epts in Healthcare

rriculum Alignment and Evaluation

undation / Millennium Conference • Preclinical M1/M2 • November 2026

ument maps Session 1 learning objectives to teaching, practice, and assessment. It also provides an evaluation plan across Ki –4, a session survey, and an annotated reading list.

culum Alignment Matrix

slide numbers refer to the Session 1 deck. PBL, Active Learning (AL), and MCQ references refer to the Session 1 deliverables. ck level of the primary assessment appears in the final column. Reading the matrix: every objective is taught (lecture), practice L), and assessed at Level 2 or higher unless noted; the two rows marked “introduced here” are owned in depth by later session s-session concept map.

e (abbreviated)	Bloom's level	Lecture (slides)	PBL	Active Learning	MCQ	Evaluation
e AI families and match clinical task	Remember / Understand	5–9	Case 1 Q2; Case 2 context	Step 0 orientation	1	L2: MCQ pre/post
e/internal vs. external compliant use	Apply	17	Case 1 Q4	Setup rule; Prompt C tier	4	L2: MCQ; L3: self-audit
al applications and history , pathology, CDS)	Recognize	6	—	—	—	L1: survey item 1
models are trained, deployed	Understand	22–25	Case 2 Q3	—	5	L2: MCQ
different problems need approaches	Analyze	5–9	Case 2 Q2	—	1	L2: MCQ
anics of large language	Understand	9–11	Case 1 Q2	Rounds 1–3	2	L2: MCQ
fy and mitigate ions	Analyze	11–13	Case 1 Q1, Q2, Q4	Round 3 + spot-check	2	L2: MCQ + worksheet

e (abbreviated)	Bloom's level	Lecture (slides)	PBL	Active Learning	MCQ	Evaluation
ive prompting with privacy d	Create	14–17	Case 1 Q3	Entire lab; scored artifact	4	L2: Prompt Artifact Rubric
quality (history & physical) utput (introduced here; Session 2)	Understand	16	—	Round 2 vs. 3 contrast	—	L1; S2 assesses
cs beyond accuracy y, specificity, PPV, recall)	Apply / Analyze	18–21	Case 2 Q1, Q2	—	3	L2: MCQ pre/post
tting, underfitting, ation	Understand	22	Case 2 Q3	—	5 (distractors)	L2: MCQ
thmic drift (introduced ed by Session 3)	Understand	23–26	Case 2 Q4	—	5	L2: MCQ

clinical decision support; PPV — positive predictive value. MCQ numbers refer to the question set (answers D, A, E, B, C).

pt Artifact Rubric (Level 2 artifact)

ch pair's Round 3 prompt and exit reflection (worksheet) 0–2 per criterion; 8 points total. Collected at the end of the lab (paper ing-management system (large classes). Suggested competence threshold: 6/8 with no zero. Below-threshold artifacts identify for brief follow-up; item-level patterns identify weaknesses in the teaching.

	0 — Absent	1 — Named	2 — Specified
ructure	Prompt is a bare question with no design elements	Prompt uses at least one taught element (role, task/format, constraint)	Prompt deploys role, bounded tas at least one constraint, coherently at the study goal
n anchor	No mechanism for checking output	Asks for sources or flags in general terms	Requires per-claim anchors with a explicit [UNVERIFIED] fallback, ar spot-check was actually performed

	0 — Absent	1 — Named	2 — Specified
hoice &	No awareness of where the prompt may be run	States the public/private distinction	States the distinction and the gove rule (no patient data outside sanct tools), unprompted
ction	Missing or restates the instructions	Habit named but vague (“be careful with AI”)	Habit in observable behavior langu classmate could verify, tied to wha student saw in the rounds

ation Plan — Kirkpatrick Levels 1–4

y tiers accompany the higher levels so all eleven consortium schools can implement at least the core plan. Levels 1 and 2 are y school, day of session); Level 3 requires modest follow-up infrastructure; Level 4 indicators are consortium-level context, mo e data already exist.

— Reaction

ession survey (below), about five minutes, administered immediately after the lab. Core signals: perceived ability on each obje alue of the lab and the embedded lecture interactions, pacing, and the two open-ended items.

