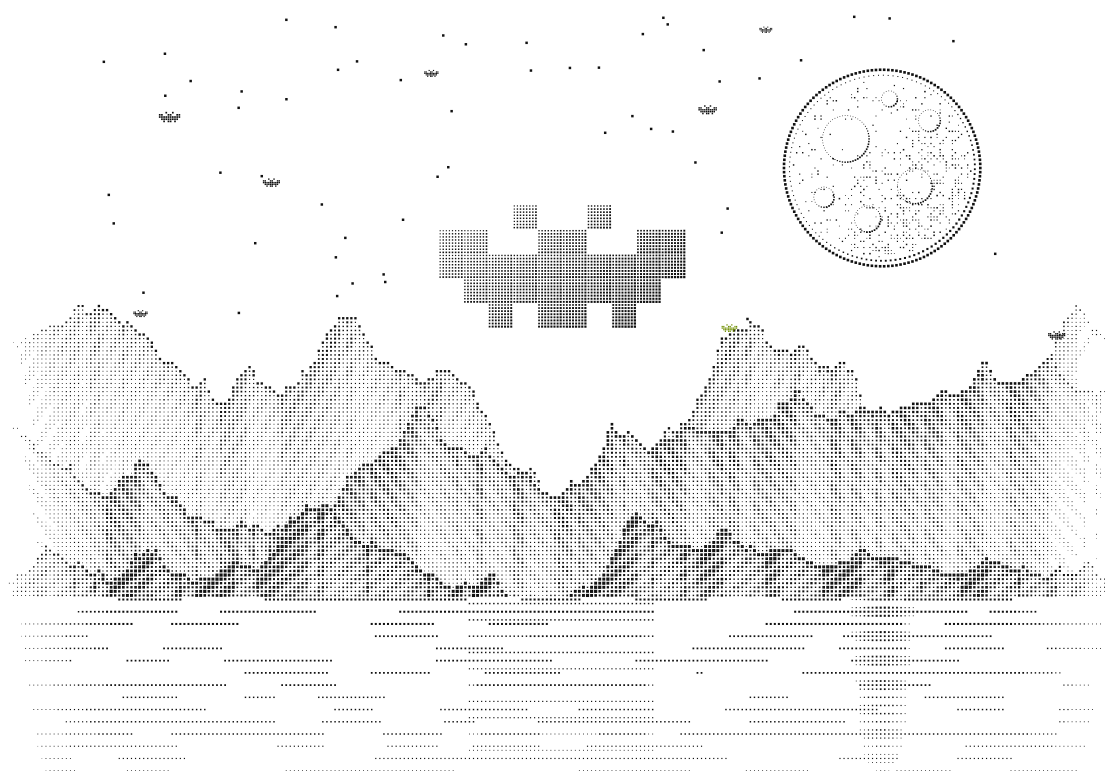




Moth

Private Inference Below List Price

Moth, October 2026, draft for discussion

Abstract

Calling AI models from several makers means one account, one credential and one invoice per maker, a request that carries the caller's identity to every host, and the published list price for anyone below negotiating volume. Moth replaces these with one wallet, one prepaid balance and one key across 34 chat, image, video and music models from 11 makers. Every call passes one address, the Mothpoint, where its worst-case cost is held before dispatch and every field that identifies the caller is removed, so the host receives the question under a credential shared by all callers. No prompt or output is stored. Prices are set below each maker's published list and shown with the exact saving. A planned token, \$MOTH, would divide a fixed 3% swap tax on chain and set a share of it aside each hour as credit for inference. This paper states the design principles. The system detail is specified in the docs.

Not an offer. This paper is not an offer to sell, or a solicitation of an offer to buy, \$MOTH or any other token or security, in any jurisdiction. \$MOTH has not launched, and nothing here promises a price, a return or a profit. Holder Credit is a share of swap tax set aside as platform credit or USDG. It is not interest, a dividend or a return, and in any hour it can be zero.

Notice

This paper is a draft for discussion. It describes Moth as built in October 2026, and parts of it will change. Moth Chat, Moth Console and the API are in operation and are paid in USDG on Robinhood Chain. The Sealed tier, the \$MOTH token, Holder Credit and paying in \$MOTH are planned and not live, and the contract that would divide the token's tax is built and tested but not deployed.

