



Strategies for Managing OpEx for Individual and Workflow AI Usage with Open Weight Models

2026





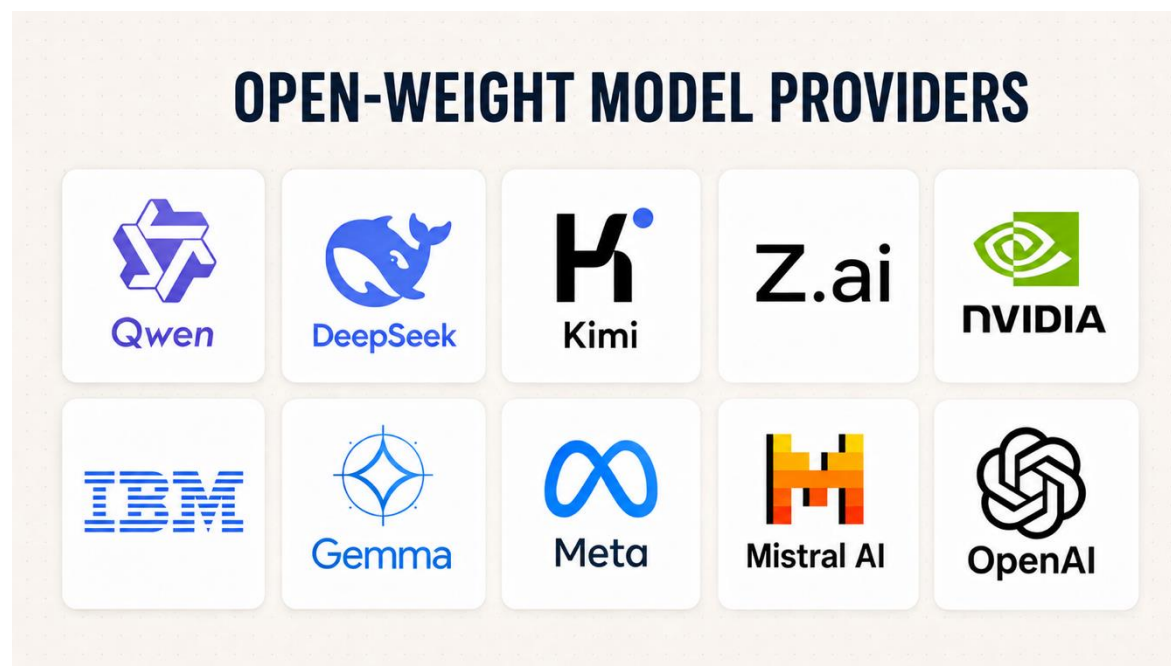
Agenda

- Super Quick primer on open weight models
- 3 different cases for replacement:
 - Cost control
 - Quality improvement
 - Data access
- Decision-making framework (when/why do we replace)



Quick Primer on Open Weight Models

- Open Weight $\neq$ Chinese models (though they are very over-represented)
- Open weights means that both the architecture, training methodologies, and the actual weights are public and have a permissible license
- Open weights can be run locally but not necessarily.





