Designing Scalable and Transparent Interfaces for Multi-Criteria Evaluation of LLM Outputs

Amin El Asery
masery3@gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

Lakshya Sharma
lsharma697@gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

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

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

Justin D. Weisz
jweisz@us.ibm.com
IBM Research
Yorktown Heights, NY, USA

Abstract

This paper investigates the challenges and opportunities in evaluating outputs generated by large language models (LLMs) at scale. With LLMs increasingly integrated into applications ranging from customer service to content creation, organizations face significant hurdles in assessing qualitative aspects such as accuracy, coherence, bias, and compliance with brand guidelines. Through a comprehensive literature study and comparative analysis of existing evaluation tools—including both graphical interfaces and code-driven systems—this research identifies critical challenges in scalability, multi-criteria support, aggregation of results, and transparency. Complementing the literature review, contextual inquiries with professionals from diverse technical backgrounds provided insights into user preferences and practical challenges in evaluating extensive datasets of LLM outputs. Based on these findings, we propose design recommendations for next-generation LLM evaluation tools, emphasizing advanced filtering and drill-down capabilities, multi-level aggregated insights that combine quantitative and qualitative analyses, iterative refinement of evaluation criteria to adapt to evolving requirements, and interactive visualizations that elucidate the underlying scoring processes. The recommendations aim to enhance the reliability and trustworthiness of evaluation systems, ultimately supporting more efficient and nuanced assessments of LLM performance across varied real-world applications.

1 Introduction

Large Language Models (LLMs) have become integral to a wide range of applications, from customer service to content generation. Organizations are increasingly adopting LLMs to enhance productivity, streamline workflows, and support human decision-making. However, as these models are deployed in real-world scenarios, critical questions arise about their ability to understand and adhere to the qualitative criteria required for generating high-quality, contextually appropriate outputs. Evaluating the performance of LLMs at scale—particularly across dimensions such as accuracy, coherence, ethical alignment, and compliance with brand guidelines—remains a significant challenge [1]. Professionals such as AI developers, data scientists, and consultants are often tasked with assessing thousands of LLM-generated outputs, balancing the need for efficiency with the demand for nuanced, multi-criteria evaluations. While numerous tools and frameworks exist to support this process, many fall short in addressing key challenges: scalability, flexibility in criteria definition, meaningful aggregation of results, and transparency in evaluation processes [1]. These limitations hinder the ability of organizations to trust and effectively utilize LLMs in high-stakes applications. In this paper, we probe into this problem space, to address the following research questions:

- What are the challenges in evaluating large datasets of LLM outputs, and how can tools better support scalability and efficiency?
- How can evaluation tools accommodate diverse, customizable criteria to reflect the nuanced needs of different applications and stakeholders?
- What visualization and aggregation techniques are most effective for interpreting multi-dimensional evaluation results at scale?
- What design features are necessary to enhance the reliability, transparency, and trustworthiness of LLM evaluation tools?

To answer these questions, we conducted a comprehensive literature review and comparative analysis of existing LLM evaluation tools, identifying gaps and opportunities for improvement. We complemented this analysis with contextual inquiries involving professionals from diverse technical backgrounds, gathering insights into their workflows, pain points, and preferences. Based on these findings, we propose design recommendations for next-generation evaluation tools that prioritize scalability, flexibility, and transparency, ultimately enabling more effective and trustworthy assessments of LLM performance.

