

A Fairness–Utility Evaluation Framework for Assessing Large Language Models (LLMs)

Samah Alhazmi

Computer Science Department, College of Computing and Informatics,
Saudi Electronic University, Riyadh 11673, Saudi Arabia

Abstract—Large language models are increasingly used in contexts where their outputs can affect people directly, including hiring, admissions, and lending. This growing role makes it important to consider not only how well these models perform, but also whether their behavior is fair. Although many fairness metrics, bias benchmarks, and mitigation techniques have been proposed, there is still limited guidance on how fairness and predictive performance can be evaluated together through a consistent procedure. This study presents a five-stage Fairness–Utility Evaluation Framework that brings together dataset selection and preparation, model assessment, fairness evaluation, utility evaluation, and trade-off analysis and reporting. Rather than introducing a new metric, the framework organizes these elements into a structured and reproducible evaluation process. Its application is illustrated using LLaMA-2-7B, CrowS-Pairs, and WinoBias to show how the five stages can be followed in practice; the example is intended to demonstrate the procedure rather than provide new experimental results. Documentation is incorporated throughout the process so that evaluation choices, measures, and outcomes can be traced and reviewed. The resulting framework provides researchers and professionals with a structured way to consider fairness and performance together when evaluating an LLM for potential deployment.

Keywords—Large language models; fairness–utility evaluation; AI governance; natural language processing

I. INTRODUCTION

Large language models are no longer limited to research environments. They are increasingly used in education, employment, and other contexts where their outputs may affect people directly [1]. Their broad language capabilities have supported this wider use.

Fairness has consequently become an important concern in the evaluation of LLMs. LLMs are trained on large collections of text that may contain historical stereotypes and demographic biases [2], [3]. Models can learn these patterns during training and reproduce them in their outputs, which may lead to biased outcomes when the models are applied.

Several approaches have been developed to reduce bias in LLMs. These methods can be applied to the training data, during model training, or to the outputs after generation. However, improving fairness may also affect predictive performance [4]. For this reason, evaluating a model requires considering fairness together with its performance on the intended task.

Recent studies have examined fairness from several perspectives. A survey of 121 studies [5] examined bias sources as well as methods for detecting, evaluating, and mitigating bias. Fairness–utility relationships have also been examined in text-

to-image generation [6], while another study investigated how LLMs respond to fairness-related questions in resource-allocation tasks [7]. Despite this progress, studies still vary in how fairness and utility are evaluated and reported. A consistent and reproducible procedure for assessing both within the same evaluation remains limited.

This study addresses that gap. Much of the existing LLM fairness literature focuses on individual metrics, mitigation techniques, or specific models. Less attention has been given to combining these elements within the same evaluation process. Documentation also receives limited attention, even though it is needed to understand how an evaluation was conducted and how its findings were reached. Bringing fairness, utility, and documentation into one framework could provide a clearer basis for evaluating an LLM before deployment.

To address this need, this study proposes a Fairness–Utility Evaluation Framework. The framework follows a common sequence that begins with benchmark selection and preparation, then moves through model assessment, fairness and utility evaluation, and trade-off analysis and reporting. The procedure can be applied across different LLMs and application contexts.

Documentation is maintained throughout the framework and is not limited to the final reporting stage. Recording the evaluation procedure, key choices, and supporting evidence makes it possible to review how the assessment was conducted. Considering fairness and predictive performance within the same process also helps evaluators use both types of evidence when making decisions about LLM deployment.

A. Research Contributions

The main contributions of this study are as follows:

- A Fairness–Utility Evaluation Framework organized into five stages, covering the evaluation process from dataset preparation to trade-off analysis and reporting.
- An approach that evaluates fairness and utility together, allowing their results to be interpreted within the same assessment.
- A model- and benchmark-independent structure that can be used with different LLMs, benchmarks, and application contexts.
- An illustrative example using established fairness benchmarks to demonstrate how the framework can be applied in practice.

These contributions provide a common structure for evaluating fairness and utility while incorporating documentation relevant to AI governance.

B. Research Questions

Three research questions guide the study:

- RQ1: What components should a comprehensive fairness–utility evaluation framework include?
- RQ2: How can fairness assessment and predictive–performance evaluation be integrated into one unified, reproducible workflow?
- RQ3: How can documentation and governance-oriented reporting be built into that workflow to support transparency and accountability in LLM deployment?

The remainder of the study is organized as follows. Section II reviews prior work on bias, fairness evaluation, and governance in LLMs. Section III presents the framework and its five stages, followed by the illustrative example in Section IV. Section V discusses the implications and limitations of the proposed framework, and Section VI concludes the study.