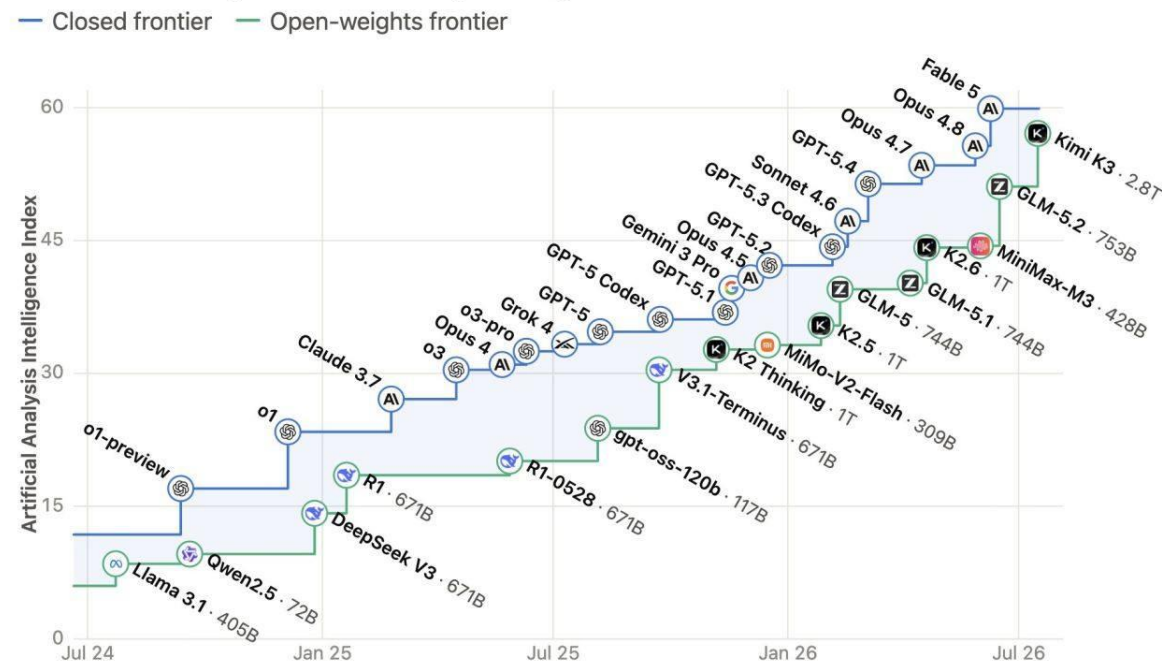
Cost Control

With a clean Eval for “good enough”, you can start to experiment with cheaper models

Many “simple” tasks are already workable on the edge (ie device) like STT, summarization, reformatting (within reason).

Best ROI: library of use cases to model matching

Frontier AI intelligence over time: open weights vs closed models



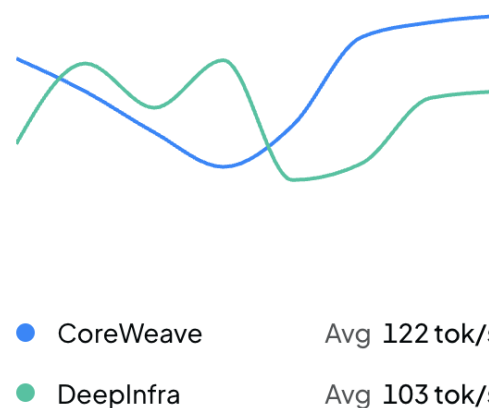


Quality Improvement

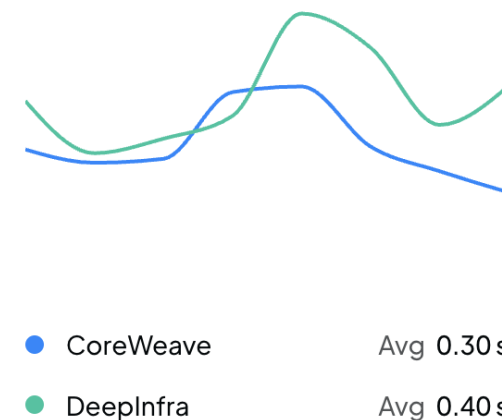
Quality is in the eye of the beholder: but we'll talk in 2 workflow categories: speed (throughput and/or latency) and output quality.

- Speed because token generation speed can be a limiting factor (eg for time-sensitive operations)
- Quality because some operations still have no defined stable/upper bound

