Teacher Answer Key

Case Study

1. A lie requires the speaker to know the truth and deliberately say something false. Hallucination, by contrast, occurs when the AI has no internal awareness that it is producing inaccurate information — it generates statistically plausible text without any mechanism to verify whether that text is factually correct. The AI is not deceiving; it is failing to distinguish between accurate retrieval and confabulation.

2. AI models generate outputs based on statistical patterns learned during training — they produce text that would be plausible in the context of the prompt, whether or not that text corresponds to reality. Because the model does not have a separate 'fact-checking' module, it cannot identify when it is generating invented content. This means users must independently verify any specific factual claims, citations, or statistics produced by AI, rather than relying on the AI to flag its own errors.

Multiple Choice

1. D

2. A

3. D

Matching

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

Fill-in-the-Blank

1) token, 2) context, 3) Act

Short Answer (Sample Responses)

Q1: When an AI 'understands' a question, it converts the input into numerical embeddings, processes them through billions of parameters that were adjusted during training to recognize patterns, and generates the most statistically likely response given those patterns. There is no comprehension or meaning in the human sense — it is a mathematical operation over learned statistical correlations between words and concepts.

Q2: Bias comes from inequities or distortions in the training data — if the data reflects historical social inequalities, the model learns to replicate them. Hallucination comes from the model's statistical generation process itself — it produces plausible-sounding text regardless of factual accuracy. Bias is a training

data problem; hallucination is an architectural property of how language models generate outputs. They have different causes and require different interventions.

Q3: Jailbreaking works through novel, creative prompt construction that safety guidelines have not yet encountered. Since there are infinite possible phrasings and framings a user could try, developers cannot write rules for every one in advance. Safety guidelines are trained patterns, not perfect logic rules, and new jailbreaks can find gaps in those patterns. This reveals that AI safety is probabilistic, not absolute — systems are trained to be unlikely to produce harmful outputs, not mathematically guaranteed to never do so.