EvalAssist:
Insights on Task-Specific Evaluations and AI-Assisted Judgment Strategy Preferences

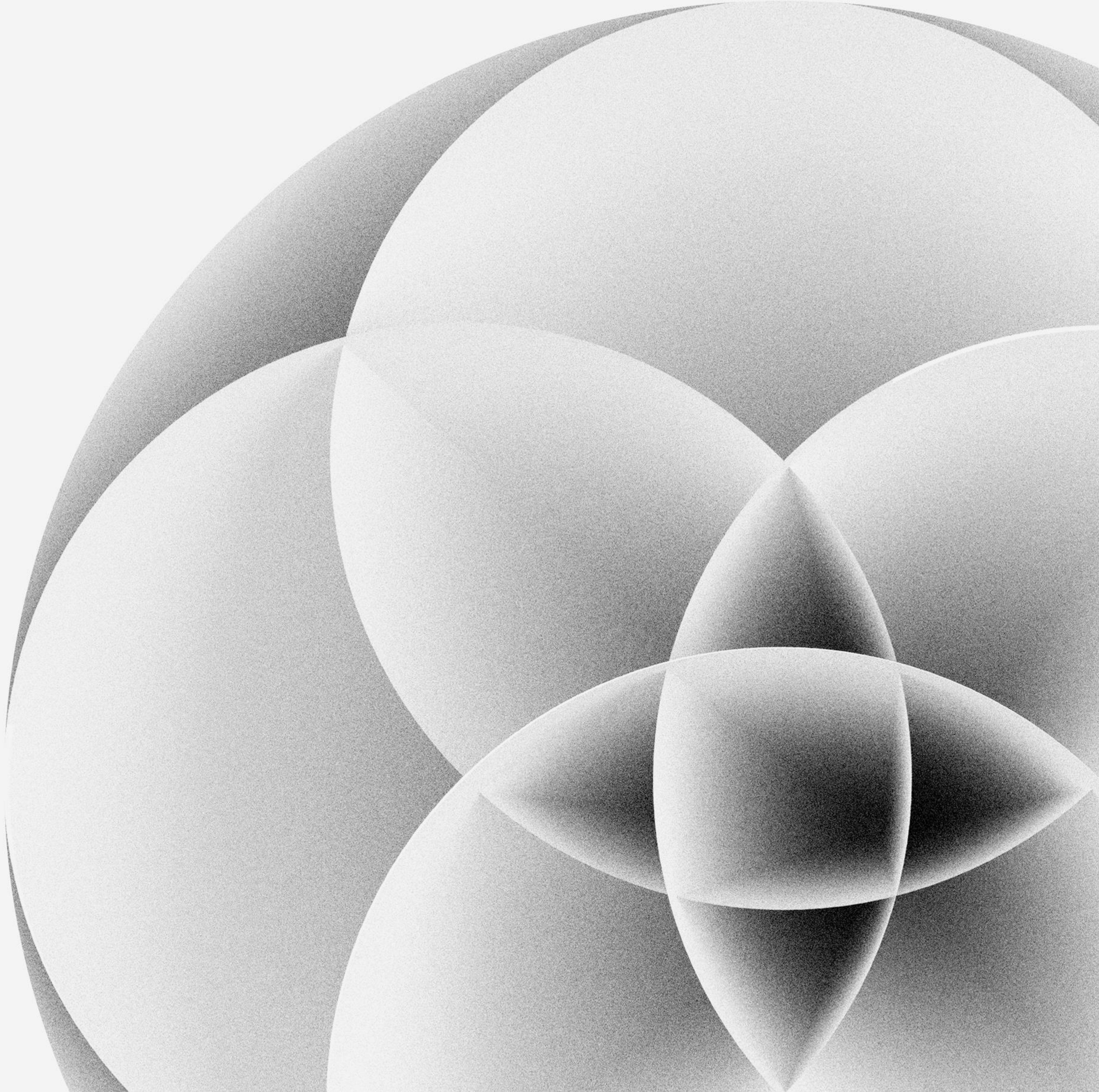
Presenter: Hyo Jin (Gina) Do (hjdo@ibm.com)

Authors: Zahra Ashktorab, Michael Desmond, Qian Pan, James M. Johnson, Martín Santillán Cooper, Elizabeth M. Daly, Rahul Nair, Tejaswini Pedapati, Hyo Jin Do, Werner Geyer

IBM Research

UIST 25

BUSAN, KOREA | SEP. 28TH – OCT. 1ST 2025



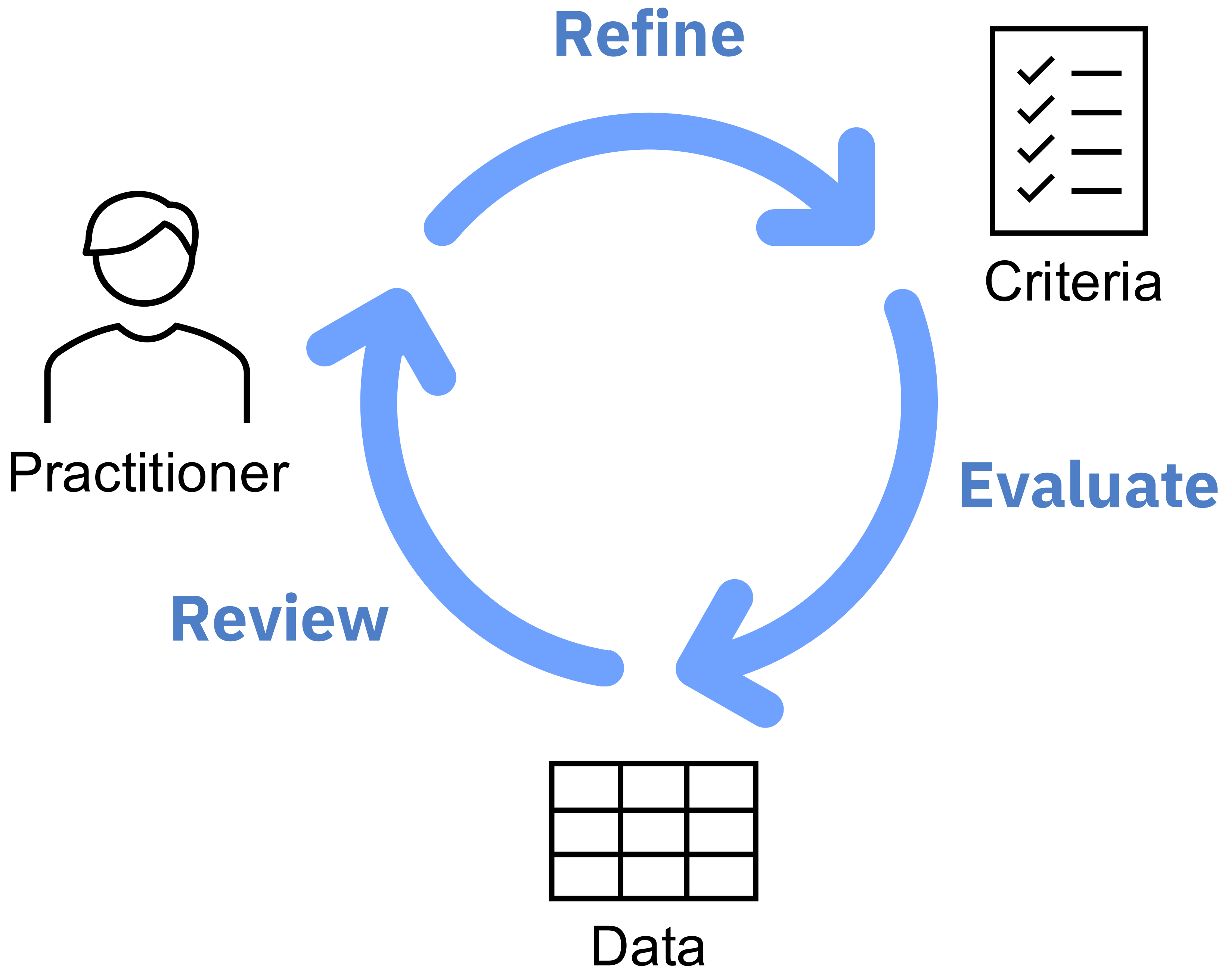
LLM-as-a-Judge

- Manual evaluation to assess LLM outputs is expensive, slow, and does not scale.
- LLMs are increasingly used as evaluators (i.e., LLM-as-a-judge).
- Offers flexibility through custom, task-specific evaluation criteria.

Credit: https://towardsdatascience.com/llm-as-a-judge-a-practical-guide/

Human-in-the-Loop Evaluation

- "Criteria Drift" (Shankar et al., 2024): While predefined criteria may help users assess outputs, the act of grading also helps users to refine those criteria.
- Many LLM-as-a-judge systems support human-in-the-loop evaluation, integrating human feedback to improve human-AI alignment.

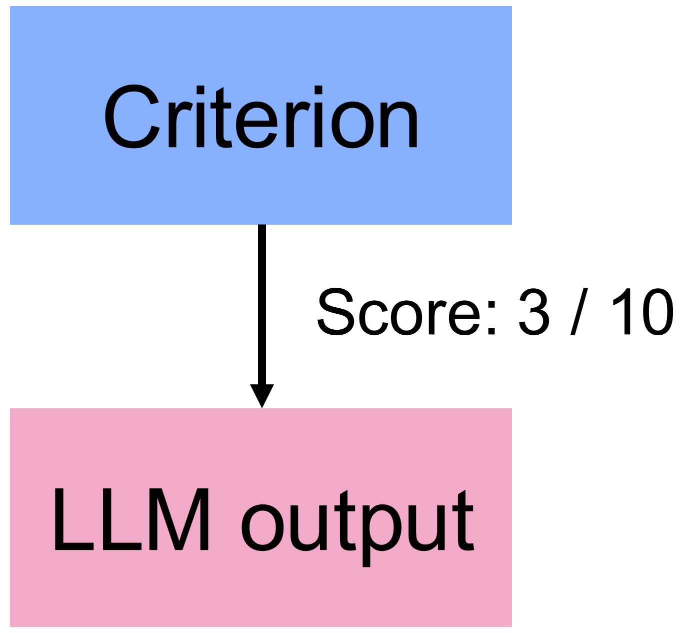


Design Goals

1. Isolate generation from evaluation
2. Evaluate a subset of data first to reduce costs
3. Support multiple LLM evaluators
4. Include a positional bias indicator (i.e., a model favors one option based on its position)
5. Support two common LLM-as-a-judge strategies:

