

Prompting, Cost Effective Use of AI, & Model Choice

Sanjeev Pulapaka

Generative AI lead Public Sector

Amazon Web Services

Sujatha Dantuluri

Sr Solutions Architect

Amazon Web Services

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

WHO THIS IS FOR

Your bill grew faster than your *traffic.*

— and you don't know which lever to pull.

FIVE SYMPTOMS

Bill outpaces traffic

Output tokens are 3-5× input. 20% user growth → 50%+ bill growth.

Agent feels slow, root cause unclear

TTFT, TPS, tool-loop depth — decompose before you optimize.

No quality regression gate

One accidental prompt edit silently degrades quality for a week.

Caching / routing — when and how?

Each technique has a break-even, a failure mode, a measurement.

Prototype → production

Notebook code doesn't survive 50k req/month.



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

TIER · LOW EFFORT

01 One-hour changes. Quick wins.



01

Prompt design

Clarity beats cleverness. Headers, structure, no hedging — 20-30% fewer tokens.

02

Prompt tricks

Chain-of-Draft, Verbalized Sampling, Self-Refine — one-line changes.

03

Prompt caching

Up to 90% off on repeated static content (system, tools, RAG docs).

04

Model selection

Match model class to task. Start small, escalate on eval evidence.

05

Parameter tuning

max_tokens, temperature, stop_sequences, structured output.

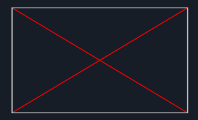
06

Adaptive thinking

Claude 4.6/4.7: set thinking:adaptive, model picks budget.



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.



“Prompt engineering is the practice of designing inputs for generative AI tools that will produce optimal outputs.”

McKinsey & Company

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

An introduction to prompt engineering

- What is **prompt engineering**?
  - ✔ Using natural language processing techniques to craft prompts that steer Foundation Models (FMs) or Large Language Models (LLMs) towards desired responses

- Why it is **important**
  - ✔ Enables fine-grained and strategic control over models' behavior
  - ✔ Targets desired capabilities
  - ✔ Mitigates risks



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

(*Typical*) prompt structure



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

PROMPT DESIGN

Same intent. Half the tokens.

BEFORE · VAGUE

~85 tokens

```
Could you please help me by telling
me about the return policy? I was
wondering if you could maybe explain
it in detail if that's okay. Please
be polite and thorough. Make it
engaging. Thanks!

[product category]: laptop
```

