SINCE 1994

Choosing the *right model.*

AND KNOWING WHEN TO SWITCH · REX BLACK, INC.

"Use the *biggest* model"

is the default.

And it's *wrong*.

Safe at prototype. Expensive at production.

FOUR AXES

Score every choice on the *same four axes.*

AXES

Capability. Score on *your* golden set. Public benchmarks have no predictive value.

Latency. P50 and P95 against the call site budget.

AXES (CONT.)

Cost. Per-decision, fully loaded. Include retry rate and fallback rate.

Reliability. Single-provider 99.5% = four hours of degraded ops per month.

The rule that matters most.

If a smaller model scores **within 2 to 3 points** of a larger model on *your* rubric, that is within the noise floor.

The smaller model is capable enough.
The cost gap is a pure economic win.

ROUTING

Cascade routing with a *confidence gate.*

SHAPE

First pass, small model. ~80% resolved here.
Confidence score on the output.
Promote below threshold to flagship.

TUNING METRIC

Promotion rate. 15-20% after tuning is the target.
Above 30% = gate is wrong.
Below 10% = quality is bleeding.

The economics at *three tiers.*

10K/DAY

Flagship: **~$9K/month.**
Cascade: **~$3K/month.**

100K/DAY

Flagship: **~$90K.**
Cascade: **~$28K.**

1M/DAY

Flagship: **~$900K/month.**
Cascade: **~$240K/month.**

PAYBACK

Fastest engineering work on the backlog.

PROMPT · FINE-TUNE · SWITCH

Three capability moves. *In order.*

PROMPT FIRST

Reversible. Cheap. Always first.

FINE-TUNE

Only when the task is **stable and narrow.**
Be honest about the re-train obligation.

SWITCH TIERS

When the gap is **capability, not data.**

No amount of prompting closes a capability gap.

Switching cost is *higher* than you think.

TAXES

Prompt tuning. Prompts rarely port across models.
Eval re-run. Full golden set, new scores.

TAXES (CONT.)

Regression re-verify. Known bugs may not still be bugs.
Operational memory. Two weeks of on-call recalibration.

Multi-provider fallback.

Not optional for revenue-touching AI.

Every provider has **at least one significant outage per quarter.**

ARCHITECTURE

The provider *abstraction layer.*

Fifteen lines of code. Build it *before* you need it.

INPUT

Request payload.
Model hint.
Latency budget.

OUTPUT

Response.
Provider used.
Cost incurred.

This week.

Instrument per-decision cost on the hottest feature. If the answer is "we would need to calculate," that is the first gap.
Run the cascade prototype. Two-tier, naive gate. If quality holds and cost drops, you just funded the next two quarters of AI work.
Audit switching cost. If migrating off the current model takes more than five engineering days, your lock-in is bigger than the team realizes.

REX BLACK, INC.

Route *before* you upgrade.

REXBLACK.COM/RESOURCES/WRITING/AI-MODEL-SELECTION-FRAMEWORK