Throughput



Latency





Data Access

Share button is the devil

Data access exists beyond just copy/paste

Data egress concerns

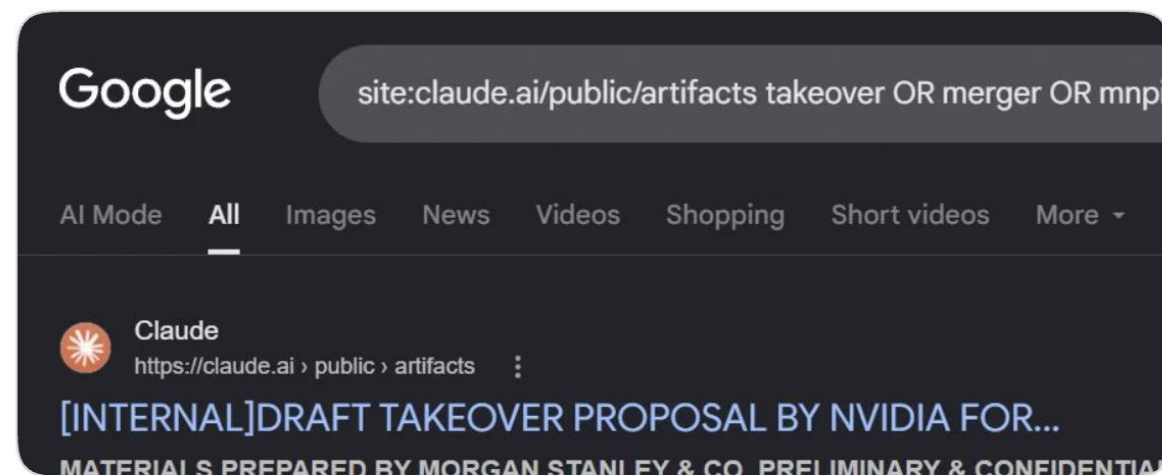


illiquidity providoooooor

@skyquake_1



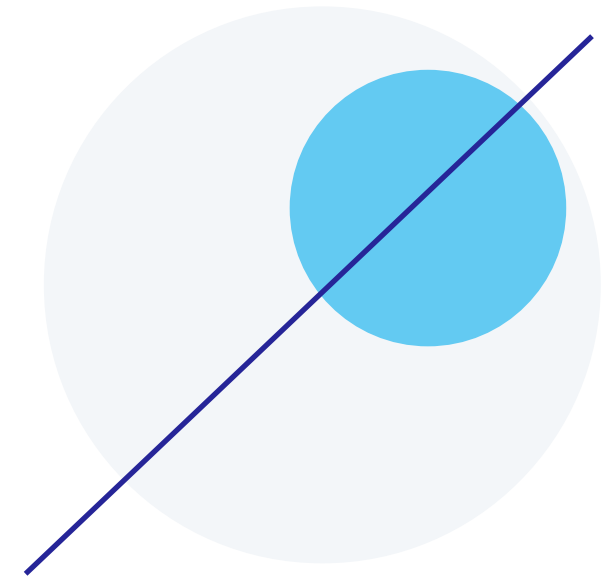
you, an imbecile, learning to prompt LLMs better
me, also an imbecile, learning to prompt google better