2 Related Work

2.1 Challenges in LLM Evaluation

The evaluation of Large Language Model (LLM) outputs has evolved significantly, moving from traditional reference-based metrics such as ROUGE [2] and BLEU [3] to more flexible frameworks that support qualitative and customizable criteria. Tools like EvalLM [4] exemplify this shift, offering interactive criteria definition, LLM-assisted refinement, and libraries of predefined criteria. Despite these advancements, several critical challenges remain in effectively evaluating LLM outputs at scale. Our research builds on these insights to address the following key challenges:

- **Scalability:** While manual evaluation by experts can provide detailed insights, it is time-consuming and resource-intensive, making it impractical for large datasets. Automated evaluation using LLMs offers a faster alternative, but it introduces new challenges in visualizing and interpreting results across thousands of outputs. [1]. Key questions include: How can users efficiently identify patterns, outliers, and areas of poor performance? And how can they validate automated evaluations without manually reviewing every output?
- **Multi-Criteria Support:** Organizations often require LLM outputs to adhere to diverse qualitative criteria, such as accuracy, coherence, ethical alignment, and brand compliance. However, the definition and weighting of these criteria can vary significantly across use cases. Our contextual interviews revealed that users value tools that allow for customizable and adaptable criteria, enabling them to tailor evaluations to specific organizational needs.
- **Aggregation and Visualization:** As evaluation datasets grow in size and complexity, users need tools that can aggregate results across multiple criteria and present them in an interpretable manner. Effective visualization techniques are essential for enabling users to quickly identify trends, compare performance across dimensions, and drill down into specific outputs for deeper analysis.
- **Transparency and Trust:** A recurring theme in our interviews was the need for transparency in the evaluation process. Users expressed skepticism about automated evaluations and emphasized the importance of understanding how scores are generated. Features such as explainable scoring mechanisms, detailed feedback, and the ability to trace evaluation logic were identified as critical for building trust in LLM evaluation tools.

These challenges highlight the need for next-generation evaluation tools that balance scalability, flexibility, and transparency, ultimately enabling more reliable and actionable assessments of LLM performance.

2.2 Existing Tools and Their Limitations

A range of frameworks has emerged to evaluate LLM outputs. Broadly, these frameworks can be grouped into two categories. The first category comprises user-friendly graphical interfaces, which allow professionals to evaluate LLM outputs with minimal coding. Examples include EvalLM and ChainForge [5], both of which feature intuitive dashboards for prompt generation, criteria definition, and model comparisons. The second category consists of code-driven systems, such as OpenAI Evals [6] and LM Evaluation Harness [7] , which rely on scripting in Python to perform evaluations, typically appealing to data scientists and engineers comfortable with coding workflows.

Despite the recent progress, scalability remains a leading challenge [1]. Although certain platforms facilitate structured evaluations, many still lack robust mechanisms for filtering, grouping, or dynamically navigating extensive datasets. As a result, practitioners must often hunt for localized errors or anomalies by manually sifting through large result sets. This absence of integrated segmentation and sampling features becomes especially problematic for iterative evaluation scenarios, where teams need quick, detailed feedback on performance across diverse prompts or topics.

Multi-criteria support is another area of concern. Evaluating LLM outputs often requires assessing multiple dimensions, such as accuracy, coherence, bias, and compliance with brand guidelines. However, many tools lack the flexibility to support diverse or customizable evaluation criteria. Our study illustrates that while some tools allow users to define and refine criteria, others rely on fixed or pre-defined metrics, limiting their applicability to specific use cases. This lack of flexibility can be restrictive when stakeholders need to evaluate nuanced factors, such as ethical considerations or domain-specific stylistic constraints. Furthermore, the ability to iteratively refine criteria and adjust their weighting is often absent, restricting users' ability to prioritize certain dimensions based on context.

Moreover, there is significant variation in the presentation and aggregation of results. Several tools only return raw numbers or textual feedback, leaving it to users to build custom dashboards for deeper insights. Others integrate visual reports—such as bar charts or aggregate summaries—but may lack the ability to overlay multiple evaluation criteria or group outputs by additional contextual factors. Without robust tools to segment and visualize performance, users can struggle to identify the precise conditions under which an LLM fails or excels, a shortfall that can slow optimization and deployment decisions.

