Designing and Evaluating an "LLM-as-a-Judge" Experiment

In this exercise, you will design and run a small-scale LLM-as-a-Judge experiment. You will use one language model to evaluate outputs generated by another model (or by a human baseline).

The purpose is to test:

- whether the judge model produces consistent scores
- how closely its judgments align with human evaluation
- whether prompt wording changes the evaluation outcome

Your report must be formal, evidence-based, and reproducible.

Experimental Scenario

Assume you are evaluating the quality of scientific abstracts generated by an AI system.

You are given three candidate abstracts for the same research topic.

For example:

Topic: The role of retrieval-augmented generation in scientific literature review

Use either:

- AI-generated abstracts from different models / prompts, or
- one AI-generated and one human-written abstract

Minimum: 3 candidate outputs

Part A — Design the Evaluation Rubric

Create a structured rubric that the judge model will use.

Your rubric must include at least four criteria, for example:

- factual accuracy
- clarity
- academic writing style
- coherence
- citation awareness
- relevance to topic

Each criterion must be scored on a 1–5 or 1–10 scale.

Present the rubric as a table.

Part B — Prompt the Judge Model

Design a formal evaluation prompt.

Example structure:

You are an expert academic reviewer.

Evaluate the abstract on factual accuracy, clarity, coherence, and academic style.

Provide a score from 1–5 for each criterion and a short justification.

Run the same evaluation for all candidate abstracts.

Include the full prompt in the appendix.

Part C — Human Comparison

Independently perform a human evaluation of the same outputs using the exact same rubric.
This may be:

- your own evaluation, or
- evaluation by a peer

Compare:

- human scores
- LLM judge scores
- agreement / disagreement

A comparison table is required.
Suggested format:

Output Human Score LLM Score Difference Notes

Part D — Sensitivity Experiment
Modify the judge prompt in one meaningful way.
Examples:

- change the order of criteria
- change wording
- request stricter judgment
- switch from pointwise to pairwise comparison

Re-run the experiment.
Analyze whether the scores changed.
This section tests prompt sensitivity and evaluation robustness.

Part E — Critical Analysis
Write a short discussion addressing:

1. Did the LLM judge align with human judgment?
2. Where did disagreement occur?
3. Was the judge biased toward longer or more fluent outputs?
4. Did prompt wording affect the result?
5. Would you trust this method in a real research workflow?

This is the most important section.

Submission Requirements

- 3–5 pages
- formal report structure
- rubric table required
- appendix with prompts and outputs
- PDF format

For small LLM ChatGPT-3 might be used via
https://stablediffusion.fr/chatgpt3