Not an offer. This paper is not an offer to sell, or a solicitation of an offer to buy, \$MOTH or any other token or security, in any jurisdiction. \$MOTH has not launched, and nothing here promises a price, a return or a profit. Holder Credit is a share of swap tax set aside as platform credit or USDG. It is not interest, a dividend or a return, and in any hour it can be zero. Claimed credit is prepaid access to the platform. This paper is not a prospectus, and no regulator has reviewed it.

No advice. Nothing here is financial, legal, tax or accounting advice.

Figures. Every number is read from Moth's model catalog, its published rules and the splitter contract on the date above. List prices are those each maker publishes on its own pricing page, each stored with the date it was read. No speed or accuracy figure is given, because no benchmark of Moth's own supports one.

Names. Model and company names identify the models Moth offers. Moth is not affiliated with or endorsed by their makers.

Contents

1	Introduction	3
2	Design	4
3	Privacy tiers	5
4	Pricing and billing	6
5	\$MOTH	7
6	Risks and limits	8
7	Future work and conclusion	9
	Contract and links, references	10

The docs specify the system in full, including the request lifecycle, the ledger, the API endpoints, the Sealed protocol and the contract interface. This paper does not repeat them.

1 Introduction

1.1 Fragmentation

Model quality is uneven across tasks, so a team that uses AI in earnest calls several models from several makers. Each maker issues its own credential, applies its own rate limits, reports usage in its own units and sends its own invoice. Credentials multiply with the number of makers and are stored in environment files, CI secrets and developer machines, where each is a separate point of leakage. Moving a workload from one maker to another requires a new account, a new credential, new limits, a client change and a new billing relationship. That switching cost keeps workloads on a model after a better or cheaper one is available, and it is rarely accounted for, since it appears as engineering time.

1.2 Identity in the request

A model requires only the prompt. A request to a host also carries an account credential that names the caller, the caller's IP address, a user agent and, in many applications, a user identifier. A single request discloses little. A host that receives every request under one credential can associate them over time into a record of who asked, from where, about what and how often, held under that host's retention policy. A sender working with several makers is subject to several such records and several policies, any of which can change, and has no means of sending a prompt without also sending an identity.

1.3 List price

Published prices are set for individuals and small teams. Buyers with sufficient volume negotiate terms. A buyer below that volume pays list, and a buyer who spreads usage across several makers fragments the volume that would otherwise support negotiation. Aggregating the demand of many small buyers is the only route by which most of them obtain terms below list.

1.4 Common cause

The three costs have one cause. Each maker is reached directly, one relationship per maker, and no intermediary separates the caller's identity, payment and history from the request. The design in Section 2 introduces such an intermediary and constrains what it may retain.

1.5 Requirements

- **R1.** One key and one balance reach every catalogued model, at one price per model in the chat and the API.
- **R2.** No field that identifies the sender is forwarded to a host.
- **R3.** No prompt or output is written to storage by Moth.
- **R4.** Billing cannot take a balance below zero.
- **R5.** Each price is published beside the maker's list price, its source and the date read, with the saving rounded down.
- **R6.** If the token launches, its tax is divided on chain by a contract whose every change emits an event.

2 Design

A sender signs in once with a wallet, using a signed message [7] that costs nothing and sends no transaction. The account is stored under a keyed hash of the address. The account holds one prepaid balance in micro-dollars, funded in USDG [4], and one or more keys with the prefix `moth_`. One key reaches every catalogued model, and the price of a model is the same in Moth Chat and in the API.

Every call is sent to one address, the Mothpoint. There the key is checked, the worst-case cost of the call is held on the balance, fields that identify the sender are removed, and the request is reissued to a host under a credential that all Moth callers share. The host receives the prompt and no identifier of the sender. The reply returns through the Mothpoint, which charges the call's actual cost and returns the remainder of the hold. Moth stores no prompt or output. Billing keeps the model, the token counts and the cost for 90 days, after which only daily totals remain.