→ hedging, filler, no structure

AFTER · STRUCTURED

~38 tokens · −55%

```
# TASK
Return policy for {category}.

# FORMAT
2-3 sentences. Plain text.

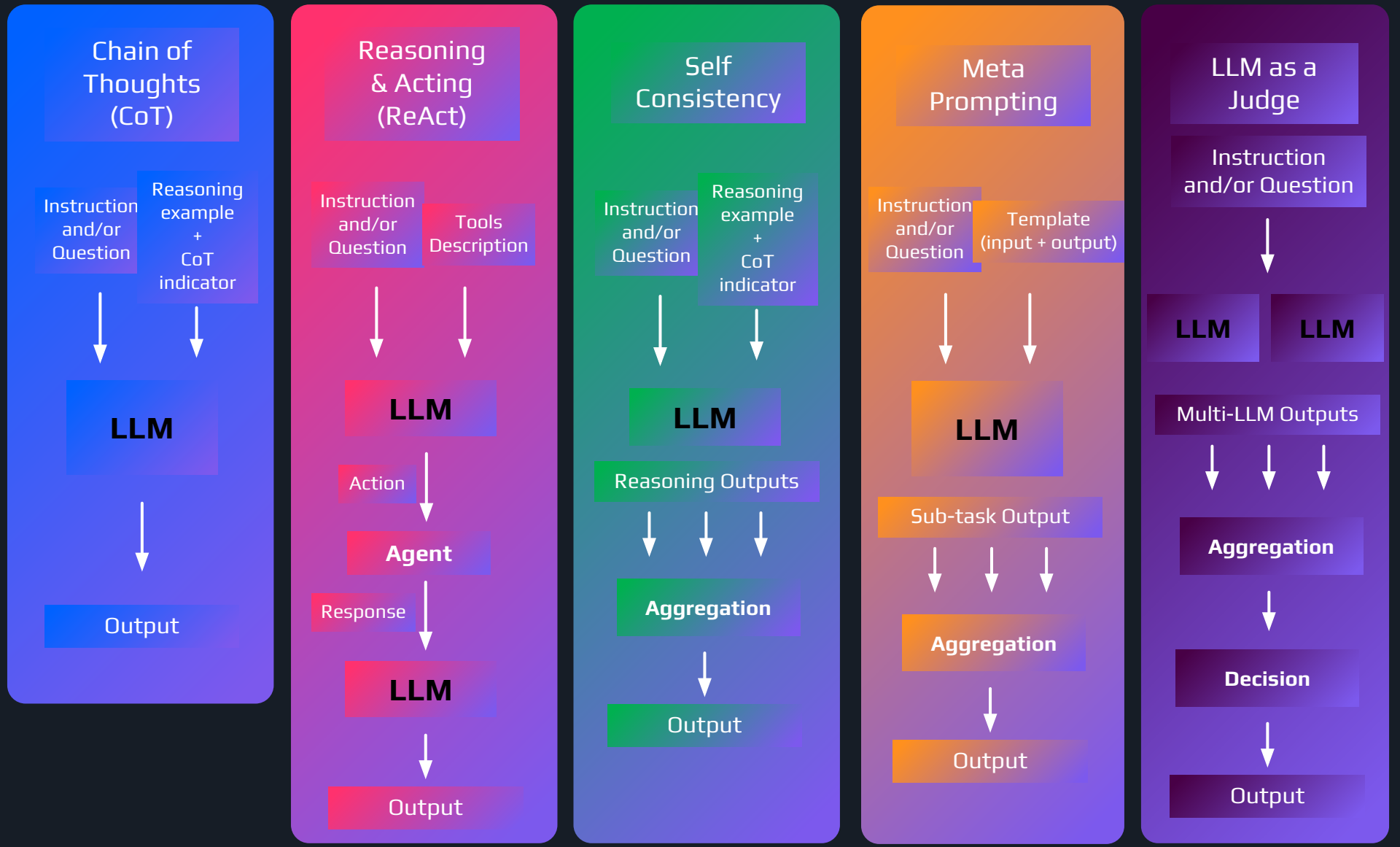
# INPUT
category: laptop
```

→ directive, sectioned, parsable



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Prompt engineering techniques



Among others like:

- Generate Knowledge Prompting
- Automatic Prompt Engineer
- Active-Prompt
- Directional Stimulus Prompting
- Augmented Language Models
- Multi-Modal CoT
- Graph Prompting

Plus all the possible combinations, e.g.:

- Zero-shot + CoT
- Few-shot + CoT
- Few-shot +ReAct
- etc.

Plus other factors like:

- Aggregations
- Chaining
- etc.



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Quick wins on prompt engineering

CHAIN-OF-DRAFT

CoT on a diet

"Limit each reasoning step to 5 words." 80-92% token reduction at similar accuracy.

Xu et al · Zoom · Feb 2025

VERBALIZED SAMPLING

Real diversity

"Emit N responses with explicit probabilities." Beats temperature alone, fixes mode collapse.

Zhang et al · Oct 2025

SELF-REFINE

Critique before reply

"First critique your draft, then revise." Largely subsumed by adaptive thinking; still a fine prompt-only fallback.

Madaan et al · 2023

CHAIN-OF-VERIFICATION

Verify the facts

Generate → list verification questions → answer each → revise. 4-5× single-call cost; reach for fact-critical work.

