AI Limitations and Biases

Assessment Questions

Macy Foundation / Millennium Conference • Preclinical M1/M2 • November 2026

Eight NBME-style multiple-choice items plus two short-answer items, with answer key, explanations, and scoring guides. Suggested use: pre/post administration for the Level 2 evaluation (see the Alignment Matrix & Evaluation Plan).

Question 1

Objective 1 — verification must extend to content, not just existence

A resident drafts a teaching summary on anticoagulation reversal using her hospital's sanctioned, retrieval-enabled clinical LLM. The tool's output cites a real, current specialty-society guideline, and the citation resolves correctly. One sentence of the output attributes to that guideline a specific reversal-agent dosing recommendation. When the resident opens the guideline itself, it contains no such recommendation.

Which of the following BEST characterizes what has occurred?

A. A hallucination is excluded, because retrieval-enabled tools ground every generated statement in the documents they retrieve and therefore cannot misstate them.

B. An error of omission, because the resident failed to detect content the guideline contained.

C. A persuasive fabrication: the model attributed invented content to a real source, so verification must confirm what a source says, not merely that it exists.

D. Dataset shift: the guideline was updated after the model's training window, so the model's summary reflects an earlier version rather than a fabrication.

E. Automation bias on the part of the resident.

Question 2

Objective 1 — why self-confirmation is not verification

A student asks a general-purpose LLM whether a particular enzyme deficiency causes a metabolic finding. The model says yes. Unsure, the student replies, "Are you certain? Please double-check." The model reaffirms its answer and adds two supporting physiologic details. The student concludes the fact is now verified.

Which of the following BEST explains the flaw in the student's reasoning?

A. The model's reaffirmation shows the fact appeared frequently in its training data, and frequently repeated facts are reliable enough for preclinical use.

B. Asking the same question twice is a valid test–retest reliability check for a stochastic system, but the student needed to repeat the identical phrasing rather than paraphrase it, so the check was performed incorrectly rather than being conceptually flawed.

C. The flaw is minor: reaffirmation with added detail is meaningfully stronger evidence than reaffirmation alone.

D. The model generates the most plausible continuation of the conversation: having just asserted the claim, confirmation with supporting detail is the likeliest continuation, produced by the same process being questioned.

E. The student should have asked the model to rate its own confidence numerically before accepting the answer.

Question 3

Objective 2 — proxy bias with a superficially objective label

A health system builds a model to allocate scarce post-discharge follow-up calls. The model is trained to predict 30-day emergency-department (ED) return visits, using a training set that is demographically representative of the region and a model that is well calibrated overall. A validation committee member notes that patients facing transportation barriers and out-of-pocket cost pressures often do not return to the ED even when their condition worsens.

If the committee member is right, which of the following is the MOST important implication?

A. None — an ED return visit is an objectively recorded event, and a label that is objectively recorded cannot introduce systematic bias into a model that was trained on demographically representative data.

B. The label is a proxy for deterioration that fails differentially: patients with access barriers deteriorate without returning, so the model under-allocates calls to them.

C. The model is overfit and should be regularized more aggressively before deployment.

D. The concern describes dataset shift and is addressed by scheduled recalibration after deployment.

E. The concern is resolved because the training set was demographically representative of the region.

Question 4

Objective 2 — what equity evidence to demand before deployment

A vendor offers a deterioration-prediction model reporting an overall area under the receiver operating characteristic curve of 0.91, externally validated at three academic medical centers. Your hospital serves a patient population with a substantially different demographic, language, and insurance profile from those centers.

Which of the following additional pieces of evidence MOST directly addresses the equity risk of deploying this model?

A. Performance reported by demographic subgroup, from a validation population comparable to your hospital's.

B. A higher overall discrimination figure from a fourth academic medical center.

C. Documentation of the model's regulatory clearance pathway and predicate devices.

D. Evidence that the training set was larger than that of competing products, since larger training sets reduce all forms of bias.

E. An integrated explainability dashboard displaying the top features behind each individual prediction.

Question 5

Objective 3 — internal versus external validity at a new site

A 30-day readmission model is developed and internally validated at a quaternary academic center, with strong discrimination and calibration. A rural community hospital adopts it. Integration testing confirms all input variables map correctly, and clinicians use the tool as designed. Over the first year, discrimination is substantially worse than the published figures.

Which of the following is the MOST appropriate interpretation?

A. The published internal validation was most likely fraudulent, and the developer should be reported to the journal.

B. The performance drop reflects clinician documentation errors at the community hospital, since the model itself is unchanged.

C. The model has become overfit through repeated use in production and must be retrained on its original development data to restore performance.

D. The difference most likely reflects random year-to-year variation and requires no action beyond continued observation.