Finally, user experience and trust remain top concerns. Code-based frameworks grant powerful customization and advanced functionality but are less accessible to non-technical users, potentially siloing the evaluation process. Meanwhile, graphical tools with simpler interfaces often have narrower capabilities for advanced analytics. Many practitioners also question the reliability and transparency of LLM-driven evaluation processes, particularly when these methods are used to judge subtle qualities like bias or ethical alignment. Furthermore, features such as historical performance tracking and reliability analysis are often absent, limiting users' ability to assess consistency over time.

In summary, the existing ecosystem of LLM evaluation tools remains fragmented. Professionals face trade-offs between user-friendliness and depth of customization, as well as between scalability and clarity of results. These limitations reinforce the need for solutions that can handle large datasets, incorporate flexible and context-specific criteria, provide actionable visualizations, and offer transparent, reproducible assessments to stakeholders with varying levels of technical expertise.

3 Methodology

3.1 Literature Study

3.1.1 Comparative Analysis. The comparative analysis of existing LLM evaluation tools was designed to address the challenges outlined in the problem statement. The tools examined include EvalLM, Constitution Maker[8], LLM Comparator[9], EvalGen[10], Deepchecks[11], and Robustness Gym[12]. Specifically, the analysis focused on identifying tools that effectively handle large-scale evaluations, support multi-dimensional assessments, and provide meaningful aggregation of results. These parameters were selected because they directly align with the needs of professionals—such as AI developers, data scientists, and consultants—who require tools capable of efficiently processing extensive datasets, evaluating outputs across diverse criteria, and delivering actionable insights through intuitive and customizable interfaces. The analysis was structured around the following key parameters:

Scalability. Our problem statement underscores the need for evaluating large volumes of LLM-generated text without overburdening the evaluation workflow. Consequently, we examined whether each tool offered robust mechanisms—such as filtering, grouping, or dynamic navigation—to efficiently manage extensive datasets. We assessed the extent to which these features enabled professionals to isolate anomalies, identify systematic errors, and conduct iterative evaluations quickly. Scalability is particularly critical for organizations deploying LLMs in production environments, where thousands of outputs must be evaluated daily.

Multi-Criteria Support. Professionals frequently judge LLM outputs on multiple dimensions, such as accuracy, bias, coherence, and style. Therefore, the ability to support diverse evaluation criteria is essential. We explored whether each tool accommodates various metrics concurrently and how easily users can introduce or switch between new criteria, such as ethical considerations or domain-specific guidelines. Special attention was given to whether the tools allow re-weighting or prioritizing certain metrics, which is crucial for real-world decision-making. For example, a tool that enables users to iteratively refine criteria based on evolving business needs can significantly enhance the relevance and accuracy of evaluations.

Aggregation Mechanisms. One of the core limitations identified in our problem statement is the complexity of synthesizing multi-dimensional results into clear, actionable insights. We investigated how effectively each tool aggregates evaluations across multiple criteria—whether it relies on raw data alone or offers meaningful visual summaries, charts, or trend analyses. Tools capable of displaying comprehensive scorecards or time-series plots can help teams identify emerging issues, compare different model versions, and track performance shifts over time. Effective aggregation mechanisms are particularly important for stakeholders who need to derive high-level insights from large, complex datasets.

User Interface and Usability. Professionals evaluating LLMs span a wide range of technical expertise, from highly specialized AI developers to consultants with minimal programming experience. Given this diversity, we assessed each tool's interface design and overall accessibility. Key factors included how quickly a new user could navigate the features, whether dashboards were clearly structured, and to what extent the platform minimized cognitive load—especially under tight iteration cycles. A well-designed interface not only expedites the evaluation process but also ensures that diverse stakeholders can interpret complex data effectively.

Customization and Flexibility. Finally, we examined each tool's capacity for customization, reflecting the widespread need to adapt evaluations to unique organizational or project requirements. We evaluated whether users could introduce new metrics (e.g., "ethical fairness" or "sentiment bias") or tailor existing workflows. This aspect is critical because real-world LLM deployments often require constant refinements of both prompts and evaluation rubrics to meet shifting business or regulatory constraints. Tools that offer flexibility in defining and adjusting evaluation criteria can better support the dynamic needs of organizations.

