DRAFT - November 2026

DEEPER Framework - Guidance Document

Deep R**e**search **E**valuation **P**rotocol for **E**vidence **R**eviews (DEEPER)

About DEEPER

DEEPER is a framework for evaluating the quality of AI-generated deep research reports. The table below gives an overview of key DEEPER features.

Structure	The tool uses signalling questions to reach a judgement for each domain. The tool has five domains leading to overall quality assessment for each report.
Domains	The framework comprises of five domains covering the full end-to-end evidence synthesis pathway: ● Domain 1: Selection of Evidence ● Domain 2: Evidence Quality ● Domain 3: Citations & Fidelity of Data to Source ● Domain 4: Synthesis and Conclusions ● Domain 5: Structure and Usability
Basis of judgement	Signalling questions answered: 1. Yes 2. Probably yes 3. Probably no 4. No 5. No information For simplicity and clarity, signalling questions are phrased such that a response if 'Yes' may be indicative of either a low or high concern, depending on the most natural way to ask the question. Responses of 'Yes' and 'Probably yes' have the same implications for quality, as do responses of 'No' and 'Probably no'. The definitive versions (i.e. 'Yes' and 'No') would typically imply firm evidence is available in relation to the signalling question. When calculating measures of agreement for answers to signalling questions, we recommend treating 'Yes' and 'Probably yes' as the same response, and 'No' and 'Probably no' as the same response. The 'No information' response should be used only when insufficient details are reported to permit a response of 'Probably yes' or 'Probably no'.

Judgement options	Responses to individual items provide the basis for domain-level judgements. Domain-level judgements form the basis of an overall quality judgement for the specific report being assessed. The possible judgements are: 1. Low concern 2. Some concerns 3. High concern Domain-level judgements have the same implication for each of the five domains with respect to concern about the quality of the report. Judging a report to be at a particular level of concern for an individual domain implies that the report has an overall concern at least this severe. Therefore, a judgement of 'High concern' within any domain should have similar implications for the overall judgement, irrespective of which domain is being assessed. 'Some concerns' in multiple domains may lead to an overall judgement of 'High concern'.
Decision	Users are advised not to use information from reports graded as 'High concern' for decision-making. However, some potential mitigations to improve quality will be provided at the end of the assessment.

DEEPER Framework

Preliminary Considerations

Report and Query

Original prompt: _____
AI system, model and version: _____
Date run: _____

Question type

Specify the kind of question being asked (e.g. effectiveness, diagnostic accuracy, prevalence, mechanism, policy etc) and whether it is broad/exploratory or narrow/specific.

_____

What matters for this question

Specify what you need from the answer (e.g. population, setting, comparator, or decision being informed).

_____

User and intended use

Which user type best describes you?

- ☐ Academic/researcher
- ☐ Policy-maker or decision-maker
- ☐ Practitioner (e.g. clinician, humanitarian worker)
- ☐ General public/lay audience
- ☐ Other

Please indicate your AI/evidence competence level.

- ☐ Beginner
- ☐ Intermediate
- ☐ Advanced

How will the AI-generated report be applied (e.g. to directly inform a decision, or as a starting point for future reading)?

_____

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	**Signalling Questions**	**Elaboration**
A1	Relevance to the question and context	Do the included sources match the user's specific question and context (e.g. the population, setting and/or decision being informed)?	Consider the population, setting and/or decision recorded in the preliminary considerations (i.e. 'what matters for this question'). Answer 'Yes' if the included sources are directly relevant to that context. Answer 'No' if sources are only tangentially related, or answer a different question than the one asked.
A2	Appropriateness of the evidence base	Are the included sources of an appropriate study type/design to answer the question?	The appropriate design depends on the question type (e.g. RCTs for effectiveness questions, qualitative methods for experience-based questions). Answer 'Yes' if the report draws on study designs (e.g. RCTs, qualitative, observational studies) that are appropriate for the question, and excludes inappropriate designs (e.g. retracted studies, or designs with no bearing on the question. Answer 'No' if the study designs included are clearly mismatched to the question type, if inappropriate designs are included without justification, or if it is unclear why some

