Transparent AI Tool Review Rubric

A reproducible, evidence-first framework for comparing AI tools without fabricated results, hidden scoring or pay-to-play influence.

Prepared by
Danny Rasul
AI News & Updates
hello@ainewsandupdates.com

Included in this edition	Review principles Seven-part scorecard Five benchmark test cases Standard research prompt Evidence and disclosure rules

Companion editable workbook: AI News & Updates - Transparent AI Tool Review Rubric.xlsx

No benchmark scores are included in this blank protocol. Results must come from dated, preserved observations.

1. Review principles

The goal is not to make every product look comparable. The goal is to make every conclusion traceable to the same disclosed process.

Principle	Rule
Test before scoring	No product receives a score without a dated observation and supporting evidence.
Keep tasks identical	Use the same prompt, eligibility criteria, trial count, plan tier and time limit across products.
Separate evidence types	Distinguish hands-on tests, vendor documentation, third-party evidence, inference and unverified claims.
Preserve failure	Record outages, usage caps, broken citations, lost state and failed exports. Do not silently retry until the tool looks good.
Verify material claims	A human checks citation identity, relevance and support before publication.
Disclose influence	Vendor access, corrections, affiliate relationships and sponsorship are logged; none may alter the protocol or score.
Correct visibly	Material corrections are dated, explained and reflected in the public update log.

Evidence labels

Label	Meaning
Hands-on test	Observed directly; preserve export or screenshot.
Vendor documentation	Supplier claim with current URL and access date.
Third-party evidence	Independent source with author, publication and URL.
Inference	Reasoned conclusion explicitly labelled as inference.
Unverified	Could not be corroborated; never presented as fact.

2. Seven-part scorecard

Each criterion is scored from 0 to 10. The weighted score is calculated only from completed observations. Blank means untested, not zero.

Criterion	Weight	A high score means	Guardrail
Task Completion	25%	Completes the benchmark accurately within stated constraints.	Polished prose alone does not earn points.
Evidence Traceability	20%	Material claims link to matching, accessible sources.	Broken or mismatched citations cap the score.
Reliability & Recovery	15%	Behaves consistently, preserves state and recovers clearly.	Log outages and usage caps separately.
Privacy & Data Handling	10%	Provides clear controls and retention terms for uploaded material.	Score only documented and tested controls.
Security & Control	10%	Offers suitable account, sharing and administration controls.	Do not assume enterprise controls exist on free plans.
Cost Value	10%	Produces useful, verifiable outcomes at transparent total cost.	Record currency, billing period and limits.
Maintenance	10%	Exports cleanly, communicates changes and needs manageable upkeep.	Include correction and update burden.

Scoring anchors

Score	Interpretation
0-2	Fails or produces materially unreliable output.
3-4	Partially works but needs significant correction or manual rescue.
5-6	Completes the core task with meaningful limitations.
7-8	Strong, verifiable performance with manageable limitations.
9-10	Excellent, repeatable performance with clear controls and evidence.

Weighted score formula: sum of each completed criterion score multiplied by its published weight. Publish no ranked winner until every product has equivalent completed trials.

3. Benchmark test cases

Run the cases in order. Preserve evidence immediately after each case so later tool state cannot overwrite the observation.

ID	Test	Task	Success criteria	Min
T01	Search coverage	Run the standard prompt and collect candidate studies.	Traceable, plausibly relevant records; identifiers resolve; duplicates are identifiable.	8
T02	Evidence traceability	Trace five material claims or summaries to their cited sources.	Every sampled claim links to a matching paper whose content supports it.	6
T03	Screening consistency	Apply eligibility rules to ten borderline records.	Decisions are reproducible, explained and consistent with the protocol.	6
T04	Synthesis and disagreement	Request separate supporting, null and conflicting findings.	Uncertainty is surfaced without flattening results into one unsupported conclusion.	6
T05	Recovery and export	Correct one constraint, recover from one failed action and export.	Correction applies; state remains clear; export preserves citations and notes.	4

Required observation fields

Observation ID	Product
Test case	Trial
Test date	Plan / mode
Status	Seven criterion scores
Time	Cost
Failure?	Evidence URL or export
Observation notes	Reviewer

Publication gate

Do not publish a ranked winner until every product has the same number of completed test cases and trials, material citation checks are human-verified, and account-tier limitations are disclosed.

4. Research-assistant benchmark

Products: ResearchRabbit, Elicit, Scite, Consensus and Undermind

Field	Specification
Question	For university students, does retrieval practice improve delayed retention compared with rereading?
Population	Students enrolled in higher education.
Intervention	Retrieval practice, practice testing or active recall.
Comparator	Rereading or restudy of the same material.
Outcome	Delayed retention measured at least 24 hours after learning.
Eligible studies	Randomised or quasi-experimental direct comparisons; peer-reviewed full papers or traceable preprints.
Trials	At least three independent runs per product when usage limits permit.
Time limit	30 minutes per full product run, excluding account setup and vendor outages.
Stop rule	Time limit, documented usage limit or completion of all five test cases.

Standard prompt

Find studies that directly compare retrieval practice with rereading in higher-education students and measure retention at least 24 hours later. Identify the study design, sample, delay, outcome and direction of effect. Include persistent identifiers or direct paper links. Separate claims supported by a cited paper from your own synthesis, surface conflicting findings, and state what may have been missed.

Controls

Control	Requirement
Account tier	Record exact plan and promotional access.
Model / mode	Record model, search or agent mode and material settings.
Clean start	Begin each trial in a new project, collection or conversation.
Input parity	Use the same prompt and approved seeds, if required.
Duplicate control	Flag duplicates and never count them twice.
Human screening	Verify relevance and identifiers before any gold-set claim.

5. Disclosure, corrections and publication checklist

Done	Publication check
[ ]	Every published score maps to at least one completed, dated observation.
[ ]	Every material product claim is labelled by evidence type.
[ ]	Five sampled AI-generated claims per product have been traced to matching sources.
[ ]	Plan, mode, usage limits, currency and test date are disclosed.
[ ]	Vendor contact or correction requests are logged without granting scoring influence.
[ ]	Affiliate links, free access, sponsorship and paid placement are disclosed.
[ ]	Outages, retries and failed exports remain in the record.
[ ]	Comparisons use equivalent trials and the published weights total 100%.
[ ]	A named human reviewer has checked relevance, identifiers and publication wording.
[ ]	Material corrections include a date, explanation and affected conclusion.

Correction log fields

Field	Entry
Date	
Page or review	
Original wording or score	
Corrected wording or score	
Reason and evidence	
Editorial impact	

Editorial independence statement

No paid placement, affiliate relationship or vendor correction may alter the protocol or score. Vendor feedback can correct factual product details, but the test record and editorial judgement remain independent.

Contact: hello@ainewsandupdates.com **Website:** https://ainewsandupdates.com/

This PDF is the reference edition. Use the companion Excel workbook for live observations and formula-driven summaries.