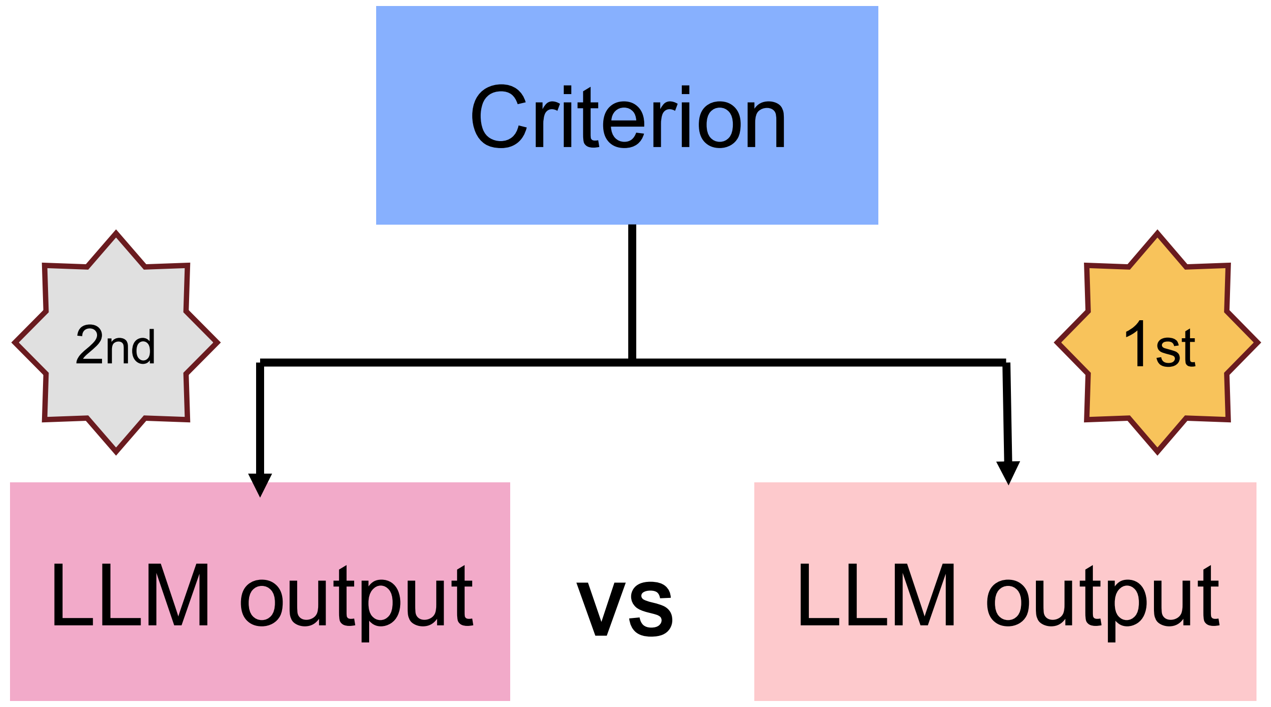
Direct Assessment

Outputs are graded based on a rubric



Pairwise Comparison

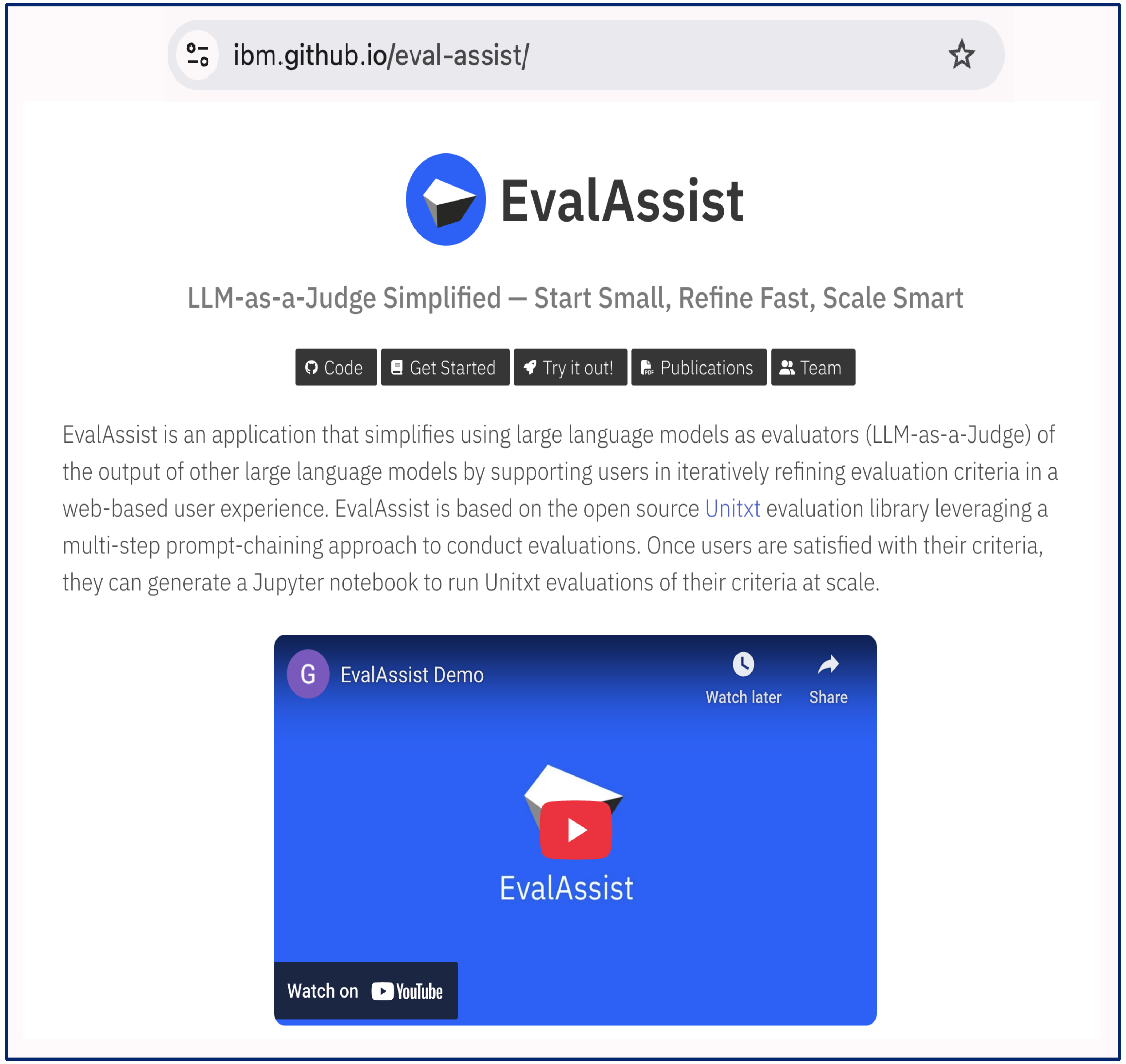
Outputs are ranked by comparing outputs



System Design

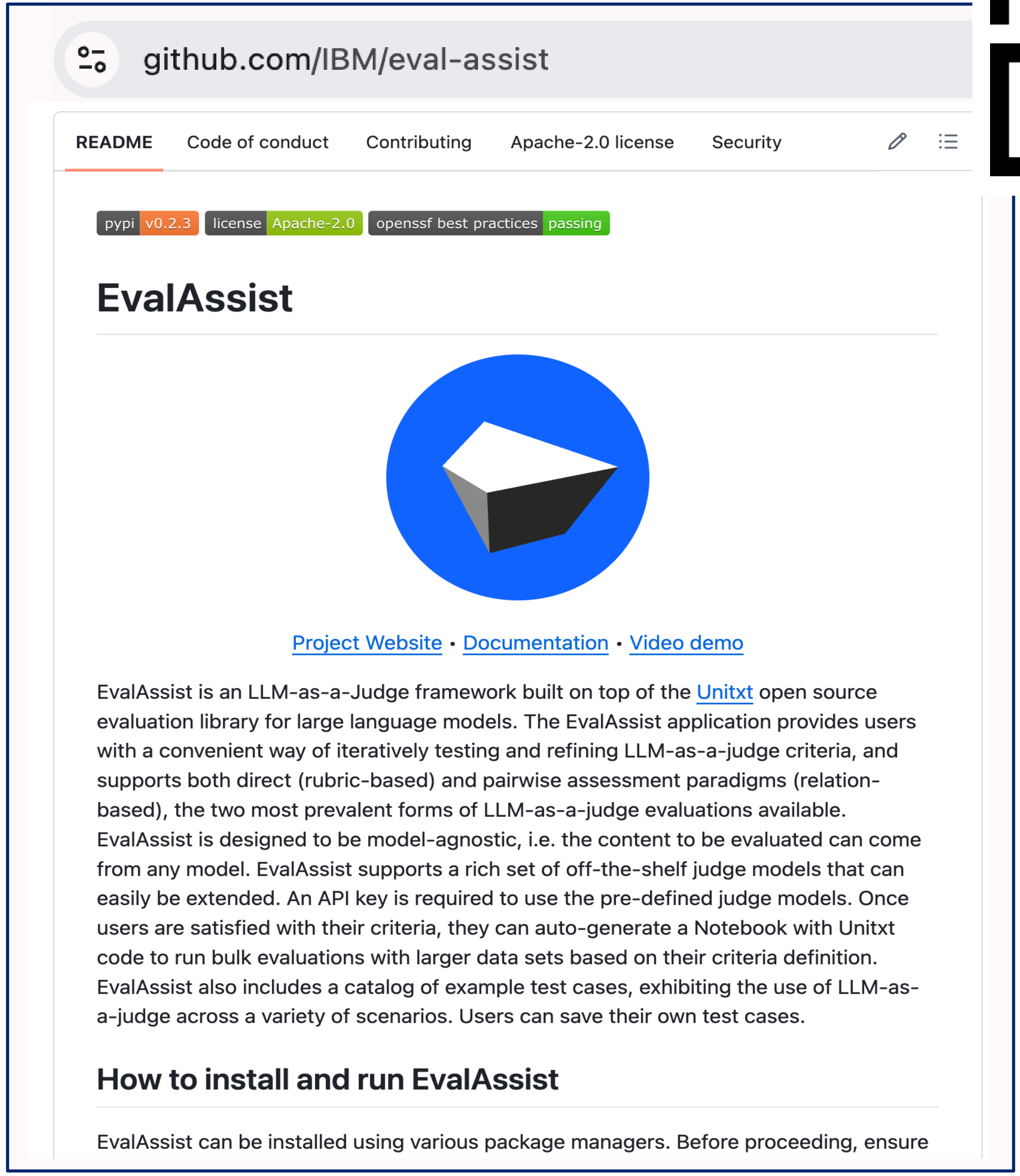
EvalAssist

Try out!



