AI Concepts in Healthcare

Active Learning

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

Students design, run, and score prompts of increasing structure for the same medical study task. Run this 25-minute activity after the Session 1 lecture. Tier A uses an institutionally sanctioned large language model. Tier B uses the complete scripted transcripts in the appendix and requires no live AI tool.

Formal learning objective

By the end of this activity, the student will be able to construct and iteratively refine a structured prompt — specifying role, task and format, constraints, and verification anchors — for a medical learning task, and evaluate the resulting outputs for accuracy, usefulness, and verifiability, using only systems appropriate to the data involved.

Bloom's taxonomy level: Create — students generate and revise original prompt designs and evaluate their products.

Alignment. The lab practices the prompting objective presented on lecture slide 3 and developed on slides 14–17. MCQ 2 addresses generation and verification; the scored worksheet and Prompt Artifact Rubric directly assess prompt construction and evaluation.

Student-Facing Instructions

Premise. You and your partner are second-year medical students (M2s) studying for an examination on acute kidney injury (AKI). You decide to use a large language model (LLM) to make a one-page study aid. You will try three prompt designs and measure what each buys you.

Scoring. After each round, you and your partner each score the response 1–5 on three criteria: (a) accuracy — is it right, as far as you can tell?; (b) usefulness — could you study from it for this exam?; (c) verifiability — could you check it without re-deriving everything? Record scores on the worksheet grid before discussing them with each other.

Step 0 — Commit (2 minutes, individually, before touching the tool)

On your worksheet, privately predict which round will score highest on each criterion — and write one sentence saying why. Then write your own best one-sentence improvement to the basic prompt “Tell me about acute kidney injury.” You will check both predictions against reality at the end.

Step 1 — Round 1 (3 minutes, in pairs)

Submit Prompt A: “Tell me about acute kidney injury.” Read the response together. Score it on the three criteria.

Step 2 — Round 2 (5 minutes, in pairs)

Submit Prompt B: “You are a medical school tutor preparing an M2 student for a Step-1-style exam. Explain three etiologic categories of acute kidney injury (pre-renal, intrinsic, post-renal) in 4 bullets each. Use plain physiology language and provide one classic Step-1 association per bullet.” Score the same way. (These etiologic categories are distinct from KDIGO severity staging, which uses creatinine and urine-output criteria.)

Step 3 — Round 3 (5 minutes, in pairs)

Submit Prompt C: Prompt B, plus: “After each bullet, name the standard review-resource section that supports it. If you cannot identify a source, mark the bullet [UNVERIFIED] and say why.” Score the same way, then actually spot-check one cited bullet and one uncited claim against your course resource.

Step 4 — Pair-square (4 minutes, two pairs together)

Compare score grids with a neighboring pair. Where your scores disagree by 2 or more points, each pair defends its score in one sentence. Then agree, as a four, on the single verification habit from today that is most worth keeping — written in behavior language (“I will…”, checkable by a classmate).

Step 5 — Class harvest (5 minutes, facilitator-led)

Each square reports its one habit (large classes: a sample of squares — see facilitator guide). Before leaving, each student writes on the worksheet: (a) which prompt actually won, and whether Step 0's prediction survived contact with the evidence; (b) the one verification habit you are personally adopting this week.

WORKSHEET — score grid (each partner scores independently, then compare)

Step 0 predictions: highest accuracy: Round ___ highest usefulness: Round ___ highest verifiability: Round ___ Why (one sentence): _____

My improved one-sentence prompt:

_____

Round 1 (Prompt A): accuracy ___ / 5 usefulness ___ / 5 verifiability ___ / 5

Round 2 (Prompt B): accuracy ___ / 5 usefulness ___ / 5 verifiability ___ / 5

Round 3 (Prompt C): accuracy ___ / 5 usefulness ___ / 5 verifiability ___ / 5

Spot-check result (Round 3): claim checked: _____ verdict: confirmed / wrong / could not verify

Exit reflection: (a) Which prompt won, and did your Step 0 prediction hold? _____ (b) The verification habit I am adopting this week (behavior language): _____

FACILITATOR GUIDE — PROMPT ENGINEERING LAB

Educational purpose. Anchor the learning objectives on prompt engineering and hallucination mitigation in a lived experience. Students should leave with a concrete, self-chosen verification habit and a felt sense of how dramatic the difference between a one-line prompt and a structured prompt is. Bloom's level: create — students design and refine prompts and evaluate their products against criteria.

Setup. Tier A: confirm institutional LLM access in advance; students work in pairs at one shared device. If a public consumer tool must be used, instruct students to enter no patient data of any kind (this activity intentionally involves none). Tier B: print one appendix transcript packet per pair; students score the pre-run transcripts as if they had run the prompts. The commit step, the squares, and the harvest are identical in both tiers.

