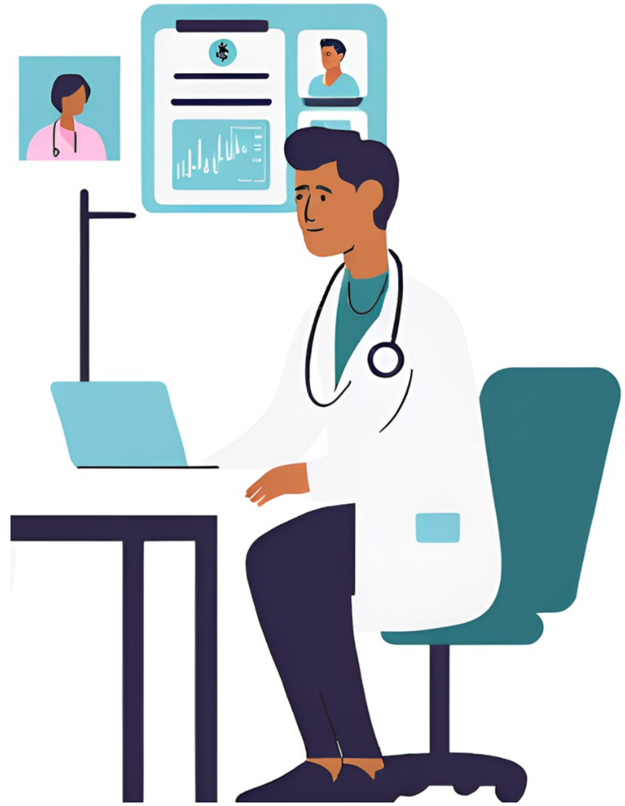
The Provider Documentation Summarization Quality Instrument (PDSQI-9) for AI Generated Text

The Problem

- You are a specialist pulmonologist, 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
  - Current state of "do you like it" won't cut it

Current State of Evaluation

Limitations of Current Methodology

- Traditional methods measure overlap in words or meaning to some reference text but....
  - Heart Attack ≠ Hear Attak
  - Myocardial Infarction ≠ Heart Attack
  - Bacterial ≠ Viral

- Most rubrics were designed to assess clinical documentation quality
  - All designed to be used by humans, for human-authored notes

Development of a Human Evaluation Framework and Correlation with Automated Metrics for Natural Language Generation of Medical Diagnosis. AMIA Symp Proceed. 2024

PDQI-9

- **P**hysician **D**ocumentation **Q**uality **I**nstrument
- Nine criteria assess documentation quality
- Validation on human-authored admission notes, progress notes, and discharge summaries

Attribute	Score	Description of Ideal Note
1. Up-to-date	Not at all 1 2 3 4 Extremely 5	The note contains the most recent test results and recommendations.
2. Accurate	Not at all 1 2 3 4 Extremely 5	The note is true. It is free of incorrect information.
3. Thorough	Not at all 1 2 3 4 Extremely 5	The note is complete and documents all of the issues of importance to the patient.
4. Useful	Not at all 1 2 3 4 Extremely 5	The note is extremely relevant, providing valuable information and/or analysis.
5. Organized	Not at all 1 2 3 4 Extremely 5	The note is well-formed and structured in a way that helps the reader understand the patient's clinical course.
6. Comprehensible	Not at all 1 2 3 4 Extremely 5	The note is clear, without ambiguity or sections that are difficult to understand.
7. Succinct	Not at all 1 2 3 4 Extremely 5	The note is brief, to the point, and without redundancy.
8. Synthesized	Not at all 1 2 3 4 Extremely 5	The note reflects the author's understanding of the patient's status and ability to develop a plan of care.
9. Internally Consistent	Not at all 1 2 3 4 Extremely 5	No part of the note ignores or contradicts any other part.
Total Score:		

(Version 1: 11/21/2011)

Assessing Electronic Note Quality Using the Physician Document Quality Instrument (PDQI-9). Appl Clin Inform. 2012

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

Current and Future State of Evaluation of Large Language Models for Medical Summarization Tasks. NPJ Health Systems. 2025