8:20 PM · Jul 26, 2026 · **1.7M** Views

From token pricing to an AI operating model

CapEx, OpEx, and the economics of enterprise AI workloads

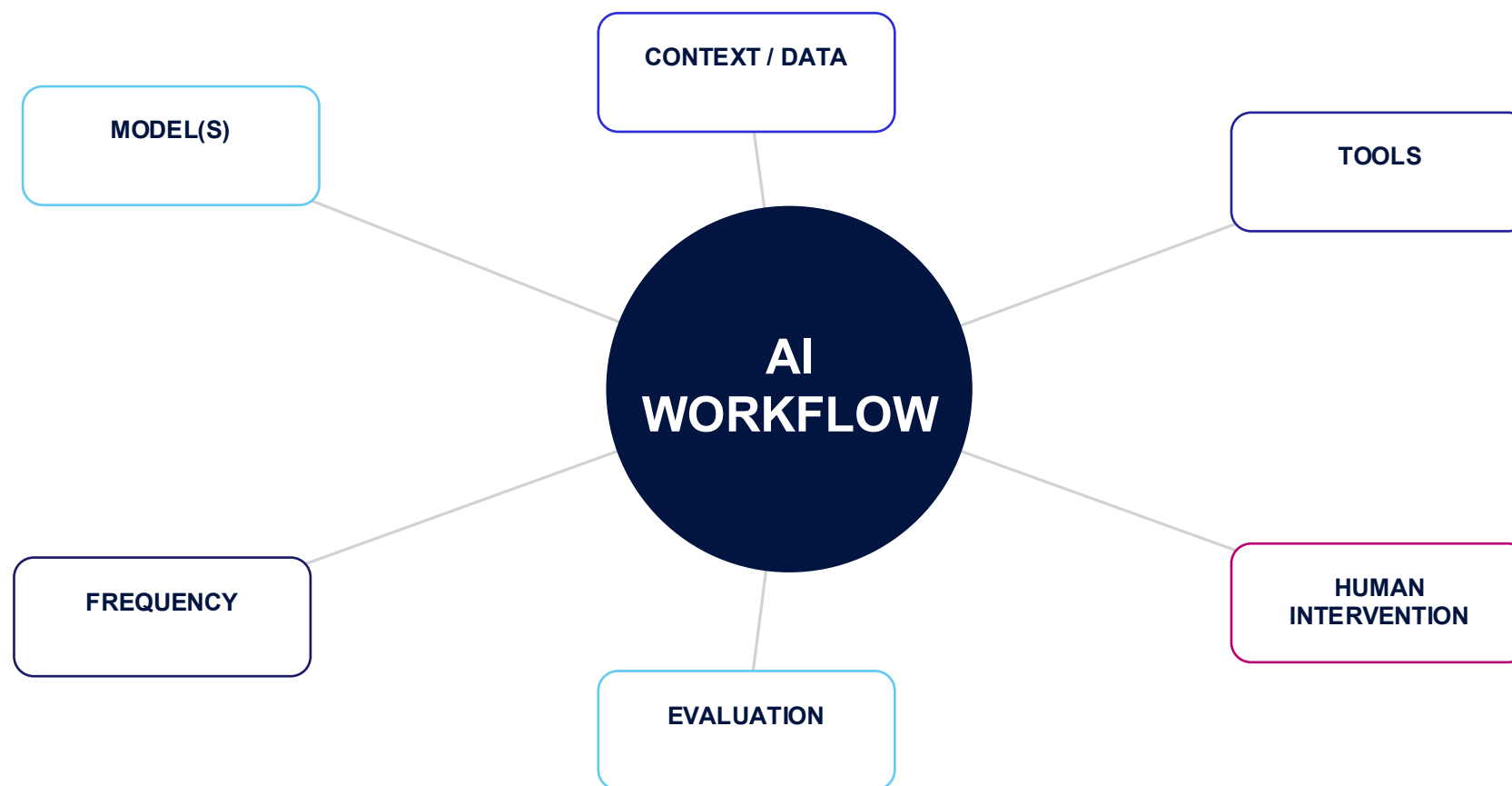


The question is not “which model is cheapest?” It’s “what’s the cost per acceptable business outcome?”





The workflow is the unit of management



Cost per successful execution × **Useful executions** → **Business outcome**



Experiments and production need different economics

EXPERIMENTAL WORKFLOWS

The ROI is still a hypothesis.

- Define a spend ceiling
- State the hypothesis
- Measure evidence and learning
- Build reusable capability
- Kill weak ideas quickly

ESTABLISHED WORKFLOWS

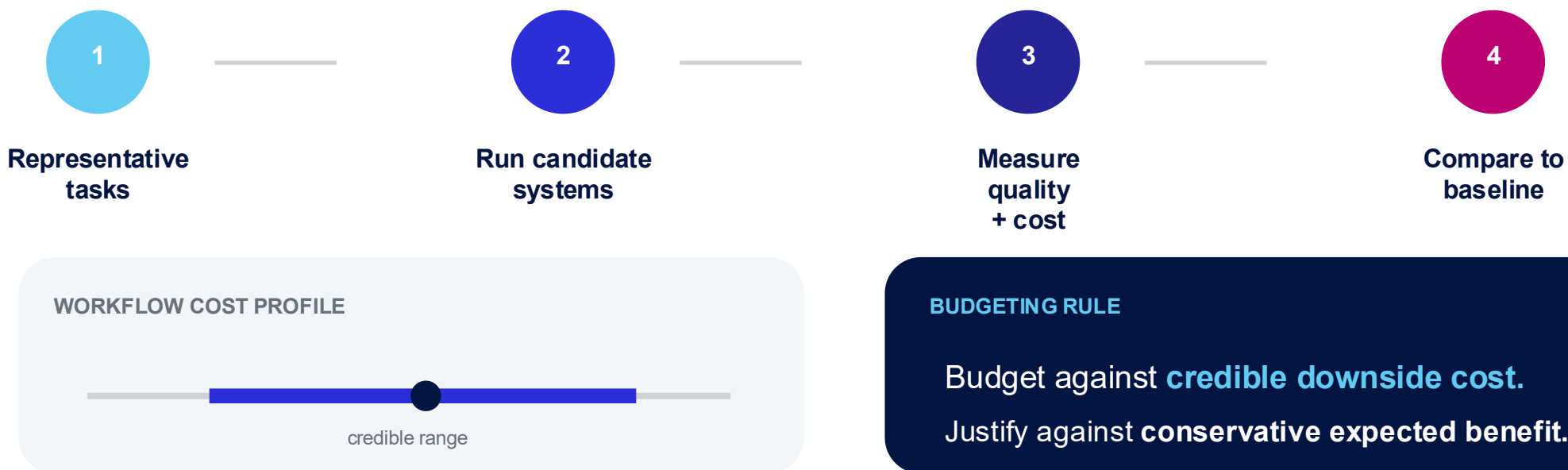
The economics should be increasingly measurable.

