AI Concepts in Healthcare

Problem Based Learning Cases

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

Two problem-based learning (PBL) cases for small groups. Case 1 addresses the session's learning objectives on large language model (LLM) mechanics, hallucinations, and prompt quality. Case 2 addresses the learning objectives on model performance metrics and the training–validation–deployment lifecycle. A single small group can complete both cases in 45–55 minutes; programs running a single 30-minute block should pick one. Sequencing: the cases always follow the lecture — the discussion assumes its vocabulary.

Case 1: “The Confident Tutor”

Scenario

Maya is a first-year medical student (M1) studying for her cardiovascular pathophysiology exam. She is preparing the topic of cardiogenic shock. To save time, she opens a publicly available chatbot built on a large language model (LLM), pastes the lecture learning objectives, and asks: “Explain the mechanism of decreased cardiac output in cardiogenic shock as if I were an M1 student.”

The chatbot returns a clear, well-organized two-paragraph answer. The writing is polished. One sentence reads: “In cardiogenic shock, the failing left ventricle leads to decreased preload, which is the primary driver of decreased cardiac output.”

Maya is reassured. The explanation matches the structure of her notes and uses the right vocabulary. She copies the paragraph into her study guide. The next day, her small-group tutor asks her to explain cardiogenic shock at the whiteboard. Maya recites the chatbot's explanation. Her tutor's facial expression changes.

Discussion Questions

1. **Identify what the chatbot got wrong.** In one or two sentences, describe the actual relationship between left-ventricular failure and preload in cardiogenic shock.
2. Explain why a large language model can produce a confident, well-structured answer that is mechanistically incorrect. What does the model actually “know”?
3. Maya's prompt was: “Explain the mechanism of decreased cardiac output in cardiogenic shock as if I were an M1 student.” Re-write the prompt to reduce — but not eliminate — the risk of a fluent-but-wrong answer. Justify each change.
4. Suppose Maya had pasted the same prompt into a private, institutionally sanctioned LLM that has access to her medical school's vetted lecture notes through retrieval-augmented generation (RAG). Would you expect the answer to be more reliable? Explain what RAG does and what it does not do.
5. Maya's tutor is now considering whether to ban LLM use for studying. Critique that policy. Propose an alternative policy that addresses the risk this case illustrates without forfeiting the legitimate study benefits of LLMs.

FACILITATOR GUIDE — CASE 1

Educational purpose. This case anchors the learning objectives on hallucinations, on LLM mechanics, and on prompt engineering. It is designed to make the abstract concept of probabilistic text generation felt rather than just told: students experience the seductiveness of a fluent answer in a context where they can evaluate it.

Group size and timing. Groups of 5–7 students. 25–30 minutes total: 5 minutes individual reading, 15–20 minutes discussion, 5 minutes facilitator-led debrief. If running both cases in one 45–55 minute block, hold each case to 22–25 minutes.

Question 1 — Identify the error

Anticipated responses. Strong: Most students with an accurate understanding of physiology will recognize that left-ventricular failure in cardiogenic shock is often associated with elevated filling pressures, rather than low preload as its primary mechanism, and that the fall in cardiac output is driven primarily by decreased contractility and forward stroke volume. Partial: Some students will sense that "something is off" but locate the error in the wrong place (e.g., quibbling with "primary driver" while accepting the preload claim) — ask them to draw the pressure–volume loop to force specificity. Misconception: A few students will defend the chatbot's answer because it "sounded right" or matches a half-remembered mnemonic. That confusion is the pedagogical point — surface fluency is a poor proxy for accuracy. Let the group correct them rather than doing it yourself.

Model answer. In the typical left-ventricular pump-failure scenario described here, impaired contractility lowers forward stroke volume and cardiac output, often with elevated left-sided filling pressures and pulmonary congestion. Reduced preload is not the canonical primary mechanism. Cardiogenic shock is heterogeneous, so avoid teaching that filling pressures are invariably elevated in every phenotype.

Question 2 — Why LLMs hallucinate

Anticipated responses. Strong: "It predicts the next word from patterns; its learned representations can support useful reasoning, but fluent generation does not guarantee correct physiology." Partial: "Its training data contained errors" — true but incomplete; push further: even on error-free data, generation is pattern-completion, not fact-retrieval, so novel wrong statements can still be produced. Misconception: "The model is lying / making things up." Push past the anthropomorphism — fluency and self-confidence do not establish factual correctness. "Lying" obscures the mechanism and suggests the wrong safeguard (trusting it when it seems sincere).

Model answer. An LLM is a next-token prediction engine: it generates tokens based on learned probabilities and the available context. When the common pattern of words aligns with the correct mechanism, the answer is right. The output can be wrong even when fluent; mistakes may reflect learned patterns, missing context, retrieval problems, or unsupported extrapolation.

