

1.2.15 Live Lab: Prompt Design and Optimization

Jeffrey Jordan (ghostinthemaking@gmail.com)

Date: 11:14:01 PM 02/24/2026

Time Spent: 01:24:51

Score: 100%

Passing Score: 80%



Task: Create storage and search service

✓ Correct

Task: Deploy gpt-4.1-mini model

✓ Correct

Question 1

✓ Correct

In the chat runtime, raising the strictness setting mainly...

- ☐ Expands results to include looser matches
- ☒ Tightens filtering so only high-similarity documents are used ✓ Correct
- ☐ Changes sampling temperature
- ☐ Increases max tokens

Question 2

✓ Correct

Which parameter primarily reduces verbatim repetition?

- ☒ frequency_penalty ✓ Correct
- ☐ presence_penalty
- ☐ max_tokens
- ☐ temperature

Question 3

✓ Correct

Which metric represents the total time from request sent until the final token/byte is received?

- ☐ Time to First Token
- ☐ Tokens per second
- ☒ Time to Last Byte ✓ Correct
- ☐ Prompt tokens

Question 4

✓ Correct

What's a common trade-off when switching to a smaller model?

- ☒ Lower latency and cost; weaker reasoning ✓ Correct
- ☐ Higher latency and cost; stronger reasoning
- ☐ Same quality at lower cost
- ☐ Automatic decrease in max tokens

