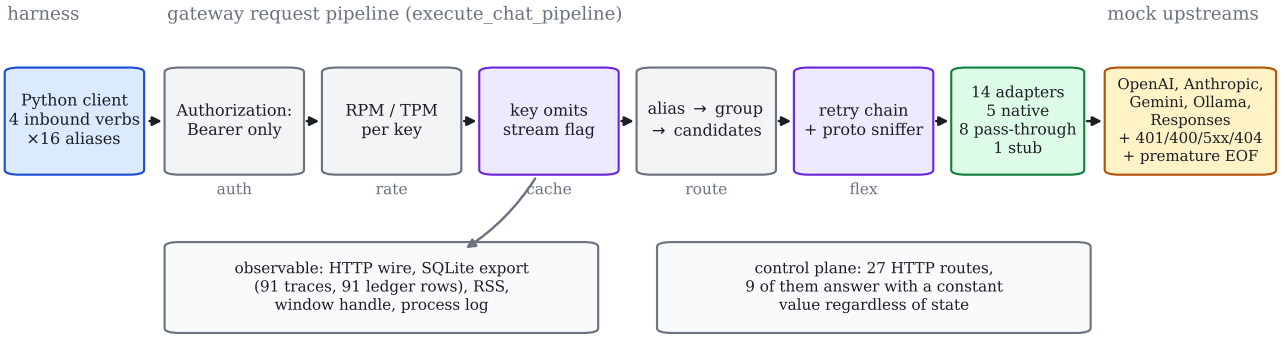

harness
gateway request pipeline (execute_chat_pipeline)
mock upstreams
Python client
4 inbound verbs
×16 aliases
Authorization:
Bearer only
auth
RPM / TPM
per key
rate
key omits
stream flag
cache
alias → group
→ candidates
route
retry chain
+ proto sniffer
flex
14 adapters
5 native
8 pass-through
1 stub
OpenAI, Anthropic,
Gemini, Ollama,
Responses
+ 401/400/5xx/404
+ premature EOF
observable: HTTP wire, SQLite export
(91 traces, 91 ledger rows), RSS,
window handle, process log
control plane: 27 HTTP routes,
9 of them answer with a constant
value regardless of state