This comparative analysis not only highlights the strengths and limitations of existing tools but also informs the design recommendations for next-generation LLM evaluation systems, ensuring they meet the evolving needs of professionals and organizations.

3.1.2 *Feature Matrix.*

Following our comparative analysis of existing LLM evaluation tools, we developed a comprehensive *feature matrix* (**Table 1**) to illustrate how each shortlisted platform addresses the challenges detailed in our problem statement. We expanded five key parameters—scalability, multi-criteria support, aggregation, user interface and usability, and customization—into nine features that capture critical facets of large-scale LLM evaluation.

- **User Interface & Usability:** Examines how intuitive and user-friendly the tool is for diverse stakeholders, including those with limited technical expertise. A well-designed interface can expedite evaluations and reduce cognitive load.
- **Customizable Evaluation Functions:** Determines whether the tool supports user-defined criteria or scoring methods, allowing teams to adapt evaluations to specific organizational needs and contexts.
- **LLM-Generated Criteria:** Investigates whether the tool harnesses large language models to automatically propose or refine evaluation dimensions, thus offloading some of the manual effort in criteria definition.
- **Scalability:** Focuses on whether the tool has robust filtering, grouping, or dynamic navigation capabilities that enable professionals to quickly isolate anomalies, identify systematic errors, and conduct iterative evaluations.
- **Interactive Criteria Refinement:** Looks at whether evaluators can iteratively adjust or improve metrics in response to ongoing findings, updated requirements, or newly discovered biases.
- **Flexible Filtering & Exploration:** Reviews how the tool supports the segmentation of data—such as filtering, grouping, or searching—enabling evaluators to pinpoint problematic outputs, thematic clusters, or domain-specific concerns.
- **Integration of Quantitative & Qualitative Analysis:** Considers the extent to which the tool combines numerical metrics (e.g., accuracy, coherence scores) with contextual insights or textual explanations, providing a more nuanced view of model performance.
- **Historical Performance Tracking:** Determines if the tool allows longitudinal comparisons, enabling teams to track performance trends or regressions across multiple runs or model versions.
- **Reliability and Consistency at Scale:** Addresses built-in features for ensuring trustworthy results when working with large datasets. This may include inter-rater reliability checks,

confidence intervals, or other mechanisms that highlight scoring consistency over time.

3.2 Contextual Inquiry

To build on the findings from our literature review and comparative analysis, we conducted a contextual inquiry to better understand the real-world usability and perceived value of LLM evaluation tools in practice.

3.2.1 Participant Recruitment. To gather diverse perspectives on large-scale LLM evaluation, we recruited six participants representing a broad range of technical proficiencies and organizational roles. Our sample included AI developers, data scientists, and PhD students specializing in machine learning. This variety allowed us to capture differing priorities and pain points when evaluating LLM outputs at scale.

3.2.2 Tool Selection. We selected *EvalLM* [4] as the core platform for our contextual inquiry based on three key factors: usability, flexibility, and transparency. Below, we outline the rationale for this choice and its alignment with our study's objectives.

- User-Friendly Interface: *EvalLM* allows participants to interact with the system through an intuitive, graphical dashboard rather than through code. This was essential for gathering feedback from users with varied technical backgrounds, ensuring the tool's usability could be assessed more holistically.
- Flexible Criteria Definition: *EvalLM* supports both predefined criteria and user-defined metrics, entered in natural language. Participants can introduce custom dimensions—e.g., ethics, brand alignment, or style—and then refine or split them iteratively via an integrated LLM assistant. This aligns with our study's emphasis on multi-dimensional, adaptable evaluation criteria.
- Open-Source Tool: Being open-source, *EvalLM* was readily available for our study. It also enabled us to inspect how prompt generation and scoring mechanisms are implemented, which is central to our focus on transparency and trustworthiness in LLM evaluations.

