§ MULTI-PROVIDER AI PIPELINES

Route the work. Not just the model.

Combining Claude, OpenAI, and Google Gemini for cost-optimized, quality-maximized enterprise workflows.

§ 01 OVERVIEW

Stop picking one model. Start routing the stages.

Single-provider routing guides assign each use case to the cheapest viable model within one provider's lineup. Multi-provider pipelines go a step further: they decompose a workflow into distinct stages, then match each stage to the best-fit model from any provider based on what that stage actually requires: classification volume, context window size, generation quality, or response latency.

That produces three routing patterns, each suited to a different workflow shape. The three pipelines in this guide each demonstrate a different pattern, chosen because they map to high-value use cases where the cross-provider cost advantage is most significant.

THE THREE ROUTING PATTERNS

01

Volume funnel

Run cheap classification on everything. Qualify aggressively. Deploy the premium model only on the fraction of inputs that cleared the threshold. Savings compound with volume.

02

Structural advantage

Use one provider because of a capability no other matches at that price tier: context window, multimodal input, retrieval integration. Switch providers once that structural need is satisfied.

03

Real-time triage

Every input passes through a fast, cheap classifier first. The routing decision determines which model tier handles the response. Effective cost reflects the distribution of input complexity.

§ 02 CROSS-PROVIDER MODEL REFERENCE

Five models. Three pipelines.

Five models appear across the three pipelines, all sourced from the current model routing guides for each provider. Prices are per million tokens (MTok) for input and output.

PROVIDER	MODEL	INPUT $/MTOK	OUTPUT $/MTOK	CONTEXT	APPEARS IN
GOOGLE GEMINI	**Gemini 2.5 Flash-Lite**	$0.10	$0.40	1M	01·S1 03·S1
OPENAI	**GPT-4.1 Nano**	$0.10	$0.40	1M	02·S2
OPENAI	**GPT-5.4 Nano**	$0.20	$1.25	400K	01·S2 03·S2
GOOGLE GEMINI	**Gemini 3.1 Pro**	$2.00*	$12.00*	1M	02·S1
ANTHROPIC	**Claude Sonnet 5**	$3.00†	$15.00†	1M	01·S3 02·S3 03·S3

* Gemini 3.1 Pro rate applies to requests ≤ 200K tokens; $4.00 / $18.00 above that threshold.
† Claude Sonnet 5 standard pricing; introductory pricing of $2.00 / $10.00 applies through August 31, 2026.

§ READING THE TABLE

The tier gap matters more than the absolute numbers. The cheapest classifier is ~30× less expensive per output token than the premium generator. That gap is what lets a well-structured pipeline concentrate premium spend on the outputs that actually change the outcome.

§ 03 · PIPELINE 01 VOLUME FUNNEL ROUTING

Outbound prospecting at scale.

Score the entire prospect universe cheaply. Research the qualified slice. Spend the premium budget only on the top of the pyramid.

ARCHITECTURE

STAGE 01
Signal detection & scoring
GOOGLE GEMINI
Gemini 2.5 Flash-Lite
$0.10 / $0.40 per MTok

→

STAGE 02
Account research & synthesis
OPENAI
GPT-5.4 Nano
$0.20 / $1.25 per MTok

→

STAGE 03
Personalized outreach drafting
ANTHROPIC
Claude Sonnet 5
$3.00 / $15.00 per MTok

STAGE BREAKDOWN

01 Signal detection & scoring

GOOGLE GEMINI
Gemini 2.5 Flash-Lite
$0.10 / $0.40 per MTok

Every account in the prospect universe passes through this stage: potentially thousands of signals per day. Gemini 2.5 Flash-Lite's $0.10 input cost makes this economically viable at scale. The task is well-defined classification (job change, funding round, tech stack shift) with clear output categories that do not require premium reasoning.

In: job posting data, LinkedIn activity, funding signals. · Out: prospect score + signal tags per account.

02 Account research

OPENAI
GPT-5.4 Nano
$0.20 / $1.25 per MTok

Only accounts that cleared the Stage 1 threshold reach this point, typically 5 to 15% of the total prospect universe. GPT-5.4 Nano synthesizes RAG-retrieved account intel (recent news, product usage signals, CRM history) into a structured brief. Synthesis against retrieved context, not premium generation, so the Nano tier fits.

In: retrieved account intel + Stage 1 tags. · Out: structured brief (context, trigger, personas, angle).

03 Personalized outreach drafting

ANTHROPIC
Claude Sonnet 5
$3.00 / $15.00 per MTok

Claude Sonnet's writing quality and natural fluency is where the pipeline's premium spend is concentrated. This stage runs only for the top 1 to 5% of prospects with high signal scores and a qualified research brief. Personalization quality is a direct lever on conversion, so the higher cost is justified as a fraction of total pipeline spend.

§ 04 · PIPELINE 02 STRUCTURAL ADVANTAGE ROUTING

Legal contract review & redlining.

Use Gemini's 1M-token window for full contract-suite ingestion in one pass, then hand off to Claude for the high-value legal output.

