

ModelScale API cost and evidence cheatsheet

Guide reviewed: 2026-09-13

KEEP UNITS EXPLICIT

Token prices use USD per 1M tokens.

Monthly tokens = calls/month x tokens/call, allocated by route share.

Cache-read and cache-write prices are separate facts.

COMPARE COSTS CONSISTENTLY

3:1 blended token price = $(3 \times \text{input price} + \text{output price}) / 4$.

Unknown price is unavailable, never zero.

Round micro-USD only after exact integer arithmetic.

CHECK THE EVIDENCE

Keep observation, verification, and publication clocks separate.

Retained last-good values keep their original age and stale label.

Source-scoped records are not unique canonical products.

DECIDE SAFELY

Separate entitlement and capacity evidence from price.

Qualify runtime by metric, unit, region, and cohort or mark unavailable.

Use the ModelScale calculator for current published values.