Dhuliawala et al · 2023

DEFAULT STANCE

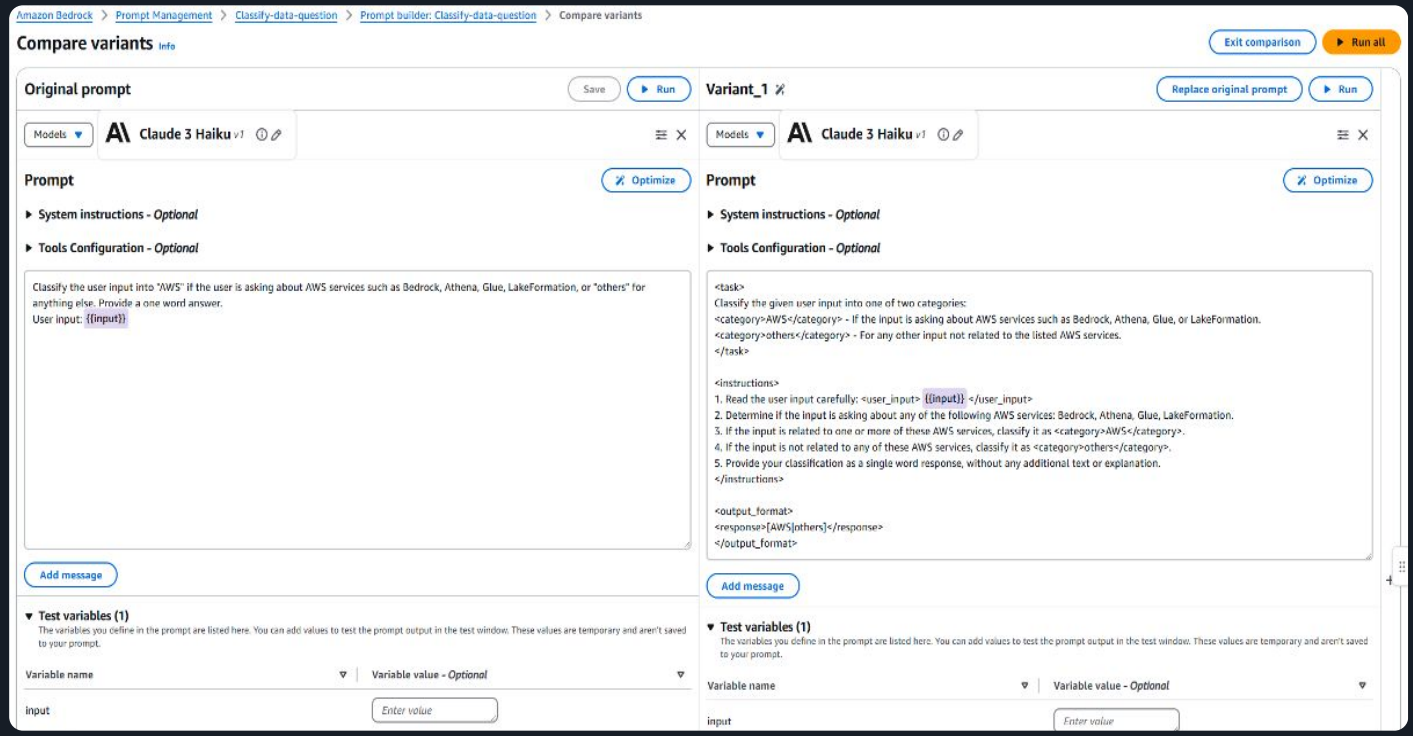
Reach for these only when the base prompt + caching + routing aren't enough. They're tactics, not architecture.



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock Prompt Optimization

Improve up to 35% in model accuracy, and reduce up to 63% in output tokens



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

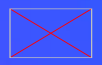
Best practices for prompt engineering

- Prompt engineering – Context Objective Style Tone Audience Response
- Prompt templates
- Chaining prompts for complex tasks
- Dynamic prompt generation based on content type
- A/B testing different prompt versions
- Incorporating user feedback for continuous improvement
- Implement system prompts



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Select the right model



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Many requests don't need the workhorse.

DIY · PATTERNS

01 · Static rules	Token count, language ID, code-fence presence, JSON-mode flag → bucket → model. Sub-ms. Built into LiteLLM, Portkey.
02 · LLM Classifier	Cheap LLM reads the prompt and emits an intent or complexity label → map to model. ~100–300ms. Easiest; quality depends on classifier prompt.
03 · Cascade (FrugalGPT) *(Chen et al., 2023)*	Cheapest model first; quality scorer escalates on low confidence. Best-case = small-model only; worst-case = sum of all stages.
04 · Preference-trained classifier (RouteLLM) *(Ong et al., LMSYS, 2024)*	Learn P(strong wins | query) from human preference data. ~95% GPT-4 quality at 26% cost on MT-Bench. OSS, MIT.
05 · Generative router (Arch-Router / Router-R1) github.com/ulab-uiuc/Router-R1	Small LLM emits route token from natural-language model descriptors. Add new models with no retraining. SOTA on multi-turn agents.

