Clinical Practice Enhanced by Artificial Intelligence (AI) Grand Rounds 2025-2026

AI Evaluation in the Era of Large Language Models and Natural Language Generation

Interprofessional
Continuing Education Partnership
UNIVERSITY OF WISCONSIN–MADISON

AI Evaluation in the Era of Large Language Models and Natural Language Generation

Clinical Practice Enhanced by AI Grand Rounds

Guanhua Chen, PhD
Associate Professor
Department of Biostatistics & Medical Informatics
University of Wisconsin School of Medicine and Public Health

Frank Liao, PhD, ACHIP
Senior Director, Digital Health and Emerging Technologies,
UW Health
Adjunct Clinical Assistant Professor,
BerbeeWalsh Department of Emergency Medicine,
University of Wisconsin-Madison

Disclosure: Guanhua Chen, PhD; Frank Liao, PhD, ACHIP

We do not have relevant financial relationships with ineligible companies to disclose.

Learning objectives

- Explain pros/cons of human-proposed, pre-specified metrics vs using an LLM as a judge
- Discuss how evaluation choices impact adoption of AI-generated clinical summaries
- Describe PDSQI-9 as a framework for clinical summary evaluation
- Recognize common pitfalls: data shift, metric gaming, and judge bias

Roadmap

- Introduction
- Part I: LLM as predictor/classifier (metrics + calibration)
- Part II: Evaluating generated text (human vs automatic vs LLM-as-judge)
- Case study: PDSQI-9 + LLM-as-judge in healthcare summarization
- Takeaways

Two evaluation regimes

LLM as classifier / predictor

Discrete/continuous targets (e.g., risk, labels)

Accuracy, AUROC, AUPR

Calibration + clinical utility

External validation & drift monitoring

LLM as generator (text/NLG)

Free-form summaries, notes, messages

Human evaluation (rubrics, pairwise)

Formula-Based Automatic metrics (limited)

LLM-as-judge (scalable, but risky)

Principle: evaluate in context

- Define intended use and decision: who uses it, when, and to do what?
- Specify time zero, eligibility, inputs available at decision time
- Choose evaluation set that matches deployment environment
- Pre-specify metrics and success thresholds to reduce “metric shopping”

A model can be “good” here and “bad” there

- Shift in population, workflow, documentation style, or measurement can change performance
- LLMs are especially sensitive to prompt context and data sources
- External validation > internal CV for deployment decisions
- Monitor drift post-deployment (inputs + outputs + outcomes)

Part I

LLM as classifier / predictor

Evaluate like other ML models

Core classification metrics

Confusion matrix (at chosen threshold)

	Predicted +	Predicted −
Actual +	**TP**	**FN**
Actual −	**FP**	**TN**

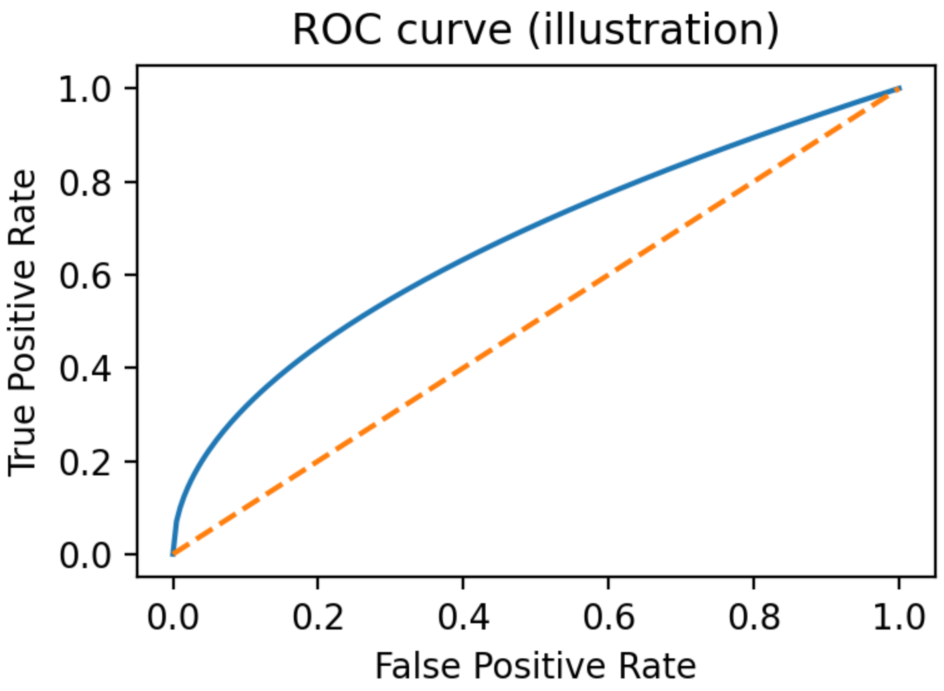
TP=true positive, FP=false positive, FN=false negative, TN=true negative

