AI Limitations and Biases

# Problem Based Learning Cases

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

## Running the cases: two implementation tiers

- **Tier A (live verification).** Groups verify citations live in PubMed (and, where available, query an institutionally sanctioned AI tool). Any device with a browser is sufficient; no patient data is used at any point.
- **Tier B (no technology).** Groups use the pre-printed search results and AI outputs included in this document. The two tiers reach the same teaching points.

Timing: Case 1 runs in 30–35 minutes, Case 2 in 35–40 minutes, with groups of 5–7. Programs running a single PBL block should prefer Case 2, which touches four of the session's five objectives; Case 1 is the stronger standalone if the program's priority is generative-AI literacy.

# Case 1 — "The Confident Citation"

Objective 1 (fabrication and verification), with a closing bridge to Objective 5 • Groups of 5–7 • 30–35 minutes

## The setting

Renee, an M2, is preparing a five-minute evidence summary for her small group on the management of community acquired pneumonia (CAP). To anchor the talk in evidence, she asks a publicly available large language model (LLM): "Summarize the three most important randomized controlled trials and guidelines supporting current first line antibiotic recommendations for outpatient community acquired pneumonia in adults."

## Part 1:  The output in hand (5 minutes)

The LLM returns three references, each with a one paragraph summary. Read them as a group. Do not look anything up yet. On paper, have each group member privately mark each reference: looks real / not sure / looks fabricated and note what feature drove the judgment.

- **Metlay JP, Waterer GW, Long AC, et al.** "Diagnosis and treatment of adults with community-acquired pneumonia. An official clinical practice guideline of the American Thoracic Society and Infectious Diseases Society of America." Am J Respir Crit Care Med. 2019;200(7):e45–e67.
- Postma DF, van Werkhoven CH, van Elden LJR, et al. "Antibiotic treatment strategies for community-acquired pneumonia in adults (CAP-START)." N Engl J Med. 2015;372(14):1312–1323.
- **Patel R, Morrison K, Hernandez T, et al.** "Comparative effectiveness of azithromycin monotherapy versus combination beta-lactam therapy in outpatient adults with community-acquired pneumonia: the AMPLE-CAP trial." Lancet Respir Med. 2021;9(4):412–423.

## Part 2: Verify it yourselves (5-10 minutes)

Tier A: split the group three ways; each subgroup takes one reference and verifies it live,  PubMed search, journal archive, DOI or PMID lookup. Write down, for your reference: (a) verdict, (b) exactly which check settled it, (c) how long it took.

### Tier B: use the printed verification results below.

| Reference | PubMed / journal archive result |
|---|---|

| Reference | PubMed / journal archive result |
|---|---|
| 1 — ATS/IDSA guideline (2019) | Found on first search. PMID resolves; full text on the journal site; matches the LLM's summary. |
| 2 — CAP-START (2015) | Found on first search. PMID resolves to the NEJM trial; the LLM's summary is broadly accurate. |
| 3 — AMPLE-CAP (2021) | Zero results in PubMed for the trial name, the author list, or the title. The cited volume and pages of Lancet Respiratory Medicine contain unrelated articles. The trial does not exist. |

Before moving on, compare Part 2's verdicts with Part 1's paper-only guesses. The expected finding, and the first teaching point, is that the fabricated reference was not reliably identifiable by inspection. It has the same formatting, the same rhythm, and a plausible design. Verification, not vigilance-by-eyeball, is the skill.

## Part 3: The double-down (2 minutes)

Renee, having found nothing in PubMed, returns to the LLM:

Renee: "Are you sure that AMPLE-CAP is a real trial? I can't find it."

LLM: "Yes, AMPLE-CAP was published in Lancet Respiratory Medicine in 2021. The principal investigator was Dr. R. Patel. It may not appear in some search indexes; try searching the journal archive directly."

The model did not just repeat the error, it added new confirming detail, and offered an explanation for why Renee's check failed. Discuss questions 1 and 2 below.