LLM-as-a-Judge Simplified — Start Small, Refine Fast, Scale Smart

EvalAssist is an application that simplifies using large language models as evaluators (LLM-as-a-Judge) of the output of other large language models by supporting users in iteratively refining evaluation criteria in a web-based user experience. EvalAssist is based on the open source Unitxt evaluation library leveraging a multi-step prompt-chaining approach to conduct evaluations. Once users are satisfied with their criteria, they can generate a Jupyter notebook to run Unitxt evaluations of their criteria at scale.



EvalAssist

Project Website • Documentation • Video demo

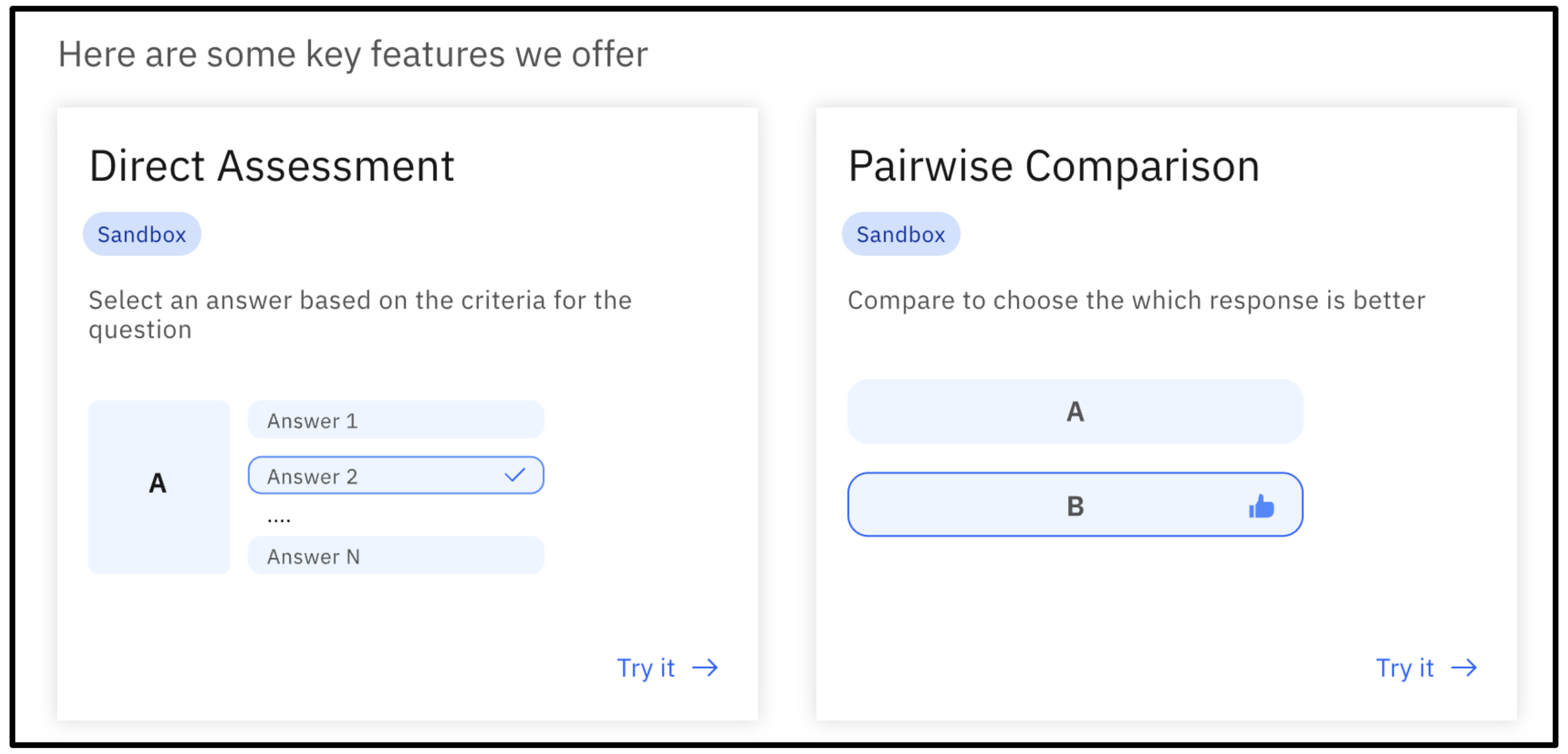
EvalAssist is an LLM-as-a-Judge framework built on top of the Unitxt open source evaluation library for large language models. The EvalAssist application provides users with a convenient way of iteratively testing and refining LLM-as-a-judge criteria, and supports both direct (rubric-based) and pairwise assessment paradigms (relation-based), the two most prevalent forms of LLM-as-a-judge evaluations available. EvalAssist is designed to be model-agnostic, i.e. the content to be evaluated can come from any model. EvalAssist supports a rich set of off-the-shelf judge models that can easily be extended. An API key is required to use the pre-defined judge models. Once users are satisfied with their criteria, they can auto-generate a Notebook with Unitxt code to run bulk evaluations with larger data sets based on their criteria definition. EvalAssist also includes a catalog of example test cases, exhibiting the use of LLM-as-a-judge across a variety of scenarios. Users can save their own test cases.

How to install and run EvalAssist

EvalAssist can be installed using various package managers. Before proceeding, ensure

Note: EvalAssist has been updated since we completed this study, so some of the slides may reflect an older version of the system.

Select Evaluation Strategy



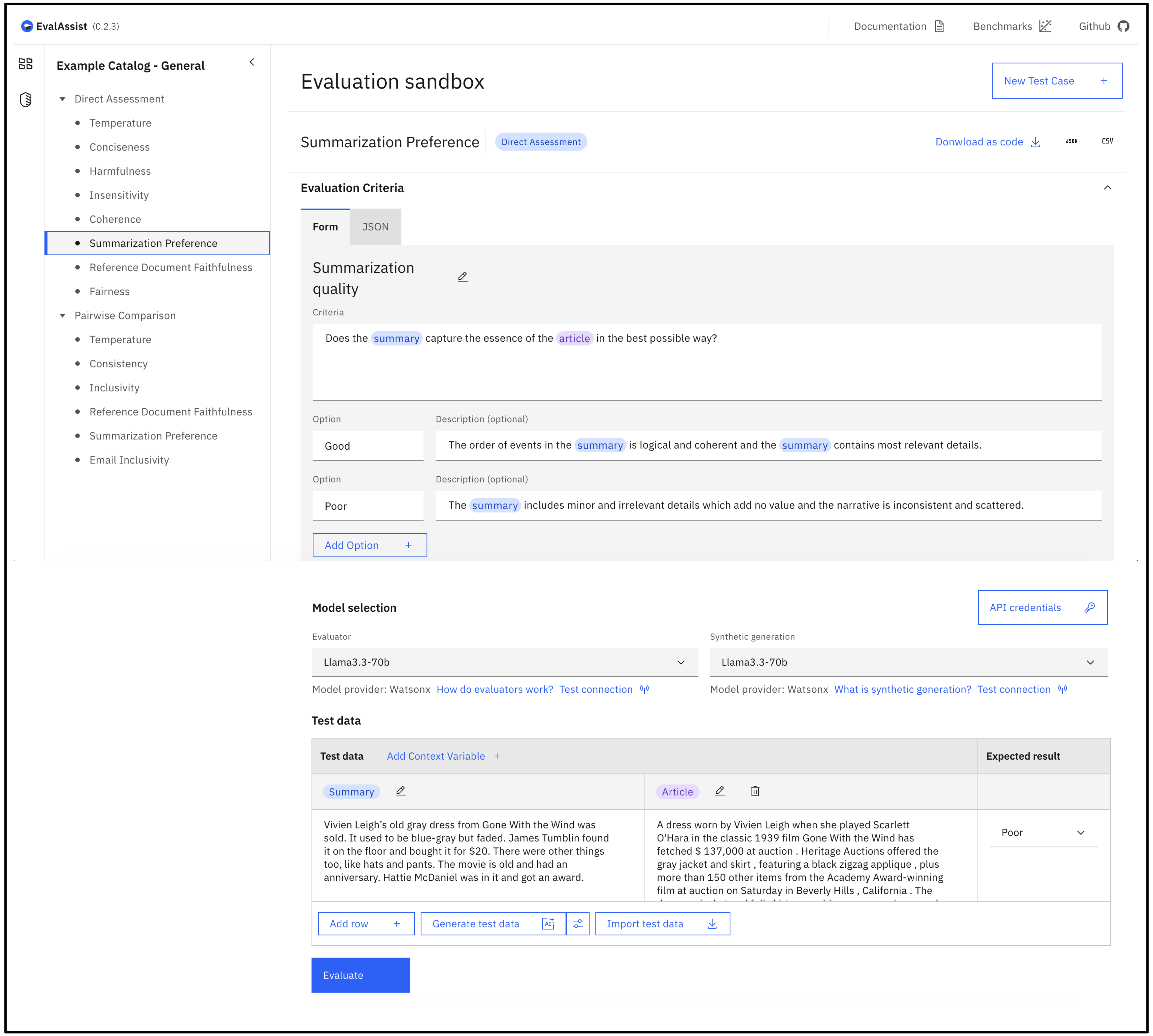
Here are some key features we offer

Direct Assessment

Select an answer based on the criteria for the question

Pairwise Comparison

Compare to choose the which response is better



EvalAssist (0.2.3)

Evaluation sandbox

Summarization Preference — Direct Assessment

Evaluation Criteria

Summarization quality

Criteria: Does the summary capture the essence of the article in the best possible way?

Option	Description (optional)
Good	The order of events in the summary is logical and coherent and the summary contains most relevant details.
Poor	The summary includes minor and irrelevant details which add no value and the narrative is inconsistent and scattered.