- **Metric definitions (at chosen threshold):**
- S(ample size) = TP + FP + FN + TN
- Accuracy = (TP + TN) / S
- Sensitivity / Recall = TP / (TP + FN)
- Specificity = TN / (TN + FP)
- Precision (PPV) = TP / (TP + FP)
- F1 = 2TP / (2TP + FP + FN)
- Report the full 2×2 table at the operating point.

[Source] Steyerberg et al. (2010) PMC3575184: https://pmc.ncbi.nlm.nih.gov/articles/PMC3575184/

AUROC (discrimination)

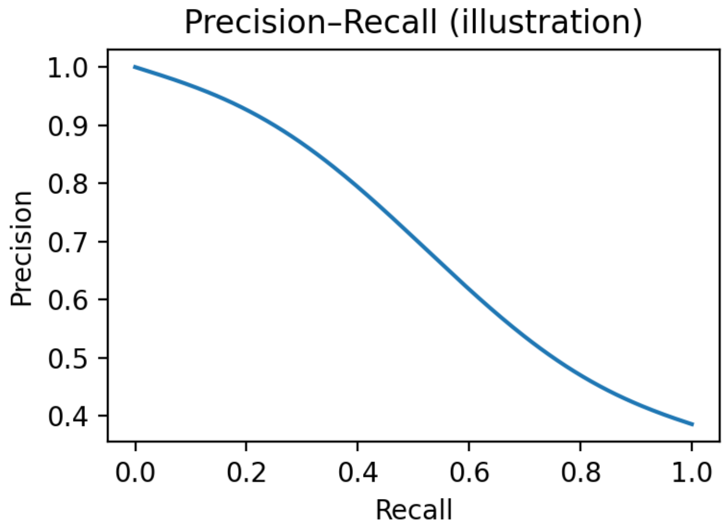
- Probability a random positive is ranked above a random negative
- Threshold-free summary of ranking performance
- Can look “good” even when PPV is low in rare events



Interpretation depends on prevalence and operating point.

AUPR (rare-event focus)

- More informative than AUROC when outcomes are rare
- Baseline AUPR equals prevalence
- Use when you care about PPV/precision at clinically feasible recall

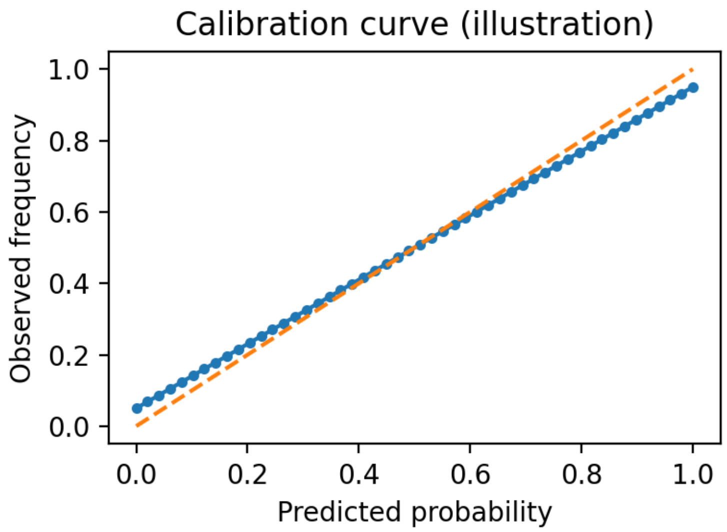


Calibration: the missing half

- Discrimination: “who is higher risk?”
- Calibration: “are predicted risks numerically correct?”
- Clinical decisions often require calibrated probabilities
- A model can have high AUROC but be poorly calibrated

Calibration diagnostics

- Reliability diagram (calibration curve)
- Calibration intercept (overall bias) and slope (over/under-confidence)
- Brier score; expected calibration error (ECE)
- Check calibration by subgroup



Calibration resources: PubMed 29854127; PubMed 28379439.

