A PM’s Operating Guide to Comparing Models

A practical no nonsense guide to choosing a model strategy for a real product feature.

THE PROMISE

We will start with a business problem, choose use cases, compare models, run a pilot and present a decision an executive team can act on.

For product leaders

A decision with evidence, owners and operating limits.

For product managers

A working method for comparing routes by use case.

02

TL;DR

This guide follows a PM through a support copilot decision. Each page records the question, the evidence and the action that follows.

WHY THIS GUIDE EXISTS

A model decision changes how agents work and what customers are told while assigning new owners for policy, monitoring and rollback. Use this guide to make those decisions explicit before launch.

The steps we will follow

STEP	PM OUTPUT
1. Frame the problem	Business reason to act now
2. Define the job	User outcome and human role
3. Segment the work	Use cases with different risk
4. Build the pool	Candidates with a reason to enter
5. Select dimensions	Metrics tied to each route
6. Run the pilot	Evidence with stop gates
7. Present the decision	One-slide executive brief

HOW TO READ THE EXAMPLE

The support reply copilot is illustrative. The facts, baseline numbers, model names and results are invented for teaching. Follow the PM's reasoning from one decision to the next then replace the example with evidence from your own product. Before presenting the decision, replace each number with its source, sample size and date.

The output is a model strategy tied to the product's users, risks and operating limits.

03

Start with the product decision

Trace the business pressure to a user problem before you compare models.

ILLUSTRATIVE COMPANY SITUATION

A B2B software company has seen support demand rise for six months after a pricing change and a new enterprise packaging model. Billing questions and policy exceptions now take longer to resolve.

SIGNAL	WHAT WE SEE
Volume	Support contacts are up 28%.
Response time	Average response time grew from 4 hours to 19 hours.
Customer impact	Renewal conversations include complaints about inconsistent answers.
Business impact	Account teams offer more concessions while support spend exceeds the quarterly plan.
Root cause review	Agents search several policy sources before drafting, reconcile conflicting language manually, rewrite the response to fit the customer and escalate routine work when they cannot verify the answer, while some difficult cases still reach customers without the right handoff.

The CPO question

Can we reduce response time while protecting policy accuracy, agent control and customer trust?

My product decision

Create a support reply workflow that gives agents a useful first draft, shows the approved policy behind it and sends uncertain cases to the right reviewer while keeping high-consequence decisions with a human because the cost of a wrong answer exceeds the value of faster drafting.

04

Build the pool

Build the candidate pool from models the company can buy, deploy or maintain, then run a small screen before committing engineering time to the pilot.

Map the market before you name the candidates

Start with models the company can access and support. Major provider catalogs supply hosted frontier models for difficult reasoning and hosted smaller models for routine high-volume work, extraction or classification. Public model hubs supply open-weight models, which give the company more control over hosting but also make it responsible for more of the serving, security and evaluation work. The company's own ML team may supply an internal SLM when the task is narrow, the training examples are proprietary and the team can maintain the model. Use provider documentation, public benchmarks, comparable launches and internal constraints to decide which candidates deserve the screen.

SOURCE	WHERE IT COMES FROM AND WHY IT MATTERS
Hosted models	Provider catalogs and documentation. Compare frontier models for difficult reasoning and smaller models for routine high-volume work.
Open-weight and internal SLMs	Public model hubs or the organization's own training work. Consider these routes when data control or a narrow task justifies added ownership.
Benchmarks and studies	Public evaluations and independent research. Use them to form a hypothesis, then test the actual product job.
Comparable launches	Published product examples. Study the workflow, human review and failure controls rather than copying the model choice.
Internal constraints	Security, privacy, retention, integration and support requirements. Remove candidates that the organization cannot govern.

Shortlist with a reason

CANDIDATE	WHY IT ENTERS
Model B	Hosted smaller model with a lower-cost execution profile for high-volume routine work.
Model A	Hosted frontier model with a higher-cost reasoning profile for difficult policy work.
Model C	Alternate hosted model with a different context or integration profile, such as tool-calling reliability, structured output support or data-residency options.
Internal SLM	Organization-owned candidate for a narrow workflow if data and maintenance capacity exist.
Open-Weight Option	Candidate for stronger deployment control if the quality gate holds.

SCREEN BEFORE THE PILOT

Before the 30-ticket screen, remove candidates that fail data residency or retention rules, cannot support required tools, lack a stable version or cannot meet the latency target. Keep the rest only when each candidate has a distinct quality hypothesis to test.

05

Segment the work

A support queue contains different jobs with different consequences. Define the segments before you decide which model or metric belongs in each route.