E. An expected generalization gap: internal validity does not establish performance at a different site, and local validation is required.

Question 6

Objective 3 — recognizing dataset shift from its causes

Two years after a kidney-injury prediction model is deployed with acceptable local performance, monitoring shows steadily worsening calibration. Over the same period the hospital switched its creatinine assay vendor and rolled out new admission documentation templates. Patient demographics and the model's code are unchanged, and clinicians' use of the tool is stable.

Which of the following BEST explains the deterioration?

A. Automation bias: clinicians have gradually stopped scrutinizing the model's predictions.

B. Computational degradation: prediction models gradually lose numerical precision when executed repeatedly in production, and slowly worsening calibration is the expected signature of that decay.

C. Dataset shift: the assay and documentation changes altered the model's inputs, so its learned relationships no longer hold.

D. Proxy bias introduced by the new documentation templates' demographic fields.

E. An adversarial attack on the model's input pipeline, which monitoring should be redesigned to detect.

Question 7

Objective 4 — a workflow that blunts errors of commission

On overnight shifts, a radiology resident reads studies with an AI triage overlay that marks suspected findings. Reviewing her quality data, she notices a pattern: when the overlay asserts a finding she did not independently suspect, she usually defers to it — and several of those deferrals proved wrong.

Which change to her workflow would MOST directly reduce these errors?

A. View the overlay only for high-acuity studies, where the pretest probability of a real finding is higher.

B. Record her own written interpretation before displaying the overlay, then investigate discrepancies deliberately.

C. Ask the tool to display a calibrated confidence score alongside each finding, and defer to the overlay only in cases where its stated confidence exceeds 90 percent.

D. Route studies flagged by the tool to a second AI system, deferring only when both systems agree on the finding.

E. Disable the overlay entirely until she completes fellowship training.

Question 8

Objective 5 — where responsibility and liability exposure actually sit

For a patient with atrial fibrillation and a high stroke-risk score, a clinical AI tool recommends withholding anticoagulation. The applicable specialty-society guideline clearly recommends anticoagulation for this profile. The physician, deferring to the tool, withholds it; the patient suffers a stroke. The tool is FDA-cleared, and the hospital's purchase contract disclaims vendor responsibility for clinical decisions.

Under one useful but context-dependent legal framing of liability, which statement is MOST accurate?

A. The vendor bears primary responsibility, because the algorithm generated the erroneous recommendation that caused the harm.

B. The FDA clearance substantially shields the physician, because following a cleared device's recommendation is presumptive evidence of reasonable care.

C. The hospital bears primary responsibility, because procurement and deployment decisions determined which tool the physician saw.

D. The physician bears primary responsibility, and exposure is heightened because the AI advice followed deviated from the standard of care.

E. Responsibility is shared equally among vendor, hospital, and physician, which proportionally reduces each party's exposure.

Short-Answer Question 1

Objectives 2 and 3 — appraising a tool before deployment (suggested time 8 minutes; 6 points)

Your clerkship site is evaluating a commercial model that flags inpatients at risk of clinical deterioration. The vendor's one-page summary states: "AUROC 0.89 in validation. Trained on 1.2 million admissions from a national database, 2016–2019. FDA-cleared. Seamless EHR integration."

List three distinct questions you would insist the vendor or your institution answer before this model is used on your patients. For each, state in one or two sentences why the answer changes whether — or how — the model should be deployed.

Short-Answer Question 2

Objective 4 — the human-in-the-loop procedure and its mechanism (suggested time 6 minutes; 4 points)

Describe the three steps of the Read–Record–Reconcile procedure for working with a predictive tool or AI interpretation. Then explain the specific cognitive mechanism by which the second step protects against errors of commission — that is, why writing the impression down matters, rather than merely intending to think independently.

Answer Key & Explanations

Key: 1-C, 2-D, 3-B, 4-A, 5-E, 6-C, 7-B, 8-D.

Q1 — Correct: C

The citation is real but the attributed content is invented — the harder-to-detect form of persuasive fabrication that retrieval-enabled tools make more common, because the resolving citation passes the existence check. (A) states the misconception this item tests: retrieval reduces but does not prevent misstatement of sources. (B) misuses omission, which describes missing real content, not detecting invented content. (D) fails because a version-lag error would still require the recommendation to have existed in some guideline version. (E) is wrong because the resident did not defer — she verified.

Q2 — Correct: D

Generation, not verification: the likeliest continuation of a challenge to a prior claim is confirmation, and the "supporting details" are produced by the very process in question. (A) is false — reaffirmation demonstrates nothing about training-data frequency, and frequency would not establish truth. (B) dignifies self-consistency with psychometric language; test–retest reliability presumes an instrument measuring something external. (C) is exactly backwards — added generated detail is not added evidence. (E) swaps one internal signal for another; model self-confidence is not calibrated verification.

