Amazon Bedrock Pricing - Simple Guide

Amazon Bedrock offers AI models, but how much does it cost? Let's break it down in simple terms.

Three Ways to Pay for Amazon Bedrock

Think of these like different phone plans - each works differently depending on your needs.

1. On-Demand (Pay-As-You-Go)

What it is: Like a prepaid phone - you only pay for what you use, no contracts.

How you're charged:

- Text models: Every word you send in + every word you get back
- Image models: Every image created
- Embedding models: Every word you send in

Best for: When you don't know how much you'll use or want flexibility.

2. Batch Mode (Bulk Processing)

What it is: Like ordering in bulk from Costco - process many requests together for a discount.

How it works:

- Send many requests at once
- Get all results in one file (stored in Amazon S3)
- Wait a bit longer for results
- Save up to 50% on costs

Best for: When you can wait and want to save money.

3. Provisioned Throughput (Reserved Capacity)

What it is: Like renting a dedicated server - you pay upfront for guaranteed performance.

How it works:

- Buy capacity for 1 month, 6 months, etc.

- Get guaranteed speed (tokens per minute)
- Usually costs more, but performance is guaranteed

Best for: When you need consistent, fast performance and have custom models.

Different Ways to Improve Your AI Model

From cheapest to most expensive:

1. Prompt Engineering 💰 (Cheapest)

What it is: Writing better instructions to get better results. **Cost:** Almost free - no extra training needed. **Example:** Instead of "Write something," write "Write a professional email to a customer about a delayed order."

2. RAG (Adding External Knowledge) 💰💰

What it is: Connecting your model to external information (like a database). **Cost:** Need to pay for the database and connection system. **Example:** Connecting your AI to your company's product catalog.

3. Instruction-Based Fine-Tuning 💰💰💰

What it is: Teaching the model specific ways to respond. **Cost:** Requires some training computation. **Example:** Training the model to always respond in a formal business tone.

4. Domain Fine-Tuning 💰💰💰💰 (Most Expensive)

What it is: Completely retraining the model on your specific data. **Cost:** Very expensive - lots of computation needed. **Example:** Training a model specifically for medical diagnoses using hospital data.

How to Save Money

Smart Choices:

- **Use On-Demand** if you're not sure about usage
- **Use Batch Mode** if you can wait for results (50% savings!)
- **Choose smaller models** when possible (usually cheaper)

The #1 Money-Saving Tip:

Watch your words! The main cost driver is the number of words (tokens) you use.

To save money:

- Write shorter, clearer prompts
- Ask for shorter responses when possible
- Be specific about what you want

Things That DON'T Affect Cost:

- Changing temperature, Top K, or Top P settings
- These adjust how creative/focused the AI is, but don't change the price

Quick Decision Guide

Choose On-Demand if: You're testing, learning, or have unpredictable usage.

Choose Batch Mode if: You have lots of data to process and can wait for results.

Choose Provisioned Throughput if: You have custom models or need guaranteed fast performance.

Remember: Start simple with On-Demand and prompt engineering, then upgrade as needed!