InferenceGate — Pilot & Commercial Overview

Cut LLM costs without breaking UX. *No upfront cost pilot.*

What it is

InferenceGate is an outcome-first inference gateway deployed in front of your existing LLM provider. It reduces LLM costs by 65–75% through intelligent routing while preserving identical response shape and UX. No SDK or client-side changes are required.

Why it matters

Most LLM stacks send every request to the most expensive model, with poor cost visibility and high optimisation risk. InferenceGate introduces control, auditability, and deterministic cost reduction without quality regression.

How it works

SS gates obvious cases early. AP resolves easy requests using a low-cost probe. FULL models are used only when required. Every decision is captured via structured telemetry to enable governance and optimisation.

Pilot program

Objective: Validate cost reduction on real production traffic without harming quality.
Includes: Drop-in integration · Live KPI report · Clear go / no-go decision.
Timeline: Week 1 integration · Week 2 validation.
Cost: No upfront cost.

Target FULL rate	15–25%
PROBE_ACCEPT	60–75%
NO_INFERENCE	5–15%
Expected savings	65–75%

Commercial model

Post-pilot pricing is usage-based and aligned with verified savings. There is no public pricing because value depends on traffic mix and acceptance thresholds. Enterprise contracts are available.

Next steps: Apply for the pilot, run InferenceGate on production traffic, review the KPI report, and decide go / no-go.

SSAP / InferenceGate · marko@ssap.io · https://ssap.io