2 — Learning

pre/post. The five MCQs, administered before the session (or asynchronously as pre-work) and after. Paired scores give an ing measure per objective cluster; item-level analysis flags weak instruction.
pt Artifact Rubric. The create-level artifact from the lab, scored 0–8 as above.
ksheet score-grid check. A one-point screen: did the pair's verifiability score rise from Round 1 to Round 3, and was the spot- rded? Captures whether the lab's central contrast landed, at near-zero grading cost.

3 — Behavior

y-habit self-audit (core tier). Four to six weeks after the session, students log one real instance of LLM use in their own study three sentences: the prompt design they used, the verification step they performed, and what the check caught (or confirmed) ewed for completion, not graded; a 10% sample is read against the Prompt Artifact Rubric's verification criterion to estimate ha tion.
ctioned-tool behavior item (moderate tier). Where the institution operates a sanctioned LLM sandbox, usage telemetry (aggr ns, not individual surveillance) before versus after the session provides an unobtrusive adoption signal for the private-versus-pu ctive.

nstream OSCE hook (integrated tier). Session 2's evaluation plan includes an objective structured clinical examination (OSC
station; one checklist line in that station — “verifies an AI-sourced claim against a primary source before acting” — is attributal
y to Sessions 1 and 2 and requires no new station.

4 — Results

icated-citation signal. Where student scholarly output is already screened (research blocks, capstone papers), track the rate
rifiable or fabricated references per cohort. A falling rate is the most direct institutional trace of this session's verification habit.
ic-tool incident signal. Trend privacy or academic-integrity incidents involving patient data or coursework entered into unsanc
ols, where the institution already logs them. Expect small numbers; direction, not magnitude, is the signal.
eracy carry-forward. Where later courses or clerkship examinations include AI-related items (e.g., interpreting a model perfor
), trend cohort performance year over year as a durability signal for objectives 3a–3c.

on caveat. Level 4 outcomes are influenced by far more than one preclinical session. These indicators are monitored as conse
text, not claimed as session effects.

on Evaluation Survey

tems (1–5, strongly disagree to strongly agree)

name the major families of AI methods used in healthcare and give a clinical example of each.
explain, in plain language, how a large language model produces text and why hallucinations occur.
write a structured prompt — role, task, constraints, verification anchors — and explain what each element buys me.
explain why a high headline accuracy figure may be misleading for a rare disease, and name the metrics I would ask for instea
state when AI use requires a secure, sanctioned tool and when a publicly available tool is acceptable.
Prompt Engineering Lab was a valuable use of class time.
nteractive moments during the lecture (poll, spot-the-problem, commit vote) helped me engage with the material.
ecture content was clearly presented and appropriately paced.

ence self-assessment (retrospective pre/post, 1–5)

nguishing generative from predictive AI systems, and naming the characteristic failure mode of each.
tructing a prompt whose output I can actually verify.
preting a vendor's performance claim (accuracy, sensitivity, positive predictive value) at a stated prevalence.

ended

t was the single most valuable thing you learned in this session?
t is one thing about this session that could be improved before it is taught again?

ed Reading List

owing readings support the session. Distinguish empirical findings from conceptual frameworks, and check current local poli
guidance.

omar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine. 2019;380:1347–1358. — Access
tation to the principal families of machine learning and where each fits clinically; the natural first read for preclinical students ar
on behind lecture slide 5.

ol EJ. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books, 2019 (instructor-curated ex
iagnostic-AI chapters). — A physician-friendly historical anchor for the four-era story on slide 6; remind students the field has m
since 2019.

ler EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? FAcc
critique of risks from language-model training and deployment, supporting the discussion on lecture slides 10–12. [Note: techn
anding — assign alongside a one-page reading guide; the conceptual sections suffice for preclinical learners.]

X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases
cal imaging: a systematic review and meta-analysis. Lancet Digital Health. 2019;1:e271–e297. — How computer-vision perfor
rted and why head-to-head comparisons are hard to interpret; pairs with MCQ 3 for a discussion of metric selection.

Food and Drug Administration. Artificial Intelligence and Machine Learning in Software as a Medical Device (web resource). —
uage overview of the lifecycle-based regulatory approach behind lecture slide 24.

ber A, Rodman A. AI and medical education — a 21st-century Pandora's box. N Engl J Med. 2023;389:385–387.
0.1056/NEJMp2304993. — An accessible perspective on generative AI and medical education.