

QSI Public API

Version 1.0 · quantumsi.in · 16 October 2026 · Private preview

What it is

The QSI Public API lets your product use AI through QSI. You call **one endpoint** with a QSI API key; QSI's server runs the model and returns the answer. You never buy, hold, or manage any model provider key — no Moonshot, no OpenAI, nothing. QSI handles the model, the limits, the uptime and the bill upstream; you get one simple monthly plan.

For developers: one POST request, one header, JSON in → JSON out. OpenAI-style messages, so most existing code connects in minutes.

Quick start

Endpoint: `https://quantumsi.in/api/v2/pub.php`

Auth: send your key in the header `X-QSI-Key: qsi_live_<your-key>` (or `Authorization: Bearer <your-key>`).

```
curl -X POST https://quantumsi.in/api/v2/pub.php \
-H "Content-Type: application/json" \
-H "X-QSI-Key: qsi_live_XXXXXXX" \
-d '{"action":"chat","messages":[{"role":"user","content":"Hello!"}]}'
```

```
{"ok":true,"answer":"Hello! How can I help?","finish":"stop"}
```

Actions

Action	Who	What it does
<code>chat</code> (POST)	any key	Send OpenAI-style <code>messages</code> ; get the AI answer. This is the product.
<code>status</code> (GET)	any key	See your plan, calls used today/this month, and quota remaining.
<code>issue</code> (POST)	owner only	Create a new API key for a plan. The full key is shown once .
<code>revoke</code> (POST)	owner only	Disable a key immediately.

How usage is counted (simple, predictable)

- **One call = one completed AI answer.** Failed or timed-out calls are never counted.
- Per call: at most 50 messages, each ≤ 20,000 characters; request ≤ 100 KB. Answers are capped by the server.
- Speed limit: 20 calls per minute per key.
- Counts are kept on the server per key (`used_today` , `used_month`), reset automatically every month. Both sides can check them any time via `status` .
- When the monthly allowance ends, the plan either stops (`monthly_cap`) or continues at the overage price per call — set per plan by QSI.

Plans & pricing (launch offer)

Plan	Price / month	Included calls	Beyond allowance
Starter	₹199	400 calls	Overage price per call (set per plan), or service stops at the allowance — your choice per plan.
Pro	₹499	1,500 calls	

Business	₹1,999	8,000 calls	
----------	--------	-------------	--

Prices are launch placeholders, in Indian Rupees, billed monthly per key. GST extra as applicable. Custom volumes and annual billing: contact QSI.

Reliability & rules

- Every call passes automated guards: key valid → not revoked → rate limit → quota → spend ceiling → model. Failures return a clear JSON error code (`bad_key` , `disabled` , `rate_limited` , `monthly_cap` , `spend_ceiling` , `upstream_error`).
- A global daily spend ceiling protects the service for everyone.
- QSI's margin floor: overage pricing is enforced so every call stays profitable to serve — the service cannot lose money on its own plans.
- CORS is enabled; browser-based apps may call the API directly.
- Keys are stored as hashes on QSI's server — a server leak cannot expose them.

Getting a key today

During the private preview, QSI issues keys manually: choose a plan, complete payment, and QSI emails you the key (shown once) with the docs. Self-serve key creation in the QSI dashboard arrives next.

Contact: [quantumsi.in](mailto:sarthak@quantumsi.in) · sarthak@quantumsi.in

QSI Public API v1.0 · This document describes the private preview. Pricing, quotas and features may change at launch. © 2026 QuantumSI.