Example: Stroke risk (JAMIA)

- Concrete example (JAMIA 2024): predict 2-year stroke risk in atrial fibrillation using EHR data (All of Us)
- Real-world validation: temporal split + subgroup (race) fairness checks; thresholds may differ by subgroup
- Key result: discriminative performance improved from ~0.70 ($CHADS_2$/CHA_2DS_2-VASc) to >0.80 (LightGBM)
- What to report beyond AUROC: AUPR + calibration + operating-point metrics (PPV/recall)

[Sources]

Gao J, Mar P, Tang Z-Z, Chen G. Fair prediction of 2-year stroke risk in patients with atrial fibrillation. J Am Med Inform Assoc. 2024. https://doi.org/10.1093/jamia/ocae170

Example: stroke risk (JAMIA)

Rare-event anchor: If 2-year stroke risk prevalence is 3%, then in 1,000 patients ≈ 30 strokes.

What to report	What it answers in practice (using the 1,000-patient anchor)	Common trap -> better practice
AUROC	Does the model rank higher-risk patients above lower-risk ones? (ranking quality, not workload)	AUROC can look strong while PPV is poor under class imbalance -> always pair with an operating point
AUPR + baseline	Among the top-ranked patients, do true strokes concentrate? Baseline AUPR ≈ prevalence = 0.03; report lift above baseline	Reporting AUPR alone -> report prevalence + fold-lift over baseline
Calibration plot + slope/intercept	Are predicted probabilities numerically correct? (If we say 10% risk, is it ~10%?) Miscalibration drives over/under-treatment	Transport drift -> overconfidence -> recalibrate on validation; show subgroup calibration
Operating point (choose one policy) (e.g., Top 5% flagged)	If we alert 50/1,000, how many strokes caught (recall) and how many false alerts (alert burden)? Also report NNE (number needed to evaluate)	Tuning threshold on test set -> pre-specify threshold rule (capacity/cost), choose on validation, lock for test/prospective

Uncertainty & reproducibility/Subgroup performance & fairness

- Report confidence intervals for key metrics (bootstrap)
- Avoid leakage (e.g., note timestamps, future labs, post-treatment info)
- Use locked test sets; document preprocessing

- Stratify by clinically relevant subgroups (age, sex, race/ethnicity, unit)
- Compare both discrimination and calibration across groups
- Beware small-n instability and multiple comparisons
- Use subgroup results for targeted remediation, not “p-hacking”

Part II

Evaluating text generation•

Human evaluation, automatic metrics, and LLM-as-judge

- This portion of the slides is adopted from **"Scaling Medical Evaluation of LLM Summaries From PDSQI-9 to LLM-as-a-Judge"**, and is reproduced with permission from the original authors: **Majid Afshar, MD, MSCR** (Associate Professor, Department of Medicine), **Brian Patterson, MD, MPH** (Associate Professor, Department of Emergency Medicine), and **Emma Croxford** (PhD student, Department of Biostatistics and Medical Informatics), **University of Wisconsin–Madison**.

What are we evaluating?

- Factuality / groundedness (no hallucinations)
- Completeness (critical facts included)
- Relevance (no irrelevant copy-paste)
- Clarity, organization, and clinical usability
- Safety, privacy, and bias

Human evaluation (gold standard)

- Rubric-based rating (e.g., 1–5 per dimension)
- Pairwise preference comparisons (often more stable)
- Measure inter-rater reliability and adjudicate disagreements
- Costly and slow, but critical for high-stakes clinical use

Automatic metrics: families of LLM evaluation metrics (cheat sheet)