MANAGED · BEDROCK INTELLIGENT PROMPT ROUTING

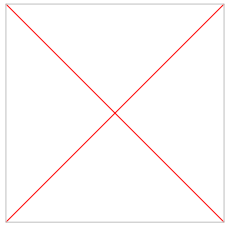
AWS routes within a model family based on prompt complexity. Up to 30% cost reduction with quality maintained.

- \+ Zero infra, transparent fallback
- \+ Single endpoint, single billing
- − Routing logic is opaque — verify on your eval
- − Only supports specific models (Nova, Claude 3/3.5 Haiku, Meta)

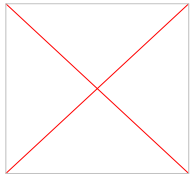


© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

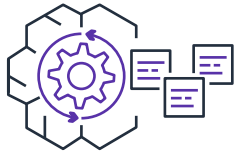
Model choice and evaluation are essential



Discover industry-leading, emerging, and domain-specific FMs



Evaluate models based on specific use case requirements



Implement systematic testing frameworks



Design architectures that allow for model switching



Leverage multiple models

24 +

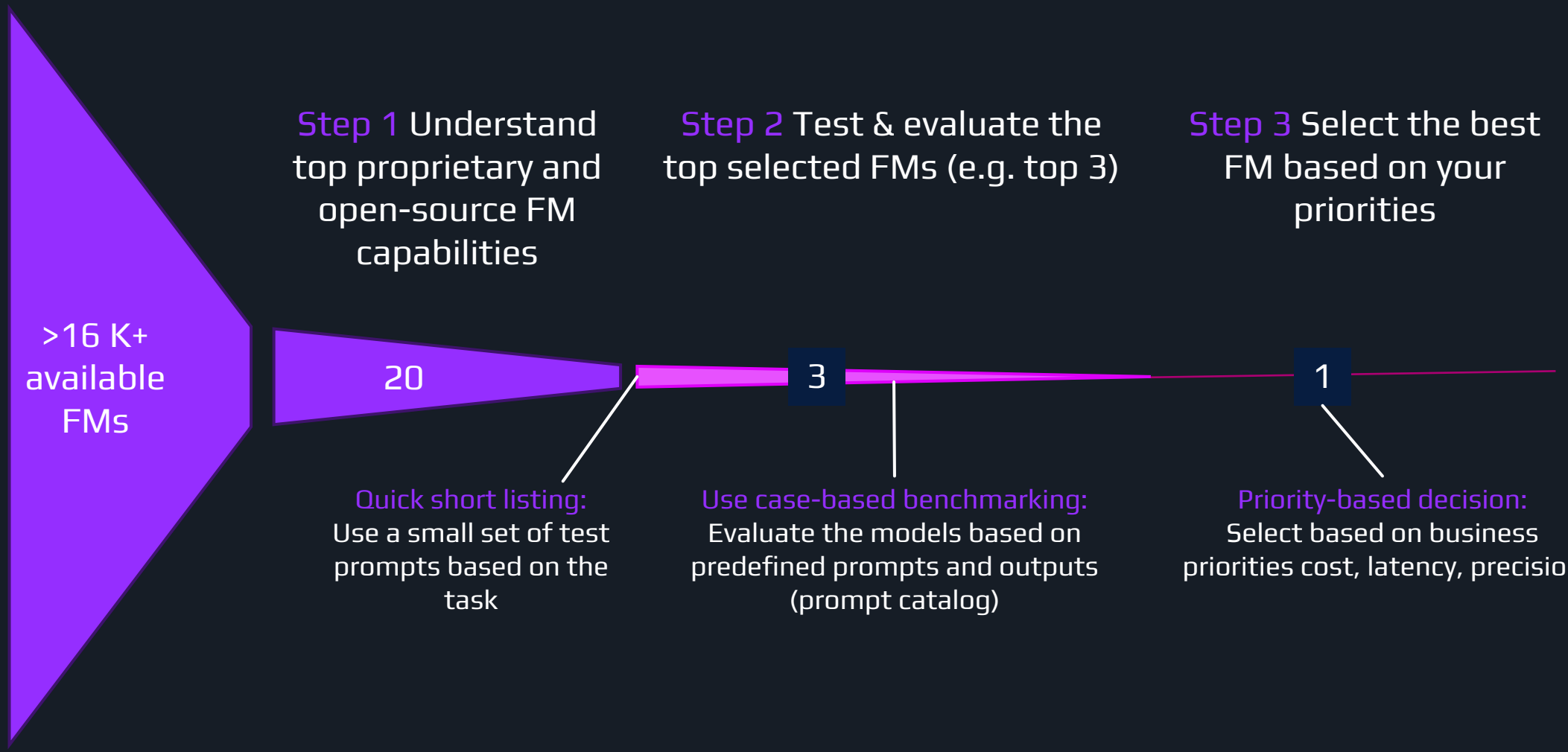
Gen AI models

By 2028, the majority of global enterprises will have more than two dozen gen AI model implementations from more than three major vendors plus additional smaller vendors, up from 10% today.



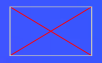
© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Foundation model selection



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Parameter Tunning



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

PARAMETER TUNING

Four config values worth getting right.

01

max_tokens

Cap at expected + small buffer. Frees TPM headroom, prevents runaway responses.

02

temperature

0 for extraction / classification. 0.3-0.7 for generation. Don't ship the default.

03

stop_sequences

5-15% faster TTLT. List 1-3 known terminators. Free latency win.

04

Thinking (extended/adaptive)

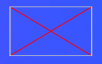
Let the model spend extra tokens reasoning before answering.

- Anthropic 4.6+, set effort: low | medium | high and the model picks the budget.
- Nova 2 Lite uses reasoningConfig.



