INTELLISYNC SOLUTIONS

STRUCTURE - CLARITY - BETTER DECISIONS

ECONOMICS

AI Cost and Token Budget Planner

A worksheet for estimating the operating cost of model routing, retrieval, tools, evaluation, and cached context.

WHAT THIS TEMPLATE HELPS YOU DECIDE

Build a practical cost model across usage scenarios, model tiers, token volume, tool calls, evaluation passes, caching, and monthly operating budget.

BEST FOR

- SMBs deciding whether AI can scale affordably
- Teams comparing model routes
- Operators budgeting production AI workflows

OUTPUTS

- Usage assumptions
- Cost model
- Routing and cache opportunities

STEP 1

Frame the operating need

Start with the workflow, decision, owner, and business pressure. The template is useful only when it is grounded in a real operating moment.

Operating frame

Workflow volume
Estimate requests, users, documents, calls, or sessions per day.

Model need
Reasoning, extraction, summarization, voice, classification, or tool use.

Cost owner
Name who reviews monthly spend and routing changes.

Readiness check

- [ ] Usage scenarios are estimated
- [ ] Model capability is matched to workflow need
- [ ] Retrieval and tool costs are included
- [ ] Evaluation and retry costs are visible
- [ ] Caching or routing opportunities are identified

STEP 2

Map the architecture questions

Use this page to separate the parts of the system that need design before anyone jumps to tools, prompts, or implementation details.

Design map

Volume	How many requests, sessions, or documents will this workflow process?
Context	How much retrieved or uploaded context is needed per run?
Routing	Which requests need stronger models and which can use smaller models?
Tools	Which deterministic calls add cost, latency, or operational value?
Evaluation	How many review, scoring, or regression passes are required?

AI cost is an architecture question. Spend should reflect workflow value, model need, context size, and operating cadence.

STEP 3

Turn the answers into a brief

A strong brief makes the next decision easier: proceed, defer, redesign, govern, or assess more deeply before implementation.

Decision brief

Monthly estimate	What range should leadership expect?
Cost guardrail	What threshold triggers routing, caching, or scope review?
Optimization	Where can caching, summaries, or smaller models reduce cost safely?
Decision	Proceed, reduce scope, redesign, or defer?