Step 0 — Commit: what to expect

Most students predict Round 2 or 3 will win everything. The interesting minority predict Round 1 wins usefulness (“more comprehensive”). Do not correct any prediction — the commit exists so the evidence

can do the correcting. Collect a quick show of hands (“who picked Round 3 for accuracy?”) to create stakes. Watch for: students skipping the written why — the sentence is what converts a guess into a testable belief.

Round 1 — anticipated responses

Expected output: a long, generic, encyclopedia-style overview. Expected scores: usefulness 2–3 (too diffuse to study from), verifiability 1–2 (no anchors), accuracy “seems fine” at 3–4 — students cannot actually tell, and a few will notice that and say so. Facilitator move: when a pair rates accuracy 4 or 5, ask “how do you know?” — the honest answer (“it sounds right”) is the session's central problem, surfaced by the students themselves.

Round 2 — anticipated responses

Expected output: structured, exam-aligned, and rated highly — usefulness 4–5. This is the seduction round: the response may contain plausible-sounding errors students will not notice, and verifiability stays low (2–3) though few notice that at first. Facilitator move: challenge any pair scoring accuracy 5: “Which of those twelve bullets would you bet your exam score on, and how would you check?” Misconception to catch: “better formatting means better accuracy.” Round 2 improves scope and format; it adds no verification mechanism at all.

Round 3 — anticipated responses

Expected output: visibly slower, occasionally refusing to anchor a bullet, sometimes marking [UNVERIFIED]. Scores: verifiability jumps to 4–5; usefulness may dip (less fluent); accuracy becomes checkable rather than guessed — the spot-check makes at least a few pairs discover a citation that does not hold up. Facilitator move: reframe frustration at refusals: a model declining to fabricate an anchor is the safeguard working. Watch for: pairs treating a generated citation as automatically real — the spot-check step exists precisely because citation-shaped text can be fabricated; that is the bridge to Session 3's citation-verification content.

Pair-square and harvest — anticipated responses

Score disagreements of 2+ points are common on usefulness (comprehensiveness versus focus) — that argument is productive; let squares resolve it themselves. Habit statements will arrive vague (“be more careful with AI”). Require behavior language: “Before any LLM bullet enters my notes, I check it against the course syllabus and tag it with the source.” The quality bar: a classmate could observe whether you kept it.

Scaling the debrief

Seminar (≤ 30 students). Every square reports its habit aloud; facilitator boards them, groups duplicates, and closes by naming the two most observable ones. Full sequence fits in 25 minutes.

Mid-size (30–100). Squares write their habit on an index card or shared document; facilitator samples 4–5 aloud, chosen for contrast (one prompting habit, one verification habit, one system-choice habit). One roving co-facilitator per ~40 students keeps rounds on time. Budget 28–30 minutes.

Lecture hall (100+). Pairs still run all rounds (Tier B packets make this device-independent). Replace the open harvest with: (1) a show-of-hands recount of Step 0 predictions versus actual winners — fast and surprisingly dramatic; (2) two or three squares pre-recruited during Step 4 by roving facilitators to report; (3) all habit statements submitted via the learning-management system or a QR-coded form, with the best three read back at the start of the next session (or posted with the session materials). The submitted habits double as the Level 2 artifact for the evaluation plan.

Common pitfalls

Debriefing the nephrology instead of the prompting — the medicine is scaffolding; defer clinical questions to the relevant renal session. Letting Round 1 run long because the output is boring to read — cap it at 3 minutes; boredom is the data. Skipping Step 0 to save time — the commit is what makes Step 5 land; cut a minute from Rounds 1 and 2 instead.

Closing debrief prompts

- “Which round felt most efficient? Which round will produce the fewest exam-day surprises? Why are those answers different?”
- “Where would your verification habit break down in clinical practice — and what would replace it there?” (Bridge to Session 2.)
- “Was there a moment where the model said something you would have copied into your notes if Round 3 hadn't been required?”

Appendix — Tier B (Dry-Lab) Transcripts

Scripted teaching responses for programs without live LLM access. Print the student transcripts separately from the faculty key. These are constructed examples, not outputs attributed to a particular model. Students should assess all claims and verify the proposed source anchors.

Transcript A

Acute kidney injury is a rapid decline in kidney function, recognized by changes in creatinine or urine output. It can occur during acute illness, after fluid loss, with obstruction, or after exposure to drugs that affect the kidneys. Some patients have reduced urine output, swelling, or nonspecific symptoms, while others are identified through laboratory testing.

Causes are often organized into reduced renal perfusion, injury within the kidney, and obstruction to urine flow. These groups can overlap. Evaluation depends on the clinical history, examination, medication exposure, laboratory results, and sometimes imaging. Finding a cause requires more than the creatinine number alone.

The consequences can include fluid overload, electrolyte disturbances, and acid-base changes. Management depends on the cause and severity. Recovery varies, and some patients need follow-up after the acute episode. This overview does not provide the focused associations or source anchors requested in a structured study aid.