USE CASE	PRODUCT RISK	INITIAL ROUTE
Account or order status	A wrong fact creates rework for the agent and a bad promise for the customer.	Model B
Routine billing	High volume makes small errors expensive to correct.	Model B
Refund eligibility	An unsupported policy claim can create a concession.	Model B then Model A when uncertain
Billing dispute	A wrong answer creates financial and trust risk.	Model A
Ambiguous policy	False confidence can prevent the right escalation.	Model B then Model A
Conflicting policy	Conflicting sources require a decision owner.	Model A then human
Multilingual low-risk	Meaning or tone can shift. Use a language-specific eval set and review.	Model B with review
Sensitive complaint	A tone failure can increase escalation. Keep a human reviewer in the route.	Model A with human review
Privacy or regulated	The data boundary and consequence require human control.	Human only
High-consequence decision	The cost of a critical failure is unacceptable.	Human only

Choose the metric from the user action and the consequence of failure

Start with the user action, the consequence of failure and the threshold that would change the route. The support copilot uses this mapping.

USE CASE	USER ACTION	METRIC	WHY IT MATTERS / ROUTE
Account status	Agent sends a factual answer.	Correctness	A wrong status creates a false promise. Keep Model B only when the quality gate holds.
Refund or billing dispute	Agent verifies a policy answer or reviews a financially sensitive draft.	Groundedness and edit burden	An unsupported claim can create a concession and rework can hide a wrong answer. Add Model A or human review.
Ambiguous policy	Agent decides whether to answer or escalate.	Escalation recall	A missed escalation leaves uncertainty with the wrong person. Send it to Model A or a qualified reviewer.
Live queue and repeated use	Agents need the draft quickly and often.	p95 latency, cost per completion and reliability	Slow or costly routes strain capacity. Inconsistent behavior reduces adoption. Use the route that meets all limits.

ROUTE TEST

Define the pass threshold for each row before testing begins. The threshold must tell the team whether to keep the route, add another model or return the case to a human.

06

Turn the metrics into evals

Convert each product question into a repeatable test. Run the tests before live traffic changes the customer or agent experience.

DIMENSION	FORMULA	EVAL CHECK
Correctness	Correct replies / Evaluated tickets	Check that required facts are present and ask a support lead to review the answer.
Groundedness	Supported claims / Total claims	Check that each material claim matches an approved source.
Edit burden	Material edited words / Draft words	Compare the agent's material edits with the baseline workflow.
Escalation recall	Correct escalations / Required escalations	Check that every required case reaches the right owner.
p95 latency	95th percentile response time	Check that the reply arrives within the agent workflow.
Cost per completion	Total route cost / Completed tickets	Track model, review and infrastructure cost over time.
Reliability	Stable results / Repeated runs	Compare repeated runs against the pinned baseline.

Build the eval set

Turn each metric into an eval by defining the ticket set, expected result, pass rule and failure owner. Correctness needs required facts and disallowed claims. Groundedness needs approved sources and a claim-to-source check. Edit burden needs an edited draft and baseline formula. Escalation recall needs labeled cases that require a human. p95 latency needs one response-time record per run. Cost per completion needs model calls, retries, review time and infrastructure cost. Treat the retrieval set, citation format and policy conflict rule as part of the product under test. Run these checks before the pilot so the team can change the prompt, retrieval or route while exposure is limited. Reuse this set as the pilot baseline and regression suite.

EVAL LAYER	WHO OR WHAT SCORES IT	DECISION USE
Automated	Format, citation presence, source match, latency, cost, regression	Compare candidates with repeatable checks
Human	Correctness, usefulness, tone, uncertainty, escalation judgment	Resolve product and policy questions
Product signal	Acceptance, edits, handling time, completion, satisfaction	Decide whether the route improves work

EVAL OUTPUT

Store the model version, prompt version, input ticket, retrieved context, output and scores together. When a result changes, identify whether the cause was the model, prompt, policy source, retrieval step or reviewer.

07

Model the pilot before you run it

Treat the pilot as a controlled product experiment. Define what you expect to learn at each stage and limit exposure while the team learns.

STAGE	QUESTION	EVIDENCE	GATE
Screening	Which candidates deserve more work?	Test 30 representative tickets.	Remove poor fits.
Shadow traffic	How does each route behave on real inputs?	Compare correctness, groundedness and escalation.	Zero critical failures.
Assisted use	Does the draft reduce agent work?	Measure acceptance, edit burden and handling time.	Agent effort improves.
Limited launch	Does the route improve the product?	Measure completion, response time and trust signals.	All stop conditions hold.

Pilot design

Write the hypothesis before the pilot starts. For this example, I expect Model B to handle routine tickets at the required quality and speed. I expect Model A to reduce failures on ambiguous or conflicting policy cases. I will test Model B followed by Model A because a second pass may catch uncertainty without sending every ticket to the more expensive model. Human review remains the baseline for high-consequence work. Name the eval owner, policy owner, groundedness reviewer and rollback owner. Record the sample, gates and decision date with each run, then schedule a regression check after every model, prompt or policy-source change.