Contract review typically hits a hard constraint: context window size. A full suite (MSA, SOW, exhibits, and prior amendments) can exceed 500K tokens, which makes single-pass review impossible for models capped at 200K. Gemini 3.1 Pro's 1M-token window covers this comfortably: one API call, no chunking, no lost cross-document context, no stitching artifacts.

ARCHITECTURE

STAGE 01
Full document ingestion & extraction
GOOGLE GEMINI
Gemini 3.1 Pro
$2.00 / $12.00 per MTok

→

STAGE 02
Clause classification & risk tagging
OPENAI
GPT-4.1 Nano
$0.10 / $0.40 per MTok

→

STAGE 03
Redline generation & deal memo
ANTHROPIC
Claude Sonnet 5
$3.00 / $15.00 per MTok

STAGE BREAKDOWN

01 **Full ingestion & extraction**

GOOGLE GEMINI
Gemini 3.1 Pro
$2.00 / $12.00 per MTok

The 1M-token window holds most full contract suites in a single API call, eliminating chunking logic, cross-chunk reconciliation, and context bleed. A defined term in the MSA that modifies an obligation in a SOW exhibit is visible in the same context as the clause it affects.

In: full suite (up to 1M tokens). · Out: JSON of all clauses with type, parties, position, extracted text.

02 **Clause classification & risk tagging**

OPENAI
GPT-4.1 Nano
$0.10 / $0.40 per MTok

Extraction is complete. Stage 2 is pure classification against a risk taxonomy (liability caps, IP ownership, data processing, termination, indemnification) with high/medium/low tags. A well-defined structured task at the cheapest available rate. The premium context has already been consumed in Stage 1.

03 **Redline generation & deal memo**

ANTHROPIC
Claude Sonnet 5
$3.00 / $15.00 per MTok

The premium spend goes where it matters: drafting the redlines and the deal memo a lawyer will actually read. Writing quality, legal reasoning, and the ability to hold nuance across a long output are the point of the tier.

§ 05 · PIPELINE 03 REAL-TIME TRIAGE ROUTING

Customer support with escalation routing.

Every message hits the cheapest classifier first. Premium reasoning activates only on escalation, keeping effective cost per ticket close to the Nano rate.

Most AI support pipelines apply the same model to every ticket. A complex escalation and a simple FAQ get identical treatment: expensive and slow. Adding a sub-200ms classification layer routes each message to the right tier before any response is generated. Effective cost per ticket then reflects the actual distribution of complexity, not the worst case.

ARCHITECTURE

STAGE 01

Real-time intent & sentiment classification

GOOGLE GEMINI

Gemini 2.5 Flash-Lite

$0.10 / $0.40 per MTok

→

STAGE 02

Tier-1 response generation

OPENAI

GPT-5.4 Nano

$0.20 / $1.25 per MTok

→

STAGE 03

Complex escalation resolution

ANTHROPIC

Claude Sonnet 5

$3.00 / $15.00 per MTok

COST MATH

§ PIPELINE EFFECTIVE COST / MESSAGE

~$0.615 per message

Per 1,000 support messages:
1,000 × $0.10 first-look → $100
~700 routine replies × $0.20 → $140
~125 escalations × $3.00 → $375
Total $615 → $0.615 avg / message

§ CLAUDE SONNET ON EVERYTHING

~$3.00+ per message

Roughly 5× cost reduction, while improving routing quality for the complex cases that need it.

§ WHEN TO USE THIS PATTERN

High ticket volume where the majority of queries are routine but a meaningful minority require genuine reasoning. Stage 1 classification accuracy is the critical investment: if routing precision is low, escalation leakage inflates Stage 3 costs and erodes the advantage.

§ 06 IMPLEMENTATION NOTES

The plumbing under the pipeline.

Each pipeline requires an orchestration layer to pass context between stages, handle errors at each provider, and enforce stage-gating logic. The following notes apply to all three pipelines.

Orchestration

Each pipeline maps directly to an n8n workflow. Stage 1 is an HTTP Request node hitting the Gemini API; its output feeds a conditional branch (signal threshold, risk tag, intent classification); qualifying outputs route to Stage 2 and Stage 3 nodes. Stage-gating logic lives in the branch node, not in the model prompts.

Context passing

The structured output from each stage becomes part of the next stage's input prompt. Use JSON for all inter-stage context, not free-text summaries, so each stage can reference specific fields (`signal_score`, `clause_type`, `intent_tag`) rather than re-parsing prose.

Error handling

Each provider's API can fail independently. Wrap each stage in try/catch with exponential backoff. If Stage 1 fails, do not route to Stage 2: fail the pipeline and retry from the beginning to avoid partial outputs reaching the premium generation stage.

Provider API keys

Each provider requires its own API key and rate-limit management. Do not share rate-limit budgets across pipelines. Set per-pipeline daily spend caps in each provider's dashboard and alert before limits are hit.

§ THE TAKEAWAY

The cheapest model is rarely the right answer, and neither is the best one. The right answer is a pipeline where each stage runs on the model that fits that stage, and the premium spend concentrates where it moves the outcome. That is the shape of production AI economics for the next few years.