BRIDGE GenAI Lab

BIDMC–DFCI Radiology & Imaging Generative AI Hub

Beth Israel Deaconess Medical Center • Harvard Medical School

Supplementary Information

Large Language Models in Inflammatory Arthritis: A Systematic Review Across Clinical Tasks

Yosef Adiniaev, Mahmud Omar, Tohar M Timor, Yiftach Barash, Olga R Brook, Mohammad E Naffaa, Alon Gorenshtein, † Eyal Klang†

Table of Contents

Supplementary Figures

S1: Risk of bias summary using adapted QUADAS-2 framework.

Supplementary Tables

S1: Search strategies across databases.

Database 1: PubMed

(
"large language model"[Title/Abstract] OR "large language models"[Title/Abstract] OR LLM[Title/Abstract] OR LLMs[Title/Abstract]
OR "generative AI"[Title/Abstract] OR "generative artificial intelligence"[Title/Abstract]
OR "ChatGPT"[Title/Abstract] OR "GPT"[Title/Abstract] OR "GPT-4"[Title/Abstract] OR "GPT-3"[Title/Abstract] OR "GPT-4o"[Title/Abstract]
)

AND

(
"rheumatoid arthritis"[Title/Abstract] OR "psoriatic arthritis"[Title/Abstract] OR "ankylosing spondylitis"[Title/Abstract]
OR "axial spondyloarthritis"[Title/Abstract] OR "gout"[Title/Abstract] OR "gouty arthritis"[Title/Abstract]
OR "spondyloarthritis"[Title/Abstract] OR "inflammatory arthritis"[Title/Abstract]
)

Filters: English language; Date: 2022/01/01 – Present

PubMed Total: 53

Database 2: Scopus

TITLE-ABS-KEY(
(
"large language model" OR "large language models" OR "LLM" OR "ChatGPT" OR "GPT-4" OR "GPT-3" OR "GPT-4o"
OR "generative artificial intelligence" OR "generative AI"
)

AND

(
"rheumatoid arthritis" OR "psoriatic arthritis" OR "ankylosing spondylitis" OR "axial spondyloarthritis"
OR "gout" OR "gouty arthritis" OR "spondyloarthritis" OR "inflammatory arthritis"
)
AND LANGUAGE(english) AND PUBYEAR > 2022

Scopus Total: 99

Database 3: PubMed Central (PMC)

(
"Large Language Models"[Title/Abstract] OR "generative AI"[Title/Abstract] OR ChatGPT[Title/Abstract]
)

AND

(
"Arthritis, Rheumatoid"[MeSH] OR "Arthritis, Psoriatic"[MeSH] OR "Spondylitis, Ankylosing"[MeSH]
OR "Spondylarthritis"[MeSH] OR "Gout"[MeSH]
)

Filters: Date range 2022–2026

PubMed Central Total: 7

Summary

PubMed: 53 results

Scopus: 99 results

PubMed Central: 7 results

Total records before duplicate removal: 159

S2: Adapted QUADAS-2 checklist for risk of bias assessment.

The QUADAS-2 framework was adapted with AI-specific modifications to address methodological considerations unique to LLM evaluation studies. Modifications included redefining the index test domain to capture LLM-specific factors (model version, access date, prompt design, reproducibility) and adjusting the reference standard to account for comparators used in LLM studies (expert panels, clinical guidelines, validated scoring tools). Applicability concerns were assessed for the first three domains. Each domain was rated as low risk, high risk, or unclear.

Domain	Signaling Question / Criterion	Low Risk	High Risk	Unclear
Domain 1: Data Selection (Risk of Bias)				
Data Selection	Were the clinical questions or vignettes representative of the target clinical domain?	Questions derived from validated sources, clinical guidelines, or patient-generated queries	Small, non-representative, or non-validated question set	Insufficient detail to judge representativeness
Data Selection	Was the question set large enough to draw meaningful conclusions?	Sample size justified or consistent with similar studies	Fewer than 10 questions/cases with no justification	Sample size not reported or borderline
Data Selection	Were selection criteria for questions/cases clearly described?	Clear inclusion/exclusion criteria reported	No criteria reported; questions appear arbitrary	Partial description of selection process
Domain 1: Data Selection (Applicability Concerns)				
Applicability	Do the selected questions match the clinical tasks and patient populations relevant to this review?	Questions address diagnosis, treatment, or education in RA, PsA, AS, or gout	Questions focus on a different disease, non-clinical task, or unrelated population	Questions partially relevant or conducted in a language limiting generalizability
Domain 2: Index Test — The LLM (Risk of Bias)				
Index Test	Was the LLM model version clearly specified?	Model name and version reported (e.g., GPT-4-turbo, Gemini 2.0 Flash)	Model described only generically (e.g., 'ChatGPT') without version	Version partially specified or ambiguous
Index Test	Was the date of LLM access reported?	Specific access date(s) documented	No access date reported	Approximate time frame given but not exact dates

Index Test	Was the prompting strategy described and reproducible?	Prompts provided verbatim or described in detail; zero-shot or few-shot specified	No description of prompts; unable to reproduce	General description without exact prompts
Index Test	Was more than one LLM evaluated to allow comparison?	Two or more LLMs compared under the same conditions	Single LLM tested without justification	Single LLM tested with stated justification
Domain 2: Index Test (Applicability Concerns)				
Applicability	Are the LLM(s) tested relevant and currently accessible for clinical use?	Widely available models tested in publicly accessible versions	Discontinued models or custom fine-tuned models not available to clinicians	Models tested in a specific language or interface that limits generalizability
Domain 3: Reference Standard (Risk of Bias)				
Reference Standard	Was the reference standard clearly defined?	Clinical guidelines, expert consensus panels, or validated scoring tools (e.g., DISCERN, Likert) used as comparator	No reference standard; LLM output evaluated in isolation	Reference standard partially described
Reference Standard	Were evaluators blinded to the LLM source?	Blinded evaluation explicitly stated	Evaluators knew which responses were LLM-generated	Blinding not described
Reference Standard	Was inter-rater reliability assessed?	Kappa, ICC, or similar reliability metric reported with adequate agreement	Single evaluator with no reliability check	Reliability metric reported but agreement was poor (e.g., kappa < 0.40)
Domain 3: Reference Standard (Applicability Concerns)				
Applicability	Does the reference standard reflect the clinical decision-making process relevant to this review?	Comparator reflects real clinical practice (e.g., specialist judgment, current guidelines)	Comparator does not align with standard clinical evaluation (e.g., non-clinical framework)	Comparator partially relevant or applied inconsistently
Domain 4: Flow and Timing (Risk of Bias)				
Flow and Timing	Were all LLM queries included in the analysis?	All generated responses analyzed; no selective exclusion	Some responses excluded without justification; selective reporting	Exclusions described but rationale unclear

Flow and Timing	Was the evaluation conducted within a consistent time frame?	All LLM queries performed within a defined period; model version stable	Queries performed across model updates with no version control	Time frame not reported but likely consistent

Abbreviations: LLM, large language model; RA, rheumatoid arthritis; PsA, psoriatic arthritis; AS, ankylosing spondylitis; DISCERN, quality criteria for consumer health information; ICC, intraclass correlation coefficient.