Metrics that decide the route

METRIC	FORMULA	STOP OR EXPAND
Correctness	Correct replies / Evaluated tickets	Expand routine route at 90% or higher.
Edit burden	Material edited words / Draft words	Expand when below baseline.
Escalation recall	Correct escalations / Required escalations	Require 100% on the escalation set.
Cost per completion	Total route cost / Completed tickets	Make cost a gate for high-volume routes.
Reliability	Stable results / Repeated runs	Stop on material regression.

COST NOTE

Cost stays in the operating record even though it does not decide this launch recommendation. The team still needs cost per completion for capacity planning, provider negotiations and future pricing decisions.

08

Define rollback before launch

Write the stop condition before you increase exposure. A reversible route lets the team learn while keeping the customer and the business protected.

REAL-WORLD TRUST LESSON

Air Canada’s customer service chatbot gave incorrect bereavement-fare information. The British Columbia Civil Resolution Tribunal found that the chatbot was part of Air Canada’s website and that the company was responsible for the information it presented. The customer experience treated the system as one product.

What this changes in the product

Air Canada treated its chatbot as a customer service channel. The tribunal treated the answer as part of the company’s service. Give the copilot visible sources, human review, clear escalation and a stop condition before you increase traffic.

Rollback triggers

TRIGGER	METRIC AND STOP THRESHOLD	ACTION
Unsupported policy claim	Groundedness = Supported claims / Total claims. Stop on any critical unsupported claim.	Stop expansion. Review retrieval and policy source.
Missed escalation	Escalation recall = Correct escalations / Required escalations. Require 100%.	Route the segment to Model A or human review.
Agent rework rises	Edit burden = Material edited words / Draft words. Stop when above baseline for two review periods.	Reduce traffic. Inspect prompt, model and workflow.
High-consequence failure	Critical failure rate = Critical failures / High-consequence tickets. Target zero before expansion.	Move to human-only handling.
Slow response or cost shift	p95 latency = 95th percentile response time. Cost per completion = Total route cost / Completed tickets.	Contain the route. Recheck capacity and pricing model.
Model or provider change	Drift = Current eval score - Pinned baseline. Stop on material regression.	Freeze expansion. Rerun the eval set.

REVERSIBLE DECISION

Start with shadow traffic and move to assisted use only after the evidence clears the gate. Keep the prior route available so the team can restore service quickly if the new route fails.

Sources: Moffatt v. Air Canada, 2024 BCCRT 149. Google AI Overviews update, May 2024.

09

Present the decision in one slide

Give the executive team a decision they can approve. Connect the business problem to the evidence, the route and the controls.

DECISION

Approve a controlled pilot for a routed support reply workflow.

WHY NOW	EVIDENCE	RECOMMENDATION
Support demand is up 28%. Response time is 19 hours. Agents report inconsistent policy answers.	Illustrative screening shows Model B at 94% on routine tickets with 280 ms p95 latency. Model A reaches 92% on edge cases while Model B reaches 68%. The results support a split route rather than one model for every ticket.	Use Model B for routine tickets because it clears the quality gate at the required speed. Use Model A for ambiguity and policy conflict because it performs better on those cases. Keep high-consequence cases in human review.

APPROVAL REQUEST	PILOT DESIGN	STOP GATES
Limited traffic, named policy owner and named rollback owner.	Same ticket, policy context, prompt and scoring rules across routes.	Zero critical failures. 100% escalation recall on the required set.

Measures that decide the route

USER OUTCOME	BUSINESS OUTCOME	CONTROL
Acceptance and edit burden	Completed tickets and response time	Groundedness and escalation recall
Customer answer quality	Capacity and cost per completion over time	Critical failure rate and rollback

Executive decision

Approve the controlled pilot. The two-week readout will return with segment results, route economics, user outcomes and a launch recommendation.

DECISION STANDARD

Expand only when routine correctness clears 90%, edit burden improves against baseline, escalation recall reaches 100% on the required set and high-consequence tickets have zero critical failures.

Illustrative facts are labeled. Attach the source, sample size and date to every number. Cost remains in the operating model.

About Ascendvent

Better continuity for consequential human work.

Ryan K. McDonald

Founder of Ascendvent.

Throughout my career I have watched organizations lose context between moments of action, forcing professionals to rebuild understanding instead of building momentum.

Ascendvent focuses on the guidance infrastructure that helps people keep context, make better decisions and stay at the center of the work.

Visit ascendvent.life

Use this guide to start a conversation

If your team is exploring agent workflows, secure repository access, internal developer tools or a practical AI operating model, this reference can be adapted into a working discovery and build plan.

ASCENDVENT | ASCEND AND REINVENT