

16. Tokens on a Budget

You have a budget of 50 cents for one call. The model is priced at 2 cents per 100 tokens. How many tokens can the call use?

Step 1: How many 100-token blocks the budget buys

$$50/2 = 25 \text{ blocks}$$

Step 2: Convert the blocks to tokens

$$25 \times 100 = 2,500 \text{ tokens}$$

Practice 1

You have a budget of 90 cents for one call. The model is priced at 3 cents per 100 tokens. How many tokens can the call use?

Step 1: How many 100-token blocks the budget buys

Step 2: Convert the blocks to tokens

Practice 2

You have a budget of 60 cents for one call. The model is priced at 4 cents per 100 tokens. How many tokens can the call use?

Step 1: How many 100-token blocks the budget buys

Step 2: Convert the blocks to tokens

17. How Long Will It Take?

A call sends 2,000 input tokens and gets back 200 output tokens. The model reads input at 1,000 tokens per second and writes output at 40 tokens per second. How long does the call take?

Step 1: Time to read the input

$$2,000 / 1,000 = 2 \text{ seconds}$$

Step 2: Time to write the output

$$200 / 40 = 5 \text{ seconds}$$

Step 3: Add the two times

$$2 + 5 = 7 \text{ seconds}$$

Practice 1

A call sends 3,000 input tokens and gets back 300 output tokens. The model reads input at 1,000 tokens per second and writes output at 50 tokens per second. How long does the call take?

Step 1: Time to read the input

Step 2: Time to write the output

Step 3: Add the two times

17. How Long Will It Take? (continued)

Practice 2

A call sends 4,000 input tokens and gets back 120 output tokens. The model reads input at 2,000 tokens per second and writes output at 30 tokens per second. How long does the call take?

Step 1: Time to read the input

Step 2: Time to write the output

Step 3: Add the two times

18. The Split Matters

A model reads input at 500 tokens per second and writes output at 50 tokens per second. Task A summarizes a long document: 2,000 input tokens, 500 output tokens. Task B expands a short outline into a full article: 500 input tokens, 2,000 output tokens. Both tasks process 2,500 tokens in total. Which finishes first?

Step 1: Time for task A

$$2,000/500 + 500/50 = 4 + 10 = 14 \text{ seconds}$$

Step 2: Time for task B

$$500/500 + 2,000/50 = 1 + 40 = 41 \text{ seconds}$$

Step 3: Compare

$$14 < 41: \text{task A finishes first.}$$

Practice 1

A model reads input at 500 tokens per second and writes output at 50 tokens per second. Task A answers a question about a long report: 2,500 input tokens, 500 output tokens. Task B drafts a detailed report from a short brief: 500 input tokens, 2,500 output tokens. Both tasks process 3,000 tokens in total. Which finishes first?

Step 1: Time for task A

Step 2: Time for task B

Step 3: Compare

18. The Split Matters (continued)

Practice 2

A model reads input at 500 tokens per second and writes output at 50 tokens per second. Task A classifies a batch of messages: 1,500 input tokens, 500 output tokens. Task B writes replies from short descriptions: 500 input tokens, 1,500 output tokens. Both tasks process 2,000 tokens in total. Which finishes first?

Step 1: Time for task A

Step 2: Time for task B

Step 3: Compare

19. Pricing Per Million Tokens

A model charges \$3 per million input tokens and \$15 per million output tokens. A task processes 2,000,000 input tokens and 1,000,000 output tokens. What does the task cost?

Step 1: Cost of the input tokens

$$2,000,000 / 1,000,000 \times \$3 = 2 \times \$3 = \$6$$

Step 2: Cost of the output tokens

$$1,000,000 / 1,000,000 \times \$15 = 1 \times \$15 = \$15$$

Step 3: Add for the total cost

$$\$6 + \$15 = \$21$$

Practice 1

A model charges \$4 per million input tokens and \$12 per million output tokens. A task processes 3,000,000 input tokens and 2,000,000 output tokens. What does the task cost?

Step 1: Cost of the input tokens

Step 2: Cost of the output tokens

Step 3: Add for the total cost

19. Pricing Per Million Tokens (continued)

Practice 2

A model charges \$2 per million input tokens and \$10 per million output tokens. A task processes 4,000,000 input tokens and 1,000,000 output tokens. What does the task cost?

Step 1: Cost of the input tokens

Step 2: Cost of the output tokens

Step 3: Add for the total cost

20. Tokens From the Bill

An agent task's bill shows \$12 for input and \$30 for output. Input is priced at \$3 per million tokens, output at \$15 per million. How many tokens did the task process in total?

Step 1: Input tokens from the input charge

$$12/3 \times 1,000,000 = 4,000,000 \text{ input tokens}$$

Step 2: Output tokens from the output charge

$$30/15 \times 1,000,000 = 2,000,000 \text{ output tokens}$$

Step 3: Add for the total tokens

$$4,000,000 + 2,000,000 = 6,000,000 \text{ tokens}$$

Practice 1

An agent task's bill shows \$35 for input and \$45 for output. Input is priced at \$5 per million tokens, output at \$15 per million. How many tokens did the task process in total?

Step 1: Input tokens from the input charge

Step 2: Output tokens from the output charge

Step 3: Add for the total tokens

20. Tokens From the Bill (continued)

Practice 2

An agent task's bill shows \$6 for input and \$50 for output. Input is priced at \$2 per million tokens, output at \$10 per million. How many tokens did the task process in total?

Step 1: Input tokens from the input charge

Step 2: Output tokens from the output charge

Step 3: Add for the total tokens