DRAFT - November 2026

DEEPER Framework - version 3

Pre-Assessment - Context

Original prompt	What was the original prompt asked?
AI system and version	What AI system, model and version was used?
Date run	When was the query run?
Question type	What kind of question is this (e.g. effectiveness, diagnostic, etc)? Is it broad or specific?
What matters for this question	What does the end-user need from the answer (e.g. population, setting, decision being informed)?
User and intended use	Who is the end-user (e.g. academic, policy-makers, humanitarian worker etc)? What is their level of AI/evidence competence (e.g. beginner, intermediate, advanced)? How will the AI-generated report be applied (e.g. to inform decision-making)?

Before you start, please select a random sample of five references from the reference list and try to obtain the original documents. This source verification sample will be used for Domain C assessment.

Domain A - Selection of Evidence

Note. *It is possible to assess what evidence has been included in a report, but it is more difficult to assess what has been excluded without expert knowledge and access to systematic evidence sources such as recent systematic reviews and national/international guidelines (such as those from the World Health Organization)*

Item number	Item	What to assess	Assessment
A1	Relevance to the question and context	Do the included sources address the user's question and context (e.g. the	Report only

		population, setting and/or decision being informed)?	
A2	Appropriateness of evidence base	Are the included sources of an appropriate study type/design to answer the question? Is the rationale for evidence selection apparent? **Note.** for some queries, (a single large, definitive RCT may suffice).	Report only
A3	Contextual specificity of populations, settings and locations	Is the evidence relevant to the population you are making a decision about, and their setting? (e.g. sex, age groups, urban/rural populations, country/region)?	Report only
A4.1	Recency of evidence	Does the evidence in the report include recent sources, where recency matters for the question?	Report only
A4.2	Timeliness of evidence	Is the time period covered by the evidence appropriate for the question?	Report only
A5	Basis on systematic sources	Are the key conclusions consistent with, or explicitly drawn on, systematic sources, e.g. a recent systematic review or national/international guidelines (e.g. WHO)?	Deeper - expertise/inspectability of methods

<u>Domain B - Evidence Quality</u>

For item B4, please verify that a quality/risk-of-bias assessment has genuinely been performed, where one is reported, requested or warranted by the question type, by checking whether the methods and results of the appraisal are reported with enough detail to be inspected and confirmed, rather than simply claimed.
It is important to recognise that most AI-generated reports will not include a formal quality/risk-of-bias assessment, especially if it is not requested in the prompt.

Item number	Item	What to assess	Assessment

B1	Source quality discussed	Does the AI-generated report discuss the quality/risk-of-bias and limitations of sources?	Report only
B2	Conclusions reflect source quality	Is the assessed quality/risk-of-bias reported and reflected in the strength of the conclusions?	Report only
B3	Weighting of evidence	Were sources weighted or ranked by quality, avoiding poor proxies (e.g. impact factor, citation count)?	Report only
B4	Quality assessment genuinely performed	Does it appear that a quality/risk-of-bias assessment has been performed?	Deeper/inspectability of methods

Domain C - Citations & Fidelity of Data to the Source

For domain C, please take your random selection of five source documents from the reference list, and:

1. **verify them against their original sources - confirming that each source genuinely exists and that it accurately supports what the report claims**
2. **try to obtain the original documents, and check how the information in the report compares to the original source**

Item number	**Item**	**What to assess**	**Assessment**
C1 stepwise	C1.1. In-text citations correspond to reference list	C1.1. Does each in-text citation correspond to a source in the reference list?	Deeper - cross-checking
	C1.2. References exist and retrievable	C1.2. Can the original source be located and retrieved?	
	C1.3. Supported claim	C1.3. Are they key claims genuinely supported by cited evidence, without misquoting, overstating or misrepresenting it?	

C2	Unsupported statements	Are any key empirical claims, findings, or conclusions presented without evidential support?	Deeper - cross-checking
C3	Extraction accuracy	Are key study characteristics (e.g. number of sources included, sample size, study design, population), quality-relevant data, results and conclusions reported and accurately represented without error?	Deeper - cross-checking
C4	Extraction completeness	Are key study characteristics (e.g. number of sources included, sample size, study design, population), quality-relevant data, results and conclusions described in sufficient detail?	Deeper - cross-checking

Domain D - Synthesis & Conclusions

Item number	**Item**	**What to assess**	**Assessment**
D1	Answers the question	Does the synthesis answer the question or problem that needs to be addressed?	Report only
D2	Heterogeneity and uncertainty	Does the AI-generated report discuss conflicting findings, outliers and counterfactuals?	Report only
D3	Comparability of synthesised sources	Are the settings and populations across the synthesised sources comparable enough to combine?	Deeper - cross-checking to confirm comparability?
D4	Conclusions match the evidence	Do the conclusions directly follow from the synthesised sources, free of spin and optimism?	Deeper - cross-checking

Domain E - Structure & Usability

Item	**Item**	**What to assess**	**Assessment**

number			
E1	Process transparency	Does the AI-generated report clearly outline the origin of the data reported, and what was included/excluded and why?	Report only
E2	Structure and proportionality	Is the AI-generated report appropriately structured, easy to navigate, and proportionate in length for the use case?	Report only (informed by intended user and use case)
E3	Clarity of language	Is the language clear, accessible and suited to the audience, avoiding needless jargon and emphatic and emotive statements?	Report only
E4	Date and temporal transparency	Does the AI-generated report state when the search was run, the period covered by the included evidence, and the date/version of the tool/model used?	Report only

Rating scale:
Yes (Y)
Probably Yes (PY)
No (N)
Probably No (PN)
No Information (NI)