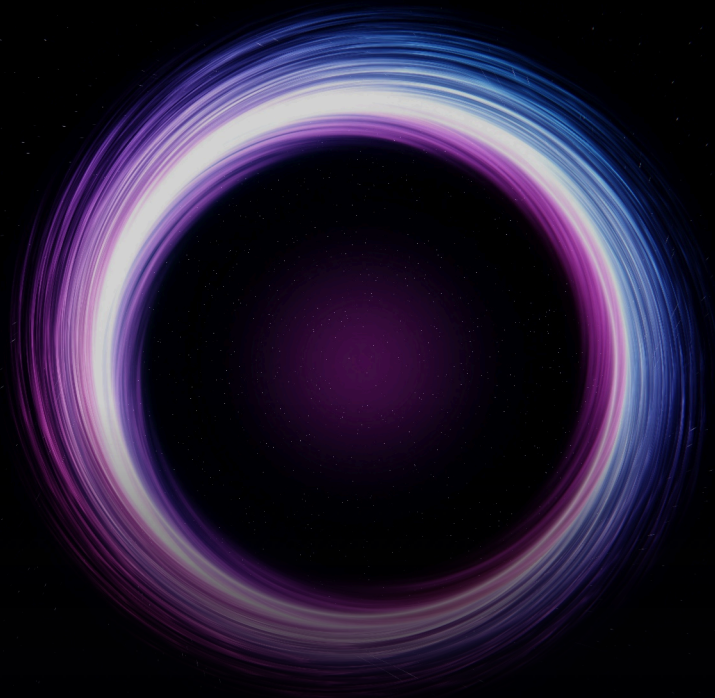
— Export and erasure

Export gives you everything held against your wallet as one JSON file, from the dashboard. Secrets are the only omission, because they cannot be read back.

Erasure deletes the account, through `DELETE /api/account` with a confirmation phrase. Some things outlast it on purpose. Ledger rows and usage records lose their link to you but stay, because the books have to add up and a deposit must never be claimable twice. The free allowance already spent, a freeze or an open abuse flag, and a referral are remembered under a one way key of your address, so deleting an account does not buy a fresh one. Erasing an account that still holds credit forfeits that credit: the answer names the amount, and nothing is erased until you confirm the forfeit.



google/gemini-3.8-flash



09

The

\$ANTI on Sable, on Robinhood Chain.

\$ANTI is the Antimatter token, on Sable, a launchpad on Robinhood Chain. It has not launched. There is no contract address yet, and until one is published on the site, no address is it.

— What it does

Three things at most, and only once the operator switches them on. An account whose own wallet holds \$ANTI on Robinhood Chain reaches a holder tier, and a tier can grant a larger lifetime free allowance, a higher request rate for the account's API keys, and a share off the margin. That share comes off the margin only: it never takes a price below what the upstream charges, and with no margin set there is nothing for it to take. The wallet you sign in with is the one that is read; there is nothing to link. The tiers, the holding each one needs and what each one grants are the operator's settings. Nothing else changes: routing is the same for everyone.

On this deployment the tiers are off, because the token has not launched.

— What it is not

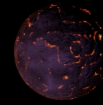
It is not a way to pay. Antimatter is funded with stablecoins only, and \$ANTI buys nothing here. It is not a share of anything: it carries no claim on revenue, on ownership or on a vote. There is no promise about its value.

This paper states no price, holder count or market cap, and the site states none that a source did not measure.

A token is a risk you take alone. A token can lose all of its value, and [Sable's own page](#) says so. Read it before you do anything. [The token page](#) has the holder tiers as the deployment sets them, and the measured numbers once there are any.

09 THE TOKEN

A token is a risk



deepseek/deepseek-v3.2

10

What is

Said plainly, so nobody has to find it out.

- The token has not launched, and the holder tiers stay off until it has and the operator switches them on.
- There are no card payments and no fiat. Funding is stablecoins, claimed by transaction.
- Credit cannot be withdrawn to a wallet. It is spent on requests.
- Antimatter hosts no models. Every answer comes from an upstream provider, and a world is only as available as the provider behind it.
- There is no benchmark of its own and no model judging another. Quality is what the probes and real traffic measured, and nothing more.
- The API does not remember conversations. Each request carries its own messages, and the Responses wire refuses `previous_response_id` and `store`; only chat, on the site, keeps threads.
- There is no realtime voice: it needs a socket the gateway does not open. There is no request priority, on any plan.
- Antimatter's own SDKs, command line client and MCP server are not yet published to npm or PyPI.

This paper describes what ran on the date on its cover. It makes no promise about what comes next.



Antimatter

Every model is a planet.

USEANTIMATTER.COM

Antimatter Compute, whitepaper. As of 3 October 2026. The margin, the deposit networks and their stablecoins, the free allowance and every switch it mentions were read from the deployment's configuration when this paper was built, by the same code the site reads them with. It states no price, no count of models, and nothing about the token that no source measured.

The pictures are live renders, not illustrations: the cover is the black hole of the post cards, its light integrated through the

Schwarzschild field; the token's chapter is the front page's own wormhole; every world is drawn by the chart's own material from the model id written beneath it, so each is the world you find on the chart. The ids were checked against OpenRouter's public model list on 3 October 2026.

Set in Space Grotesk and Inter. Nothing in this paper is an offer, and \$ANTI is not a way to pay for anything here.