## Part 4: Discussion (10–12 minutes)

- **Define hallucination as it applies to LLMs.** Define persuasive fabrication. Apply both terms to this scenario.
- Explain, in mechanistic terms, why an LLM may confidently confirm a false fact when asked "Are you sure?" and why the added detail ("the PI was Dr. R. Patel") makes the situation worse, not better.
- **Renee verified by searching PubMed.** Identify three additional verification habits for clinical literature work involving an LLM, and rank them by effort-to-value for a busy student.
- Suppose Renee had not verified and had cited AMPLE-CAP in her talk. Identify three categories of harm, ordered from least to most serious.
- Some LLM tools now retrieve from live literature databases before answering (retrieval-augmented generation, RAG). Does this fully solve the problem AMPLE-CAP demonstrates? State precisely what RAG fixes and what it does not, including what your verification work looks like when the citation is real.

### FACILITATOR GUIDE — CASE 1

**Educational purpose.** Objective 1 in applied form: students experience that fabricated citations are not detectable by inspection and that verification must be external to the model. Session 1 introduces the hallucination concept; this case is where students practice the detection skill. The Part 1 commit-then-check design is deliberate: the private paper votes create the before/after contrast that makes the lesson stick.

**Suggested group size and timing.** Groups of 5–7; 30–35 minutes. Keep Part 1 strict, no devices out until Part 2

**Part 1/2 expected pattern.** Most groups flag nothing, or flag the wrong reference. If a student does flag AMPLE-CAP, ask what feature drove the call; the honest answer is usually a guess. Highlight that: a correct guess is not a verification method.

**Q1 — model answer.** Hallucination: content an LLM presents fluently and confidently that is not grounded in any real source. Persuasive fabrication: a hallucination dressed in the surface form of authoritative content — here, a complete citation with author list, journal, volume, and pages, plus a methodologically plausible design. AMPLE-CAP is both: fabricated, and persuasive precisely because its form recruits trust.

**Q2 — model answer.** Repeating a question can reproduce the same error, and asking for reassurance may elicit confident supporting detail without independent evidence. Self-consistency does not establish accuracy. Some systems can retrieve and examine external sources; if that occurs, inspect the source and confirm that it supports the claim. Distinguish an actual external check from another unsupported assertion.

**Q3 — model answer (any three).** (a) PubMed/DOI/PMID lookup of every citation before use, highest value, minutes of effort. (b) Cross-check the claim against a tertiary source (UpToDate, BMJ Best Practice) before relying on it. (c) Prefer purpose-built evidence-retrieval tools over general chatbots for literature work. (d) Ask the LLM for DOIs — absent or non-resolving DOIs are a red flag (necessary but not sufficient). (e) For high-stakes questions, involve a librarian.

**Q4 — model answer (escalating).** (a) Reputational/academic: Renee's preceptor and peers lose trust in her work. (b) Educational: classmates encode a fictional trial; the false fact propagates through study materials. (c) Clinical: if the fictional trial supported a divergent recommendation, a later decision could be built on it; harm by misinformation. The escalation from embarrassment to patient risk is the point.

**Q5 — model answer.** The clinician retains duties to assess the patient and exercise reasonable clinical judgment. A learned-intermediary framework helps explain that role, but responsibility is not universally assigned to the clinician alone. The institution may be responsible for selection, validation, integration, training, and monitoring; vendors may have responsibilities related to defects or representations. Liability depends on the facts, applicable law, and the standard of care. Neither following nor overriding an AI output is automatically protective. Document the clinical reasoning, not a claim that clearance transfers responsibility. See Mello and Guha, 2024.

**Equity extension (optional, 3 minutes).** Ask: for which clinical questions would you expect fabrication risk to be highest? Guide toward: questions where the real literature is thin — rare diseases, under-studied populations, non-English-language evidence. Where training text is sparse, plausible invention fills the gap. The verification habit is therefore also an equity habit.

**Common pitfalls.** Students may conclude "never use LLMs for literature." Redirect to the session's stance: the skill is calibrated trust and external verification, not abstinence

**Suggested debrief prompt.** "What is the one verification habit you will adopt this week, and what will it cost you — honestly — in time?"