Transcript B

Pre-renal AKI

- Reduced effective kidney perfusion can follow fluid loss or low cardiac output. Classic association: the BUN:creatinine ratio is typically less than 10:1.
- With preserved tubular function, sodium retention may increase. Classic association: FENa below 1% can support a pre-renal mechanism, although it is not definitive.
- **Water retention may produce concentrated urine.** Classic association: a bland sediment can accompany a pre-renal process.
- Prolonged ischemia can progress to structural injury. Classic association: persistent dysfunction requires reconsidering a purely pre-renal explanation.

Intrinsic AKI

- Tubular injury may follow ischemia or nephrotoxic exposure. Classic association: muddy brown granular casts support acute tubular injury.
- Glomerular inflammation may cause blood and protein in the urine. Classic association: red-cell casts suggest glomerulonephritis.
- **Interstitial injury may be drug-associated.** Classic association: new kidney dysfunction after a relevant medication exposure warrants review.
- **Vascular disorders can injure the kidney.** Classic association: a compatible systemic vascular syndrome changes the differential.

Post-renal AKI

- **Obstruction impairs urine drainage.** Classic association: urinary retention may arise from bladder-outlet obstruction.
- The effect depends on the level and extent of blockage. Classic association: bilateral obstruction, or obstruction of a solitary functioning kidney, can reduce total filtration.
- Imaging may reveal dilation upstream from obstruction. Classic association: hydronephrosis supports obstruction but can be absent early.
- The clinical context matters even when urine output persists. Classic association: partial obstruction can present without complete anuria.

Transcript C

Pre-renal AKI

- Reduced effective kidney perfusion can follow volume depletion or low cardiac output. Source to check: Merck Manual Professional, “Acute Kidney Injury (AKI),” Etiology.
- Increased urea reabsorption can produce a BUN:creatinine ratio above 20:1 in conventional units; this is suggestive, not diagnostic. Source to check: Merck Manual Professional, “Urinary Diagnostic Indices in Prerenal Acute Kidney Injury and Acute Tubular Injury.”
- FENa below 1% may support sodium retention, but diuretics, kidney disease, sepsis, and other conditions limit its interpretation. Source to check: the same urinary-indices table and the AKI evaluation discussion.
- [UNVERIFIED] No source was supplied for a precise percentage of all hospitalized AKI cases that are pre-renal. I will not add that statistic.

Intrinsic AKI

- Ischemic or toxic tubular injury can produce muddy brown granular casts. Source to check: Merck Manual Professional, “Acute Kidney Injury (AKI),” Evaluation and urine sediment.
- Glomerular inflammation can cause hematuria, proteinuria, and red-cell casts. Source to check: Merck Manual Professional, “Acute Kidney Injury (AKI),” Evaluation.
- Interstitial injury can follow medication exposure; the history helps identify a possible trigger. Source to check: Merck Manual Professional, “Tubulointerstitial Nephritis,” Etiology.
- Vascular disease is another intrinsic cause and requires a compatible clinical context. Source to check: Merck Manual Professional, “Acute Kidney Injury (AKI),” Etiology.

Post-renal AKI

- Obstruction can occur anywhere along the urinary drainage system. Source to check: Merck Manual Professional, “Acute Kidney Injury (AKI),” Etiology.
- Bilateral obstruction or obstruction of a solitary functioning kidney can cause AKI. Source to check: the same source, post-renal causes.
- Hydronephrosis supports obstruction, but a normal ultrasound does not exclude every early or nondilated obstruction. Source to check: the same source, Evaluation.
- Partial obstruction may allow continued urine output. Source to check: the same source, post-renal causes and Evaluation.

Faculty Key for the Scripted Transcripts

Do not distribute this page until after students have scored and checked the transcripts. The examples are constructed to support comparison; live model outputs may behave differently.

Transcript A

The answer provides a broad orientation but little exam-focused organization and no source anchors. Ask students to separate plausibility from verified accuracy.

Transcript B

The first pre-renal bullet contains the planted error: a BUN:creatinine ratio below 10:1 is not the classic pre-renal association. In conventional units, a ratio above 20:1 is a common teaching association, with important limitations. FENa, urine concentration, and sediment findings are clues rather than stand-alone diagnostic rules. The rest of the response still requires checking; useful formatting does not establish correctness.

Transcript C

The response corrects the ratio, acknowledges important limitations, and omits an unsupported percentage. Students must still open the cited resources and confirm that each supports its claim. Source names are verification targets, not proof. A refusal or [UNVERIFIED] label should be evaluated on its merits, not rewarded automatically.

Preparation for the dry lab

Provide an authorized course resource or access to Merck Manual Professional's AKI article and urinary-indices table. Confirm the resource sections before teaching, since web headings and course editions can change. Ask each pair to check one physiologic claim and one source attribution. KDIGO severity staging is separate from the etiologic categories used in this exercise.