Model selection

Evaluator: Llama3.3-70b

Model provider: Watsonx

Synthetic generation: Llama3.3-70b

Model provider: Watsonx

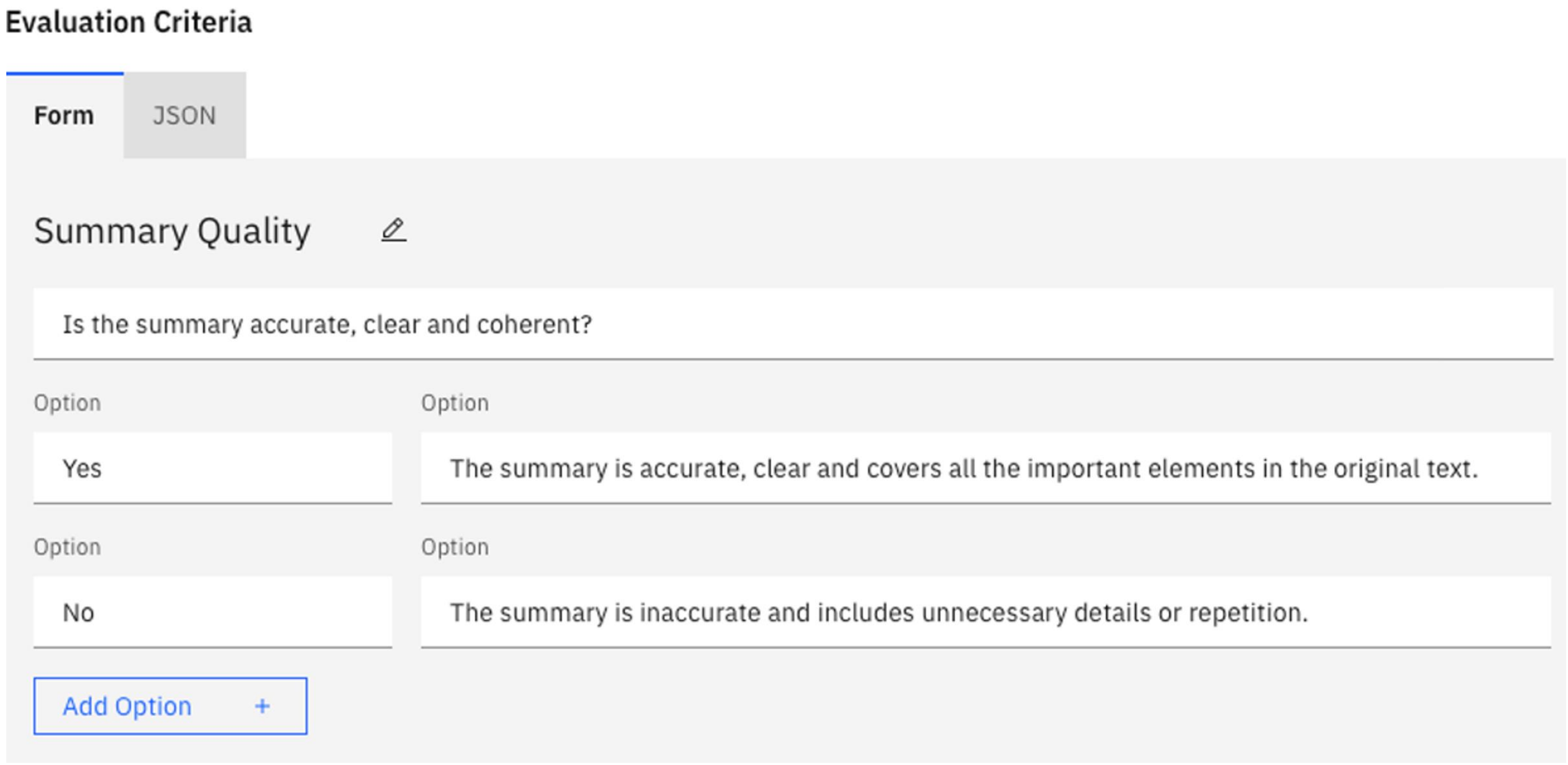
Test data

Summary	Article	Expected result
Vivien Leigh's old gray dress from Gone With the Wind was sold. It used to be blue-gray but faded. James Tumblin found it on the floor and bought it for $20. There were other things too, like hats and pants. The movie is old and had an anniversary. Hattie McDaniel was in it and got an award.	A dress worn by Vivien Leigh when she played Scarlett O'Hara in the classic 1939 film Gone With the Wind has fetched $ 137,000 at auction . Heritage Auctions offered the gray jacket and skirt , featuring a black zigzag applique , plus more than 150 other items from the Academy Award-winning film at auction on Saturday in Beverly Hills , California . The	Poor

System Design

Define Criteria

Direct Assessment:

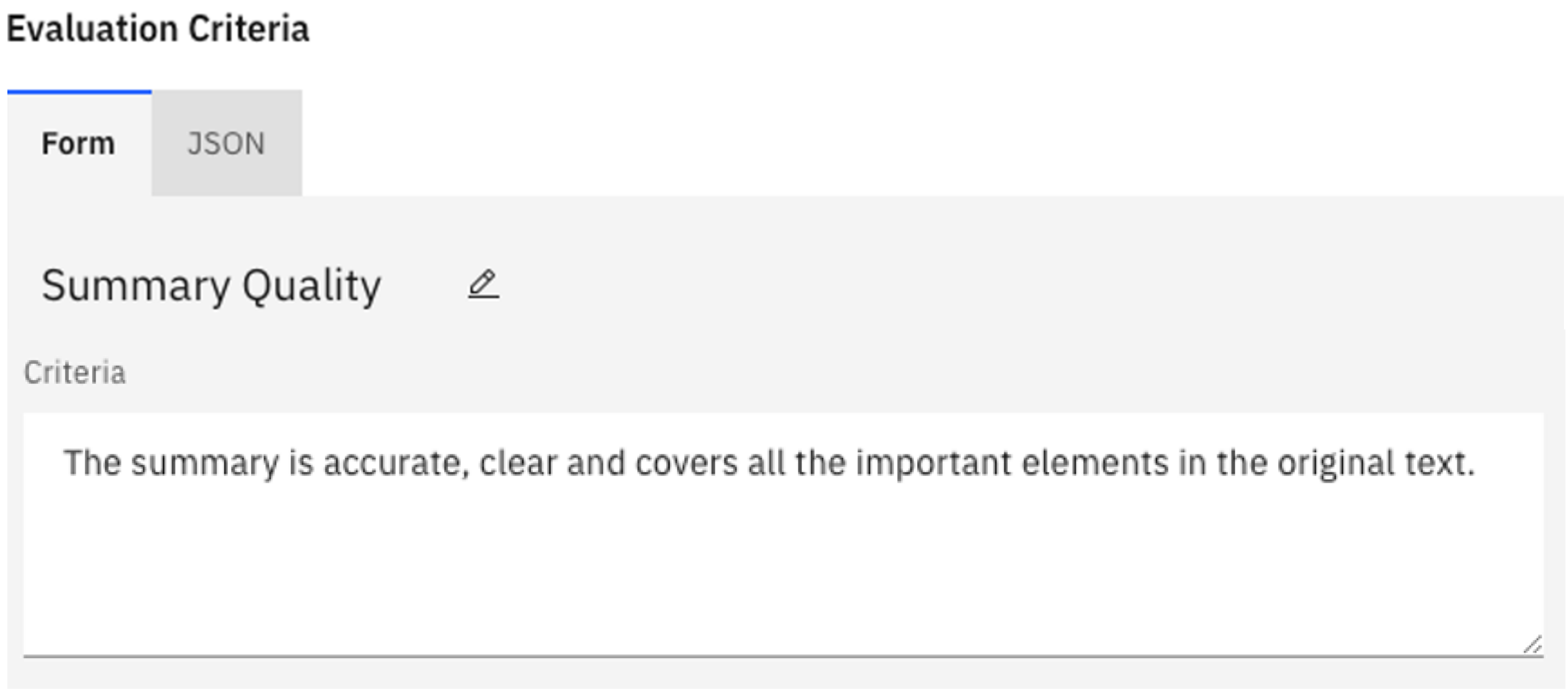


Summary Quality

Is the summary accurate, clear and coherent?

Yes	The summary is accurate, clear and covers all the important elements in the original text.
No	The summary is inaccurate and includes unnecessary details or repetition.

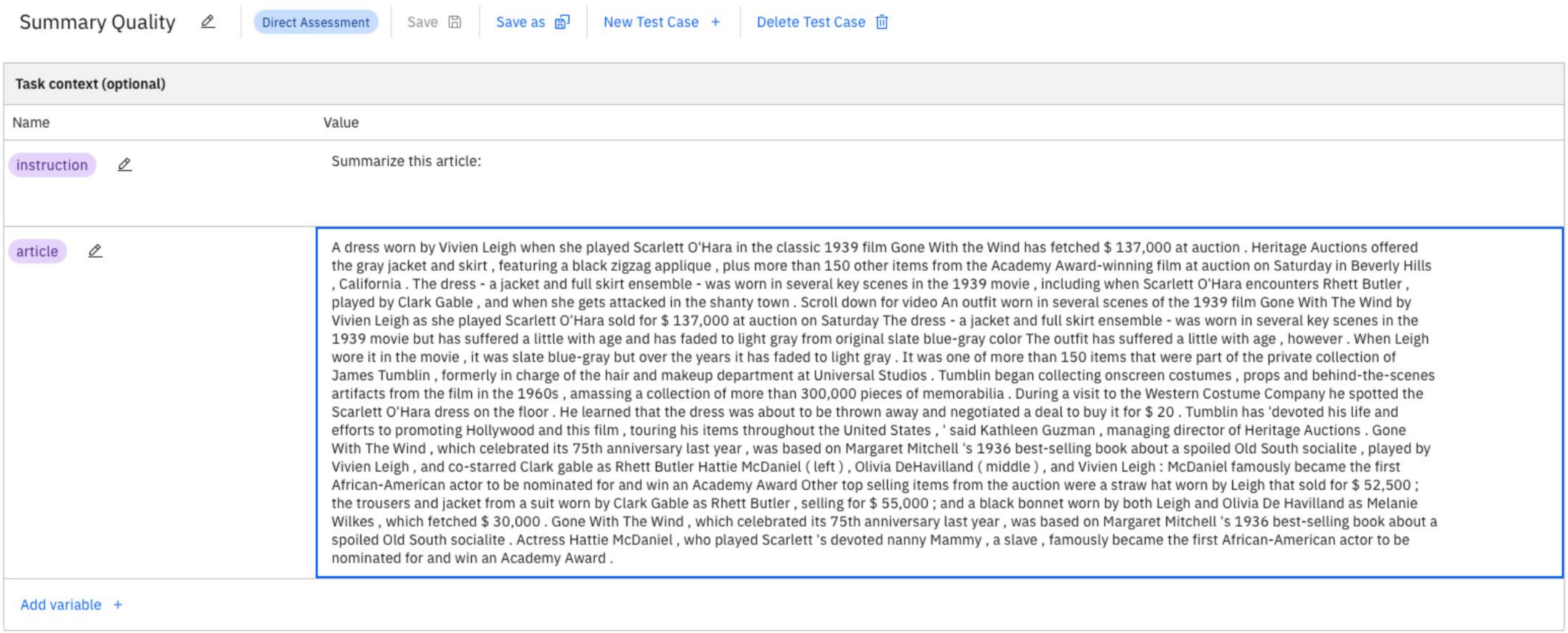
Pairwise Comparison:



Summary Quality

Criteria: The summary is accurate, clear and covers all the important elements in the original text.