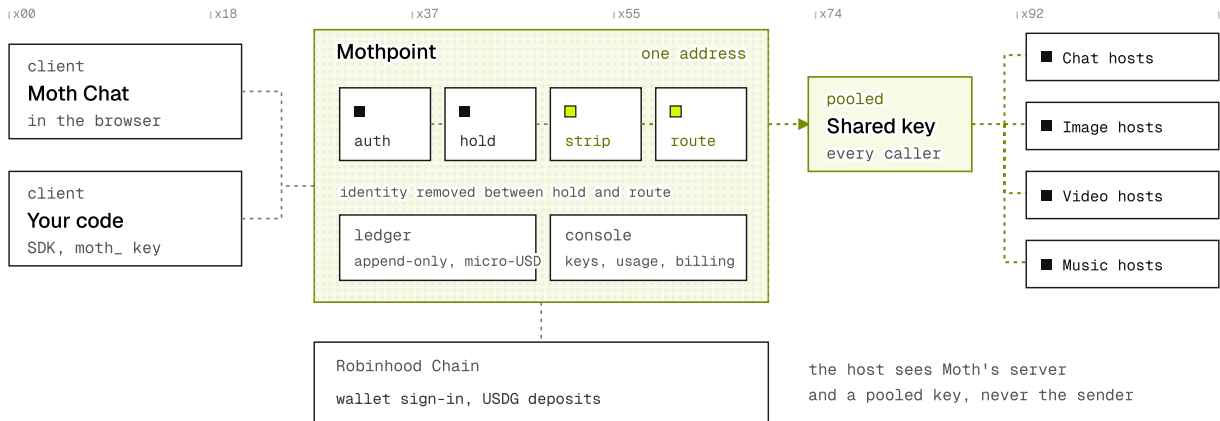


Figure 1. The system, end to end. Grey connectors carry the sender's identity. Lime connectors begin after the strip stage and carry only the question.

Existing SDK code changes in two values, the base URL and the key, as in Listing 1. The model is named by its catalog id.

Listing 1. The base-URL change, with a `moth_` key.

```
1 from openai import OpenAI
2 client = OpenAI(base_url=API_URL + "/v1", # the Mothpoint, published at launch
3                 api_key="moth_...")      # one key for every model
4 client.chat.completions.create(model="claude-sonnet-4.6", messages=[...])
```

Endpoints, request states and error codes are specified in the docs at the address on page 10.

3 Privacy tiers

A privacy statement is meaningful only when it names the party it excludes. Moth handles every request in one of two tiers, and each model in the catalog is labelled with its tier. All 34 models are in the Anonymized tier today.

Anonymized excludes the host from learning who is asking. The wallet, IP address, user agent and user fields do not reach the host, the request arrives under the pooled credential from Moth's servers, and Moth keeps no copy. The host reads the prompt under its own policy, and Moth's servers handle it in transit. The tier suits work in which the asker is sensitive and the question is not.

Sealed excludes the host and Moth from reading the content. It is planned for a small set of open-weight models in attested hardware. Before sending, the browser verifies the hardware's attestation against a measurement pinned in the application, then encrypts the request with HPKE [8] to a key bound to that hardware. Moth relays ciphertext. A request that cannot be verified fails, and there is no fallback to an unsealed route.

Table 1. The two tiers.

	Anonymized, live	Sealed, next
Hidden from the host	the sender's identity	the sender's identity and the content
Readable by	the host, under its own policy	the attested hardware only
Moth sees	content in transit, no copy kept	ciphertext and billing counts
Models	all 34 in the catalog	a few open-weight models
Relies on	identity removal and the pooled credential	the hardware and its attestation
Does not cover	the host's handling of content	flaws in the hardware or its attestation

Neither tier makes prompt text anonymous. A prompt that names its author, or contains details sufficient to identify the author, is forwarded as written. Neither tier conceals payments, which are transfers on a public chain. Terms such as end-to-end encrypted, confidential, verifiable and zero retention are not applied to Anonymized models, because the host can read their content. Moth adds no moderation layer of its own, and its terms prohibit illegal use.

Table 2. What each party can learn from a request.