# Case 2 — "The Algorithm That Stopped Catching Sepsis"

Objectives 2, 3, 4, and 5 • Groups of 5–7 • 35–40 minutes

## The story

The study background is historical; the bedside scenarios and clinician behaviors below are constructed for teaching. Wong and colleagues externally evaluated a proprietary sepsis model at one academic center using 38,455 hospitalizations involving 27,697 patients. The retrospective analysis found poorer discrimination than vendor-reported internal metrics and estimated a substantial alert burden. No model alerts were delivered to clinicians as part of this validation study.

**Note: Wong, et.** a. JAMA Intern Med. 2021;181(8):1065–1070. doi10.1001/jamainternmed.2021.2626.

**Study results.** Sepsis occurred in 2,552 of 38,455 hospitalizations (7%). The hospitalization-level AUROC was 0.63 (95% CI, 0.62–0.64). At a threshold of 6, the model missed 1,709 sepsis hospitalizations (67%) and would have flagged 6,971 hospitalizations (18%). These were retrospectively calculated alerts, not observed clinical notifications. This study establishes an external-validation problem; it does not demonstrate an observed chain from false alerts to ignored true alerts.

**Teaching extension.** Consider two possible responses to an alerting system: a clinician may dismiss repeated low-value alerts (alert fatigue), or defer to an alert despite contrary primary evidence (automation bias). The following scenarios let students practice recognizing those behaviors; they are not reported observations from Wong and colleagues. Differences between development and validation populations motivate questions about generalization; temporal drift is a related possibility requiring longitudinal evidence.

Epilogue: following the external validation and the attention it drew, the vendor subsequently released a substantially revised version of the model and moved toward requiring local tuning and validation. The lesson of this case is not that one vendor built one bad model, it is what the episode revealed about deploying any model without external, local validation and ongoing monitoring.

## Part 1 — Read and react (5 minutes)

**Read the story.** Before any structured discussion, go around the group once: each member names the single detail they find most alarming, in one sentence.

## Part 2 — The drill: Read, Record, Reconcile (10 minutes)

The session's lecture teaches a three-step procedure for working with any predictive tool: read the primary data first, record your own impression in writing, and only then reconcile it against the model's output. This drill runs the procedure twice. The facilitator controls the reveals; do not read ahead.

### Round 1 — the alert that fired.

**Snapshot (read first, model output hidden): Ms.** Alvarez is 71, admitted this evening with pyelonephritis. Vitals at 21:00: temperature 38.6 °C, heart rate 104, blood pressure 112/64, respiratory rate 20, oxygen saturation 96% on room air. She is alert, talking in full sentences, and asking when she can eat. Urine culture pending; first antibiotic dose given at 20:15.

Record: as a group, commit one written line — your clinical impression and your next action, before seeing anything else.

Reveal (facilitator): the EHR sepsis model fired an alert on Ms. Alvarez 40 minutes ago.

Reconcile: does the alert change your impression? What primary data would you check before acting on it and what would you document either way?

### Round 2 — the alert that didn't.

**Snapshot: Mr.** Osei is 58, post-operative day 2 from a partial colectomy. The bedside nurse calls you: "he just seems off, more confused than this afternoon." Vitals: temperature 36.1 °C, heart rate 118, blood pressure 98/60, respiratory rate 24. The sepsis model shows a low score and has generated no alert.

Record: one written line, impression and next action.

**Reconcile: the model is quiet.** Does that reassure you? What is the failure mode if it does and whose assessment governs?