			obviously relevant sources are missing while others are included.
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)?	Consider the populations, settings and locations, as relevant to the question. Answer 'No' if the evidence is drawn from populations, settings or locations that differ significantly from the context of the question, without the mismatch being recognised.
A4.1	Recency of evidence	Does the evidence in the report include recent sources, where recency matters for the question?	Consider whether the question concerns something that changes over time (e.g. a new intervention, changing policy and practice, an emerging health threat, or changing climate conditions). For some questions, older but definitive evidence remains appropriate. Answer 'Yes' if the report prioritises recent evidence where recency plausibly matters for the question. Answer 'No' if the report relies mainly on older evidence where more recent evidence is likely to matter and exist. Answer 'No information' if the dates of the included sources cannot be established.
A4.2	Timeliness of evidence	Is the time period covered by the evidence appropriate for the question?	Answer 'Yes' if the report states the period covered by its evidence, and that period is appropriate to the question. Answer 'No' if the period covered is clearly too narrow or too dated for the question (e.g. it omits years in which key evidence is likely to have been published). Answer 'No information' if the dates of the

			included sources cannot be established.
A5	Basis on systematic sources	Are the key conclusions of the report based on systematic sources - for example recent systematic review or national/international guidelines (e.g. WHO)?	Where a recent systematic review/meta-analysis, or national/international guidelines (e.g. from WHO) exist on the topic, answer 'Yes' if the report's key conclusions are consistent with, or explicitly draw on, that systematic source. Answer 'No' if the key conclusions conflict with an identifiable systematic source, or fail to draw on an existing relevant systematic source, without justification. Answer 'No information' if it cannot be established whether a systematic source exists for this question.

Domain-level judgement	**Criteria**
Low concern	The evidence addresses the user's question, is grounded in systematic sources where these exist, is of an appropriate study type/design, and is specific in population, setting and temporal coverage (A1-A5 all Yes/Probably Yes).
Some concerns	One or more of A2-A5 is answered No/Probably No, but the gap is disclosed and limited in effect. For example, no systematic review or guideline exists on the topic and this is acknowledged (A5), the population/setting or temporal mismatch is minor or transparently flagged (A3/A4), or the study-design limitation reflects a genuinely sparse evidence base rather than a selection failure (A2), such that the report still substantively addresses the user's question (A1 remains Yes/Probably Yes).
High concern	The report does not address the user's question, its key conclusions conflict with or disregard systematic evidence sources that are identifiable, the evidence type/design is fundamentally inappropriate, or sources are not relevant to the

	question asked.

Domain B: Evidence Quality

Note. ***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	Signalling Questions	Elaboration
B1	Source quality discussed	Does the AI-generated report discuss the quality/risk-of-bias and limitations of sources?	Answer 'Yes' if the report explicitly discusses the quality, risk-of-bias, or limitations of the sources it has drawn on. A formal appraisal tool is not required for a 'Yes'. Answer 'No' if the report does not discuss the quality or limitations of its sources, or does so only in generic terms.
B2	Conclusions reflect source quality	Is the assessed quality/risk-of-bias reported and reflected in the strength of the conclusions?	Answer 'Yes' if conclusions drawn from lower-quality or higher risk-of-bias evidence are appropriately presented (e.g. limited evidence suggests…), and conclusions from higher-quality evidence are stated with corresponding confidence. Answer 'No' if the strength of language used is unrelated to, or inconsistent with, the quality and limitations described elsewhere in the same report.
B3	Weighting of evidence	Were sources weighted or ranked according to quality, avoiding poor	Answer 'Yes' if higher-quality or more directly relevant sources are visibly given more weight in the report than lower-quality ones.

		proxies (e.g. impact factor, citation count)?	Answer ‘No’ if all sources appear to be treated as equally informative regardless of their quality, or if quality is equated to poor proxies such as journal impact factor or citation count.
B4	Quality assessment genuinely performed	Does it appear that quality/risk-of-bias assessment has been performed?	Please verify that a quality/risk-of-bias assessment has genuinely been performed by checking whether the methods and results of the appraisal are reported with enough detail to be inspected and confirmed, rather than simply claimed. Answer ‘Yes’ if the quality/risk-of-bias assessment is described with enough methodological detail (e.g. a named tool or criteria, reported results per source) to be verified. Answer ‘No’ if the report claims a quality/risk-of-bias assessment was performed without reporting the method or results in enough detail to inspect, or if an assessment was requested in the prompt or is warranted by the question type but does not appear to have been performed.