3.2.3 Protocol. We prepared a JSON file containing short news articles from diverse topics (e.g., technology, sports, local events). Each participant loaded this dataset into *EvalLM*, where the model first generated responses based on the following instruction and prompts:

- **Instruction:** *"Given a piece of news article, write an example that would help a young child understand the concept."*
- **Prompt Context 1:** *"You are the narrator of a story. Narrate the news article in an exciting story-like format."*
- **Prompt Context 2:** *"You are a kindergarten teacher. Provide a concise, simple example to help grasp the news article's concept."*

Using an OpenAI API key, the tool generated a tailored response for each article according to these prompts. Participants then evaluated the outputs using criteria they defined within the *EvalLM* interface—such as clarity, style, or age-appropriateness for young children. *EvalLM*'s interactive features allowed them to refine, split, or merge criteria as necessary.

Throughout and after the evaluation process, we asked participants about:

- **Evaluating Large Datasets:** Participants described the ease or difficulty of handling large volumes of responses, including how effectively they could locate relevant examples, identify outliers, or apply bulk operations.
- **Assessing Outputs Against Diverse Criteria:** We explored how participants defined, weighted, or refined multiple criteria. They discussed whether *EvalLM*'s interface supported the complexity of real-world scenarios, including the need for continual adjustment of standards (e.g., ethical considerations or brand guidelines).
- **Interpreting Aggregated, Multi-Dimensional Results:** We probed participants' experiences in reviewing compiled scores and visual summaries. They offered feedback on how they interpreted these metrics for patterns, anomalies, or direct comparisons across different criteria or prompts.
- **Ensuring Reliability and Transparency:** We investigated how participants gauged the trustworthiness of automated evaluations, whether the system's scoring logic was understandable, and what additional validation or explanatory features might bolster confidence in the results.

Through a combination of direct observation and semi-structured interviews, our protocol uncovered participants' experiences with the tool's interface as well as broader insights into the challenges and opportunities of conducting large-scale, multi-criteria LLM evaluations.

On one hand, participants found percentage scores vague and wanted more interpretable metrics like intervals and standard deviation. There was a clear need for more rigorous validation, including larger sample sizes and involvement from stakeholders to build trust. Participants also expressed the desire for clearer explanations accompanying the scores to better understand the evaluation results.

On the other hand, they appreciated summary metrics that helped them understand data without reviewing every detail. Filtering mechanisms were seen as valuable for efficiently navigating through responses. Participants liked having access to a criteria library with editable definitions to fit their use case. Multiple visualizations that make it easy to spot trends were considered useful for gaining insights.

3.3 Limitations

Recruiting participants with direct, hands-on experience in large-scale LLM evaluation proved challenging, due in part to the emerging nature of this research area. Furthermore, several of the tools identified in our comparative analysis were not freely accessible; some required commercial licenses, and others were only partially open source. In addition, updates described in tool-specific publications were sometimes not reflected in the publicly available versions, leading to discrepancies between the literature and actual tool performance. Certain implementations we tested also contained bugs that disrupted data processing, limiting our ability to conduct fully consistent evaluations.

Table 1: Comparative Feature Matrix for Large-Scale LLM Evaluation Tools

Feature	EvalLM	ConstitutionMaker	LLM Comparator	EvalGen	Deepchecks	Robustness Gym
1. User Interface & Usability	✓	✓	✓	✓	✓	✓
2. Customizable Evaluation Functions	✓	✓	✓	✓	✓	˜
3. LLM-generated Criteria	✓	✓		✓		
4. Scalability			✓	✓		
5. Interactive Criteria Refinement	✓	✓	˜	✓	✓	˜
6. Flexible Filtering & Exploration	✓	✓	✓	˜		
7. Integration of Quant. & Qual. Analysis	✓	✓	✓	✓	˜	
8. Historical Performance Tracking	✓	✓		✓	✓	
9. Reliability and Consistency at Scale	✓		˜		˜	

Legend: ✓ = fully supported, ˜ = partially supported or manual effort required, blank = not supported.

