

ATERNA's agent fleet on ROUTEXOR

Our own company. ATERNA AI builds and operates ROUTEXOR, so ATERNA is not a paying customer and none of this is revenue. We publish it because it is the only production workload we can show today, and because every design-partner case study will follow the same format.

The workload

ATERNA runs AI agents for itself and for its small-business clients: a voice assistant, website concierges that answer chat and calls, internal operations agents and three product apps.

They run around the clock, on schedules and in bursts, across several model providers.

How it is set up

- One ROUTEXOR key per agent deployment, so each agent's calls, tokens and estimated spend show up on their own.
- ATERNA's own provider keys sit in the encrypted vault. Providers bill ATERNA directly.
- Agents point their existing OpenAI-format or Anthropic-format clients at api.routexor.com. No SDK change.
- Fallback chains across providers are set per agent, in the agent harness or in ROUTEXOR task profiles, so a failing provider hands the call to the next model.

The numbers, as of October 12, 2026

34,597 Successful requests routed All time, ATERNA-operated accounts	835.7M Tokens routed All time	\$1,244.52 Model spend, billed by providers to ATERNA Estimated from catalog prices, not reconciled invoices	\$0 Held on deposit at ROUTEXOR No balance, no credits
13,510 Successful requests, last 30 days 160.7M tokens in the same window	19 ROUTEXOR keys active in the last 30 days All under ATERNA-operated accounts, including internal tool keys	\$1.49 Blended model cost per million tokens Reflects ATERNA's model mix, not a discount	\$27.38 ROUTEXOR fee at the public Starter rate 2.2% of \$1,244.52. ATERNA actually pays \$0

What changed for ATERNA

- Spend is visible per agent instead of one bill per provider.
- Each key can carry its own dollar cap, checked before every call is sent, so one runaway agent stops at its own limit.
- Moving an agent to a cheaper or newer model is a configuration change, not a code change.
- When a provider rejects a key, the agent gets a 502 and its harness moves to the next model instead of locking out every fallback.

What this does not show

- No outside customer. Every figure here is ATERNA's own traffic.
- No savings figure. We did not keep a measured before-and-after baseline, so this reports usage and control, not dollars saved.
- Spend is estimated from catalog prices at the time of each call. It is not a reconciled provider invoice.
- At this volume the platform fee is small either way. Fee differences only matter at higher spend.