## Part 3 — Discussion (15–18 minutes)

- Differentiate internal validation from external validation. Why does this case illustrate the importance of external validation specifically and of local validation before deployment?
- **Define dataset shift (model drift).** Identify three plausible mechanisms that could degrade this model's performance over the years between training and the validation study.
- Two opposing patterns of automation bias appeared among the clinicians in the story and in your two drill rounds. Identify the error of omission and the error of commission, and explain which each drill round was designed to provoke.
- **Define alert fatigue.** Propose two design or workflow interventions that could reduce it without abandoning prediction models entirely.
- A patient's sepsis is recognized late because a clinician dismissed an alert that was, this time, correct. Who bears legal and ethical responsibility? Use the term learned intermediary and state what the vendor's responsibility is and is not.
- Equity question: the model was trained on other institutions' historical patients. If your hospital serves a population that differs from the training population, in demographics, language, insurance mix, or patterns of prior access to care, what specifically would you want reported before trusting the model here? Under what conditions could a locally validated model become an equity asset rather than a risk?

### FACILITATOR GUIDE — CASE 2

**Educational purpose.** Connect the historical external-validation findings to constructed scenarios involving automation bias, alert fatigue, accountability, and equity. The Read–Record–Reconcile drill turns independent appraisal into an observable practice. Distinguish evidence from the validation study from hypotheses about behavior and temporal drift.

**Suggested group size and timing.** Groups of 5–7; 35–40 minutes. The drill only works if the reveals are actually sequenced, print the two rounds on separate cards or control them verbally.

**Drill — Round 1 expected pattern.** Most groups' response is appropriate reassurance with continued monitoring (infection identified, treatment started, hemodynamics acceptable). The alert reveal should not overturn that, the correct reconcile is to re-examine primary data (trend the vitals, look at lactate if drawn, eyeball the patient) and document the reasoning for not escalating. Escalation should follow reassessment of the clinical situation, not the alert alone; a different plan may be reasonable if supported by additional findings.

**Drill — Round 2 expected pattern.** The student should be concerned for early sepsis or another post-operative deterioration (new confusion, tachycardia, hypotension, tachypnea, possibly hypothermic) and immediate bedside evaluation. The quiet model must not reassure, a normal score is the omission trap. The nurse's concern and the clinician's own read govern. State the general rule: the model's silence is not evidence of absence; it is one more output to reconcile.

**Q1 — model answer.** Internal validation: performance on held-out data from the same source, era, and pipeline as training. External validation: performance somewhere genuinely different — another site, population, or time. This model's internal metrics were strong; its external performance at a real deploying site was much worse — the generalization gap. Local validation before deployment is the operational consequence: the only external validation that fully answers "will it work here?" is one run on your patients.

**Q2 — model answer (any three).** (a) Population shift — demographics and case mix at deploying sites differ from training data. (b) Coding and documentation drift — EHR templates, order sets, and coding practices change the meaning of input features. (c) Practice drift — screening protocols, lab ordering patterns, and treatment thresholds evolve. (d) Laboratory changes — new assays and reference ranges move the numbers. (e) True epidemiologic shift — sepsis prevalence and pathogen mix change.

**Q3 — model answer.** Automation bias by omission occurs when an absent or reassuring model output discourages an otherwise indicated assessment, as in Round 2. Commission occurs when incorrect automated advice is accepted despite contrary primary evidence. Habitual dismissal of frequent alerts is alert fatigue, a distinct mechanism; it should not automatically be labeled automation bias. The scenarios illustrate risks rather than observed behaviors in the cited validation study.

**Q4 — model answer (any two).** (a) Raise the alerting threshold — trade sensitivity for specificity deliberately and transparently. (b) Tier alerts by severity; reserve interruptions for high-probability cases. (c) Replace pop-ups with passive risk displays on the patient list. (d) Require a brief structured response rather than one-click dismissal. (e) Audit alert performance by unit and retire the alert where it consistently misfires. (f) Track override rates as a patient-safety metric, a 96% override rate is a design finding, not a clinician failing.

**Q5 — model answer.** The clinician retains duties to assess the patient and exercise reasonable clinical judgment. A learned-intermediary framework helps explain that role, but responsibility is not universally assigned to the clinician alone. The institution may be responsible for selection, validation, integration, training, and monitoring; vendors may have responsibilities related to defects or representations. Liability depends on the facts, applicable law, and the standard of care. Neither following nor overriding an AI output is automatically protective. Document the clinical reasoning, not a claim that clearance transfers responsibility. See Mello and Guha, 2024.

