Design a professional AI evaluation

Use original records and a separate scoring key. This example tests whether a model can apply a date rule and cite the evidence.

Fictional evidence	Prompt and expected answer	What to score
Rule V2 requires a valid certificate on the activation date.	Which rule applies? Use the activation date under Rule V2.	Identifies the rule version and governing date.
ORCHID activation is June 10. Its certificate expires June 7.	Can activation proceed? Keep it pending. Cite both records.	Correct date comparison, status and citations.
A replacement was requested. No replacement is in the file.	What evidence is missing? Request a certificate covering June 10.	Separates a request from a valid replacement.

Plan your task and scoring

Role, decision and allowed evidence _____

Expected answer and exact source _____

Unresolved case and reviewer criterion _____

Independent reviewer and held-out variation _____

Before sharing

Keep the scoring key and held-out cases separate from training inputs. Author your own documents. Confirm the permitted AI use and exclude client, privileged, confidential and third-party material unless appropriate authority is documented.

Scale Labs rubric research · OpenAI and Ironclad research · Describe a professional workflow