Teacher Answer Key

Case Study

1. Amara's experience illustrates the over-restriction problem — AI content guidelines applied so broadly that they block legitimate educational and research uses. The reading notes that content policies sometimes cannot distinguish between studying a harmful phenomenon and enabling it, and that the voices most likely to be silenced are those whose legitimate speech touches on topics flagged as sensitive.

2. The reading's claim is partially supported but complicated. Amara is not among the most marginalized users — she has a teacher, access to academic papers, and the persistence to find alternative sources. The reading's point is most acute for users who lack those resources. Amara's case illustrates that even relatively privileged users are affected by over-restriction, but the harm falls hardest on those with fewer fallback options.

Multiple Choice

1. D
2. C
3. D

Matching

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

Fill-in-the-Blank

1) feedback, 2) United, 3) open

Short Answer (Sample Responses)

Q1: RLHF relies on human contractors, often based in countries like Kenya, the Philippines, or Pakistan, who are paid relatively low wages to review and label large volumes of content — including graphic violence, hate speech, and child sexual abuse material. Labor advocates have criticized the exposure of these workers to disturbing content without adequate psychological support or fair compensation, raising ethical concerns about how the AI safety infrastructure is built.

Q2: AI systems trained predominantly on English-language content and calibrated to Western cultural norms carry those norms with them when deployed in culturally different contexts. What counts as appropriate speech, respectful religious discussion, acceptable political commentary, or dangerous content varies enormously by culture. A system calibrated for one context can produce outputs that are

offensive or harmful in another, and communities in the Global South typically have little to no input into the guidelines that affect them.

Q3: Applied too loosely, AI can generate harassment, disinformation, and abusive content — real harms that have real victims. Applied too aggressively, AI blocks legitimate speech: a mental health researcher cannot get information about suicide, a student studying propaganda cannot research how it works, communities discussing sensitive cultural topics find their speech flagged as dangerous. The challenge is that the same policy cannot be both protective and permissive — every rule that prevents one harm also risks blocking legitimate uses.