Question 3 — Improve the prompt

Anticipated responses. Strong: Prompts that add a role, cap the length, and demand verification anchors (citations or [UNVERIFIED] flags) — the structure taught on lecture slides 15–16. Partial: Prompts that add detail but no verification mechanism ("explain step by step, at M1 level, using Guyton") — better scoping, still unverifiable; ask "how would you check bullet three?" Misconception: Adding "please be accurate" or "do not hallucinate." That instruction does not change the generation mechanism. The right moves change the structure of what is asked for, not the politeness of the request.

Model answer (one acceptable version). "You are a teaching assistant who only restates information that has been confirmed against a citable physiology textbook. Explain the mechanism of decreased cardiac output in cardiogenic shock at the M1 level in 4 short bullets. After each bullet, give the textbook chapter you would point me to. If you cannot identify a source for a bullet, mark it [UNVERIFIED] and do not elaborate on it." Justifications: a role with a verification disposition; a length cap so the model cannot pad with plausibility; required citations create a self-check the student can actually run.

Question 4 — RAG and its limits

Anticipated responses. Strong: "More reliable on covered topics, but the failure mode shifts rather than vanishes — it can misread or over-extrapolate from what it retrieves." Partial: "Yes, because it has the real

notes” — accept, then probe: what happens when the retrieval step returns the wrong page, or the notes don't cover the question? Misconception: “RAG means the answer is now a quotation and can be trusted.” It is still generation, merely grounded; paraphrase errors survive grounding.

Model answer. Retrieval-augmented generation gives the model access to a curated document store at query time and asks it to ground its answer in retrieved passages. With well-curated lecture notes, accuracy on covered topics may improve because the model paraphrases real text instead of generating from learned patterns alone. RAG does not eliminate hallucination: the model can misquote a passage, retrieve the wrong passage, or fluently extrapolate beyond what the passage says. Fabrication can persist, and retrieval adds risks of incorrect sources and misinterpretation.

Question 5 — Critique the ban policy

Anticipated responses. Strong: A ban is unenforceable, forfeits legitimate benefit, and teaches nothing; the alternative is permitted use plus a mandatory verification habit. Partial: Policies that only restrict (“LLMs for concepts, never for facts”) — workable but brittle; ask how the student tells a concept from a fact mid-study. Misconception: “The ban is right — the stakes are too high.” Invite the group to test it against their own behavior: who here would actually stop? A policy nobody follows is worse than a policy with teeth.

Model answer. A ban is unenforceable, removes a legitimate study aid, and forfeits the chance to teach the prompting and verification skills students will need in clinical training. A better policy: permit LLM use for studying, require students to verify any factual claim against an authoritative course resource before incorporating it into notes, and explicitly teach that verification habit. The same habit transfers to clinical decision support later — which is exactly the bridge Session 2 builds.

Suggested debrief prompt. “Where else in your study habits might fluent-but-wrong AI output be most dangerous, and what is one verification habit you could add this week?” Collect two or three habits aloud; they seed the Prompt Engineering Lab reflection.

Case 2: “The 95-Percent-Accurate Screening Tool”

Scenario

Dr. Patel is an attending physician in a small primary-care clinic in a rural community. The hospital system has just rolled out a new artificial intelligence (AI) screening tool intended to identify patients at high risk for a particular cancer in time for early treatment. The vendor's marketing materials lead with one number: “95% accurate.” During the rollout meeting, a second-year medical student (M2) named Jordan, who is on a longitudinal preclinical clinic placement, raises a hand.

Jordan: “Can I ask what 95% accurate actually means?”

The vendor representative explains that the model was trained on 50,000 patient records from a large urban academic medical center, validated on a held-out portion of that same dataset, and “correctly classifies 95 out of 100 patients.” The cancer is rare in the screened population — the underlying prevalence is roughly 2%.

Discussion Questions

6. If the underlying disease prevalence is 2%, what could a 95%-accurate model do that would still be clinically useless? Calculate or sketch the worst case.
7. Identify three additional metrics Jordan should ask the vendor for, and explain what each one tells you that overall accuracy does not.

8. The model was trained and validated on data from a single large urban academic medical center. Identify two specific risks this creates for Dr. Patel's rural primary-care population. Use the terms internal validation, external validation, and dataset shift in your answer.
9. Six months after deployment, the model's positive predictive value (PPV) drops noticeably even though no software updates have been pushed. Propose three plausible mechanisms and explain how each maps to the concept of model drift.
10. **Dr.** Patel is now deciding whether to use the tool. Identify the single most important question she should ask before her next patient encounter, and justify your choice.