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

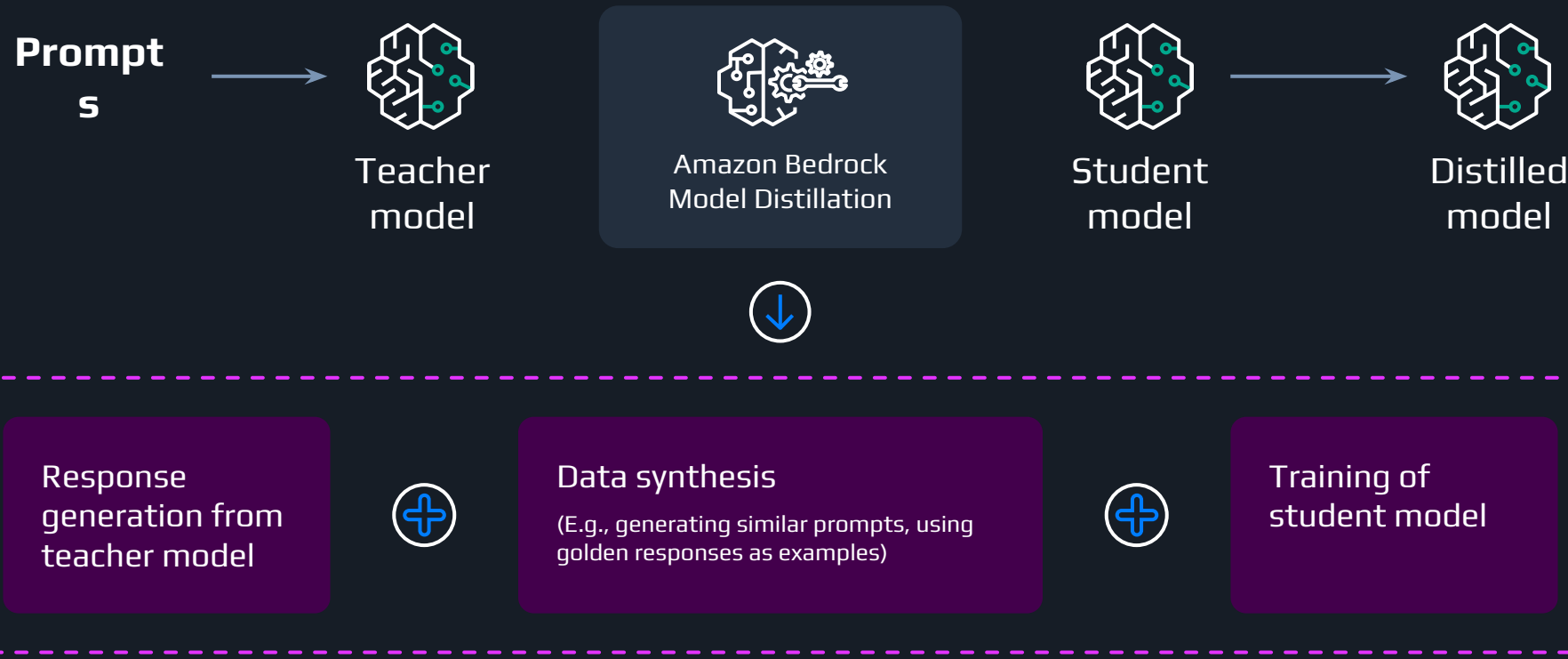
Optimizing AI & Token Usage



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Model