- Cost per run
- Success / quality rate
- Volume and frequency
- Business benefit
- Reliability and operating burden

Fund experiments to **learn**. Fund production to **return value**.



Build an eval bench before you optimize spend



inference + infrastructure + labor + evaluation + operations + risk

Consistency beats fake precision.



Not every workflow deserves the same operating model

Lifecycle



Scope



Standardization



Demand



Constraints



Who pays



Who owns



Where it runs



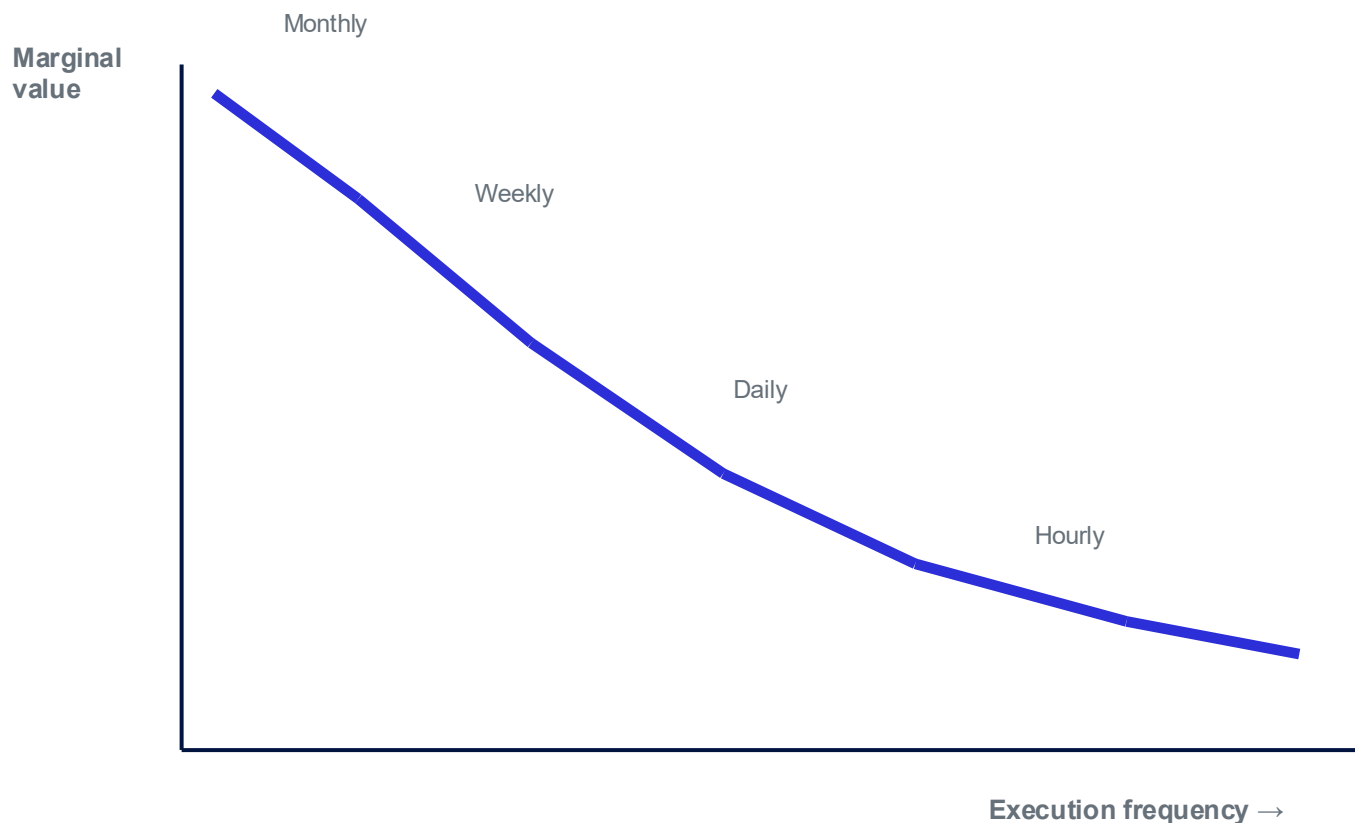
How much to optimize



When to review



Every workflow has a return curve



PROJECT STATUS EXAMPLE

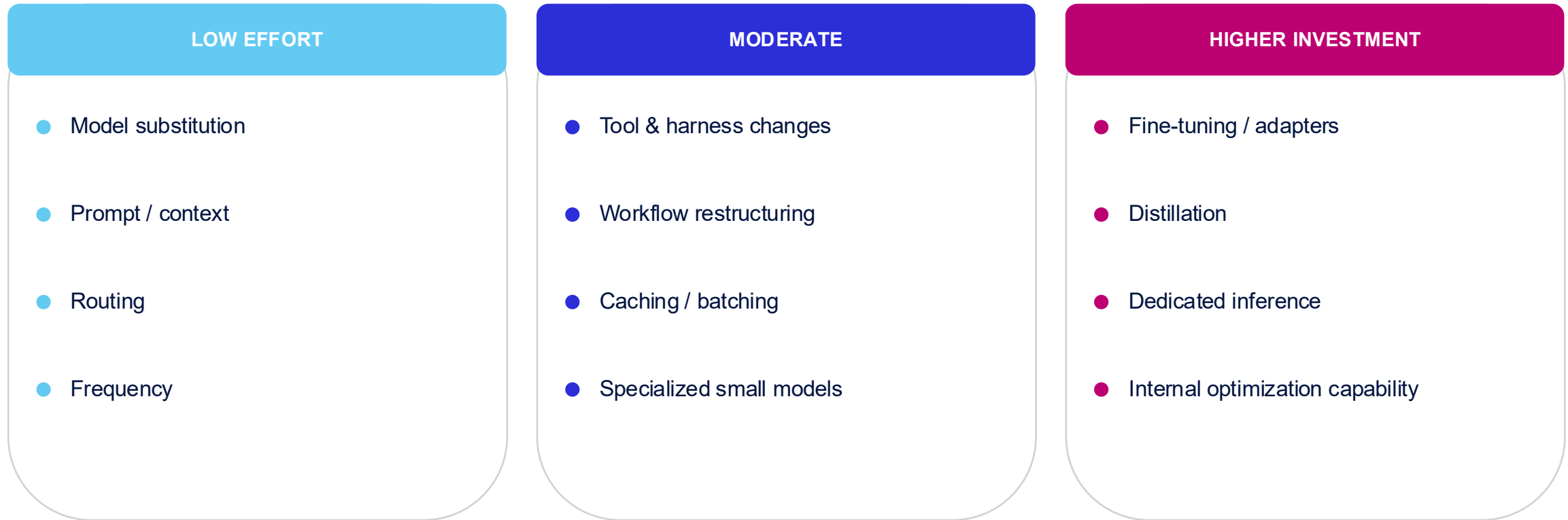
AI can regenerate a report every hour.

But if the underlying systems and people update once or twice a day, the extra runs create the perception of freshness—not new information.

Cheaper inference $\neq$ more valuable inference.



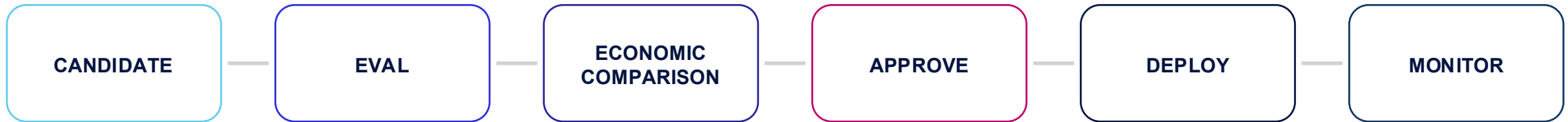
Optimize the system, not just the model



Value of improvement > cost to build + evaluate + maintain it



Treat AI changes more like releases than automatic upgrades



A cheaper or better candidate creates a capital-allocation decision, not an automatic upgrade.

Take the savings • Increase frequency • Increase quality • Fund another workflow • Do nothing

The right answer depends on the workflow's return curve and the portfolio around it.



Manage AI as a portfolio, then continuously reallocate

Workflow	Stage	Owner	Cost	Return	Hosting	Next decision
Status reporting	Established	PMO	\$	High	API	Reduce model cost
Contract review	Established	Legal	\$\$	High	Private	Test owned inference
Research assistant	Experimental	Strategy	\$\$	Unknown	API	Continue experiment
Classification	Established	Ops	\$\$\$	Moderate	Owned	Increase volume



The mature AI capability is the ability to decide where the next dollar of AI spend produces the most value.



Thank You

phil.heijkoop@valiantys.com