PDSQI-9 Human Validation

PDSQI-9 Development

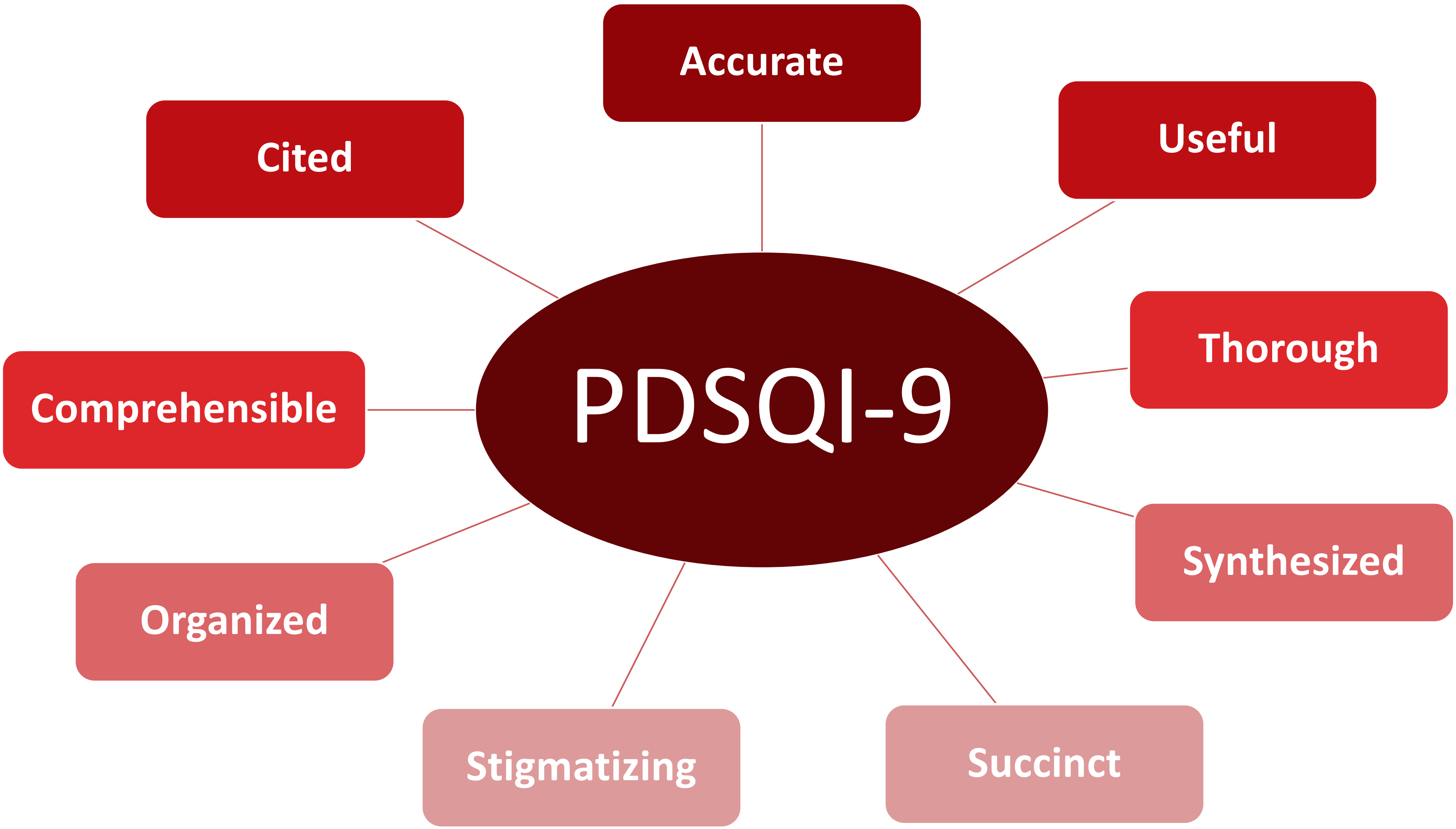
Stage 1
- Literature Review and Attributes Identification

