



AI Cost and Token Budget Planner

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

STRUCTURE - CLARITY - BETTER DECISIONS

Practical architecture for accountable AI-enabled work.

What this template is for

Build a practical cost model across usage scenarios, model tiers, token volume, tool calls, evaluation passes, caching, and monthly operating budget. Use this guide to make the operating choice explicit, test it with evidence, and record who owns the next decision.

BEST USED WHEN

SMBs deciding whether AI is affordable at production scale.

WHO SHOULD BE IN THE ROOM

Include the workflow owner, a person who performs the work, a reviewer with authority to approve or stop it, and the relevant data or tool owner.

CANADIAN CONTEXT

Confirm applicable federal, provincial, territorial, sector, professional, contractual, accessibility, and bilingual-service requirements. Use this workbook as operational preparation, not legal advice.

A five-step working rhythm

01

Volume

How many requests, sessions, or documents will this workflow process?

02

Context

How much retrieved or uploaded context is needed per run?

03

Routing

Which requests need stronger models and which can use smaller models?

04

Tools

Which deterministic calls add cost, latency, or operational value?

05

Evaluation

How many review, scoring, or regression passes are required?

Build the working record

Complete one record for one workflow. Write enough to support a decision, not a generic technology wish list.

PROMPT	YOUR WORKING NOTES
01 Workflow volume	
02 Model need	
03 Cost owner	
04 Volume	
05 Context	
06 Routing	
07 Decision, owner, evidence, and review date	

Budget proposal support

STARTING FRICTION

Long source packets make a pilot look cheap per call but expensive per approved proposal.

WORKFLOW MAP

VOLUME

ROUTE

CONTEXT

TOOLS

COST CHECK

BUDGET

AI-SUPPORTED ROLE

Estimate request routes, context size, cache use, tool calls, evaluation passes, and usage-derived cost. Flag uncertainty around volume.

HUMAN AUTHORITY

The owner sets a budget, quality floor, escalation criteria, and spend review cadence. Resolve the volume question for this case.

PILOT MEASURE

Track actual cost per completed outcome, token use, latency, retries, and quality. Record volume exceptions separately.

DISCUSSION PROMPT

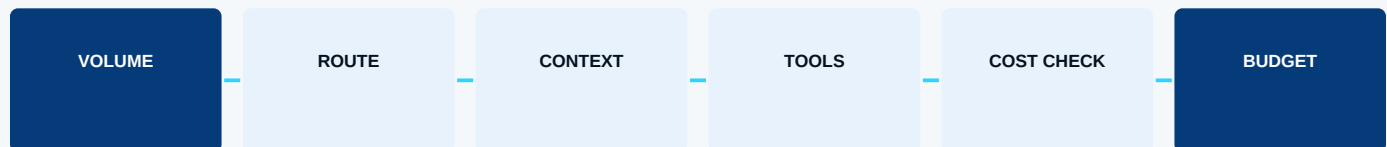
What evidence would the reviewer need to trust, correct, reject, or escalate this output?

Budget customer-service responses

STARTING FRICTION

High-frequency routine questions should not route to the most expensive model every time.

WORKFLOW MAP



AI-SUPPORTED ROLE

Estimate request routes, context size, cache use, tool calls, evaluation passes, and usage-derived cost. Flag uncertainty around context.

HUMAN AUTHORITY

The owner sets a budget, quality floor, escalation criteria, and spend review cadence. Resolve the context question for this case.

PILOT MEASURE

Track actual cost per completed outcome, token use, latency, retries, and quality. Record context exceptions separately.

DISCUSSION PROMPT

What evidence would the reviewer need to trust, correct, reject, or escalate this output?

Budget review summaries

STARTING FRICTION

Low-volume but complex cases may justify escalation only when confidence or materiality requires it.

WORKFLOW MAP

VOLUME

ROUTE

CONTEXT

TOOLS

COST CHECK

BUDGET

AI-SUPPORTED ROLE

Estimate request routes, context size, cache use, tool calls, evaluation passes, and usage-derived cost. Flag uncertainty around routing.

HUMAN AUTHORITY

The owner sets a budget, quality floor, escalation criteria, and spend review cadence. Resolve the routing question for this case.

PILOT MEASURE

Track actual cost per completed outcome, token use, latency, retries, and quality. Record routing exceptions separately.

DISCUSSION PROMPT

What evidence would the reviewer need to trust, correct, reject, or escalate this output?

Before the pilot grows

Use these prompts to test whether the proposed workflow can operate responsibly in its Canadian business context.

01

Volume

Usage scenarios are estimated. Name the owner and the operating decision before selecting a tool.

02

Context

Model capability is matched to workflow need. Record the approved sources, permissions, and missing information.

03

Routing

Retrieval and tool costs are included. Test this boundary with representative and adverse cases.

04

Tools

Evaluation and retry costs are visible. Preserve review, escalation, and correction evidence.

05

Evaluation

Caching or routing opportunities are identified. Set a measurable stop condition and the next review date.

06

Readiness to proceed

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

CANADIAN REFERENCE POINTS

Office of the Privacy Commissioner of Canada - Privacy and artificial intelligence
priv.gc.ca/en/privacy-topics/technology/artificial-intelligence/

Government of Canada - Artificial Intelligence guidance, principles, and policy
canada.ca/en/services/science/innovation/artificial-intelligence.html

General operational guidance only. Confirm applicable legal, privacy, cybersecurity, professional, contractual, and regulatory requirements with qualified advisors.