Teacher Answer Key

Case Study

1. Every AI model has a context window, the maximum number of tokens it can hold in working memory at once. A 400-page document contains far more tokens than most standard context windows can accommodate, so the model returned an error rather than processing incomplete information. The model literally could not hold the entire document in its active processing space at one time.

2. Users should research the specific capabilities and limitations of an AI tool before using it for a task, rather than assuming any AI can handle anything. Matching the right tool to the task, whether that means choosing a model with a longer context window, using a RAG-based document tool, or splitting the task into appropriate chunks, produces much better results than using a general tool incorrectly.

Multiple Choice

1. A
2. C
3. D

Matching

1-B, 2-E, 3-D, 4-C, 5-A

Fill-in-the-Blank

1) context, 2) DeepSeek, 3) scaling

Short Answer (Sample Responses)

Q1: Different tasks require different trade-offs between capability, speed, and cost. A small, fast model handles routine tasks like quick summaries or simple questions efficiently and cheaply, while a large, powerful model handles complex reasoning or nuanced tasks that require deeper capability. Offering a family of sizes lets users and developers match the right tool to the task rather than always paying for maximum power.

Q2: The Chinchilla research showed that raw parameter count is not the only factor in model performance; training data quantity matters equally or more. A 175-billion-parameter model trained on insufficient data will underfit, leaving much of its capacity unused. A smaller 70-billion-parameter model trained on four times more

data develops more robust and generalizable representations because it sees far more of the diversity of human language.

Q3: The context window limits how much text a model can actively consider at once. When a document exceeds the context window, the model cannot process it as a whole, forcing users to split it into sections where the model may miss connections between earlier and later content. In long conversations, exceeding the context window means the model loses access to earlier parts of the conversation, causing it to seem forgetful or inconsistent.