

20. Cost Forecast

Budgeting a Long Session

An AI agent has a 1,000-token system prompt, and each turn is a 200-token question and a 1,000-token answer, so k is 1,200. Input costs \$2 per million tokens, output \$10 per million. Over a 100-turn session, find what the model processes, price it, then choose a budget for the session.

Step 1: Use the equation from problem 19 for the total input

$$1,200 \times (100 \times 101) / 2 = 6,060,000 \text{ input tokens}$$

Step 2: The output is the same every turn, so it only multiplies

$$1,000 \times 100 = 100,000 \text{ output tokens}$$

Step 3: Price both halves and add

$$(6.06 \times \$2) + (0.1 \times \$10) = \$12.12 + \$1.00 = \$13.12$$

Step 4: Take the smallest budget that covers it

- ☐ \$5
 ☐ \$10
 ☒ \$15
 ☐ \$25

20. Cost Forecast (continued)

Practice 1

An AI agent has an 800-token system prompt, and each turn is a 200-token question and an 800-token answer, so k is 1,000. Input costs \$2 per million tokens, output \$10 per million. Over a 50-turn session, find what the model processes, price it, then choose a budget for the session.

Step 1: Use the equation from problem 19 for the total input

Step 2: The output is the same every turn, so it only multiplies

Step 3: Price both halves and add

Step 4: Take the smallest budget that covers it

☐ \$5 ☐ \$10 ☐ \$15 ☐ \$25

Practice 2

An AI agent has a 1,500-token system prompt, and each turn is a 500-token question and a 1,500-token answer, so k is 2,000. Input costs \$2 per million tokens, output \$10 per million. Over a 100-turn session, find what the model processes, price it, then choose a budget for the session.

Step 1: Use the equation from problem 19 for the total input

Step 2: The output is the same every turn, so it only multiplies

Step 3: Price both halves and add

Step 4: Take the smallest budget that covers it

☐ \$5 ☐ \$10 ☐ \$15 ☐ \$25