Party	Anonymized	Sealed
Network observer	that a device connects to Moth	the same
Chain observer	that a wallet funded a balance, and the amount	the same
Moth's servers	the key, the cost, the content in transit	the key, the cost, ciphertext
Host	the content, under the pooled credential	nothing outside the attested hardware

4 Pricing and billing

4.1 Priced against list

The reference for every price is the list price on the maker's own pricing page, recorded with its address and the date it was read [2]. The catalog holds 30 such list prices from 10 makers' pages, read between 13 and 17 October 2026 [1]. Moth's price is set below the list, by up to 50%. The saving is shown rounded down, and for chat models it is the smaller of the input and output savings. A model shown without a list price claims no saving. Sixteen models are at half the list, seven are between 1% and 19% below it, and four claim none (Figure 2). Prices change only by publication, never by routing, and a model that cannot be served at its price fails rather than costing more.

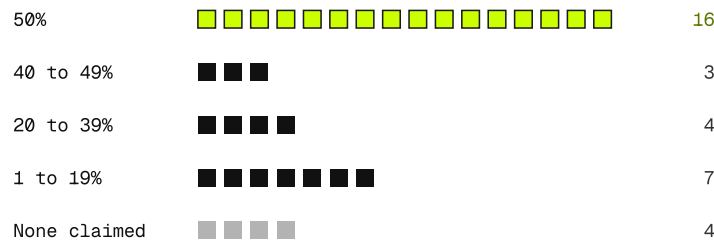


Figure 2. How the savings spread. About half the catalog sits at half price, the rest closer to list.

4.2 Hold, charge, release

A prepaid balance must not be overdrawn by a reply whose length is unknown when the call starts. Before dispatch, the Mothpoint holds the call's worst case, priced from the input and the largest output the call permits, and the hold is taken only if the balance covers it. When the call ends, the hold is released in full and the actual cost is charged, and the charge never exceeds the hold. A call that fails is charged nothing.

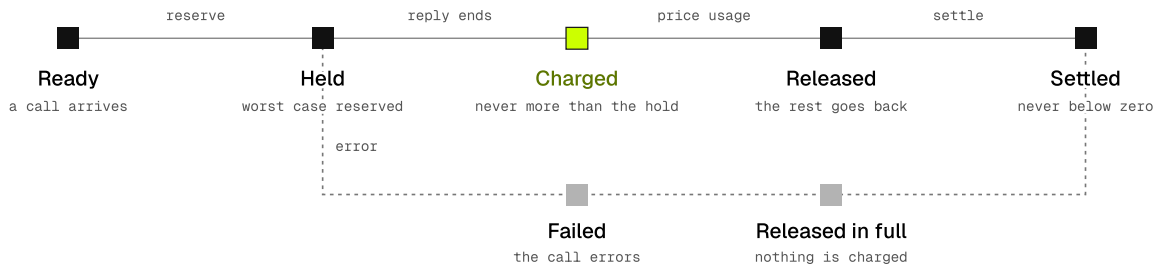


Figure 3. The billing state machine. Every call takes one of two paths from held to settled, and neither leaves the balance below zero.

Since each charge is covered by a hold already taken from the balance, billing cannot take a balance below zero. The one exception is the reversal of a provisional deposit credit when the chain drops the deposit, which is written without a hold.

5 \$MOTH

Not an offer. This paper is not an offer to sell, or a solicitation of an offer to buy, \$MOTH or any other token or security, in any jurisdiction. \$MOTH has not launched, and nothing here promises a price, a return or a profit. Holder Credit is a share of swap tax set aside as platform credit or USDG. It is not interest, a dividend or a return, and in any hour it can be zero.

\$MOTH is a planned token on Robinhood Chain [3], to be launched through Pons [5] with a fixed supply of 1,000,000,000 and no minting after launch. Pons sells the token along a bonding curve and, when the curve sells out, moves the liquidity into a Uniswap v4 pool [6] whose liquidity is locked so that it cannot be withdrawn, by Moth or by anyone. Every buy and sell in that pool pays a fixed 3% tax to a splitter contract [9] named as the Pons creator fee recipient.