- Reference-based: BLEU/ROUGE, edit distance — compare to a reference (good for narrow tasks; weak when many valid phrasings exist)
- Embedding-based (e.g., BERTScore): semantic similarity using embedding models (captures paraphrase; still not factuality)
- Rule-based / task-specific checks: keyword presence, format, entity extraction (targets failure modes)
- RAG evaluation: faithfulness, answer relevancy, context relevancy/recall (requires retrieved context)

[Sources]

Microsoft Learn AI Playbook: A list of metrics for evaluating LLM-generated content.

https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-llms/evaluation/list-of-eval-metrics

Concrete example: overlap metrics can reward wrong clinical facts

Input note (snippet)

- **"AKI stage 2 on 1/12; creatinine peaked 2.1, improved to 1.3 by discharge."**

Summary A (correct)

- **"Had AKI that improved before discharge."**

Summary B (wrong but similar words)

- **"AKI worsened and creatinine rose to 2.1 at discharge."**

Takeaway

- **Lexical overlap ≠ factuality**
- **Use targeted checks (factuality/groundedness) + human rubric in high-stakes settings**

	ROUGE	Human
Summary A (correct)	Lower	✅
Summary B (wrong)	Higher	❌

[Sources]

Microsoft Learn AI Playbook: A list of metrics for evaluating LLM-generated content (ROUGE/BLEU as n-gram overlap metrics). https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-llms/evaluation/list-of-eval-metrics

Use Case: Scaling Medical Evaluation of LLM Summaries

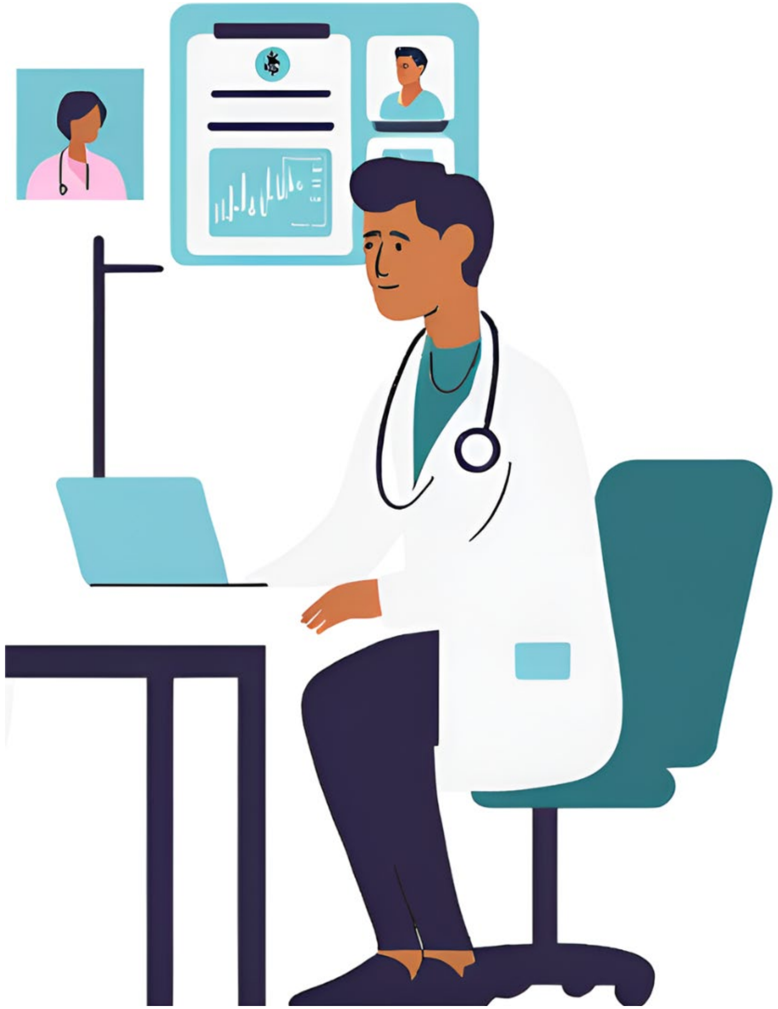
From PDSQI-9 to LLM-as-a-Judge

The Problem

- You are a specialist clinician, and a new patient shows up.

- You have a new service in your EHR that uses ChatGPT to summarize their medical history for relevant information

- How can you trust that summary is useful, accurate, and not missing important details?