FACILITATOR GUIDE — CASE 2

Educational purpose. This case targets the learning objectives on model performance metrics, internal versus external validation, generalization, and algorithmic drift. The case is intentionally light on clinical specifics so the intellectual work is in the AI reasoning, not in the medicine.

Group size and timing. Groups of 5–7. 25–30 minutes, same internal structure as Case 1.

Question 1 — The 95% trap

Anticipated responses. Strong: The all-negative model: 100 patients, 2 with disease; predict all negative → 98 correct = 98% accuracy, 0 cases caught — a 95%-accurate model could also catch no cases while adding false positives. Partial: Students who reach “accuracy isn't enough” but cannot construct the counterexample — walk the 100-patient sketch on the whiteboard; the arithmetic is the argument. Misconception: Defending 95% as good because it exceeds a familiar exam threshold. Use the calculation to make the failure visceral; this echoes the worked example on lecture slide 20, which most students will have seen hours earlier.

Model answer. Among 100 people, suppose 2 have cancer. A model that misses both cancers and falsely flags 3 of the 98 people without cancer has 95 true negatives, 3 false positives, 2 false negatives, and no true positives: 95% accuracy with 0% sensitivity and 0% PPV. An all-negative model would have 98% accuracy and still catch no cancers. Accuracy alone therefore cannot establish screening usefulness.

Question 2 — Better metrics

Anticipated responses. Strong: Sensitivity, specificity, and PPV, each tied to what it reveals (cases caught; false-alarm workload; meaning of a positive flag). Partial: Naming metrics without the interpretive clause — require the “what does it tell you that accuracy doesn't” half. Misconception: Treating the metrics as interchangeable quality scores. If it surfaces, ask which metric a hospital administrator worried about workup costs cares about, versus a patient worried about a missed cancer.

Model answer. Sensitivity (recall) — of the patients who actually have disease, what fraction does the model catch? This matters most for screening. Specificity — of the patients without disease, what fraction is correctly cleared? This drives false-positive workload. Positive predictive value — of patients flagged, what fraction truly have disease? In rare-disease screening, PPV can be devastatingly low even with impressive-sounding sensitivity and specificity. Also creditable: area under the receiver operating characteristic curve (AUROC), calibration, negative predictive value, false-discovery rate.

Question 3 — Internal vs. external validation

Anticipated responses. Strong: Held-out-same-source versus different-site testing, plus two named risks: population dataset shift and workflow/data-capture shift. Partial: “Rural patients are different” without mechanism — push for what specifically differs (demographics, comorbidity mix, labs, documentation conventions). Misconception: “50,000 records means it's well validated.” Size of the training set and breadth of validation are different axes; a huge single-site dataset is still single-site.

Model answer. Internal validation tests the model on held-out data from the same population, site, and period it was trained on; external validation tests it on a different site, population, or time. Risks for Dr. Patel: (1) dataset shift — demographics, comorbidity patterns, and case mix in a rural primary-care clinic differ from an urban academic center, so calibration and discrimination may not transfer; (2) workflow shift — different laboratories, imaging protocols, and documentation conventions mean the model's expected inputs may arrive missing or differently coded.

Question 4 — Drift mechanisms

Anticipated responses. Strong: Three mechanisms across distinct categories — population change, practice/coding change, true epidemiological change. Partial: Three variations of the same mechanism (three flavors of population change) — credit one, ask for a mechanically different second. Misconception: “Someone must have changed the software” or “the model wore out.” The premise of drift is that nothing in the model changed; the world did. Models do not degrade like machinery — their environment moves away from them.

Model answer (any three). (a) Population drift — the clinic's demographics or risk profile shifted. (b) Practice drift — a new laboratory vendor with different reference ranges changed the input distribution. (c) Coding and documentation drift — an electronic health record (EHR) upgrade, new billing codes, or new templates changed how inputs are recorded. (d) True epidemiological change — prevalence in the community shifted. All are forms of dataset shift; none requires any change to the model itself.

Question 5 — The single most important question

Anticipated responses. Strong: “What is the external-validation performance on a population like mine?” or “What is the PPV at my prevalence, and what happens to the patients I flag?” — either, well defended. Partial: Good questions about workflow or cost that skip validity — acknowledge, then ask what they'd need to know before workflow matters. Misconception: “What is the accuracy?” — the marketing question this case exists to dismantle; let the group say so.

Model answer. Accept either defended choice: external validation on a comparable population (anchors generalization), or PPV at local prevalence plus the downstream workup plan for positives (anchors actionability and false-positive harm). The justification matters more than the pick.

Suggested debrief prompt. “How would you, as a future physician, want this conversation with vendors to go in your clinic? Practice the one sentence you would say.” Have two students actually say their sentence aloud — rehearsal, not description.