One call to the splitter divides the tax by its split in force and emits an event with every amount. The default split is 833 basis points of the tax burned (0.25% of a swap), 2,500 to Holder Credit (0.75%) and 6,667 to development (2%). The owner, the wallet that creates the token, can change the split, and each change is an on-chain event.

Holder Credit is the mechanism by which Moth pays for inference. Each hour, the Holder Credit share, received as USDG, is divided among wallets holding about \$20 or more of \$MOTH in proportion to each wallet's average balance over the hour, so that a balance held briefly around the hour counts for little. A holder claims the share as credit on the balance that pays for calls, or withdraws it as USDG. Holder Credit is a share of swap tax. Nothing is lent or staked, no amount is promised, its size depends on trading volume and on a split the owner can change, and in an hour without swaps it is zero. It is not interest, a dividend or a return.

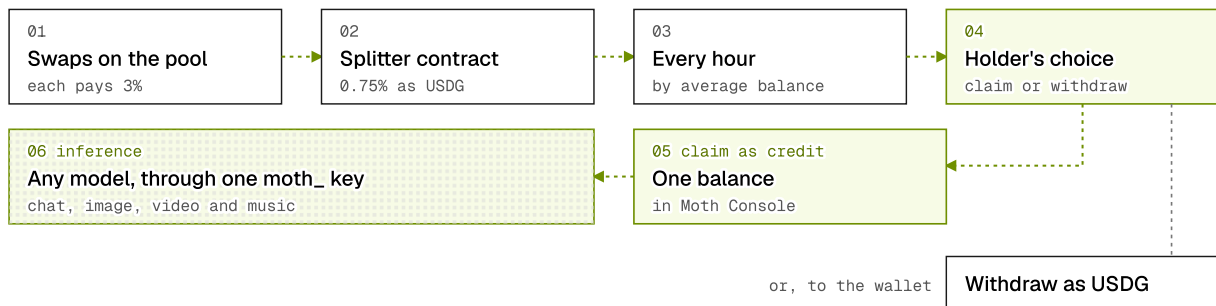


Figure 4. From a swap to a model call. The default split at launch. Nothing in this figure is a forecast of any amount.

A second use is independent of the first. After launch a balance can be funded in \$MOTH, with 10% bonus credit at a time-weighted pool price, half of the tokens received burned and half sold to pay for the calls they fund.

6 Risks and limits

6.1 Product

- **Host access under Anonymized.** The host reads content under its own policy, which Moth does not control.
- **Identifying prompts.** Text that identifies its author is forwarded as written.
- **Sealed is not live.** When it ships it depends on the soundness of the hardware and its attestation, and it covers a few open-weight models.
- **Dependence on makers and hosts.** Models, list prices and terms are set by their makers and can change or be withdrawn, and an outage at a host is an outage at Moth.
- **Moving list prices.** Each saving is measured against a list that its maker can change, and savings already range from none to half.
- **Availability.** A model that cannot be served at its price fails with an error rather than costing more.
- **Prepaid holds.** A call needs its worst case in the balance, so a large output limit on a small balance is refused.
- **Public payments.** Top-ups are public transfers, so Moth states that no account is needed rather than that payment is anonymous.
- **Early software.** The software is new, and no reproducible benchmark of Moth's own exists yet.

6.2 Token

- **Not launched.** \$MOTH does not exist. Its design can change, and the launch may not occur.
- **No guaranteed value.** \$MOTH may be worth nothing. Burns, the locked pool and Holder Credit say nothing about its price.
- **Contract risk.** The splitter, Pons's contracts and the Uniswap v4 pool and hook can contain flaws. The splitter is not deployed, and no audit report is published.
- **Holder Credit can be zero.** It comes only from swap tax, and the owner can reduce its share to nothing.
- **Not a return.** Holder Credit is credit or USDG from swap tax, not income on a holding.
- **Regulation.** Rules for tokens vary by jurisdiction and are changing, and could restrict holding, trading or use.
- **Chain and stablecoin.** A fault in Robinhood Chain, in USDG or in a bridge affects the token, top-ups and Holder Credit.