Domain-level judgement	**Criteria**
Low concern	The quality and limitations of the evidence are discussed appropriately for the query, carried through to the conclusions, and reflected in how sources are weighted, and any formal appraisal that is reported, requested or warranted has genuinely been performed (B1-B4 all Yes/Probably Yes).

Some concerns	A quality/risk-of-bias assessment is genuinely performed (B4 = Yes/Probably Yes), but is not fully tailored to the question type (B1), only partially carried through to the strength of conclusions (B2), or only partially reflected in how sources are weighted (B3), provided the shortfall does not affect the report's key claims/conclusions
High concern	The report does not discuss the quality or limitations of its sources, the strength of its conclusions is inconsistent with the quality of the evidence, sources are not weighted by quality, or are weighted by poor proxies, and no indication that a quality/risk-of-bias assessment has genuinely been performed (B1-B4 all No/Probably No).

Domain C: Citations & Fidelity of Data to the Source

Note. *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	**Signalling Questions**	**Elaboration**
C1.1	Citations correspond to the reference list	Does each in-text citation correspond to a source in the reference list?	Check a random sample of five in-text citations against the reference list. Answer 'No' if any in-text citation has no corresponding entry in the reference list.
C1.2	References exist and are retrievable	Can the original source be located and retrieved?	Verify that each reference corresponds to a real, findable source. Answer 'No' if any reference cannot be located or appears fabricated ('hallucinated').

C1.3	Supported claim	Are they key claims genuinely supported by cited evidence, without misquoting, overstating or misrepresenting it?	Answer ‘No’ if, on inspection, cited evidence does not support the specific claim attributed to it (e.g. cases where the source discusses a related but distinct point, or reaches the opposite conclusion).
C2	Unsupported statements	Are any key empirical claims, findings, or conclusions presented without evidential support?	Answer ‘Yes’ if important claims in the report have no citations at all, or are attributed only to the model’s general knowledge where a citation to a verifiable source would be expected.
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?	Cross-check the study characteristics, quality-relevant data, results and conclusions reported, using the random sample of five source documents described above. Answer ‘No’ if any of these are misinterpreted (e.g. result quoted with the wrong direction, magnitude or statistical significance, or a source’s design is mischaracterised).
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?	Answer ‘Yes’ if the level of detail given for each source is sufficient to understand what was found and how much weight to place on it. Answer ‘No’ if key details are omitted such that it is not possible to judge the sources relevance or quality.

Domain-level judgement	**Criteria**
Low concern	Citations and references are real, retrievable and supportive of the claims made,

	and no key claims are left unsupported (C1.1-C1.3 = Yes/Probably Yes, and C2 = No/Probably No). Data extraction is accurate and sufficiently complete for every source (C3-C4 = Yes/Probably Yes).
Some concerns	Citation or reference metadata is imprecise, or non-key claims lack citations (C2 = Yes/Probably Yes), but no citation is hallucinated and no cited source is found to contradict or misrepresent the claim attributed to it (C1.1-C1.3 = remain Yes/Probably Yes for all key claims). Extraction is accurate for the sampled sources (C3 = Yes/Probably Yes) but incomplete or imprecise in places (C4 = No/Probably No) (i.e. the discrepancies found are omissions of detail rather than misrepresentations of direction, magnitude or significance).
High concern	One or more hallucinated citations or references (C1.1-C1.2 = No/Probably No), or for a key claim, a cited source that does not support the claim made against it (C1.3 = No/Probably No) or no attributable evidential support (C2 = Yes/Probably Yes). Data reported in the output does not correspond to the cited sources (e.g. C3 = No/Probably No for one or more key claims).

