

FREE READINESS GUIDE

AI Evaluator Readiness Guide

A practical introduction
to human evaluation of
AI responses

Requirements → Evidence → Severity → Calibration

Original educational material. Not an official qualification test and not affiliated with any hiring platform.

How to use this guide

This guide introduces a repeatable evaluation workflow. Work through the original exercise, compare your notes with the answer key, and record where your judgment differed. The goal is not to memorize labels. It is to make decisions that another reviewer can understand and audit.

The four-part AI Work Prep framework

Step	Question	Output
1. Requirements	What did the task explicitly require?	A short constraint list
2. Evidence	Which exact words satisfy or violate each requirement?	Quoted or located evidence
3. Severity	How does the issue change usefulness, truth, or safety?	Minor, material, or critical impact
4. Calibration	Would I judge a similar response the same way tomorrow?	A consistent, reviewable decision

Readiness is more than AI knowledge

General evaluation work usually rewards careful reading, factual discipline, comparison, and concise writing. Specialist projects can additionally require verified expertise in fields such as finance, medicine, law, software, mathematics, or language.

1. Requirements: read before you judge

Convert the prompt into a checklist before reading the answer. Separate explicit constraints from preferences you personally would add. Project instructions always outrank a general method like this guide.

Requirement extraction worksheet

- ☐ What outcome must the response deliver?
- ☐ Which facts, sources, or materials may it use?
- ☐ What format, length, tone, or audience is required?
- ☐ Which actions or claims are prohibited?
- ☐ Which ambiguity would materially change the judgment?

Quick check

If you cannot state the task in one or two sentences, do not rate the response yet. Re-read the prompt and identify the decision you are actually being asked to make.

2. Evidence: show the reason

A verdict without evidence is difficult to review. Point to a specific claim, omission, contradiction, or formatting choice, connect it to a criterion, and explain the practical effect.

Weak feedback	Defensible feedback
B is more helpful.	B answers the requested recommendation in its first sentence and states the assumption behind it. A provides background but never makes the requested choice.
A is inaccurate.	A states that the deadline is Friday, while the supplied schedule lists Thursday. That changes the action date and is therefore material.

Rationale template

Verdict: [which response is stronger or which rating applies]

Criterion: [the relevant requirement]

Evidence: [the exact feature, claim, or omission]

Impact: [why it changes quality, usefulness, truth, or safety]

3. Severity: judge impact, not annoyance

Level	Working definition	Example
Minor	A real defect with limited impact on the user's decision or outcome.	A small formatting miss that leaves the answer usable.
Material	A defect that changes understanding, completeness, or the requested result.	The response omits one of the user's required comparisons.
Critical	A defect that could cause serious harm or defeats the central task.	Unsafe instructions or a fabricated source used as decisive evidence.

Severity depends on context

The same factual mistake can be incidental in one task and decisive in another. Count consequences, not merely the number of errors. Follow the project's own severity definitions whenever they exist.

4. Calibration: make quality repeatable

Calibration means aligning your decisions with the applicable rubric and using similar standards for similar cases. It does not mean suppressing uncertainty or assuming a reference answer is infallible.

Decision log

Date/task	Edge case	Decision and evidence	Feedback/change

Quality reset

- ☐ Pause if you repeatedly re-read the same instruction.
- ☐ Slow down if you have a preference but cannot state the evidence.
- ☐ Recheck the rubric after several similar tasks or after feedback.
- ☐ Stop the session when fatigue causes skipped verification steps.

Original practice exercise

User request: Recommend one of two meeting tools for a five-person nonprofit. Use only the supplied facts, give the recommendation first, and explain it in no more than 90 words.

Facts: Tool North costs \$8 per user monthly and includes live captions. Tool South costs \$5 per user monthly and does not include live captions. The nonprofit requires captions for one participant and has a monthly budget of \$45.

Response A

Tool North is the better choice. It costs \$40 per month for five users, stays within the \$45 budget, and includes the required live captions.

Response B

Both tools are popular options. Tool South is cheaper at \$25 per month, which could help a nonprofit save money. Tool North has additional accessibility features. The team should compare interfaces and consider a free trial before deciding.

Your evaluation

- ☐ List the explicit requirements.
- ☐ Choose the stronger response.
- ☐ Identify the highest-impact difference.
- ☐ Write a four-part rationale.

Answer key and self-assessment

Stronger response: A. It recommends first, uses only supplied facts, remains under 90 words, satisfies the \$45 budget, and preserves the required captions. Response B emphasizes the cheaper option but does not make the requested recommendation and leaves the accessibility requirement unresolved.

Readiness checklist

- ☐ I separate explicit requirements from personal preferences.
- ☐ I can locate evidence for each important judgment.
- ☐ I distinguish a minor defect from one that changes the outcome.
- ☐ I can explain uncertainty without hiding it.
- ☐ I revise my process when reliable feedback shows a mismatch.
- ☐ I stop or reset when fatigue causes skipped steps.

Next step

Practice additional original comparisons at aiworkprep.com/practice. Preparation does not guarantee qualification, project access, hours, or earnings.

Sources and further reading

OpenAI, Evaluation best practices: platform.openai.com/docs/guides/evaluation-best-practices

FTC, Remote job scam guidance: consumer.ftc.gov/consumer-alerts/2023/11/searching-job-work-remotely-avoid-scams-identity-theft

Deel, Global Hiring Report 2026: deel.com/global-hiring-report-2026/