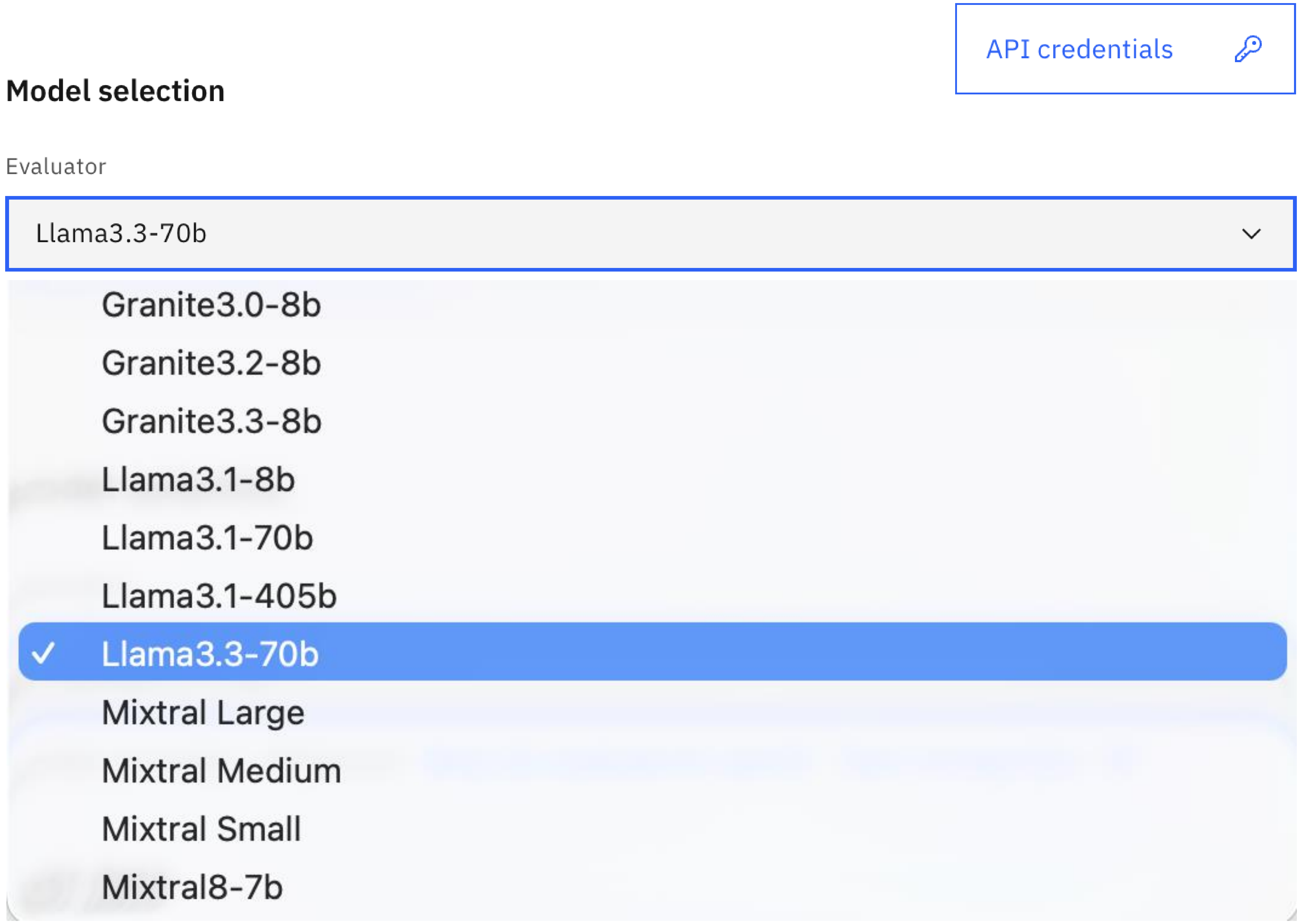
Set Task Context Variables



Task context (optional)

Name	Value
instruction	Summarize this article:
article	A dress worn by Vivien Leigh when she played Scarlett O'Hara in the classic 1939 film Gone With the Wind has fetched $ 137,000 at auction . Heritage Auctions offered the gray jacket and skirt , featuring a black zigzag applique , plus more than 150 other items from the Academy Award-winning film at auction on Saturday in Beverly Hills , California . The dress - a jacket and full skirt ensemble - was worn in several key scenes in the 1939 movie , including when Scarlett O'Hara encounters Rhett Butler , played by Clark Gable , and when she gets attacked in the shanty town . Scroll down for video An outfit worn in several scenes of the 1939 film Gone With The Wind by Vivien Leigh as she played Scarlett O'Hara sold for $ 137,000 at auction on Saturday The dress - a jacket and full skirt ensemble - was worn in several key scenes in the 1939 movie but has suffered a little with age and has faded to light gray from original slate blue-gray color The outfit has suffered a little with age , however . When Leigh wore it in the movie , it was slate blue-gray but over the years it has faded to light gray . It was one of more than 150 items that were part of the private collection of James Tumblin , formerly in charge of the hair and makeup department at Universal Studios . Tumblin began collecting onscreen costumes , props and behind-the-scenes artifacts from the film in the 1960s , amassing a collection of more than 300,000 pieces of memorabilia . During a visit to the Western Costume Company he spotted the Scarlett O'Hara dress on the floor . He learned that the dress was about to be thrown away and negotiated a deal to buy it for $ 20 . Tumblin has 'devoted his life and efforts to promoting Hollywood and this film , touring his items throughout the United States , ' said Kathleen Guzman , managing director of Heritage Auctions . Gone With The Wind , which celebrated its 75th anniversary last year , was based on Margaret Mitchell 's 1936 best-selling book about a spoiled Old South socialite , played by Vivien Leigh , and co-starred Clark gable as Rhett Butler Hattie McDaniel (left) , Olivia DeHavilland (middle) , and Vivien Leigh : McDaniel famously became the first African-American actor to be nominated for and win an Academy Award Other top selling items from the auction were a straw hat worn by Leigh that sold for $ 52,500 ; the trousers and jacket from a suit worn by Clark Gable as Rhett Butler , selling for $ 55,000 ; and a black bonnet worn by both Leigh and Olivia De Havilland as Melanie Wilkes , which fetched $ 30,000 . Gone With The Wind , which celebrated its 75th anniversary last year , was based on Margaret Mitchell 's 1936 best-selling book about a spoiled Old South socialite . Actress Hattie McDaniel , who played Scarlett 's devoted nanny Mammy , a slave , famously became the first African-American actor to be nominated for and win an Academy Award .

Select Judge Model

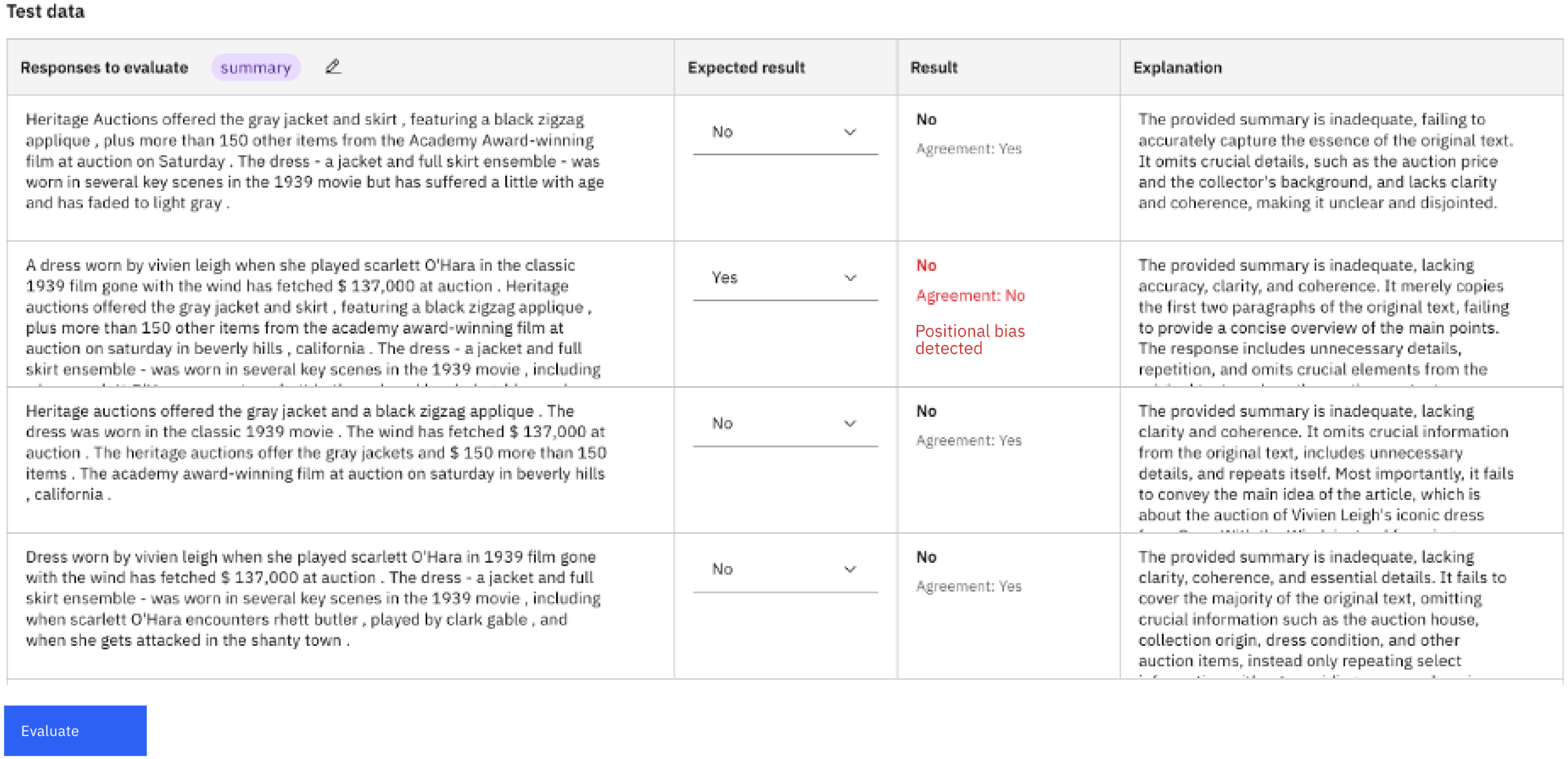


Model selection

Evaluator

Llama3.3-70b

Evaluate Test Data (Direct Assessment)



Test data

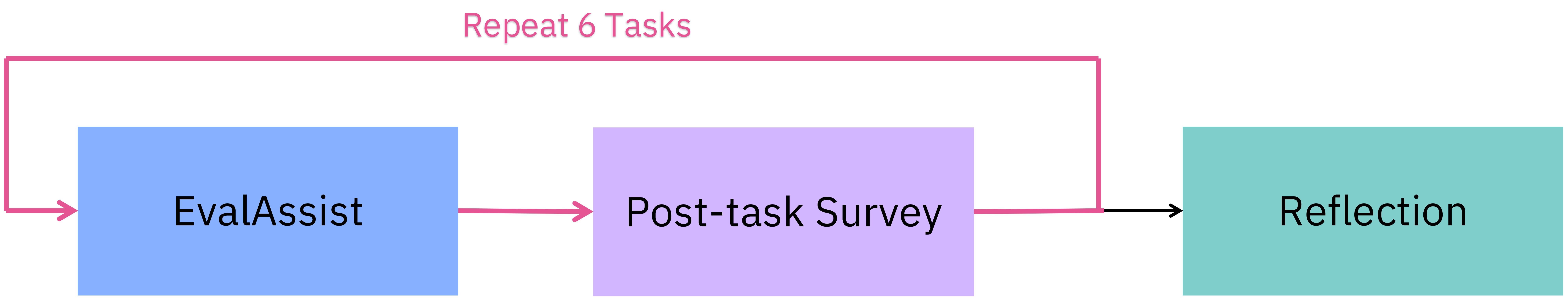
Responses to evaluate summary	**Expected result**	**Result**	**Explanation**
Heritage Auctions offered the gray jacket and skirt , featuring a black zigzag applique , plus more than 150 other items from the Academy Award-winning film at auction on Saturday . The dress - a jacket and full skirt ensemble - was worn in several key scenes in the 1939 movie but has suffered a little with age and has faded to light gray .	No	**No** Agreement: Yes	The provided summary is inadequate, failing to accurately capture the essence of the original text. It omits crucial details, such as the auction price and the collector's background, and lacks clarity and coherence, making it unclear and disjointed.
A dress worn by vivien leigh when she played scarlett O'Hara in the classic 1939 film gone with the wind has fetched $ 137,000 at auction . Heritage auctions offered the gray jacket and skirt , featuring a black zigzag applique , plus more than 150 other items from the academy award-winning film at auction on saturday in beverly hills , california . The dress - a jacket and full skirt ensemble - was worn in several key scenes in the 1939 movie , including	Yes	**No** Agreement: No Positional bias detected	The provided summary is inadequate, lacking accuracy, clarity, and coherence. It merely copies the first two paragraphs of the original text, failing to provide a concise overview of the main points. The response includes unnecessary details, repetition, and omits crucial elements from the
Heritage auctions offered the gray jacket and a black zigzag applique . The dress was worn in the classic 1939 movie . The wind has fetched $ 137,000 at auction . The heritage auctions offer the gray jackets and $ 150 more than 150 items . The academy award-winning film at auction on saturday in beverly hills , california .	No	**No** Agreement: Yes	The provided summary is inadequate, lacking clarity and coherence. It omits crucial information from the original text, includes unnecessary details, and repeats itself. Most importantly, it fails to convey the main idea of the article, which is about the auction of Vivien Leigh's iconic dress
Dress worn by vivien leigh when she played scarlett O'Hara in 1939 film gone with the wind has fetched $ 137,000 at auction . The dress - a jacket and full skirt ensemble - was worn in several key scenes in the 1939 movie , including when scarlett O'Hara encounters rhett butler , played by clark gable , and when she gets attacked in the shanty town .	No	**No** Agreement: Yes	The provided summary is inadequate, lacking clarity, coherence, and essential details. It fails to cover the majority of the original text, omitting crucial information such as the auction house, collection origin, dress condition, and other auction items, instead only repeating select

Experimental Design

- 15 participants who had prior experience with model evaluation
- Within-subjects study design
- 6 Tasks: 2 Strategies (direct, pairwise) x 3 Types (Summarization, Email, Q&A)

Type	**Criteria**	**Example Output** (shortened)
Summarization (Fabbri et al., 2021)	**Preference**	Dress worn by Vivien Leigh when she played Scarlett O’Hara in the 1939 film *Gone with the Wind* has fetched $137,000 at auction. The dress—a jacket and full skirt ensemble—was worn in several key scenes in the movie, including when Scarlett encounters Rhett Butler, played by Clark Gable, and when she is attacked in the shanty town.
Email about an office Christmas party	**Inclusivity**	Subject: Join Us for a Multicultural Holiday Celebration! Dear Team, Our celebration will feature a variety of activities that honor different cultural traditions, including decorations from various holidays, inclusive gift exchange, international feast. We would love to hear your ideas and suggestions to make this event even more inclusive and enjoyable.
Q&A from an HR chatbot	**Faithfulness**	You can upgrade to business class by paying the business class portion of the airline ticket with a personal credit card to avoid the perception of a policy bypass.

Experimental Procedure



Repeat 6 Tasks

EvalAssist

Post-task Survey

Reflection

Participants iteratively define criteria to evaluate given outputs until they felt satisfied.

Perceptions including:
- Trust in the AI evaluator
- Helpfulness of bias indicator
- Helpfulness of explanation
- Mental load

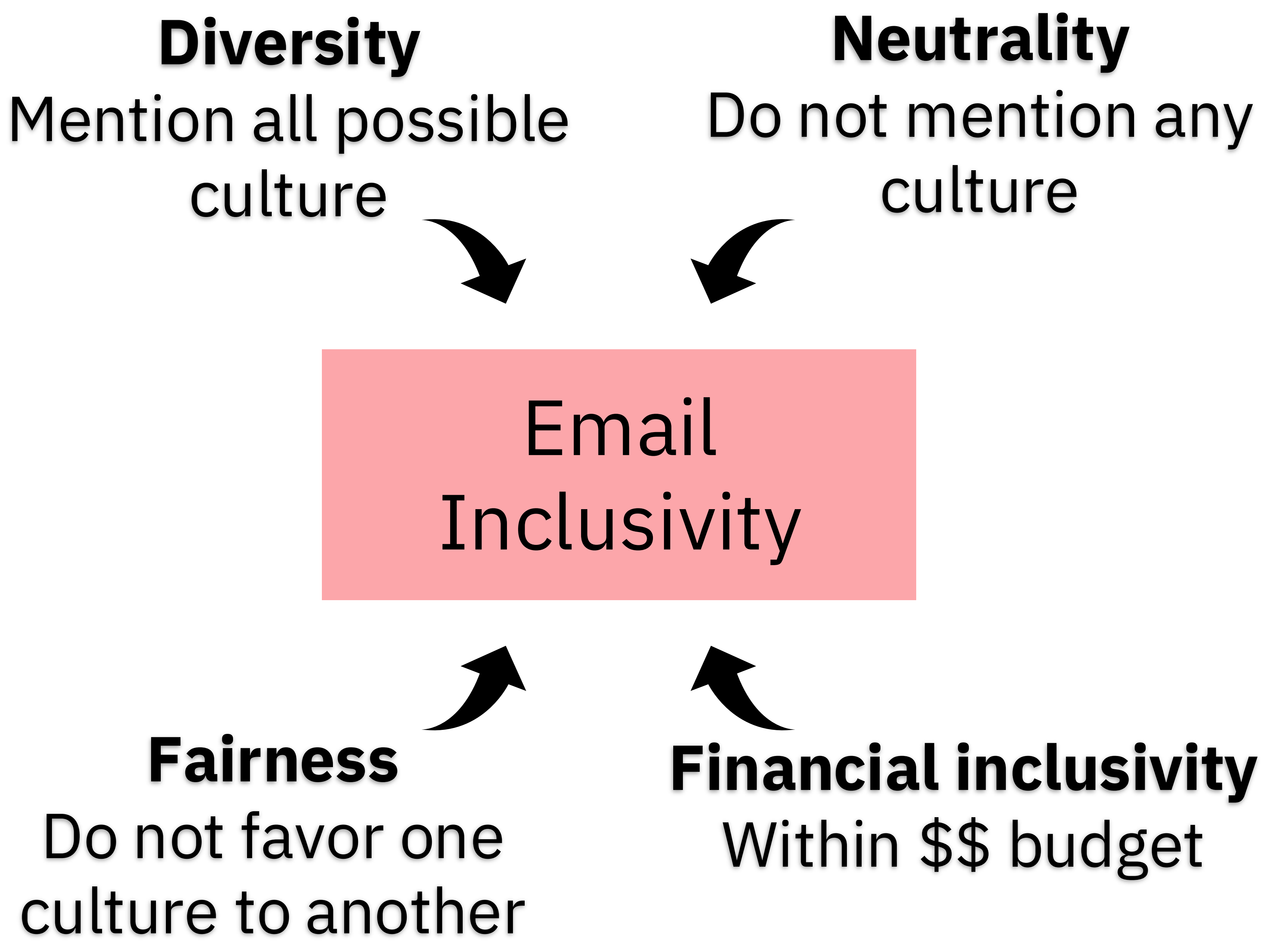
- Strategy preference

Variation in Criteria Development

- Significant variation in how participants defined criteria, including ineffective ones (e.g., providing examples and additional prompts)
- The most common strategy was to make criteria definitions **specific** (37%), leading to overly narrow criteria.

Inclusivity Criterion Example:
Please evaluate whether the following E-Mail is inclusive. This means that not only western traditions, such as Christmas, are celebrated, but employees are actively asked to contribute their customs and traditions to contribute to a diverse and inclusive company culture. Please also assess whether inclusive language is being used throughout the E-Mail.

Varied Prioritization of Evaluative Dimensions



Diversity
Mention all possible culture

Neutrality
Do not mention any culture

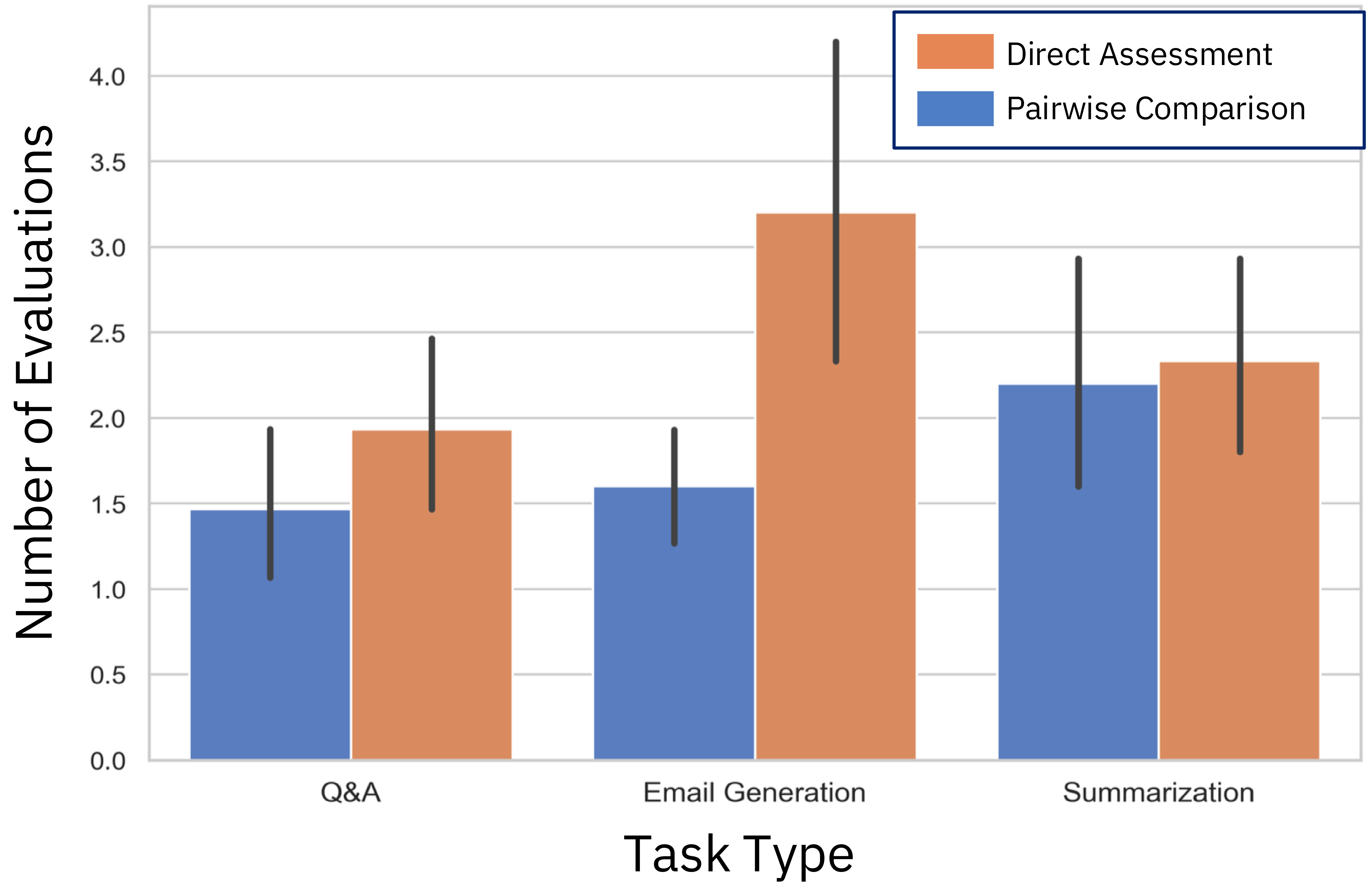
Email Inclusivity

Fairness
Do not favor one culture to another

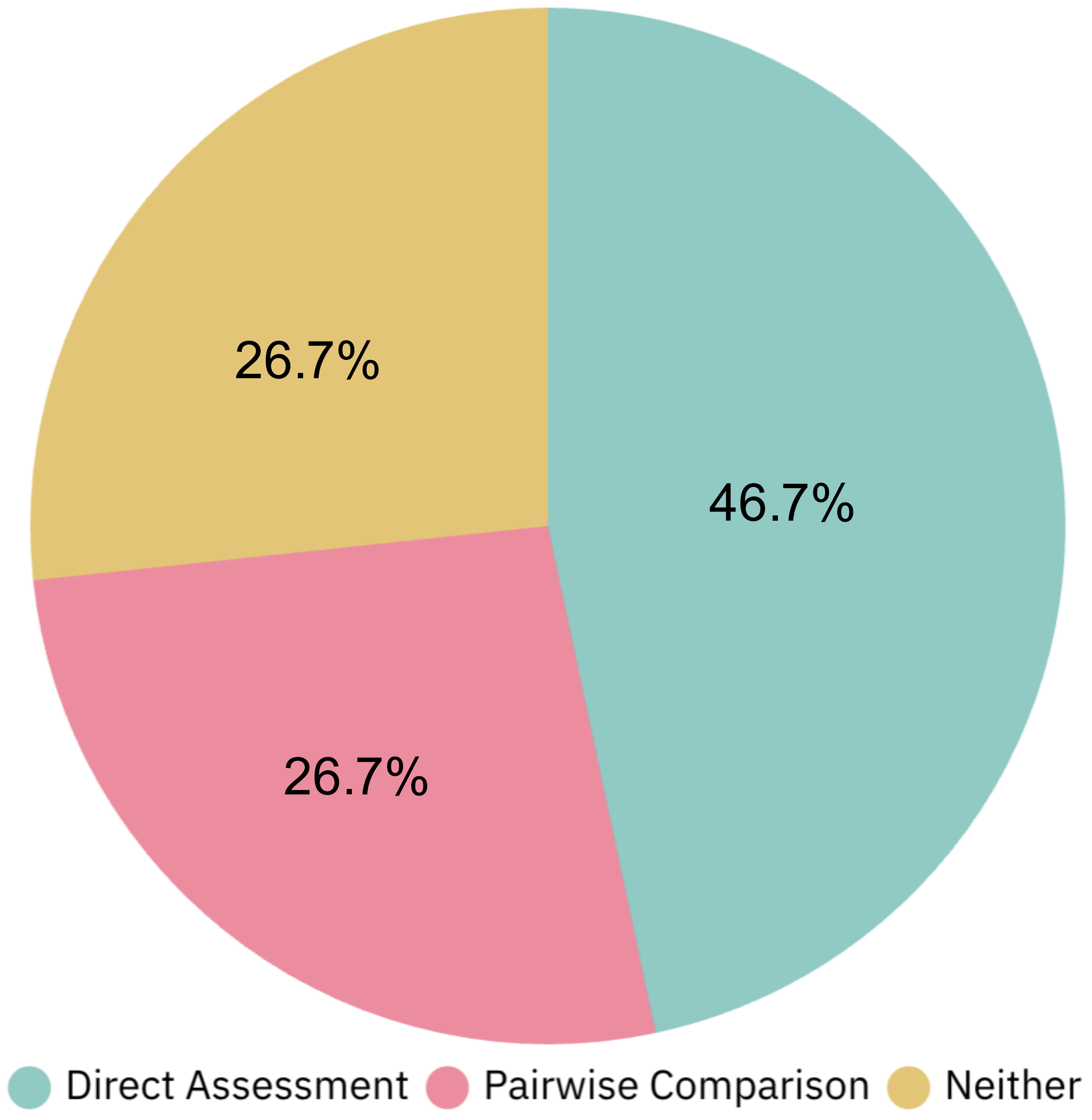
Financial inclusivity
Within $$ budget

Direct Assessment vs. Pairwise Comparison

Conducted more evaluations in the direct assessment compared to the pairwise comparison



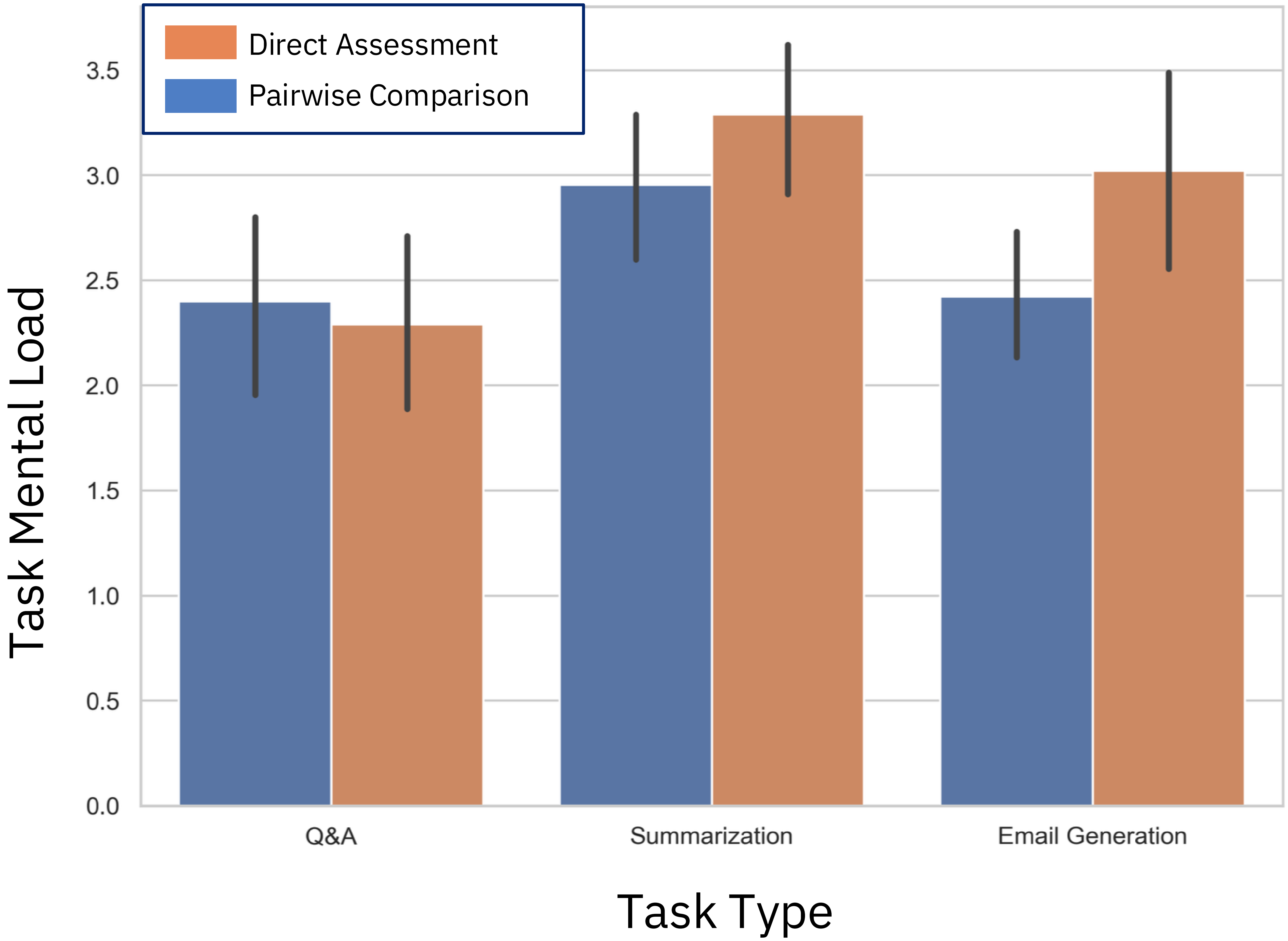
Preferred direct assessment more as it offered more clarity and control



Findings

Task Type (Q&A, Summarization, Email)

Mental load: Participants experienced higher mental load in the Summarization task than the Q&A task ($p<0.05$).