Given these constraints, our analysis draws on a blend of sources, including official documentation, academic papers, video demonstrations, and our limited hands-on trials with open-source variants where possible. Although this approach enabled us to form a broad overview of current capabilities, it also means that specific features or performance claims made by each tool's creators may not be replicable under all conditions. Consequently, the findings should be interpreted with an understanding that some tools remain under active development, and that the implementation details may evolve rapidly beyond the scope of our current study.

4 Findings

For each of our research questions, we arrived at a set of findings from our comparative analysis and contextual inquiries.

4.1 Challenges in Evaluating Large Datasets

A recurring theme among participants was the complexity of handling extensive datasets, often comprising thousands of LLM-generated outputs. Professionals emphasized the need for high-level aggregated scores—such as average accuracy or overall coherence—to quickly gauge performance trends without manually reviewing each response. At the same time, they stressed the importance of rapidly isolating outliers and potential anomalies, noting that a relatively small subset of problematic responses can reveal systemic errors or biases. Yet, few existing tools offered robust data-navigation features. Participants cited difficulties in filtering and grouping responses by topic, prompt style, or other contextual attributes. This lack of dynamic filtering often forced them to rely on manual spot checks, which can be both time-consuming and prone to oversight. Overall, participants desired a more streamlined workflow, supporting both "big picture" performance monitoring and granular error analysis.

4.2 Challenges of Multi-Criteria LLM Evaluations

Many professionals underscored the complexity of evaluating outputs along multiple dimensions—such as correctness, brand compliance, ethical considerations, and stylistic fidelity. While some tools permit defining custom criteria, others provide only fixed evaluation metrics, constraining their relevance in specific business or domain contexts. Participants who tested *EvalLM* found the ability to define or refine criteria particularly valuable, as they could align those metrics with evolving organizational standards (e.g., child-friendly language or brand tone). However, the ease of defining these metrics sometimes clashed with uncertainty about which criteria were genuinely indicative of organizational needs. Tools that automatically suggested new criteria—like *EvalLM*'s LLM-driven recommendations—intrigued participants but also raised questions about alignment with real-world priorities. Respondents mentioned that external stakeholders, such as legal teams or domain experts, often needed to be consulted to ensure the criteria matched regulatory and brand guidelines, suggesting a collaborative approach that goes beyond any single user or department.

4.3 Visualizing and Interpreting Multi-Dimensional Results

When participants examined multi-criteria evaluations at scale, the tools' presentation of aggregate metrics often felt too simplistic. They reported that raw percentages or numeric scores lacked the interpretive depth needed to discern how each output performed under different conditions. Some respondents requested confidence intervals or distribution-based metrics to account for variance within the dataset. Others advocated for color-coded visualizations or multi-layered charts that could compare performance across criteria in a single view. Additionally, participants highlighted the value of easily comparing multiple model versions or prompt variations side-by-side, noting that understanding performance trade-offs can be essential for rapid iteration. While some existing platforms support basic bar charts or table summaries, users desired more flexible, interactive dashboards that could highlight patterns, identify clusters of related errors, and facilitate quick deep dives into individual responses.

4.4 Reliability and Transparency in Large-Scale Evaluations

In general, trust remained a key concern among participants. Many expressed apprehension about relying solely on automated scoring—especially for subtle qualitative aspects like bias or ethical alignment—without clear evidence of reliability. They asked for features that would allow them to verify consistency over time, such as performance trend tracking or versioned evaluations. A few participants also pointed out that "explainable scoring," in which the system elucidates how each dimension is assessed, could alleviate skepticism and give stakeholders a clearer basis for confidence in the results. Reliability features varied across existing tools. Some included rudimentary validation checks or inter-rater reliability functions, while others lacked any built-in mechanisms for verifying accuracy. Participants consistently advocated for more extensive validation options, including peer reviews of scoring logic, larger sample sizes for evaluation, and integrated ways to annotate questionable outputs. They also noted that reporting on how each score is generated—especially when using LLM-based evaluators—would help mitigate confusion or distrust.

5 Design Recommendations

From our comparative analysis and contextual inquiries, we distilled five primary objectives that guided our design considerations. Each goal addresses a core challenge identified in our findings, offering a pathway to better meet the needs of professionals evaluating large-scale LLM outputs.

D.1 Let users fluidly explore massive datasets with advanced filtering and drill-down. Evaluating LLM outputs often involves tens of thousands of individual responses, making it impractical to manually review them all. Users therefore need intuitive mechanisms to slice and dice large datasets—such as dynamic filtering by criteria, flexible keyword searches, and cluster-based navigation. By providing these features, evaluators can quickly locate anomalies, surface frequently recurring errors, and glean high-level patterns before drilling down into individual cases when necessary.

D.2 Provide multi-level aggregated insights (both quantitative and qualitative). Participants repeatedly voiced frustration when only raw scores or dense tables were available, as this forced them to piece together trends by themselves. Our second goal emphasizes offering multiple layers of aggregation, from simple statistical overviews (e.g., mean accuracy or bias scores) to qualitative summaries that highlight key strengths and weaknesses. This multi-level approach supports rapid understanding of overall performance while also shedding light on contextual factors—such as specific topics or demographics—where the model's capabilities or limitations become most evident.

D.3 Support iterative refinement of evaluation criteria to accommodate changing needs. As organizations uncover new biases, update brand guidelines, or refine product requirements, the dimensions against which LLMs are judged can shift. Fixed, one-off criteria thus become insufficient over time. Our third design goal underscores the importance of enabling users to add, remove, or re-weight metrics on the fly, while preserving the continuity of previous evaluations. This iterative process ensures the system remains adaptable to evolving objectives and fosters more accurate, context-aware assessments of a model's outputs.

D.4 Provide clear, interactive visuals and transparent scoring processes to foster trust and confidence. Large-scale LLM evaluations can feel opaque when scores are computed by automated methods without sufficient explanation. In response, we aim to present data in an accessible manner, such as interactive bar charts or side-by-side comparisons, paired with annotations that clarify how each metric is calculated or weighted. This level of transparency not only enhances confidence in the results but also allows stakeholders—technical or otherwise—to verify how certain scores were derived and make more informed decisions about model improvements.

D.5 Enable users to efficiently view and move seamlessly between reviewing individual responses, examining aggregated results and refining criteria. To support this dynamic workflow, the system would allow the users to fluidly transition between granular response-level inspection, high-level trend analysis and ongoing refinement of evaluation criteria without losing context.

While these five design recommendations outline key opportunities for improving large-scale LLM evaluation, they represent only an initial conceptual framework. Our subsequent co-design activities will further refine each goal, translating them into actionable interface features and workflows. By iteratively testing prototypes with practitioners, we aim to validate the feasibility of these recommendations, explore edge cases, and ensure that the resulting tool effectively addresses both the scale of LLM evaluations and the complexity of multi-dimensional criteria.

6 Conclusion

This study aimed at distilling the complex challenges of evaluating large-scale outputs from large language models, focusing on issues of data aggregation, criteria flexibility, and transparency. Our findings confirm that professionals need high-level performance overviews that can be quickly parsed, alongside robust filtering and navigation features to isolate outliers or identify systemic errors. Clear and interpretable visualizations are essential for facilitating these diagnostic tasks, allowing evaluators to gain both rapid, big-picture insights and detailed breakdowns of model outputs.

Additionally, our research underscores the importance of flexible, multi-dimensional criteria. LLM deployments often span diverse domains with evolving requirements—such as ethical guidelines, brand standards, or domain-specific conventions—making it vital for evaluation tools to support easy refinement of existing metrics or the introduction of entirely new ones. Participants welcomed functionality that enables LLMs themselves to suggest or refine evaluation criteria, signaling a growing need for more adaptive, AI-assisted workflows.

A final critical dimension we identified is transparency, especially in how scores are derived and validated. Trust in automated evaluations remains fragile without rigorous methods for ensuring consistency and reliability over time. Tools that incorporate validation mechanisms—ranging from inter-rater reliability checks to the ability to track performance trends across different model versions—have the potential to strengthen user confidence.

In response to these findings, we proposed a set of design recommendations aimed at meeting the multifaceted demands of large-scale LLM evaluation. These include supporting fluid exploration of extensive datasets, providing multi-level aggregated insights, enabling iterative refinement of evaluation criteria, and fostering transparency through clear scoring and visualization processes as well as seamlessly moving between individual responses, aggregated results and criteria refinement mechanisms. Future work will focus on implementing and empirically testing these recommendations, including the development of wireframes and prototypes for next-generation evaluation interfaces. By continuing to refine these solutions in collaboration with industry practitioners, we hope to help establish more trustworthy, efficient, and context-aware workflows for organizations deploying advanced language models at scale.

References

[1] Pan, Q., Ashktorab, Z., Desmond, M., Santillan Cooper, M., Johnson, J., Nair, R., Daly, E., and Geyer, W. 2023. Human-Centered Design Recommendations for LLM-as-a-Judge. IBM Research Technical Report.

[2] Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop on Automatic Summarization*, Barcelona, Spain, July 2004, 25–31. Association for Computational Linguistics.

[3] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL '02), pages 311–318.

[4] Kim, T. S., Lee, Y., Shin, J., Kim, Y.-H., and Kim, J. 2023b. Evallm: Interactive evaluation of large language model prompts on user-defined criteria. arXiv preprint arXiv:2309.13633.

[5] Arawjo, I., Swoopes, C., Vaithilingam, P., Wattenberg, M., and Glassman, E.L. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 18 pages.

[6] OpenAI. Evals: Evaluation Framework for OpenAI Models [Internet]. GitHub repository. [cited 2025 Feb 24]. Available from: https://github.com/openai/evals?tab=readme-ov-file.

[7] EleutherAI. lm-evaluation-harness [Internet]. GitHub repository. [cited 2025 Feb 24]. Available from: https://github.com/EleutherAI/lm-evaluation-harness.

[8] Petridis, S., Wedin, B., Wexler, J., Donsbach, A., Pushkarna, M., Goyal, N., Cai, C. J., and Terry, M. 2023. ConstitutionMaker: Interactively Critiquing Large Language Models by Converting Feedback into Principles. arXiv preprint arXiv:2310.15428v1.

[9] People AI Research. n.d. LLM Comparator: A Tool for Human-Driven LLM Evaluation. Medium. Retrieved February 24, 2025 from https://medium.com/people-ai-research/llm-comparator-a-tool-for-human-driven-llm-evaluation-81292c17f521.

[10] Shankar, S., Zamfirescu-Pereira, J.D., Parameswaran, A.G., Arawjo, I., and Hartmann, B. 2024. WhoValidates the Validators? Aligning LLM-Assisted Evaluation of LLMOutputs with HumanPreferences. arXiv preprint arXiv:2404.12272v1 [cs.HC].

[11] Chorev, S., Tannor, P., Ben Israel, D., Bressler, N., Gabbay, I., Hutnik, N., Liberman, J., Perlmutter, M., Romanyshyn, Y., and Rokach, L. 2022. Deepchecks: A Library for Testing and Validating Machine Learning Models and Data. Journal of Machine Learning Research 23 (2022): 1–6.

[12] Goel, K., Rajani, N., Vig, J., Tan, S., Wu, J., Zheng, S., Xiong, C., Bansal, M., and Ré, C. 2021. RobustnessGym: Unifying the NLP Evaluation Landscape. arXiv preprint arXiv:2101.04840v1 [cs.CL], January 13, 2021.