Q3 — Correct: B

The label (ED return) is a proxy for the construct of interest (deterioration), and the proxy fails differentially with access: patients who cannot return deteriorate invisibly to the label, so the model learns to rate them lower-risk and the call allocation compounds the disparity. The stipulations do the isolating work: representative sampling rules out (E) as a rebuttal, and good overall calibration rules out (C). (A) states the target misconception — objectivity of recording does not make a label a faithful proxy. (D) mislabels a bias present at training time as post-deployment shift; recalibration on the same flawed label cannot fix it.

Q4 — Correct: A

Equity risk here is precisely the possibility that strong average performance conceals weak subgroup performance in a population unlike the validation sites — only subgroup results from a comparable population address that directly. (B) adds more of the same evidence. (C) is regulatory, not equity, evidence. (D) is false as stated — training-set size does not reduce label or proxy bias. (E) helps interrogate individual predictions but does not establish subgroup performance.

Q5 — Correct: E

Strong internal validation at one site does not establish performance elsewhere; a drop at a demographically and operationally different hospital is the expected generalization gap, and the remedy is local validation (and possibly recalibration). The stipulation that inputs map correctly and use is as designed removes (B)-type technical readings. (A) is uncharitable and unsupported. (C) misunderstands overfitting, which is a training-time phenomenon — models do not become overfit "through use," and retraining on the original development data addresses nothing about the new site. (D) is unfalsifiable comfort; a sustained, substantial drop is a finding.

Q6 — Correct: C

Two named changes to the input-generating process — assay vendor and documentation templates — with demographics, code, and usage stipulated stable, is dataset shift by mechanism. (A) is excluded by stipulated stable usage and would not degrade calibration of the model itself. (B) is fictional — models do

not decay computationally. (D) misapplies proxy bias, which concerns the training label. (E) is a exotic explanation with no supporting facts in the vignette.

Q7 — Correct: B

Her failure pattern is the error of commission: accepting asserted findings she did not independently suspect. Committing a written interpretation before exposure to the overlay is the direct countermeasure — it forces an independent read and converts disagreement into a visible discrepancy to investigate. (A) changes which studies she sees, not how she reads them. (C) formalizes the deference that is causing the errors. (D) adds a second system to defer to. (E) abandons the tool's real value against omissions and is the abstinence purism this curriculum consistently rejects.

Q8 — Correct: D

The clinician who interprets and acts on the output is the responsible decision-maker, and current analysis holds that exposure is greatest exactly here — following AI advice that deviates from the standard of care. Neither clearance (B) nor contract language nor procurement (C) displaces the front-line duty; (A) describes a possible but separate and narrower product-liability question; (E) invents a sharing rule — multiple parties can face claims, but that does not dilute the physician's standard-of-care exposure.

SAQ 1 — Model answer and scoring (6 points)

Two points per question-plus-rationale, any three of the following (or equivalent):

- External and local validation: "Has performance been evaluated outside the development database — and on our patients?" AUROC 0.89 from internal national-database validation says nothing about this site; local validation gates deployment. (Q alone: 1 pt; with rationale: 2 pts.)
- Subgroup performance: "What are sensitivity and false-alarm rates by race, ethnicity, sex, language, insurance status?" Average discrimination can conceal differential failure; equity risk is invisible in a single AUROC.
- Currency and drift: "The training window ended in 2019 — what has changed since (case mix, assays, documentation), when was performance last re-checked, and what monitoring and recalibration plan will run after deployment?"
- Label/target scrutiny: "What exactly does the model predict, and is that label a faithful proxy for deterioration in all patient groups?"
- Operational behavior: "What is the expected alert burden — alerts per 100 admissions and positive predictive value — and what override rates are observed at comparable institutions?"
- Accountability: "What does the vendor contractually represent about performance, and what does clearance actually cover?" (Award credit; note clearance itself was stipulated.)

Zero credit for questions answerable from the vendor's one-pager as printed, or for questions without a deployment-relevant rationale.

SAQ 2 — Model answer and scoring (4 points)

- 1 pt — Read: engage the primary data (image, laboratory results, the patient) before viewing any model output.
- 1 pt — Record: commit one's own impression in writing before the output is displayed.
- 1 pt — Reconcile: then view the output, treat discrepancies as findings to investigate deliberately, and document the reasoning; a silent or reassuring output receives the same treatment as an alarming one.
- 1 pt — Mechanism: an uncommitted impression is revisable without cost and silently bends toward the model's answer (anchoring/confirmation); the written record makes disagreement visible and effortful

to abandon, converting deference into a conscious, documentable decision. Full credit requires contrasting written commitment with mere intention; "it makes you think first" alone earns half.