Stage 2
- Multi-Stakeholder Input and Initial Design

Stage 3
- Study Design Setup and Training Human Expert Raters

Stage 4
- Validation with 7 Expert Raters

Development and validation of the provider documentation summarization quality instrument for large language models. JAMIA. 2025


Accurate
Cited
Useful
Thorough
PDSQI-9
Comprehensible
Synthesized
Organized
Stigmatizing
Succinct

Validation Study Design

- 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
- Seven physician raters
  - Mixture of senior, junior, and trainee physicians
- 779 summaries
- 8,329 PDSQI-9 items

Development and validation of the provider documentation summarization quality instrument for large language models. JAMIA. 2025

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

Development and validation of the provider documentation summarization quality instrument for large language models. JAMIA. 2025

PDSQI-9 LLM-as-a-Judge

Study Design

Doctor-as-a-Judge (Benchmark)

Single LLM-as-a-Judge

Customized LLM-as-a-Judge

Multiple LLMs-as-Judges

Automating Evaluation of AI Text Generation in Healthcare with a Large Language Model (LLM)-as-a-Judge. MedArXiv. 2025

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 …

Automating Evaluation of AI Text Generation in Healthcare with a Large Language Model (LLM)-as-a-Judge. MedArXiv. 2025

Results

LLM-as-a-Judge	ICC (“Inter-rater reliability”)	Median Difference (IQR)
GPT-o3-mini	**0.803**	**0 (0,1)**
Multiple LLM Judges	0.768	0 (-1,1)
DeepSeek R1	0.762	0 (0,1)
Customized Mixtral	0.746	1 (0,1)

Automating Evaluation of AI Text Generation in Healthcare with a Large Language Model (LLM)-as-a-Judge. MedArXiv. 2025

Conclusion

- Developed a novel human evaluation framework to assess LLM performance in EHR summarization tasks.

- Introduced an automated method to evaluate clinical multi-document summaries using LLMs

  - **GPT-o3-mini** achieved strong inter-rater reliability (ICC = 0.803), comparable to expert humans

  - GPT-o3-mini evaluations were **~38x faster** than human reviewers (16s vs. 600s)

Automating Evaluation of AI Text Generation in Healthcare with a Large Language Model (LLM)-as-a-Judge. MedArXiv. 2025

Acknowledgements

ICU Data Science Lab
UNIVERSITY OF WISCONSIN
SCHOOL OF MEDICINE AND PUBLIC HEALTH

Department of Biostatistics and Medical Informatics
UNIVERSITY OF WISCONSIN
SCHOOL OF MEDICINE AND PUBLIC HEALTH

- ICU Data Science Lab Members
  - Madeline Oguss (Program Manager)
  - Matthew M Churpek (PI)
  - Anoop Mayampurath (PI)
  - Majid Afshar (PI)
  - John Caskey (Data Engineer)
  - Askar Afshar (Data Engineer)
  - Jennie Martin (Data Scientist)
  - Alexandra Spicer (Data Scientist)
  - Skatje Myers (NLP/LLM Post-Doc)
  - Mahmudur Rahman (Image AI Post-Doc)
  - Emma Graham Linck (PhD Candidate)
  - Timothy Gruenloh (PhD student)
  - Sierra Strutz (PhD student)
  - James Haddad (PhD student)
- Collaborators in GenAI
  - Brian Patterson, UW Madison
  - Guanhua Chen, UW Madison
  - Elliot First, Epic
  - Karen Wong, Epic
  - Nicholas Pellegrino, Epic
  - Miranda Schnier, Epic
  - Cherodeep Goswami, UW Health
  - Frank Liao, UW Health
  - Graham Wills, UW Health
  - Dmitriy Dligach, Loyola University Chicago
  - Yanjun Gao, Univ of Colorado
- NIH/NLM training grant to the Computation and Informatics in Biology and Medicine Training Program (NLM 5T15LM007359)

National Institutes of Health

Q&A

Public GitLab Repo

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON