# Aligning Human and LLM Judgments: Insights from EvalAssist on Task-Specific Evaluations and Al-assisted Assessment Strategy Preferences

ZAHRA ASHKTORAB, IBM Research, USA
MICHAEL DESMOND, IBM Research, USA
QIAN PAN, IBM Research, USA
JAMES M. JOHNSON, IBM Research, USA
MARTIN SANTILLAN COOPER, IBM Research, Argentina
ELIZABETH M. DALY, IBM Research, Ireland
RAHUL NAIR, IBM Research, Ireland
TEJASWINI PEDAPATI, IBM Research, USA
SWAPNAJA ACHINTALWAR, IBM Research, India
WERNER GEYER, IBM Research, USA

Evaluation of large language model (LLM) outputs requires users to make critical judgments about the best outputs across various configurations. This process is costly and takes time given the large amounts of data. LLMs are increasingly used as evaluators to filter training data, evaluate model performance or assist human evaluators with detailed assessments. To support this process, effective front-end tools are critical for evaluation. Two common approaches for using LLMs as evaluators are direct assessment and pairwise comparison. In our study with machine learning practitioners (n=15), each completing 6 tasks yielding 131 evaluations, we explore how task-related factors and assessment strategies influence criteria refinement and user perceptions. Findings show that users performed more evaluations with direct assessment by making criteria task-specific, modifying judgments, and changing the evaluator model. We conclude with recommendations for how systems can better support interactions in LLM-assisted evaluations.

Additional Key Words and Phrases: human-AI interaction

#### **ACM Reference Format:**

Authors' addresses: Zahra Ashktorab, Zahra.Ashktorab1@ibm.com, IBM Research, 1101 Kitchawan Rd PO Box 218, Yorktown Heights, NY, 10598, USA; Michael Desmond, IBM Research, Yorktown Heights, NY, 10598, USA; Qian Pan, qian.pan@ibm.com, IBM Research, Cambridge, MA, 02142, USA; James M. Johnson, IBM Research, Cambridge, MA, 02142, USA; Martin Santillan Cooper, IBM Research, Capital Federal, Argentina; Elizabeth M. Daly, IBM Research, Dublin, Ireland; Rahul Nair, IBM Research, Dublin, Ireland; Tejaswini Pedapati, IBM Research, Yorktown Heights, NY, 10598, USA; Swapnaja Achintalwar, IBM Research, Pune, India; Werner Geyer, werner.geyer@ibm.com, IBM Research, Cambridge, MA, 02142, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

©~2024 Association for Computing Machinery.

Manuscript submitted to ACM

#### 1 INTRODUCTION

Large language models are foundation models that can be used for a variety of tasks such as summarization, text generation, concept extraction, analysis, or classification. Benchmarks such as Helm[30], BigBench[16], or MMLU[18] can provide guidance in what language model to pick for a certain task. However, in practice, they are insufficient when it comes to specific use cases, use case specific data, or creative tasks [43]. Often human evaluation is used to assess model fitness for specific use cases and data but it is costly and takes time given the large volumes of data being generated. Practitioners are increasingly using large language models also as evaluators of the output of large language models across various contexts including filtering training data, evaluating model performance, assessing prompt effectiveness, and assisting human evaluators with detailed assessments and explanations [6, 19, 33, 36]. This approach is often referred to as LLM-as-a-judge.<sup>1</sup>

However, relying solely on LLM evaluators is not without risks; as with many tasks, evaluator models can also hallucinate or provide explanations that lack coherence, underscoring the necessity of keeping humans in the loop. Although LLMs may not always be accurate, they can potentially reduce workload by flagging outputs that require human input due to low confidence. Additionally, LLM evaluators also offer a lot of flexibility because practitioners can customize evaluation criteria — such as conciseness, faithfulness to a context, conversational naturalness, or succinctness — enabling more targeted and effective assessments tailored to specific use cases. Many tools have emerged to help users create criteria to evaluate their outputs. The process of criteria creation for evaluation is often iterative, and the concept of "criteria drift" may arise. While predefined criteria may help users to assess outputs, in practice, the act of grading also helps users refine and redefine these criteria.

Two predominant forms of AI-assisted assessments with LLMs have emerged: Direct Assessment and Pairwise Assessment. Direct Assessment scores specific criteria (often part of a rubric) to evaluate whether outputs meet the criteria, while Pairwise Assessment compares pairs of outputs against broader, high-level criteria. Each method has its own strengths and limitations, which our study explores by examining how they affect user interaction with both the criteria and the evaluation process. To support this investigation, we developed EvalAssist, a system that allows users to view outputs and iteratively refine their evaluation criteria. Our focus was on understanding how users adjust their criteria, the changes they make, and how they ultimately find satisfaction with both the criteria and the evaluation process. To understand the how users develop and refine criteria to achieve alignment with LLM evaluators, we ran a within-subject study with 15 practitioners (data scientists, software engineers, and AI engineers) who have been involved in model performance projects. In our study we pose the following research questions:

- RQ1: What do practitioners prioritize in evaluation criteria when using LLMs as judges, and how do their priorities differ?
- RQ2: What strategies do users employ to refine their criteria for achieving human-AI alignment in LLM-as-a-judge evaluations?
- RQ3: How do task-related factors and judge strategy impact how practitioners refine criteria?
  - RQ3A How do task and judge strategy influence the total number of evaluations performed?
  - RQ3B How do task and form of assessment (direct vs. pairwise) affect the degree of human-AI alignment?
- RQ4: How do task-related factors and judge strategy impact user perceptions of the judge?
  - How do they affect users' trust in AI?
  - How do they shape users' perception of positional bias?

<sup>&</sup>lt;sup>1</sup>Note that, in this paper, we use the terms judge and evaluator interchangeably. Manuscript submitted to ACM

Manuscript submitted to ACM

- How do they influence users' perception of explanations provided by the judge?
- How do they impact users' cognitive load during the evaluation process?
- RQ5: Which assessment strategy do users prefer?

This paper makes the following contributions:

- We introduce EVALASSIST, a tool designed to help practitioners refine evaluation criteria through both direct and
  pairwise assessment strategies. It provides positional bias metrics from the AI judge, as well as an explanation
  for each judgment.
- We present results from within-subjects controlled experiment with machine learning practitioners (n=15) providing insights into how they refine evaluation criteria and uncover key differences between the two evaluation strategies.
- Based on our findings, we offer design suggestions for AI-assisted evaluation systems.

Our findings show that users conduct more evaluations under the direct assessment condition. Users refine criteria in multiple ways, such as making criteria more specific or general, adjusting their own judgments, or modifying the AI evaluator's outputs. Explanations are perceived as more helpful in the direct assessment condition. Users prefer Direct Assessment when they need clarity and control over individual item evaluations, and Pairwise Assessment when evaluating nuanced or subjective criteria.

# 2 MOTIVATION AND RELATED WORK

# 2.1 Human AI Collaboration in AI-assisted Evaluation

Recently, several AI-assisted evaluation tools have emerged, with varying focuses including: improving the iterative nature of prompt refinement or refining criteria for evaluation. Across these systems, the evolving nature of user-defined criteria is emphasized, acknowledging that such criteria are not static but adapt in response to AI outputs and user feedback. This iterative process is fundamental in tools that support criteria refinement and prompt adjustments, reinforcing the dynamic interaction between humans and AI in the evaluation process. A prominent aspect shared among these tools is the role of human-in-the-loop evaluation. Systems like EvalGen [38], ChainForge [2], EvalLM [24], and LLM Comparator [21] integrate human feedback as a key element of the evaluation loop. While AI systems assist in generating evaluations, they rely on human judgment to ensure alignment with user preferences. One of the key challenges in this interactive process is criteria drift [38], where users adjust their evaluation standards as they encounter new outputs. Prior systems illustrate this behavior, as users often modify their criteria after receiving AI-generated responses that deviate from initial expectations. This flexibility is critical, highlighting the need for evaluation systems that allow criteria adjustments throughout the evaluation process, rather than imposing rigid, predefined standards.

2.1.1 Criteria Iteration. Prior research demonstrates that users require multiple rounds of iteration to refine their criteria [38]. Users require viewing LLM outputs in order to define criteria, since there are challenges to defining criteria without seeing the range of possible outputs. Conversely, users may create criteria that are dependent on the outputs created. Allowing users to iteratively define criteria is an important consideration in the design of our tool. In EvalAssist, users can start a project by refining their evaluation criteria before scaling up to the full dataset. Effective sampling enhances learning for LLM-as-a-Judge by selecting diverse and representative outputs.

Crafting effective criteria typically requires multiple iterations. Criteria components such as name, definition, scale, and examples often need definition and refinement as users evaluate outputs. Related work [25] indicates that users

often develop new criteria during evaluations. To facilitate this, EvalAssist includes a real-time feedback system that allows users to immediately see the impact of criteria modifications. Unlike systems that rely heavily on predefined metrics or expert-labeled data, EvalAssist enables users to define evaluation criteria in natural language and to iteratively refine these criteria based on feedback from the AI model. Unlike other AI-assisted evaluation tools that combine prompting engineering with criteria definition, EvalAssist simplifies the LLM-as-a-judge process by allowing users to focus solely on defining criteria. This approach recognizes that developers often rely on external workflows to adjust configurations (e.g., model temperature) and experiment with different models and prompts to generate responses [11].

- 2.1.2 Visualization. Other tools focus on providing intuitive, interactive interfaces that facilitate complex evaluation tasks. The use of interactive and visual interfaces is another notable feature across these tools. Allowing users to visually compare model outputs in real-time, provides comprehensible evaluation experience [21]. Many existing tools share common themes such as iterative refinement, user-centered evaluation, and scalability. Together, they reflect a growing trend in human-AI collaboration, aiming to create more flexible, subjective, and adaptable evaluation systems that effectively combine human insight with AI capabilities.
- 2.1.3 Addressing Bias in Al-Assisted Evaluation. LLM Evaluators, like their human counterparts, exhibit biases. These biases include but are not limited to: positional bias, which is when judges consistently favor one side of a pair, regardless of the actual quality of the answers, self-enhancement bias when a model prefers its own responses, and verbosity bias when an LLM judge favors longer responses even if they are not a better alternative [43]. Many of the existing Al-assisted tools do not flag these kinds of biases to users. Considering the persistent challenge of bias, systems should both provide transparency when bias occurs and implement bias mitigation strategies that include swapping answer order to reduce position bias [43] and treating inconsistent results as ties, or by randomly assigning positions in large datasets [29] [43]. EvalAssist includes a check for positional bias and indicates whether positional bias exists.

# 2.2 Direct Assessment vs. Pairwise Comparison in Evaluation

Two of the most common judgment strategies in evaluation are direct assessment and pairwise comparison. Direct assessment involves outputting a scalar indicator of quality (e.g., assigning a score or rating to an item) [43], while pairwise comparison determines which of two outputs is preferred based on specific criteria. Both approaches have advantages and disadvantages depending on the context and task. One limitation of pairwise comparison is scalability. As the number of items to be evaluated increases, the number of required comparisons grows quadratically, making this method less feasible for large-scale evaluations. However, pairwise comparisons can be more effective at identifying subtle differences between outputs and according to prior research is an easier task for both humans and LLMs compared to rating a single output, often yielding higher accuracy in LLM-as-a-judge benchmarks [3, 15, 28, 44]. In contrast, direct assessment can efficiently evaluate multiple items at once, but it may struggle to detect fine distinctions between outputs.

Direct assessment often utilizes rubrics with multiple dimensions, while pairwise comparison may focus on a single dimension of preference. Additionally, the nature of the criteria—whether objective (e.g., grammatical accuracy) or subjective (e.g., creativity or tone)—plays a significant role in shaping the evaluation process. Prior work has explored LLMs' ability to make selections based on user-defined preferences across a wide range of criteria [7, 8, 43]. For example, prior work investigated how LLMs handle subjective judgments across dimensions such as brevity, formality, honesty, creativity, and political tone, highlighting the variability that emerges when criteria are subjective. EvalAssist allows users to select the evaluation strategy that best fits the task.

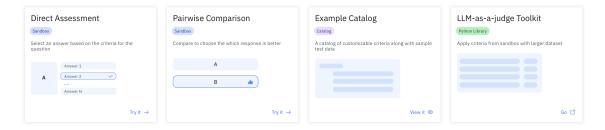


Fig. 1. EvalAssist's homepage provides users with a range of options: choose direct assessment, pairwise comparison, explore the example catalog, or utilize the toolkit to apply custom criteria from the sandbox to the entire dataset.

#### 3 EVALASSIST

EvalAssist abstracts the llm-as-a-judge evaluation process into a library of parameterize-able evaluators (the criterion being the parameter), allowing the user to focus on criteria definition. This approach acknowledges that developers often use complex external workflows to adjust configurations (e.g., model temperature) and experiment with different models and prompts to generate responses [11]. EvalAssist consists of a web-based user experience, an API, and a Python toolkit. The user interface 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 (Figure 1), the two most prevalent forms of LLM-as-a-judge evaluation available [23, 43]

Users can choose the evaluation method based on task complexity, receive AI judgment explanations, and view metrics like positional bias. Once users are satisfied with their criteria, they can use the Python toolkit to run bulk evaluations with larger data sets by exporting auto-generated JSON definitions of their criteria into predefined notebooks provided in the toolkit. We also allow users to save their test cases and provide a catalog of predefined criteria. A test case in the Example Catalog includes a criteria definition and the data being evaluated. On the landing page, users can choose between Direct Assessments and Pairwise comparisons.

#### 3.1 Direct Assessment

In this mode, users evaluate outputs based on a single criterion rubric they define. Users can define task-relevant input data through variables in the task-context (Figure 2), such as, for example, the prompt, the article to summarize, or the source data for content-grounded Q&A. The next section (Figure 3) allows them to define their criteria with a title, criteria description, and an arbitrary number of free-form options (b) the LLM evaluator will have to choose from during assessment. As such, the system supports both binary and multi-level scale assessments. In the Evaluator section, users need to select an LLM judge. Our system currently supports four judges: mixtral-8x7b-instruct-v01, llama-3-8b-instruct, llama-3-70b-instruct, and prometheus-8x7b-v2. In the Test Data Section, users enter the outputs they want to evaluate (we call them the responses; note that it is possible to edit the variable name here too so the data being evaluated can be better referenced in the criteria definition) and optionally the result they would expect for each output. After running the evaluation, the system shows the actual results next to the expected results, including agreement, positional bias if present, certainty scores, and an explanation.

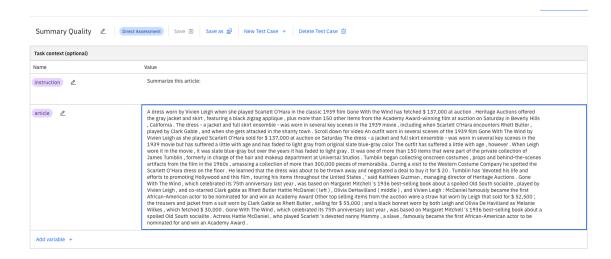
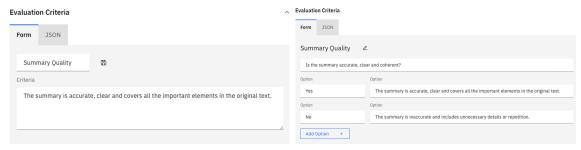


Fig. 2. Task Context for the Summarization Task. The Task Context is consistent for both Direct Assessment and Pairwise strategies. Users have the option to break down the context into variables, such as the instruction and article, to simplify reference while developing evaluation criteria.



(a) Pairwise approach: Features a concise one-sentence summary (b) Direct assessment: contains a high-level question, scale items, of each criterion.

Fig. 3. Evaluation criteria forms for pairwise assessment and direct assessment.

# 3.2 Pairwise Comparison

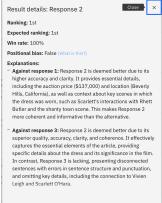
In the pairwise comparison mode (Figure 3 (a)), EvalAssist compares multiple outputs (minimum two) pairwise against one-another selecting the one that better fits their criteria. The best output is determined by computing the win rate across all pairwise output comparisons. Similar to Direct Assessment, users can provide task-relevant input data through variables, define a criteria, and select an evaluator LLM. However, options don't need to be added to pairwise comparisons. After evaluation, we display the results next to the expected results (see Figure 3), including the winner, ranking, and agreement with expected ranking. Users can click on each result to see detailed explanations including positional bias, win-rate, and explanations for the comparisons with the other outputs.

#### Test data

Responses to evaluate summary 2	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 v	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, foaturing a black gizgag applique, plus more than 150 other items from the academy award-winning film at auction on saturday in beverty hills, california. The dress - a jacket and full skirt ensemble. was worn in several key scenes in the 1939 movie, including	Yes V	No Agreement: No	The provided summary is inadequate, lacking accuracy, clarity, and coherence. It merely copies the first two paragraphs of the original text, falling 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 v	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 falls to convey the main idea of the article, which is about the auction of Vivien Leight'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 mowie , including when scarlett O'Hara encounters rhett butlor , played by clark gable , and when she gets attacked in the shanty town .	No v	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

Fig. 4. Results for Direct Assessment. Users can select their expected judgments for the output, which are auto-populated based on the criteria they define (i.e., the scale items created when setting the criteria). The results display the LLM evaluator's judgments, indicating whether there is agreement between the user and the Al, along with explanations for each result.

# Test data Responses to compare summary Expected ranking Result 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 Ranking: 3rd (33% winrate) 3rd v 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 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 skirr, leaturing a black zigzag applique, plus more than 150 other items from the academy award-winning film at auction on saturday in beverty hills ; california. The dress - a jacket and full skirt ensemble - was worn in several key seenes in the 1939 move is including when scarled to Vlara encounters right butter. 1st Ranking: 1st (100% winrate) played by clark gable , and when she gets attacked in the shanty town View Detail 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. Ranking: 4th (0% winrate) 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 thett butler , played by clark gable , and when she gets attacked in the shanty town . Ranking: 2nd (67% winrate)



- (a) Ranking results generated from pairwise comparison assessment. Users can input their expected ranking to and assess their level of agreement with the AI evaluator.
- (b) Explanations for each pairwise comparison in pairwise assessment.

Fig. 5. Results, explanations, and expected ranking generated through pairwise comparison.

#### 3.3 LLM Evaluation

When users select the "Evaluate" button, their input is sent to the chosen evaluator. Each evaluator is designed to perform either direct assessment or pairwise evaluation. The main external difference between these two lies in how the input criteria are structured. Internally, evaluators operate as a dialog with a target LLM (large language model) using a set of custom prompts specific to that LLM. First, the LLM is prompted to review the evaluation task, considering the Manuscript submitted to ACM

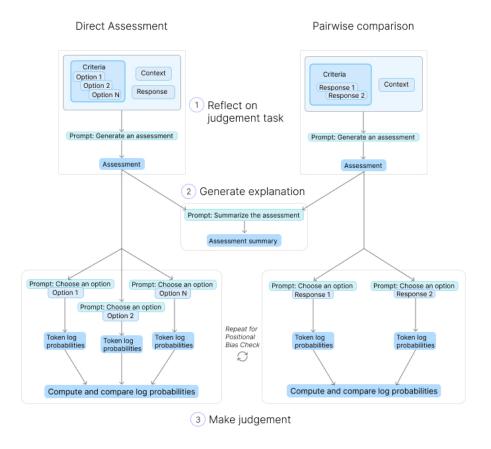


Fig. 6. Generation of judgments in EvalAssist involves initially creating an assessment, followed by summarizing the assessment and making a selection between options. Log probabilities are generated and compared during this process, which is repeated to detect positional bias. The prompts shown have been simplified for clarity in the image.

context, criteria, and subject of evaluation. Based on this, the LLM generates an open-ended assessment that explains its decision-making process. This step is inspired by Chain-of-Thought (CoT) prompting [41], encouraging the LLM to base its final judgement on its initial reasoning. This generated assessment is then added to the dialog history. Afterward, the LLM is asked to make a final judgement. Instead of having the LLM directly generate a final decision, we provide a set of options as potential completions. The system calculates the log probability of each option's tokens using a forward pass and selects the option with the highest linear probability sequence. This comparison is done by evaluating the first token of each completion, followed by the second, and so on. For direct assessment, these completions are the user-provided option strings, while in pairwise evaluation, the completions are the two responses being compared (Response 1 and Response 2). Although using log probabilities for determining the LLM's judgement is less efficient than direct generation, it is significantly more reliable. The final explanation of the judgement summarizes the initial assessment. Positional bias is checked by shuffling the order of the options presented to the LLM and verifying the consistency of its final decision. A visual representation of the algorithm is shown in Figure 6.

### 4 METHODOLOGY

The study aimed to explore several key aspects of how practitioners engage with LLMs as evaluators. Specifically, it sought to understand what practitioners prioritize in evaluation criteria when using LLMs as judges and how these priorities differ (RQ1). Additionally, it examined the strategies users employ to refine their criteria in pursuit of human-AI alignment during LLM-as-a-judge evaluations (RQ2). The study also focused on the impact of task-related factors and judge strategies on how practitioners refine criteria (RQ3). Furthermore, the research investigated how task and judge strategy affect user perceptions of the AI judge, including trust, perception of positional bias, perception of explanations, and cognitive load (RQ4). Lastly, the study aimed to identify which assessment strategy users preferred (RQ5).

We used a within-subjects study design, involving 15 participants recruited internally from our organization. Participants were recruited internally from individuals who had previously used EvalAssist and through announcements posted on the organization's Slack channels. These messages invited employees to participate in a study, with eligibility focused prior experience with model evaluation. Participants were informed about the study's purpose and provided consent before taking part. Participants were provided with detailed information about the study's purpose, procedures, and their rights as participants. We collected demographic information, including participants' roles at the company, education levels, and prior experience with model evaluation. Participants then completed a practice task to familiarize themselves with the experimental interface and procedures.

The experimental tasks involved creating criteria and evaluating output based on their criteria using either the pairwise approach or the direct assessment approach for six task contexts based on 3 tasks (Q&A, Email Generation, Summarization) each seen twice, once under the direct assessment condition and once under the pairwise condition. Each task is described in Table 1.

The Direct Assessment task involved designing a single criterion rubric with options to assess whether output complies with the criteria (as shown in Figure 4), whereas pairwise comparison (see Figure 3 involves defining the criteria and comparing pairs) of generated outputs to see which better matches criteria. While both approaches come with strengths and weaknesses, we set out to examine how they influence the evaluation of criteria and users' interactions differently.

Each task was followed by a short survey assessing participants' trust in the AI, satisfaction with the criteria, cognitive load, perception of positional bias, and the explanations provided by the AI. Throughout the study, data were logged on the number of evaluations run, final human agreement with the LLM evaluator, and the time taken to run each evaluation. The following questions were rated on a 5-point Likert scale, with 1 indicating "strongly disagree" and 5 indicating "strongly agree":

- Trust in AI Evaluator:
  - I trusted the AI evaluator to to judge the responses. (adapted from [35])
  - How confident were you in the model's judgements/evaluations? (adapted from [5])
- Perception of Positional Bias: The positional bias was helpful in completing this task.
- Perception of Explanations: The explanations were helpful in completing this task.
- Mental Load (adapted from [17]):
  - The task was mentally demanding.
  - I was successful in accomplishing what I was asked to do.
  - I had to work had to accomplish my level of performance

To mitigate order effects, we used partial counterbalancing. Participants were randomly assigned to different task orders to ensure that the sequence of tasks did not systematically bias the results. Participants were asked to reflect on the different tasks they had completed and respond to questions about their preferences between direct assessment and pairwise assessment. They also described how they interacted with two types of tasks: objective versus subjective. After completing the six tasks, participants were asked to select their preferred assessment strategy and explain the reasons for their preference.

#### 4.1 Tasks and Evaluation Criteria

Tasks and evaluation criteria were chosen to capture various levels of granularity and criteria specificity (i.e., single dimension of 'preference' vs. document groundedness) and to reflect a diversity of tasks and respective models. Despite the responses coming from various sources (various models, datasets, etc.), all of the responses were reviewed by the coauthors to ensure variability in the responses. Below we list the task descriptions. Task examples can be seen in Table 1

- Article Summarization: While the summaries can be judged on cohesiveness, consistency, fluency, relevance [4] we asked participants to define criteria based on the single dimension of "preference", as seen in work by [7, 29, 31]. For the summarization task, we presented the original reference document and corresponding generated outputs to users to be judged from [13].
- Email Generation: The email generation task was leveraged to judge inclusivity. We generated an email about an office Christmas party using various models (Gemini 1.5-Pro [37], claude-3-5-sonnet-20240620 [1], gpt-3.5-turbo-0125 [14], mixtral-8x22b-instruct-v0.1 [20]), resulting in emails with different levels of inclusivity. We asked participants to create criteria to evaluate the generated output based on the inclusiveness of the output.
- Q&A multi-turn: The Q&A task involved a context document, a multi-turn conversation, and a final response to the last question in the conversation. The output was generated using retrieval-augmented generation (RAG) [27]. RAG evaluations focused on three aspects: answer relevance (is the answer relevant to the query?), context relevance (is the retrieved context relevant to the query?), and groundedness (is the response supported by the context?) [42]. The Q&A data was selected from existing HR support documents at ABC Company, with AI-generated responses created by multiple models, including [9]. One of the question-and-answer sets involved a reference document and a customer inquiry about upgrading their business travel to business class with their corporate credit card. Participants were asked to create a criterion to evaluate the faithfulness of each response to the reference document. A response is considered faithful to a document when it only reflects the information expressed in the document [32].

# 5 RESULTS

# 5.1 Participant Demographics

We recruited 15 participants at ABC company who had prior experience with model evaluation. The job titles are listed in Table 3. Their prior experience included subjective and manual evaluation, use case specific evaluation, evaluation framework and metrics.

As a technology engineer, I interact with various LLM to perform many generative-ai tasks. However, besides manual inspection of LLM outputs, I have not found a way to consistently evaluate the performance of a LLM and see if each prompt is performing better or worse.

Task Description	Example Output		
Article Summarization	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.		
Q&A in HR Chatbot	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.		
Email Generation	Subject: Join Us for a Multicultural Holiday Celebration!  Dear Team,  As the holiday season approaches, we are thrilled to announce our annual company holiday party! This year, we are embracing the diverse cultures within our team and celebrating a variety of holiday traditions.  Details are as follows:  Date: December 20th		
	Time: 6:00 PM - 10:00 PM Location: Main Conference Hall  Our celebration will feature a variety of activities that honor different cultural traditions, including:		
	Decorations from Around the World: Help us decorate the office with symbols and ornaments from various holidays, such as Christmas trees, Hanukkah menorahs, Kwanzaa kinaras, Diwali lamps, and more. Inclusive Gift Exchange:  Participate in a gift exchange that embraces the spirit of giving across cultures. Bring a wrapped gift (under \$20)		
	or a small item that represents your cultural heritage.  - Global Holiday Music: Enjoy a diverse selection of holiday music, including Christmas carols, Hanukkah songs, Kwanzaa tunes, and traditional music from different cultures.		
	- International Feast: Indulge in a delicious array of holiday dishes from around the world. We will have a variety of foods to cater to different tastes and dietary preferences.		
	Feel free to wear traditional attire or festive clothing that reflects your cultural heritage. Whether it's a Christmas sweater, traditional clothing for Hanukkah, Kwanzaa, Diwali, or any other holiday, we welcome it all!		
	We would love to hear your ideas and suggestions to make this event even more inclusive and enjoyable. If you have any special traditions or activities you'd like to share, please let us know.  Let's come together to celebrate the season and the wonderful diversity within our team.  Happy Holidays!  Best regards,  Company CEO		

Table 1. Example Outputs for Different Task Types in the Evaluation Study. This table presents example outputs generated for three distinct tasks: Article Summarization, Q&A in an HR Chatbot, and Email Generation.

I've had to look at a number of tooling before in relation to checking the validity of models. (Cross validation, Lime, SHAP, Rouge, Bleu, Flesch-Kincaid, Flesch. I've played with others but not to a level I can say I would know enough about them (eg. SQuAD, CIDEr, SPICE, WIT, ELI5)

# 5.2 RQ1: What do practitioners prioritize in evaluation criteria when using LLMs as judges, and how do their priorities differ?

The criteria that users created varied widely across tasks, not only in terms of specificity, but also in the presence of instructions or rules, the number of items in the scale (for the direct assessment condition), the inclusion of exclusion criteria, and the use of examples. These variations reflected what aspects users prioritized when defining their criteria. To systematically analyze these differences, we generated a codebook and used it to categorize the observed criteria changes(see Table 4). We observed a range of specificity, with some users providing highly detailed and task-specific criteria, while others offered more general guidelines. In some cases, users added rules or additional prompting to guide the model. For example, some criteria included explicit if-then rules to distinguish between acceptable and unacceptable options. Additionally, some users incorporated examples within their criteria to illustrate preferred outcomes, effectively providing the model with concrete cases to guide its responses. This approach highlights the diverse ways in which

Type of Change	Original Criteria	Changed Criteria
Criteria Specified	The generated email should be inclusive. It should mention the different cultures and holidays (such as Christmas, Hanukkah, Kwanzaa, Diwali, and others) . It should not be exclusive to one culture.	The generated email should be inclusive. It should mention the different cultures and holidays (such as Christmas, Hanukkah, Kwanzaa, Diwali, and others). It should not be exclusive to one culture. The email should use inclusive terms such as holiday and festive as opposed to terms exclusive to one culture.
Criteria Generalized	The email is inclusive to all cultures, backgrounds, genders, etc.	The email is inclusive to all cultures, and backgrounds.
Scale Item Removed	Does the summary contain the main topic of the article? Scale: Absolutely, Somewhat, No	Does the summary contain the main topic of the article? Scale: Yes, No
Scale Item Added	Does the response capture the summary in the best possible way? Scale: Yes, No	Does the response capture the summary in the best possible way? Scale: Excellent, Good, Average, Poor
Scale Item Specified	Read the following email and determine if the email is inclusive or not inclusive of cultural differences.  Inclusive: The email acknowledges the different cultures  Not inclusive: The email focuses on only one culture or does not acknowledges cultural differences.  Maybe: Not sure.	Read the following email and determine if the email is inclusive or not inclusive of cultural differences.  Inclusive: The email acknowledges the different cultures.  Not inclusive: The email focuses on only one holiday, group, or culture and does not acknowledges cultural differences.  Maybe: Not sure.
Scale Item Generalized	Inclusive: The email emphasises an inclusive company culture by asking for participations from all kinds of traditions and cultural practices. It also uses inclusive language	Inclusive: The email is focusing on all kinds of traditions and cultural practices. It also uses inclusive language
Minor Edit	What is the best summary?	Which is the best summary?
Instruction to Model	Which of the following summaries best describe the article. The summary should be reflective of the key points in the article.	Which of the following summaries best describe the article. The summary should be reflective of the key points in the article. Each article must be reviewed on its own and ignore all other summaries while doing so.
Exclusion Criteria Added	The summary should be accurate and concise	The summary should be accurate and concise. Has no fake data generated outside of reference

Table 2. Types of changes made across every evaluation.

participants interpreted and operationalized the constructs in question. Moreover, the number of items included in the scale for direct assessment also varied, where some users offered detailed multi-item scales, while others opted for simpler, binary options.

5.2.1 Criteria Priority Differences within Tasks. Within each task, participants prioritized different aspects when defining and interpreting criteria, leading to varied approaches. Using inductive coding [40], two authors generated a codebook of these priorities, as shown in Table 5.

In the Email Inclusivity task, participants prioritized cultural inclusivity, neutrality, and fairness. Some emphasized mentioning all cultures to avoid bias, ensuring holidays like Christmas, Hanukkah, Kwanzaa, and Diwali were included. Others focused on maintaining a neutral tone by avoiding specific mentions of culture, gender, race, or ethnicity. Additionally, some prioritized fairness, ensuring the email did not favor one culture over another. One participant intepreted inclusivity to mean financial inclusivity and designed the criteria around adherence to budget constraints, such as keeping gift exchange mentions within a \$40 limit.

ID	Job Role
P1	AI Engineer
P2	Research Scientist Intern
P3	Software Developer
P4	Data Scientist
P5	Software Engineer
P6	Senior Research Scientist
P7	Advisory AI Engineer
P8	Senior AI Technical Architect
P9	Distinguished Engineer and Master Inventor
P10	Research Scientist
P11	Research Software Engineer
P12	Senior Technical Staff Member
P13	Platform Engineer
P14	Research Scientist Intern
P15	Technology Engineer

Table 3. Participant IDs and job roles

In the Summarization task, participants varied in their focus on factual accuracy, inclusion of key points, and conciseness. Some prioritized ensuring the summary accurately reflected the article, while others focused on covering all key points comprehensively. Conciseness was also important, with participants aiming for a summary around 20% of the original length. Grammatical correctness and the inclusion of crucial details, like dates or prices, were also emphasized. Some participants relied on the AI to decide which is the best summary by keeping their criteria general and asking for the the "best" summary.

For the Q&A Faithfulness task, participants focused on maintaining faithfulness to the source, adhering to policy, avoiding hallucinations, and ensuring correctness. They emphasized that responses should be strictly grounded in the reference document and follow company policy, with no deviation or fabrication. These differences within each task reflect the varied approaches participants took based on their interpretations and judgment of what was most important for the specific context.

# 5.3 RQ2: What strategies do users employ to refine their criteria for achieving human-Al alignment in LLM-as-a-judge evaluations?

To understand how people refine their criteria when using large language models (LLMs) as judges, we logged all metadata, including the criteria and AI evaluator used, for each evaluation across various tasks. Two authors coded the types of edits users made, identifying several ways participants modified their criteria to better align with task requirements or personal preferences. One common modification was making criteria more specific. For example, in the Email Inclusivity task, a participant refined the criterion from simply stating that "The generated email should be inclusive" to specifying that it should use inclusive terms like "holiday" and "festive." This added precision provided clearer guidance for evaluation. In contrast, some participants generalized their criteria to broaden their scope. In the same task, a criterion initially focused on specific groups—such as cultures and backgrounds—was simplified to a more general statement: "The email is inclusive to all cultures and backgrounds."

Participants also made adjustments by removing unnecessary scale items or adding more detailed options. For instance, in the Summarization task, an original scale with multiple levels ("Absolutely," "Somewhat," "No") was simplified to a binary "Yes" or "No," while in other cases, the scale was expanded to include detailed levels like "Excellent," "Good,"

Category	Definition	Taxonomy	Example
Specificity	The criteria ranges in how specific it is to the nature of the task context example	High Specificity	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.
		Low Specificity	The email is inclusive.
Additional Prompting	The criteria includes instructions beyond criteria description	Present	Which summary offers the most clear and correct answer. As well as answering anything the customer did not know what to ask for in relation to their question. You must review each summary independently of the other summaries when making your judgement.
		Absent	Which is the best summary?
Rules	If/else rules provided	Present	"Does the response refer to the Travel and Expense policy? - Good Makes exceptions to known rules (e.g., trip length different or different meeting days) - Bad Does the response say all charges should be to corporate card while picking up on the fact that upgrading/business class is a personal expensecorrectly dissociates the corporate from the personal expense -Good Including rationale or policy statements - Good"
		Absent	Is the response faithful according to the reference document?
Exclusion Criteria	Criteria includes what should not be considered	Present	'The summary does not repeat unimportant details.
		Absent	The summary covers important aspects from the text.
Examples Provided	Examples provided within cri- teria	Present	Criteria: The email should be inclusive taking into consideration all employees, 'Yes': "Happy Holidays! let's bring some gifts for all", 'No': "Merry Christmas! let's have Christmas tress!"
		Absent	Criteria: Is the email inclusive?, 'Yes': 'The email is inclusive.', 'No': 'The email is not inclusive.'

Table 4. Categories of final criteria generated by participants for each task.

"Average," and "Poor." In addition, some users specified existing scale items to clarify their meaning, such as adding definitions for "Inclusive" and "Not Inclusive" in the Email Inclusivity task. Conversely, others generalized scale items, broadening them to encompass a wider range of cultural practices. These results can be seen in Table 2.

Minor edits were also common, such as rephrasing criteria for clarity. For example, the question "What is the best summary?" was revised to include clearer instructions: "Which of the following summaries best describe the article? Review each independently of the others." Some participants added exclusion criteria to prevent irrelevant elements from being considered. One participant, for instance, added a criterion stating that "The summary should be accurate and concise. Has no fake data generated outside of reference," clearly outlining what should be excluded. These refinements illustrate the range of strategies participants used to tailor their criteria, from increasing specificity to allowing greater flexibility. Figure 8 shows the sequential changes in evaluation criteria by task type and user.

Task	Interpretation of Criteria	Example
Email Inclusivity	Email must mention all possible cultures	The generated email should be inclusive. It should mention the different cultures and holidays (such as Christmas, Hanukkah, Kwanzaa, Diwali, and others). It should not be exclusive to one culture. The email should use inclusive terms such as holiday and festive as opposed to terms exclusive to one culture.
	Email must not mention any specific culture	"The email is very formal and does not mention any specific gender, race, ethnicity, culture based terms or words."
	Email must not favor one culture to another	Which email invitation is most inclusive? An email is inclusive if it does not favour one religion over others and if it does not imply that all employees belong to a certain group or celebrate a specific holiday. An inclusive email refrains from using denominational words.
	Gift exchange budget mentioned in email should not exceed particular amount	"The price mentioned falls within the \$40 budget limit."
Summarization Preference	Summary must be factual	The summarization contains factual statements from the article
	Summary must include key points from the article	The summary accurately conveys the main points of the article.
	Summary must be succinct	The summary length is approximately 20% of the original article length.
	Summary must be grammatically correct	Is the response grammatically, lexically, semantically, and syntactically correct and common, including punctuation accuracy?
	Summary must include specific details (detailed by user) from summary Summary should be the "best"	Summary described dress worn by Vivien Leigh with details of price and date.  Which is the best summary?
Q&A Faithfulness	Response must be grounded/faithful to the reference document	"Based on the reference document, does the response answer the customer question correctly?"
	Response should adhere to company policy	"Has the employee claimed only the regular "in policy" fare and not the personal cost?"
	Response should not include hallucinations Response should be correct	"Are there some hallucinations in the answer?" "Which summary offers the most clear and correct answer."

Table 5. Variations of interpretations of criteria within each task.

# 5.4 RQ3: How do task-related factors and judge strategy impact how practitioners refine criteria?

5.4.1 What are the differences for the total amount of evaluations run? Participants were able to run multiple evaluations for each task yielding a total of 131 evaluations. We investigated whether there were differences in the strategies users employed when refining their criteria, particularly focusing on the total number of evaluations conducted. A repeated measures ANOVA was performed to compare the number of evaluations across different judge strategies (pairwise vs. rubric) and tasks (Q&A Faithfulness, Email Inclusivity, Summarization). The results indicated a significant difference in the number of evaluations between judge strategies, with more evaluations being conducted in the direct assessment (rubric) condition compared to the pairwise condition. Specifically, the ANOVA showed that the judge strategy had a significant impact on the number of evaluations run (F(1, 14) = 12.64, p = 0.003), and the interaction between task and strategy was not significant. Post-hoc analysis using Tukey's HSD further confirmed that the pairwise strategy led to significantly fewer evaluations than the rubric strategy (mean difference = 0.72, p = 0.007). The number of evaluations can be seen in Figure 9

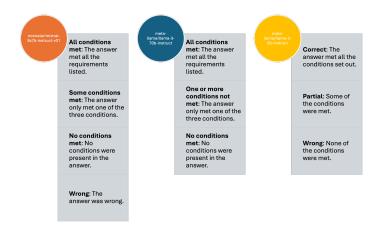


Fig. 7. Example of changes made to criteria and AI evaluator for the by Participant #1 for the Q&A HR Task. The participant changed the AI evaluator model across the evaluations and changed the number of items in the scale as well as their corresponding definitions. The criteria definition remained the same throughout the evaluations: Did the answer give the following? (1) Factually correct answer based on the document. (2) Included related information directly to the question. (3) Answer is clear and concise.

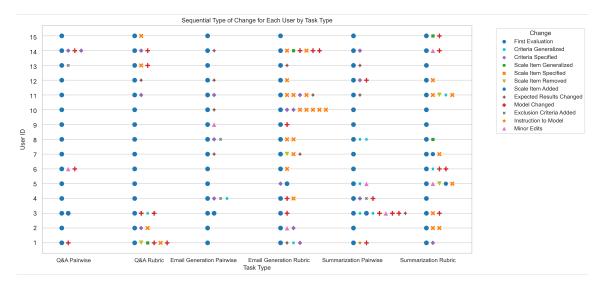
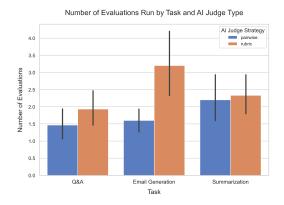


Fig. 8. Sequential changes in evaluation criteria by task type and user. Symbols represent different types of modifications, including changes to criteria, scale items, models, and exclusions, with variation across Q&A, Email Generation, and Summarization tasks.

5.4.2 What are the differences in human-Al alignment? We examined the the extent to which users agreed with the AI model's assessments across different conditions. Upon completion of each task, users indicated agreement with the model's judgement. We conducted a repeated measures ANOVA, focusing on the judge strategies (pairwise vs. rubric) and the tasks (Q&A Faithfulness, Email Inclusivity, Summarization). The analysis revealed a marginally significant main effect for task type, F(2, 28) = 2.73, p = 0.08, indicating that there was a difference between alignment achieved between the tasks. Tukey post hoc analysis further confirmed this finding, showing a marginally significant difference Manuscript submitted to ACM



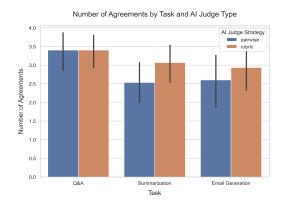


Fig. 9. Participants conducted more evaluations in the direct assessment task compared to the pairwise task (p < 0.05).

Fig. 10. There were marginal significant differences in the degree of alignment achieved across task types (p = 0.08).

in alignment between the the Q&A task and the Email Generation Task (mean difference = -0.633, p = 0.08). The number of agreements can be seen in Figure 10.

# 5.5 RQ4: How Do Task-Related Factors and Judge Strategy Impact User Perceptions?

We aimed to investigate whether task-related factors and judge strategy impact user perceptions, specifically focusing on AI trust, self-satisfaction, perception of explanations, and cognitive load. For each of these measures, we conducted a repeated measures ANOVA, and we report the results below.

# 5.6 Al Trust

For trust in the AI evaluator, a repeated measures ANOVA showed no significant differences between tasks or AI strategies (Figure 11). Despite this, we identified reasons why participants either trusted or distrusted the AI evaluator.

5.6.1 Reasons for Trust. Participants who reported higher trust (above the median score of 4) highlighted several factors, such as the alignment between AI judgments and their criteria, prior positive experiences with specific models, and cohesive explanations. Some examples of feedback include:

I didn't trust fully in the beginning, but the explanations helped. #P10 (Summarization, Rubric)

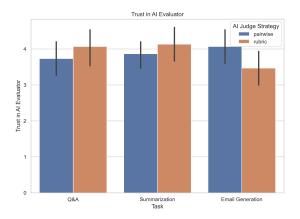
Assessment aligned with my own. #P5 (Q&A, Rubric)

The AI judge was more accurate than me on my own faithfulness metric. It helped me identify gaps in my quick assessment. #P9 (Q&A, Rubric)

5.6.2 Reasons for Distrust. Conversely, participants who expressed lower trust often felt that the AI model imposed its own criteria or failed to align with their expectations:

The model was opinionated on the way to evaluate and didn't enforce my criteria. #P4 (Email Generation, Pairwise)

5.6.3 Perception of Explanations. We conducted an ANOVA to determine whether the type of AI Judge Strategy (pairwise vs. rubric) affected participants' perception of how helpful the explanations were in completing tasks. The
Manuscript submitted to ACM



Perception of Explanations

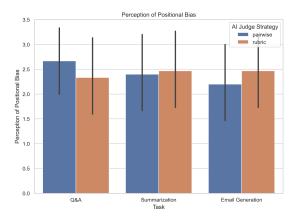
Al Judge Strategy
pairwise
rubric

1

Q&A
Summarization
Task

Fig. 11. No significant differences in user trust in the Al evaluator were found between the tasks or the Al judge strategies.

Fig. 12. Explanations were percieved as more helpful in the rubric condition than in the pairwise condition (p<0.05).



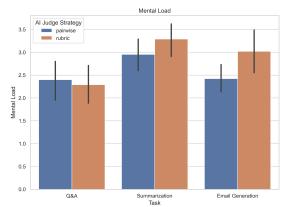


Fig. 13. No significant differences in perceived positional bias helpfulness were found across tasks or AI judge strategies.

Fig. 14. Users experienced higher cognitive load in the Summarization task than the Q&A task p<0.05).

results indicated a significant main effect of the type of AI strategy, F(1, 14) = 7.79, p = 0.014, with rubric-based explanations rated more favorably overall compared to pairwise. However, there were no significant effects of task type, F(2, 28) = 0.09, p = 0.9152, nor was there a significant interaction between task and strategy, F(2, 28) = 0.67, p = 0.521.

To further explore these differences, we performed a post hoc analysis using Tukey's HSD test. The results revealed that participants rated rubric explanations significantly higher than pairwise explanations (mean difference = 0.73, p = 0.008), suggesting that rubric explanations were perceived as more useful across all tasks (Figure 12). We also analyzed participants' feedback on when explanations were perceived as helpful versus unhelpful. Explanations were deemed helpful in two key ways: they served as tools for refining criteria or as a means of validating existing criteria.

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 (Email Generation, Pairwise)

Conversely, explanations were considered unhelpful when they were either not used or were found to be ineffective in resolving discrepancies.

Explanations helped me see the shortcoming of AI, but the shortcomings were not able to be fixed with multiple iterations of criteria refinement. #P4 (Email Generation, Pairwise)

5.6.4 Perception of Positional Bias. Of the 196 total evaluations run by the 15 users, positional bias appeared in 35% of those evaluations. We conducted an ANOVA to determine whether the type of AI Judge strategy (pairwise vs. rubric) and task type (Q&A, email generation, summarization) affected participants' perception of how helpful indication of positional bias was in completing the tasks. The analysis revealed no significant main effects, indicating that neither the type of task nor the strategy used by the judge had a significant impact on participants' perceptions of the helpfulness of positional bias (see Figure 13). This lack of effect could potentially be explained by the fact that positional bias was only present in 35% percent of evaluations. An independent samples t-test was conducted to compare the perceived helpfulness when participants results included positional bias and when they did not. The results showed a significant difference, t(28) = 2.98, t = 0.003, showing that when positional bias was present, participants reported it being more helpful than it did not appear in the results. The primary reason for positional bias not being helpful was that it did not appear in the results.

I quickly glanced to see if there was positional bias detected but it was not. Hence, I did not follow up on that. #P10 (Q&A, Pairwise)

Participants reported that the positional bias was helpful because it alerted them to rephrase their criteria.

Positional bias inspired me to simplify the criteria and rephrase them. #P2 (Summarization, Rubric)

It alerted me that some rephrasing might be needed before expanding to the whole dataset. #P2 (Q&A, Pairwise)

5.6.5 Cognitive Load. We also assessed cognitive load using three items from the NASA TLX (Task Load Index). The repeated measures ANOVA revealed a significant main effect of the task on cognitive load, F(2, 28) = 9.41, p = 0.0008, indicating that the cognitive load varied significantly depending on the task. Post hoc analysis using Tukey's HSD test showed that cognitive load was significantly higher in the summarization task compared to the Q&A task (meandiff = .779, p = 0.001) (see Figure 14). There was no significant effect of judge strategy on cognitive load, F(1, 14) = 3.02, p = 0.1044, and the interaction between task and judge strategy also did not reach significance, F(2, 28) = 2.23, p = 0.126.

# 5.7 RQ5: What do practitioners prefer: Direct Assessment or Pairwise?

Participants were asked to reflect on whether they preferred direct assessment via creating rubrics and whether they preferred pairwise, or whether it was contextual. Participant responses ranged with almost half (7 participants) reporting that they preferred direct assessment, and the remaining being split between pairwise and saying it depends/its contextual. The overview of preference reasons can be seen in Table 15. Below we report why participants preferred one judge strategy over another.

5.7.1 Preferred Direct Assessment. Participants felt that Direct Assessment gave them more control over how the model outputs were evaluated and appreciated the detailed evaluation of the responses.

Direct Assessment is more clear task for me to understand the results. Sometime ranks provided by Pairwise Assessment can be set in different order and still be correct. #P1

Preference	Code	Definition of the Code
Direct	Clarity and Control	Clearer and easier to understand, offering more
		control over the evaluation process and outcomes.
	Detailed Evaluation	Allows individual evaluation of responses, making
		it more useful for tasks requiring thorough review.
Pairwise	Flexibility and Nuance	Provides flexibility and nuance in evaluating ab-
		stract or subjective tasks.
	Ease of Criteria Formulation	Easier to create evaluation criteria, especially when
		rubrics are harder to define.
Neither	Combination Approach	A combination of both methods works better for
		complex tasks, complementing each other through
		classification (rubric) and ranking (pairwise).
	Task-Specific Use Cases	Direct Assessment is suited for classification or
		compliance, while Pairwise is better for identifying
		preferences or the best result.

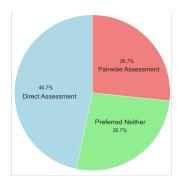


Fig. 15. Codes and Reasons for Assessment Preferences

Fig. 16. Assessment Preferences

I prefer to create direct assessments because I feel that I can control more the outcomes of the model. #P14 I would prefer the direct assessment mainly as it involves evaluating each response individually. Rankings are just relative preferences and they do not offer a detailed way of evaluating the model response. As an example in if we have all bad outputs, we would still have some ordering among them but that is essentially irrelevant to judge the response of the model which is not the case in rubric based criteria. #P3

5.7.2 Preferred Pairwise Assessment. Of the 15 participants in the study, 4 reported preferring pairwise assessment over direct assessment. The reasons included that participants felt that pairwise assessment allowed for more flexibility and nuance in evaluating subjective tasks where a binary or rigid criteria may not apply. Others reported that the pairwise was simpler and easier to user when formulating evaluation criteria, since detailed rubrics may be more complex.

"Pairwise is easier to use for crafting an evaluation criteria, especially for the summarization task where the 'goodness' of a summary is very abstract and can be broad." #P7

It's easier to formulate a sentence containing the complete criteria and not have to think about rubrics. #P5

5.7.3 Use of Al Judge Strategy is Contextual. Participants who preferred neither judge strategy reported that they would either like to use a combination of strategies to accomplish their task, or that their preference would depend on the nature of the task.

Various tasks are of various complexities, I feel that ranking and classification results coming out of the pairwise and direct assessment can help the user make better decisions which are both relative and independent.
#PA

I believe they have different use cases. Rubric can be used in a case of classification or compliance for example. I might want to consider all answers and see what classifications I can assign to them. Pairwise is about getting the best result for what you want. I don't care as much about the others, only that one is the correct. #P8

# 6 DISCUSSION

# 6.1 Variability in Subjective Criteria: Defining Stakeholder Needs in Al-assisted Evaluation

We observed significant variation in how participants defined criteria within tasks, a finding consistent with prior HCI research on diversity in subjective judgments [22]. For example, "inclusivity" had different interpretations among participants. Some participants believed it meant mentioning every possible holiday to represent all religions, while Manuscript submitted to ACM

others thought it meant avoiding specific cultural or holiday references altogether. This aligns with studies in inclusive design, where users with diverse backgrounds interpret fairness and inclusivity differently, highlighting the challenges of capturing subjective values in AI systems [10]. This variation underscores the importance of clear stakeholder definitions for criteria definitions, a need echoed in crowdsourcing research, which shows that subjective tasks can lead to inconsistent evaluations without precise guidelines [26]. When developing criteria to evaluate model outputs, it is essential to define stakeholder needs clearly, especially for subjective criteria like inclusivity.

6.1.1 Mitigating Over-Specificity in AI-Assisted Evaluation: Balancing Task Context and Generalization. When analyzing how participants adjusted their criteria to align with the AI's judgments, we observed a consistent tendency to make the criteria increasingly specific to the task at hand. This resulted in overly narrow criteria that worked well for evaluating a single article's summary but failed to generalize across multiple summaries from different articles. Other AI-assisted evaluation systems [25, 38] do not necessarily prioritize sampling diverse task contexts and outputs, especially when the evaluation heavily depends on prompts. This limits the user's exposure to varied outputs, increasing the risk of overspecifying criteria to a particular task. To address this issue, it is critical to expose users to diverse task contexts. For instance, in the case of article summarization, this would mean providing users with varied articles and summaries to encourage the development of criteria that can be applied more broadly. A designimprovement could involve enabling users to upload large datasets, with the system recommending or allowing the selection of diverse samples to help develop more robust criteria.

Conversely, we found that a subset of users kept constructs very general. For example they would rely on the AI evaluator to interpret what it means to be "inclusive" or what it means to be grounded in the document. This lack of specificity can also lead to subpar results. Striking a balance between overly specific and overly vague criteria is essential. Encouraging users to refine their criteria while keeping them general enough for broad application can improve quality. Paired with exposure to diverse outputs, this approach can create more effective evaluation processes that work well across different contexts.

6.1.2 Challenges of Natural Language Criteria Formulation. When users expressed their criteria in natural language, we observed significant variation in how they structured their inputs, particularly in the pairwise condition where no predefined form or structure was provided. Some participants introduced additional instructions, asking the AI evaluator to consider each response independently, which contradicts the purpose of pairwise comparisons where outputs are meant to be directly compared. Others created complex rule-based systems, such as: If the response includes [insert characteristic], then it is "good"; if it does not, then it is "bad". Additionally, rather than defining item scales as instructed, participants often provided examples of "good" and "bad" responses. These alternative interaction patterns did not align with the intended functionality of the evaluation process and led to suboptimal results. To address this, greater transparency should be provided regarding the prompts used in both direct and pairwise evaluation conditions, ensuring users have a clearer understanding of how their input will be processed. Providing more examples that users can customize to fit their specific needs could also improve outcomes, as well as implementing auto-correction features to guide users in entering the criteria correctly.

# 6.2 Alignment: Refining Criteria, Changing the Al Evaluator, and Altering Expected Results

We observed various user interactions throughout the evaluation process that mirror principles seen in interactive machine learning (iML), where users iteratively refine models by selecting examples, labeling data, and evaluating outputs [34, 39]. In line with iML approaches, users in our study adjusted their criteria throughout the evaluation,

Manuscript submitted to ACM

making them more general or specific as needed. This adaptive behavior reflects the need for flexible interfaces that allow seamless transitions between specifying, revising, and evaluating criteria. As noted in iML, reducing user effort while improving output alignment is crucial [12]. In some cases, midway through the evaluation, users even changed their expected outcomes to align more closely with the Al's judgments—especially after reading the explanations, which often convinced them of the Al's output. Users frequently revised their criteria, whether by adding or removing scales in the rubric condition, or rephrasing their criteria in the pairwise condition. They also experimented with different Al evaluators, switching between models like Mixtral and LLaMA to compare results. This suggests that the path to "alignment" can take many forms, from adjusting one's own expectations, to switching the Al model making the judgments, to modifying the criteria itself in search of a better fit. This process could be better supported in the future by automatically testing criteria with multiple models and showing results side-by-side.

#### 6.3 Adaptive Evaluation Strategies: Balancing Clarity and Flexibility

Participants expressed a preference for Direct Assessment due to the clarity and control it offered, allowing for detailed, individualized evaluations. On the other hand, Pairwise Assessment was valued for its flexibility, particularly in more subjective tasks. Some participants preferred a hybrid approach, selecting an evaluation method based on the specific task. The higher number of evaluations in direct assessment suggests that users felt more engaged, likely because it provided them with a greater sense of control. This aligns with the reasons participants cited for favoring direct assessment. These findings highlight the importance of adaptable evaluation strategies tailored to task type. All systems could benefit from offering multiple evaluation modes or an adaptive assessment strategy that adjusts based on task complexity. For example, users might opt for direct assessment for objective tasks and switch to pairwise ranking for more subjective ones. Additionally, some users might prefer a combination approach—starting with direct assessment to filter out responses and then using pairwise ranking to further refine the best options. Future research can explore hybrid systems that adapt dynamically based on user workflows.

# 6.4 Bias Awareness and Explanation Visibility in Evaluation Strategies

Positional bias was considered helpful when present, prompting participants to revise and improve their criteria. This result suggests that highlighting AI biases can enhance the evaluation process and improve alignment between human and AI judgments. While our study focused on positional bias, there is an opportunity to explore the representation of other biases, such as self-enhancement bias and verbosity bias. For example, is the model ranking outputs highly because they were generated by the AI evaluator, or is it favoring longer responses even if they are less accurate or clear? Research is lacking on how users perceive such bias flagging and how it impacts their criteria refinement.

Explanations were perceived as more helpful in the direct assessment condition. One reason, as reported by participants, is that in the pairwise condition, explanations often went unnoticed. Since every comparison generates an explanation, these explanations were only accessible via a modal pop-up, requiring users to click on the result. Even then, users had to sift through multiple explanations to understand the ranking. This limitation in the pairwise condition presents an opportunity to redesign how explanations are presented, making them more concise and digestible. When explanations were more visible, as in the direct assessment condition, they proved to be more effective, suggesting a need for improvement in how pairwise explanations are displayed.

#### 7 CONCLUSION

We introduced EvalAssist a tool designed to help practitioners refine evaluation criteria using both direct and pairwise assessment strategies. The tool provides positional bias metrics and explanations for each AI judgment. In a controlled experiment with 15 machine learning practitioners, we examined how users refine their criteria and identified key differences between the two evaluation approaches. Direct Assessment was preferred for its clarity and control, while Pairwise Assessment was valued for flexibility in subjective tasks. Users often refined their criteria by increasing specificity or simplifying scales to improve alignment with task needs. Human-AI alignment was stronger in the rubric-based (Direct Assessment) condition, highlighting the importance of clear evaluation criteria. Our findings suggest that AI-assisted evaluation systems should offer flexible, adaptive strategies, allowing users to switch between direct and pairwise methods. Such systems should support ongoing criteria refinement, provide clear explanations, and improve alignment between human and AI judgments to enhance evaluation reliability.

#### **REFERENCES**

- [1] Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet Accessed: 2024-09-09.
- [2] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing., 18 pages.
- [3] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2024. Benchmarking foundation models with language-model-as-an-examiner., 26 pages.
- [4] Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, E. Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. https://arxiv.org/abs/2406.18403
- [5] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. 454–464 pages.
- [6] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement biases.
- [7] Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations?
- [8] Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into using large language models for automatic evaluation. , 8928–8942 pages.
- [9] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. https://doi.org/10.48550/ARXIV.2210.11416
- [10] Sasha Costanza-Chock. 2020. Design justice: Community-led practices to build the worlds we need.
- [11] Michael Desmond, Zahra Ashktorab, Qian Pan, Casey Dugan, and James M Johnson. 2024. EvaluLLM: LLM assisted evaluation of generative outputs. , 30–32 pages.
- [12] Michael Desmond, Michelle Brachman, Evelyn Duesterwald, Casey Dugan, Narendra Nath Joshi, Qian Pan, and Carolina Spina. 2022. AI Assisted Data Labeling with Interactive Auto Label. , 13161–13163 pages.
- [13] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation., 391–409 pages.
- [14] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. , 681–694 pages.
- [15] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text., 103–166 pages.
- [16] Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, and Hans-Arno Jacobsen. 2013. Bigbench: Towards an industry standard benchmark for big data analytics., 1197–1208 pages.
- [17] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later., 904–908 pages.
- [18] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding.
- [19] Hui Huang, Yingqi Qu, Jing Liu, Muyun Yang, and Tiejun Zhao. 2024. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers.
- [20] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B.

[21] Minsuk Kahng, Ian Tenney, Mahima Pushkarna, Michael Xieyang Liu, James Wexler, Emily Reif, Krystal Kallarackal, Minsuk Chang, Michael Terry, and Lucas Dixon. 2024. Llm comparator: Visual analytics for side-by-side evaluation of large language models. , 7 pages.

- [22] Evangelos Karapanos, Jean-Bernard Martens, and Marc Hassenzahl. 2009. Accounting for diversity in subjective judgments., 639-648 pages.
- [23] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models.
- [24] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2023. Evallm: Interactive evaluation of large language model prompts on user-defined criteria.
- [25] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2023. Evallm: Interactive evaluation of large language model prompts on user-defined criteria.
- [26] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work., 1301–1318 pages.
- [27] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks., 9459–9474 pages.
- [28] Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons. arXiv:1909.03087 [cs.CL] https://arxiv.org/abs/1909.03087
- [29] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models.
- [30] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models.
- [31] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.
- [32] Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from Large Language Models faithful?, 295-337 pages.
- [33] Qian Pan, Zahra Ashktorab, Michael Desmond, Martin Santillan Cooper, James Johnson, Rahul Nair, Elizabeth Daly, and Werner Geyer. 2024. Human-Centered Design Recommendations for LLM-as-a-Judge.
- [34] Kayur Patel, Naomi Bancroft, Steven M Drucker, James Fogarty, Amy J Ko, and James Landay. 2010. Gestalt: integrated support for implementation and analysis in machine learning. , 37–46 pages.
- [35] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability., 52 pages.
- [36] Ravi Raju, Swayambhoo Jain, Bo Li, Jonathan Li, and Urmish Thakkar. 2024. Constructing Domain-Specific Evaluation Sets for LLM-as-a-judge.
- [37] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.
- [38] Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya G Parameswaran, and Ian Arawjo. 2024. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences.
- [39] Patrice Simard, David Chickering, Aparna Lakshmiratan, Denis Charles, Léon Bottou, Carlos Garcia Jurado Suarez, David Grangier, Saleema Amershi, Johan Verwey, and Jina Suh. 2014. Ice: enabling non-experts to build models interactively for large-scale lopsided problems.
- [40] David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. , 237–246 pages.
- [41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models., 24824–24837 pages.
- [42] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of Retrieval-Augmented Generation: A Survey.
- [43] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena.
- [44] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging LLM-as-a-judge with MT-bench and Chatbot Arena., 29 pages.