6.3 What Moth does not claim

Moth does not claim that any model performs better through it, that smaller models match frontier models, any speed or accuracy figure, any price, return or profit for the token, that Holder Credit is income, or anonymity beyond what its own servers do. Frontier models run as their makers ship them, beside open-weight models.

Not an offer. This paper is not an offer to sell, or a solicitation of an offer to buy, \$MOTH or any other token or security, in any jurisdiction. \$MOTH has not launched, and nothing here promises a price, a return or a profit. Holder Credit is a share of swap tax set aside as platform credit or USDG. It is not interest, a dividend or a return, and in any hour it can be zero.

7 Future work and conclusion

7.1 Future work

The next component is Sealed, on a few open-weight models, first in Moth Chat and then through a small SDK for code, since plain chat clients cannot encrypt to attested hardware. After Sealed and the token come a code mode in Moth Chat, Moth Code, a terminal agent for a desktop, a server or a cloud session, Moth Browser, a browser with every model built in, and an MCP server that exposes Moth's models to other AI applications. None of these is built. They carry no date and no feature commitment, and each uses the same wallet, balance and key.

7.2 How the design will be tested

The design is falsifiable on three counts. The first, that no content is retained, is tested in code by passing a marker string through every request path and failing if it appears in any log or table. The second, that workloads stay on one key, can be observed only from use over time. The third, that senders bring questions they would not send elsewhere, cannot be read from prompts, since none are kept, and can only be inferred from use. Results will be published, including those that contradict the design.

7.3 Status of components

Component	Status
Moth Chat, Moth Console, the API, USDG top-ups	in operation
Sealed tier	implemented in code, not live
Splitter contract	built and tested, not deployed
Holder Credit hourly computation	implemented, not running until the splitter is deployed and a time-weighted price source exists
Holder Credit withdrawal as USDG	specified, not implemented
Paying in \$MOTH	implemented, waiting for the price source
Code mode, Moth Code, Moth Browser, MCP server	planned

7.4 Conclusion

Moth places one intermediary between the caller and the hosts and limits what that intermediary may retain. One wallet, one prepaid balance and one key reach 34 models from 11 makers, priced below each maker's published list wherever one exists. Every call passes one address that holds its worst case before dispatch, removes the sender's identity before any host is selected, and reissues the request under a shared credential. No prompt or output is stored. The planned token divides a fixed swap tax on chain and sets a share aside each hour as credit for inference, a share that depends on trading and can be zero. The system is specified in the docs. This paper is a draft for discussion and is not an offer.

Contract and links

Left blank on purpose. These fields are filled in at launch from Moth's own sources. Until they are, treat any address that claims to be \$MOTH as unverified.

On chain

\$MOTH contract address

Robinhood Chain, chain id 4663

Splitter contract address

MothTaxSplitter

Uniswap v4 pool

Pons launch page

Chart

Explorer

Audit report

Moth

Website

App

Moth Chat

Console

Moth Console

Docs

the docs app, full system detail

Community

X (Twitter)

Telegram

Discord

GitHub

Filled in by

Date

References

- [1] Moth, model catalog, read 21 October 2026.
- [2] Maker pricing pages as recorded in the catalog, among them <https://claude.com/pricing>, <https://ai.google.dev/gemini-api/docs/pricing> and <https://developers.openai.com/api/docs/pricing>, read 13 to 17 October 2026.
- [3] Robinhood, Robinhood Chain documentation, <https://docs.robinhood.com/chain/>.
- [4] Paxos, USDG on mainnet, <https://docs.paxos.com/guides/stablecoin/usdg/mainnet>.
- [5] Pons, documentation, <https://docs.ponsfamily.com/v2>.
- [6] Uniswap, v4 deployments, <https://developers.uniswap.org/contracts/v4/deployments>.
- [7] EIP-4361, Sign-In with Ethereum, <https://eips.ethereum.org/EIPS/eip-4361>.
- [8] R. Barnes, K. Bhargavan, B. Lipp and C. Wood, Hybrid Public Key Encryption, RFC 9180, 2022, <https://www.rfc-editor.org/rfc/rfc9180>.
- [9] Moth, MothTaxSplitter.sol, Solidity 0.8.35, MIT license. Built and tested, not deployed.