Current Conundrum

- This technology is now available but there’s no standard to automatically evaluate the quality of the summary
- Health systems want to use new AI tools, but don’t want to put anything out that is unsafe
- Human evaluation is time and resource intensive
- Traditional automated metrics fall short in capturing the accuracy, coherence, and clinical relevance

Evaluation Needs Going Forward

- Transparent and rigorous validation

- Need criteria to assess LLM (e.g., ChatGPT) weaknesses
  - Hallucination, Omission, Revision, Faithfulness/Confidence, Bias/Harm, Groundedness, Fluency
  - Struggle with multi-document, longitudinal tasks

Provider Documentation Summarization Quality Instrument (PDSQI - 9)

PDSQI-9 Attributes

Accurate Factually correct	**Succinct** Appropriately concise	**Synthesized** Integrated content
Cited Source attribution	**Stigmatizing** Harmful language	**Thorough** Complete information
Comprehensible Easy to understand	**Organized** Logical structure	**Useful** Clinically relevant

Rubric example

2. Is the summary accurate in extraction (extractive summarization)?

Extraction-based summarization involves selecting and pulling exact phrases or sentences directly from the original text without altering the wording. The focus is on identifying the most important parts of the text and reproducing them verbatim to form the summary.

For example, in a clinical context, if the original text states, "the patient experienced shortness of breath, had an elevated white blood cell count, and showed a right lower lobe infiltrate on a chest X-ray," an extraction model would select and present these same sentences as the summary. There would be no attempt to infer or rephrase the content—just a selection of key details directly from the source.

a. The summary is true and free of incorrect information. (Example: Falsification – the provider states the last surveillance study was negative for active cancer, but the LLM summarizes the patient still has active disease.)

b. Incorrect Information can be a result of fabrication or falsification

   i. Fabrication is when the response contains entirely made-up information or data and includes plausible but non-existent facts in the summary

   ii. Falsification is when the response contains distorted information and includes changing critical details of facts, so they are no longer true from the source notes

   iii. Examples of problematic assertions: It's not in the note, it was correct at one point but not at the time of summarization, a given assertion was changed to a different status (given symptoms of COVID but patient ended up not having COVID; however, LLM generates COVID as a diagnosis).

c. Something can be an incorrect statement by the provider in the note (not clinically plausible) but if the LLM summarizes the same statement from the provider then it's NOT a fabrication or falsification.

1: Not at All	2	3	4	5: Extremely
Multiple major errors with overt falsifications or fabrications	A major error in assertion occurs with an overt falsification or fabrication	At least one assertion contains a misalignment that is stated from a source note but the wrong context, including incorrect specificity in diagnosis or treatment	At least one assertion is misaligned to the provider's source or timing but still factual in diagnosis, treatment, etc.	All assertions can be traced back to the notes

Data for Summarization

- UW Health EHR
  - March 2023- December 2023
- Perspective of Provider at Outpatient Encounter
  - 11 specialties (Gyn, Neuro, Derm, Ortho, FM, IM, Ophtho, Neurosurg)
  - Summaries over prior 3-5 encounters (real-world multi-document EHR)
  - 200 unique patients

Summarization Methods

- LLM Prompt: ”You are an expert doctor. Your task is to write a summary for a specialty of [target specialty], after reviewing a set of notes about a patient.”
- The persona and instruction were followed by two chains of thought:
  - To generate higher-quality summaries, PDSQI-9 aligned instructions were provided
  - To generate lower-quality summaries, additional variations of the prompt removed instructions or encouraged the inclusion of false information.

Model	Parameters	Context Window
GPT-4o	–	128,000
Mixtral 8x7b	7b	32,000
Llama 3-8b	8b	8,000

Validation Study Design

- Seven physician raters
  - 5 junior physicians (1-5 years post-graduate experience)
  - 2 senior physicians (10+ years experience)
- Standardized training with exemplar cases
- Statistical Power:
  - Target: 80% power required minimum 84 evaluations per rater (5 rater min)
  - Achieved: Over 100 evaluations per rater (7 raters)
- 779 total summaries evaluated
- 8,329 individual attribute ratings across all evaluations