Domain D: Synthesis and Conclusions

	Item	Signalling Questions	Elaboration
D1	Answers the question	Does the synthesis answer the question or problem that needs to be addressed?	Compare the report's stated conclusion(s) directly against the original prompt recorded in the preliminary considerations. Answer 'No' if the report answers a narrower, broader, or subtly different question than the one asked.
D2	Heterogeneity and uncertainty	Does the AI-generated report discuss conflicting findings, outliers and counterfactuals?	Answer 'Yes' if the report acknowledges disagreement or heterogeneity across sources where it exists, rather than presenting a falsely unanimous picture. Answer 'No' if conflicting evidence known to exist is omitted from the report without explanation.
D3	Comparability of synthesised sources	Are the settings and populations across the synthesised sources comparable enough to combine?	Answer 'No' if sources with materially different populations, settings, and comparators or outcome definitions have been combined or discussed together as if directly comparable, without this limitation being acknowledged.
D4	Conclusions match the evidence	Do the conclusions follow directly from the synthesised sources, free of spin and optimism?	Answer 'No' if the stated conclusion overstates what the underlying sources support (e.g. describing 'strong evidence' for a claim supported only by a single small or low-quality study), or is more favourable/confident than the evidence warrants.

Domain-level judgement	Criteria
Low concern	The synthesis directly answers the question, combines only comparable sources, engages with the context of the question, and reports heterogeneity; with conclusions that follow from sources (D1-D4 all Yes/Probably Yes).
Some concerns	One or more of D1-D3 is answered No/Probably No (e.g. a partial shortfall in comparability of sources, or discussion of heterogeneity) but the report's conclusions still directly follow from the synthesised sources (D4 remains Yes/Probably Yes), so the core report is not misleading even though the supporting synthesis process has gaps.
High concern	Conclusions reported in the AI-generated report do not follow from the synthesised sources (D4 = No/Probably No).

Domain E: Structure and Usability

	Item	Signalling Questions	Elaboration
E1	Process transparency	Does the AI-generated report clearly outline the origin of the data reported, and what was included/excluded and why?	Answer ‘Yes’ if the report distinguishes, where relevant, between information drawn from different evidence origins, and explains the basis for including/excluding evidence.
E2	Structure and proportionality	Is the AI-generated report well structured, easy to navigate, and proportionate in length for the use case?	Refer back to what was recorded in the preliminary considerations (i.e. ‘user and intended use’). Answer ‘No’ if the report is very long relative to a sparse evidence base, or poorly organised for the user.
E3	Clarity of language	Is the language clear, accessible and suited to the audience, avoiding needless jargon and emphatic and emotive statements?	Answer ‘Yes’ if the findings are clear and accessible to the intended user. Answer ‘No’ if the report obscures the findings (e.g. key results are hard to locate amid rhetorical framing), uses jargon inappropriate to the user, or uses emphatic or emotive language that conveys more certainty or importance than the evidence supports.
E4	Date and temporal transparency	Does the AI-generated report state when the search was run, the period the evidence covers, and the date/version of the tool/model?	Answer ‘Yes’ only if all stated (search date, evidence period, tool/model version). Answer ‘Probably Yes’ if some but not all are given.

Domain-level judgement	**Criteria**

Low concern	Temporal and process transparency are complete, and structure/language are appropriate and proportionate for the user (E1-E4 = all Yes/Probably Yes).
Some concerns	One or more of E1-E4 is answered No/Probably No (e.g. some but not all temporal/process details are given (E2/E4), or the structure/language is imperfect for the audience without actively obstructing use (E2/E3)), such that an expert user could still extract and correctly interpret the report's content.
High concern	Substantial failures of transparency (e.g. no indication at all of when the report was produced or what evidence period it covers) combined with structure/language that actively impedes appropriate use by the user.