**Q6 — model answer.** Ask for: subgroup performance (sensitivity and false-alarm rate by race, ethnicity, language, sex, insurance status), local validation results on the deploying hospital's own patients, and a monitoring plan with recalibration triggers. The equity asset framing: a model locally validated and monitored by subgroup can reveal deterioration earlier in exactly the patients historically under-attended — but only the validation and monitoring make that true; the deployment alone does not. This mirrors the equity framing used across the curriculum: the technology's equity effect is contingent on clinician and institutional verification habits.

**Common pitfalls.** Students may pile onto the vendor and exonerate the deploying institutions — redirect: buying a model without local validation is an institutional decision, and using one without reading primary data is a clinical one. Students may also treat the epilogue as "problem solved"; it is not, the revised model still requires exactly the local validation and monitoring this case teaches.

**Suggested debrief prompt.** "You are the medical director who has just read the validation study. What do you do this afternoon and what standing process would have caught this three years earlier?"

## Section C — Concept boxes

One glance definitions for every technical term used in the cases. Facilitators may print this page separately.

### HALLUCINATION

Content a large language model presents fluently and confidently that is not grounded in any real source, an invented fact, study, statistic, or quotation. It reads like everything else the model produces, which is why it does not look like an error.

### PERSUASIVE FABRICATION

A hallucination dressed in the surface form of authoritative content, a complete citation, correct-sounding mechanism, expert vocabulary. The form recruits the reader's trust; that is what makes it dangerous.

### RETRIEVAL-AUGMENTED GENERATION (RAG)

A design in which the system first retrieves documents from an external corpus (guidelines, a literature index) and then writes its answer conditioned on what it found. It reduces fabrication and can show sources; it does not eliminate misreading, wrong retrieval, or outdated corpora.

### INTERNAL VS. EXTERNAL VALIDATION

Internal: performance on held-out data from the same source the model was trained on. External: performance on data from a different site, population, or time, the only kind that estimates real-world behavior. Local validation is external validation run on your own patients.

### DATASET SHIFT / MODEL DRIFT

Loss of accuracy when the input distribution changes over time, patient mix, hardware, lab assays, coding practices, even though the model's code is unchanged. Often invisible until outcomes suffer.

### AUTOMATION BIAS — OMISSION AND COMMISSION

**Over-trusting an automated output.** Error of omission: missing a finding because the system stayed quiet and you stopped looking. Error of commission: accepting an incorrect output over your own correct judgment.

### ALERT FATIGUE

Desensitization from high volumes of low-value alerts; clinicians learn to dismiss everything, and the rare critical alert is lost in the noise. Documented override rates for drug-safety alerts run 49–96%.

### LEARNED INTERMEDIARY

A context-dependent legal framework in which a clinician interprets a tool's output before acting. It highlights clinical duties but does not exclude institutional or vendor responsibilities. Liability depends on the facts, jurisdiction, and applicable standard of care.

### READ – RECORD – RECONCILE

This session's named human-in-the-loop procedure: read the primary data first; record your own impression in writing before viewing the model output; then reconcile discrepancies deliberately and document the reasoning. The written commitment is intended to make the independent impression observable and reduce unexamined anchoring; it does not eliminate bias.

## Sources cited in this document


- **Wong A, Otles E, Donnelly JP, et al.** External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065–1070.
- **Mello MM, Guha N.** Understanding liability risk from using health care artificial intelligence tools. N Engl J Med. 2024;390(3):271–278.
- **Metlay JP, Waterer GW, Long AC, et al.** Diagnosis and treatment of adults with community-acquired pneumonia: an official clinical practice guideline of the American Thoracic Society and Infectious Diseases Society of America. Am J Respir Crit Care Med. 2019;200(7):e45–e67.
- Postma DF, van Werkhoven CH, van Elden LJR, et al. Antibiotic treatment strategies for community-acquired pneumonia in adults. N Engl J Med. 2015;372(14):1312–1323.
- **van der Sijs H, Aarts J, Vulto A, Berg M.** Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc. 2006;13(2):138–147.