distillation

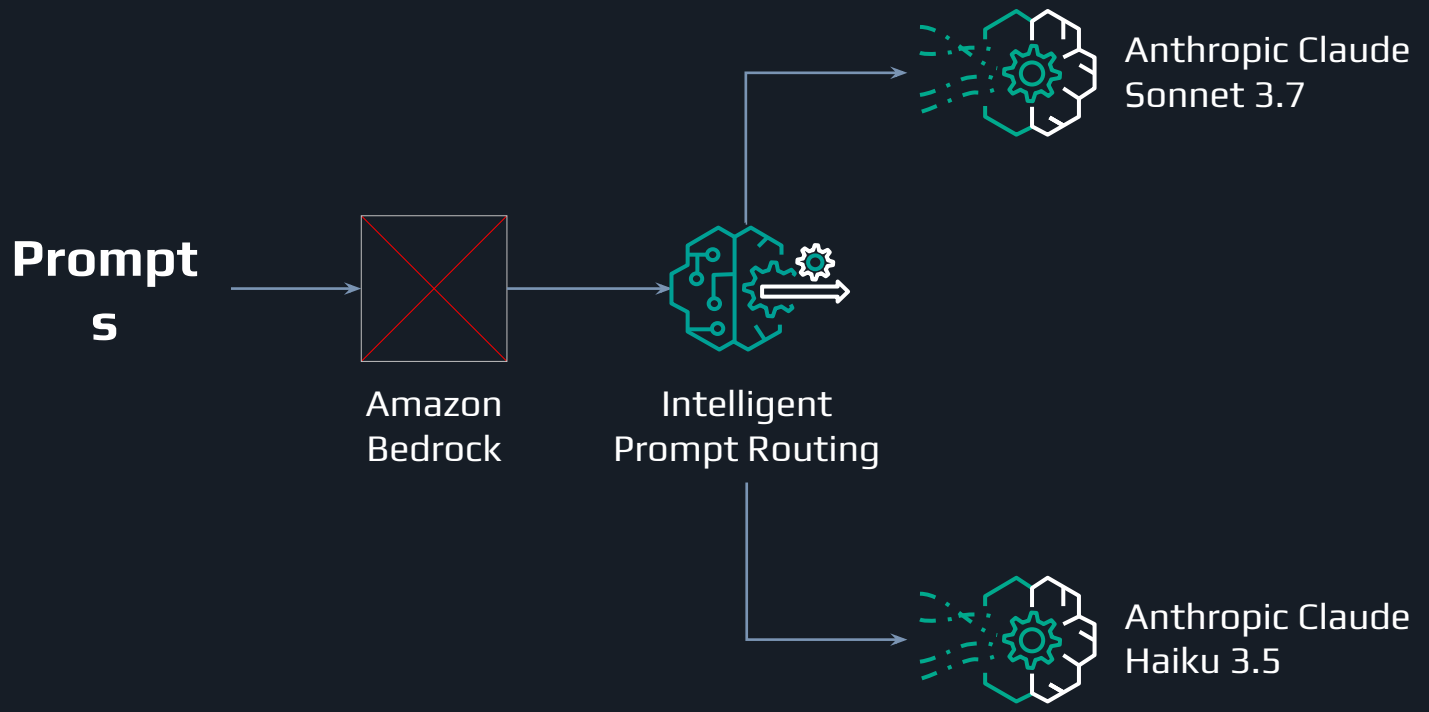
Teacher model accuracy with a smaller student model: 500% faster and 75% less expensive