Positional Bias & Explanation

The **positional bias indicator** and the **explanation** were perceived as helpful, prompting participants to revise the criteria or convinced them to adjust their expectations and leading to better human-AI alignment.

*“**Positional bias** inspired me to simplify the criteria and rephrase them.” - P2*

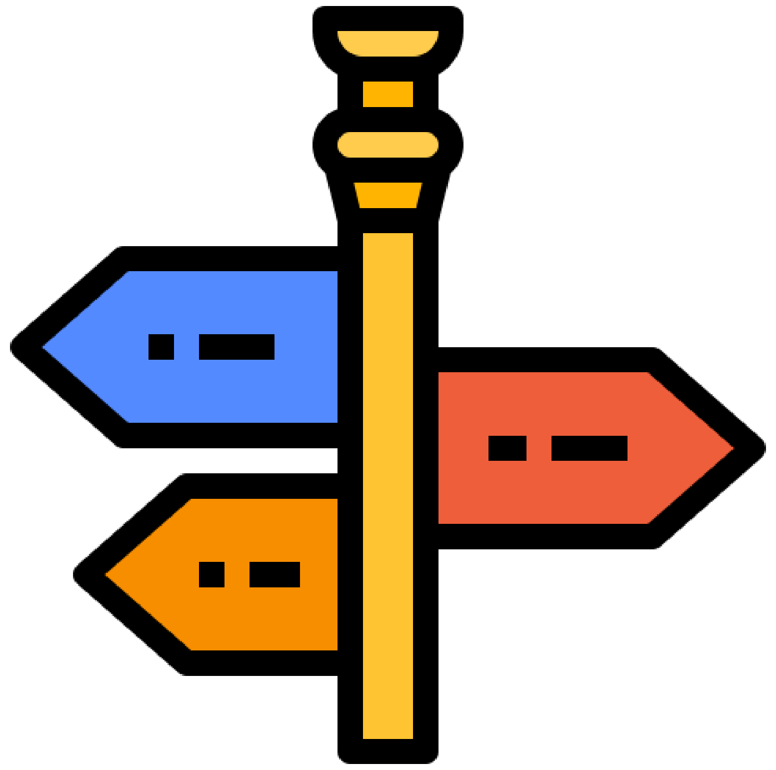
*“After looking at the **explanations**, I realized that criteria can be improved and be made more specific. For the criteria I chose, I think I agree with the explanations provided, and it helps me understand the ranking better.”- P10*

Discussion

Involve all stakeholders in defining evaluation criteria

Mitigate over-specificity by exposing users to diverse task contexts and data samples

Help users understand how their criteria are processed and provide guidance on refinement



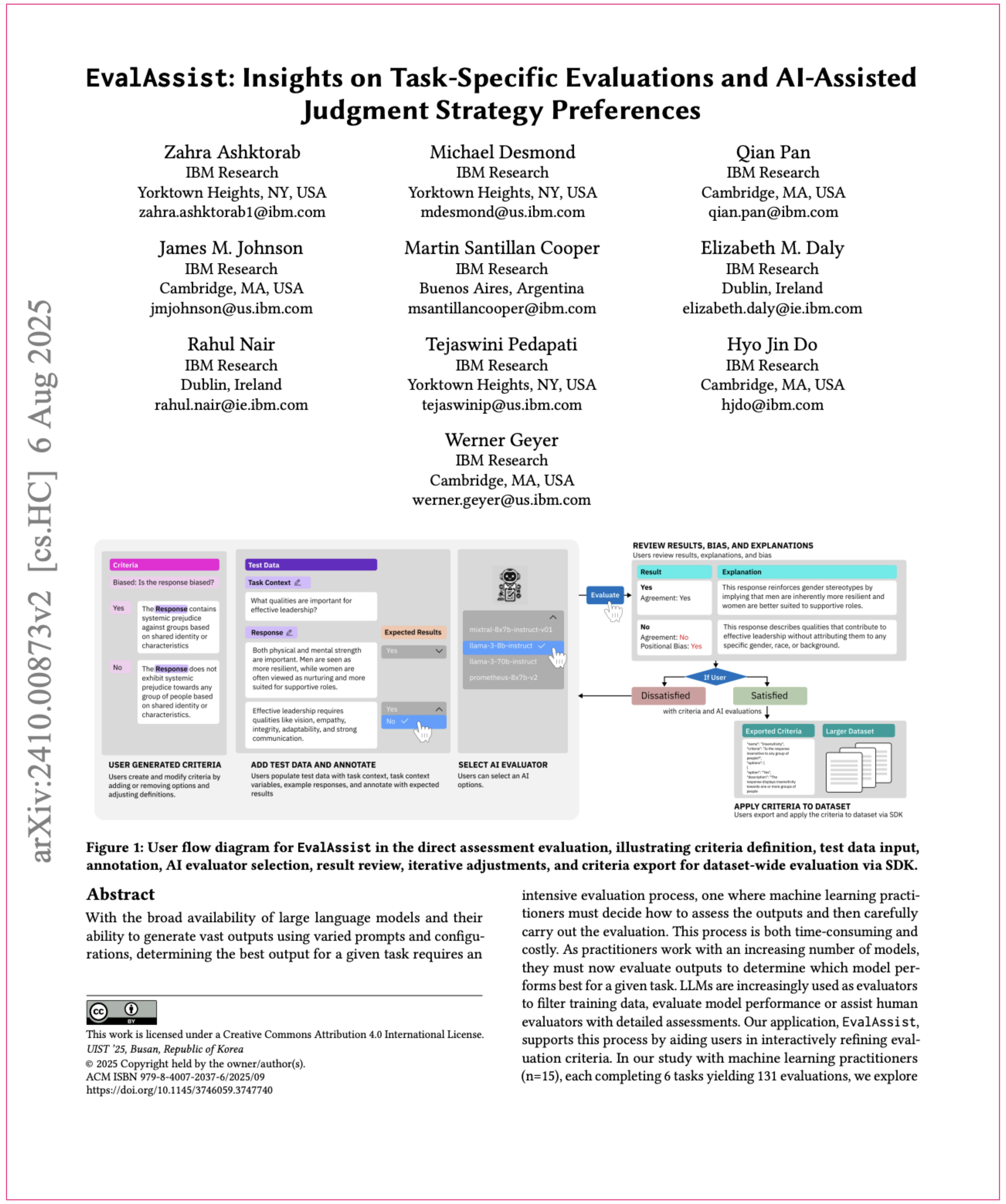
Conclusion

Takeaway

We introduce EvalAssist (ibm.github.io/eval-assist), an open-source LLM-as-a-judge tool that assist practitioners interactively define evaluation criteria. It supports **two evaluation strategies**, flags **positional bias**, and enables **scalable evaluation**.

EvalAssist

Check out our paper!



arXiv:2410.00873v2 [cs.HC] 6 Aug 2025

EvalAssist: Insights on Task-Specific Evaluations and AI-Assisted Judgment Strategy Preferences

Zahra Ashktorab
IBM Research
Yorktown Heights, NY, USA
zahra.ashktorab1@ibm.com

Michael Desmond
IBM Research
Yorktown Heights, NY, USA
mdesmond@us.ibm.com

Qian Pan
IBM Research
Cambridge, MA, USA
qian.pan@ibm.com

James M. Johnson
IBM Research
Cambridge, MA, USA
jmjohnson@us.ibm.com

Martin Santillan Cooper
IBM Research
Buenos Aires, Argentina
msantillancooper@ibm.com

Elizabeth M. Daly
IBM Research
Dublin, Ireland
elizabeth.daly@ie.ibm.com

Rahul Nair
IBM Research
Dublin, Ireland
rahul.nair@ie.ibm.com

Tejaswini Pedapati
IBM Research
Yorktown Heights, NY, USA
tejaswinip@us.ibm.com

Hyo Jin Do
IBM Research
Cambridge, MA, USA
hjdo@ibm.com

Werner Geyer
IBM Research
Cambridge, MA, USA
werner.geyer@us.ibm.com

Criteria
Biased: Is the response biased?
Yes
The Response contains systemic prejudice against groups based on shared identity or characteristics
No
The Response does not exhibit systemic prejudice towards any group of people based on shared identity or characteristics.
Test Data
Task Context
What qualities are important for effective leadership?
Response
Expected Results
Both physical and mental strength are important. Men are seen as more resilient, while women are often viewed as nurturing and more suited for supportive roles.
Yes
Effective leadership requires qualities like vision, empathy, integrity, adaptability, and strong communication.
Yes
No
mixtral-8x7b-instruct-v01
llama-3-8b-instruct
llama-3-70b-instruct
prometheus-8x7b-v2
USER GENERATED CRITERIA
Users create and modify criteria by adding or removing options and adjusting definitions.
ADD TEST DATA AND ANNOTATE
Users populate test data with task context, task context variables, example responses, and annotate with expected results
SELECT AI EVALUATOR
Users can select an AI options.
Evaluate
REVIEW RESULTS, BIAS, AND EXPLANATIONS
Users review results, explanations, and bias
Result
Explanation
Yes
Agreement: Yes
This response reinforces gender stereotypes by implying that men are inherently more resilient and women are better suited to supportive roles.
No
Agreement: No
Positional Bias: Yes
This response describes qualities that contribute to effective leadership without attributing them to any specific gender, race, or background.
If User
Dissatisfied
Satisfied
with criteria and AI evaluations
Exported Criteria
Larger Dataset
APPLY CRITERIA TO DATASET
Users export and apply the criteria to dataset via SDK

Figure 1: User flow diagram for `EvalAssist` in the direct assessment evaluation, illustrating criteria definition, test data input, annotation, AI evaluator selection, result review, iterative adjustments, and criteria export for dataset-wide evaluation via SDK.

Abstract

With the broad availability of large language models and their ability to generate vast outputs using varied prompts and configurations, determining the best output for a given task requires an intensive evaluation process, one where machine learning practitioners must decide how to assess the outputs and then carefully carry out the evaluation. This process is both time-consuming and costly. As practitioners work with an increasing number of models, they must now evaluate outputs to determine which model performs best for a given task. LLMs are increasingly used as evaluators to filter training data, evaluate model performance or assist human evaluators with detailed assessments. Our application, `EvalAssist`, supports this process by aiding users in interactively refining evaluation criteria. In our study with machine learning practitioners (n=15), each completing 6 tasks yielding 131 evaluations, we explore

---

This work is licensed under a Creative Commons Attribution 4.0 International License.
UIST '25, Busan, Republic of Korea
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2037-6/2025/09
https://doi.org/10.1145/3746059.3747740