Outcome

- Validated the instrument, demonstrating excellent validity for clinical use.
  - *Inter-Rater Reliability*
    - Intraclass correlation coefficient (ICC) = 0.867 (95% CI: 0.867–0.868)
  - *Internal Consistency*
    - Cronbach's α = 0.879 (95% CI: 0.867–0.891)

- First tool built using a semi-Delphi process on real-world, multi-site EHR data

Single LLM-as-a-Judge (Zero & Few Shot)

Input to the LLM-as-a-Judge

Patient Notes	Patient Summary	PDSQI-9 Rubric	Task Instructions
Subjective: [NAME] is a [AGE]-year old male who presents for evaluation of ...	[PATIENT NAME], a [AGE]-year-old male, presents for...	*Accurate* : Is the summary accurate in extraction? ...	Your task is to grade the summary, based on the RUBRIC_SET…

Input by the Numbers

Length of Notes (words), median (IQR)	3050 (2174, 4128)
Length of Summary (words), median (IQR)	328 (191, 498)
Length of LLM Input (words), median (IQR)	4746 (3871, 5831)

Who were the Judges?

- Microsoft Azure
  - GPT-4o
  - GPT-o3-mini
  - DeepSeek-R1 761
- Huggingface
  - Llama 3.1 8B
  - Phi 3.5 MOE
  - Mixtral 8x22B
  - DeepSeek Distilled Llama 8B
  - DeepSeek Distilled Qwen 32B

Top Results

LLM-as-a-Judge	Strategy	Intraclass Correlation Coefficient (ICC)	Median Difference (IQR)
GPT-o3-mini	Zero-Shot	0.803	0 (0,1)
GPT-o3-mini	**5-Shot**	**0.818**	**0 (0,1)**
DeepSeek-R1 761B	Zero-Shot	0.762	0 (0,1)
Mixtral 8x22B	Zero-Shot	0.733	1 (0,1)

Costs per Evaluation

LLM-as-a-Judge	Inference Time (sec)	Money ($)
Human Baseline	600	50.00*
GPT-o3-mini	16	0.02

***based on median minimum physician consulting rate of $300/hour**

Designing an LLM judge

- Start with a clear rubric aligned to clinical needs
- Provide few-shot examples (good vs bad) to anchor scores
- Use structured outputs (JSON) for reproducibility
- Use multiple judges / repeated runs; average or vote
- Continuously validate against human ratings

Judge pitfalls (and mitigations)

- Position / verbosity bias → randomize order; control length
- Style bias → focus rubric on clinical content, not prose
- Prompt injection → isolate inputs; strip instructions from evaluated text
- Non-determinism → repeated runs + confidence intervals
- Model self-preference → avoid judging its own outputs when possible

Practical recommendations

- Use a portfolio of metrics (not a single number)
- Pre-specify evaluation plan; document data and prompts
- Calibrate LLM judges against humans; monitor judge drift too
- Focus on failure modes that matter clinically (hallucinations, omissions)
- Treat evaluation as continuous: pre + post deployment

Acknowledgements & key references

- Microsoft AI Playbook: “A list of metrics for evaluating LLM-generated content” (Microsoft Learn)
- Gao J, Mar P, Tang Z-Z, Chen G. Fair prediction of 2-year stroke risk in patients with atrial fibrillation. JAMIA 2024. doi:10.1093/jamia/ocae170
- Croxford E, Gao Y, et al. Evaluating clinical AI summaries with large language models as judges. npj Digital Medicine 2025;8:640. doi:10.1038/s41746-025-02005-2
- Croxford E, et al. Development and validation of the Provider Documentation Summarization Quality Instrument (PDSQI-9). JAMIA 2025;32:1050–1060. doi:10.1093/jamia/ocaf068
- Calibration background: Steyerberg EW et al. Assessing the performance of prediction models. 2010 (PMC3575184)

- **Majid Afshar, MD, MSCR** (Associate Professor, Department of Medicine), **Brian Patterson, MD, MPH** (Associate Professor, Department of Emergency Medicine), and **Emma Croxford** (PhD student, Department of Biostatistics and Medical Informatics), **University of Wisconsin–Madison** for sharing their slides.

Please complete the session evaluation in the ICEP Learning Portal after the talk. Questions?

Thank you!