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Intelligent Prompt Routing

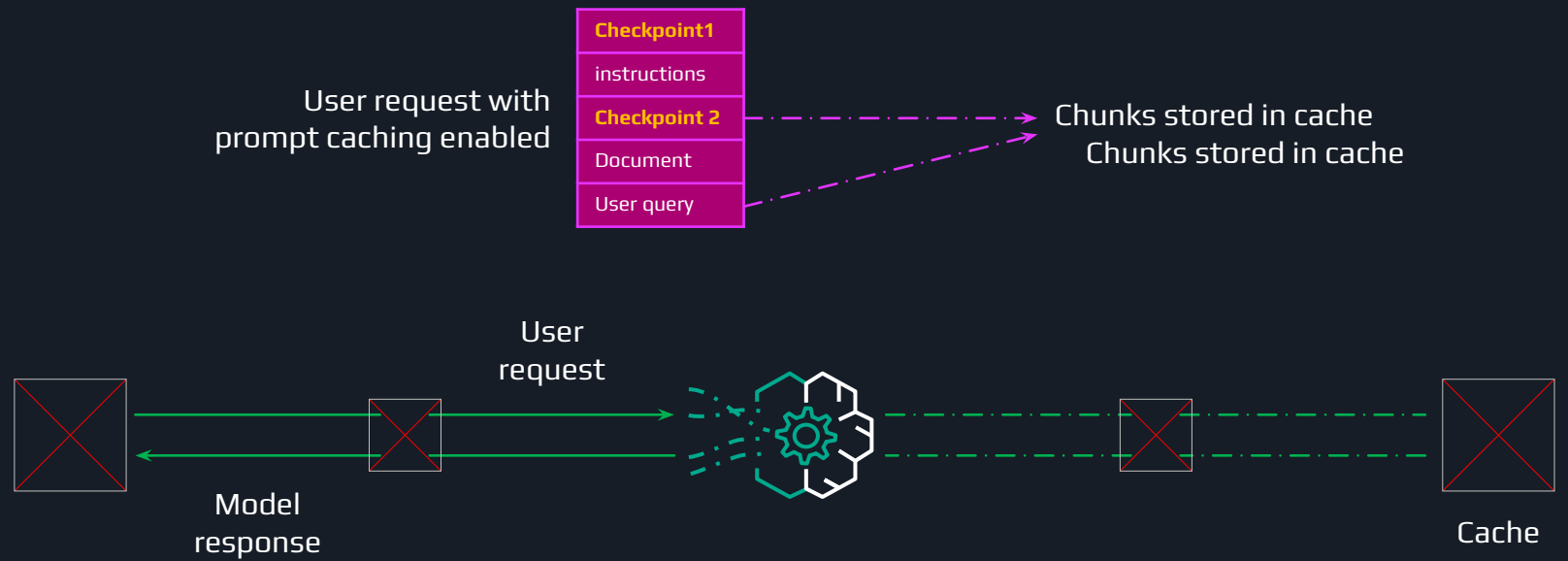
30–40% cost reduction



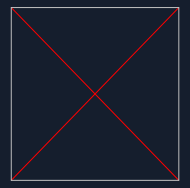
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Prompt Caching

Cache repetitive context in prompts across multiple API calls—reduce cost by up to 90%



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Amazon S3 Vectors

First cloud object store with native support to store and query vectors

Sub-second

vector query performance at scale

Up to 90%

lower overall costs for uploading, storing, and querying vectors

Billions

of vectors stored at low cost



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Batch Inference

Run multiple inference requests asynchronously with much lower cost



- **Efficiently run model inference on large volumes of data while avoid throttling**
- **50% lower price**
- **Well suited for use cases to process substantial volumes of data**
- **Enhanced security and operational features**

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Thank you!

Food Safety Prompt Patterns

Ready-to-use prompt templates you can adapt today:

• Outbreak Investigation
"Given the following epidemiological data [data], identify the most likely source using criteria: temporal clustering, geographic pattern, product commonality, and WGS phylogenetic clustering..."

• HACCP Plan Review
"Review this HACCP plan against 21 CFR Part 120. For each CCP, evaluate whether critical limits are measurable, monitoring frequency is adequate, and corrective actions are specific..."

• Regulatory Gap Analysis
"Compare our current SOP against the FSMA Preventive Controls rule. Identify gaps in: hazard analysis, preventive controls, monitoring, corrective actions, and verification activities..."

• Literature Synthesis
"Summarize the last 5 years of peer-reviewed research on [pathogen] contamination in [product]. Focus on emerging detection methods and intervention strategies..."



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.