II. RELATED WORK

A. Bias in Large Language Models

Concerns about bias in language models are not new. Early research identified social stereotypes in word embeddings. Bolukbasi et al. [8] examined gender associations in word embeddings and how they can influence later language-processing tasks. Caliskan et al. [3] reported that patterns learned from language data can reflect biases present in society.

The issue remains relevant for LLMs because they are trained on large and diverse text collections that may contain historical and demographic biases. These patterns can be learned during training and later appear in model outputs. Blodgett et al. [2] emphasized that fairness in language technology cannot be understood through technical measures alone, since judgments about bias also depend on social and contextual factors. This makes it important to consider how bias is defined, evaluated, and addressed throughout the model lifecycle.

B. Fairness Evaluation Datasets

Evaluating fairness requires benchmarks built to surface specific kinds of bias under controlled conditions. CrowS-Pairs [9] does this by pairing stereotypical and anti-stereotypical sentences across several demographic categories and checking which one a model prefers. WinoBias [10] takes a narrower angle, using occupational stereotypes in coreference resolution to test whether predictions shift with gender.

CrowS-Pairs and WinoBias examine different forms of bias and therefore provide different information about model behavior. Using multiple benchmarks can cover a broader range of demographic contexts and tasks than relying on a single benchmark.

C. Bias Mitigation Approaches

Bias mitigation methods are usually grouped according to when they are applied. Pre-processing methods modify the data before training begins [11]. In-processing methods introduce

fairness objectives during model training, often through constrained optimization or adversarial learning [12]. Post-processing methods are applied to model outputs after training. Self-Debias [13] is one example that uses prompting to identify and reduce biased language in generated text.

These methods can improve fairness, but they may also affect predictive performance [4]. Evaluating both fairness and utility helps determine how a mitigation method changes the model's overall behavior.

D. Fairness Frameworks and Governance

Fairness research has also considered broader evaluation and governance frameworks. A review of 121 studies [5] examined bias sources together with methods for detection, evaluation, and mitigation across the NLP lifecycle. The review identified considerable variation in evaluation practices and highlighted the need for more systematic and reproducible procedures.

Fairness and utility have also been studied together in other areas of generative AI. Bochernitsan et al. [6] developed a Pareto-frontier benchmarking framework to examine fairness–utility trade-offs in text-to-image models. Although the application differs from LLMs, the study provides an example of evaluating changes in fairness and utility together. Cookson et al. [7] examined fairness from another perspective by studying how LLMs respond to resource-allocation problems involving fairness considerations.

Broader benchmarking frameworks are also relevant. HELM [14] evaluates language models across multiple scenarios using seven metrics, including fairness, accuracy, robustness, and efficiency. It supports standardized comparisons across models and scenarios. The framework proposed in this study serves a different methodological purpose: rather than benchmarking models across scenarios, it structures the fairness–utility relationship itself into a decision-oriented procedure, incorporating predefined decision criteria and governance-oriented documentation that connect the evaluation to a specific deployment recommendation.

Governance also affects how model evaluations are documented and reviewed. Organizations need records of the evaluation procedure, the evidence obtained, and the basis for decisions about model use. Such documentation can support transparency, accountability, and subsequent auditing [15]. Including these records as part of the evaluation process can also help when deployment decisions are reviewed later.

Three recent studies are particularly relevant to the proposed framework. HALF [16] presents a five-stage pipeline for evaluating fairness in LLMs, with an emphasis on weighting the severity of harms across deployment domains. It does not combine fairness and utility in a dedicated trade-off stage or include governance-oriented reporting. de Lima et al. [17] evaluate fairness and utility within a benchmarking procedure that includes dataset preparation, metric selection, and statistical testing. Their study focuses on adversarial fair-learning models rather than LLMs. Templin et al. [18] propose a five-step audit workflow for LLMs in healthcare. Their approach is specific to healthcare and gives particular attention to stakeholder engagement and model calibration.

These studies share some elements with the proposed framework, but their objectives and scope differ. The framework presented here combines benchmark preparation, fairness assessment, utility evaluation, documentation, and trade-off analysis within one procedure that can be applied across different LLMs and application domains.

E. Research Gap and Framework Contribution

Existing research provides established benchmarks, evaluation measures, and mitigation methods for studying bias in language models. Recent work has also considered fairness–utility relationships and governance requirements, but these elements are often examined separately. For LLMs,

methodological guidance remains limited on how to bring them together within a consistent evaluation process. The proposed Fairness–Utility Evaluation Framework addresses this gap by connecting benchmark preparation, model assessment, fairness and utility evaluation, trade-off analysis, and documentation within one procedure. Its structure does not depend on a specific model or benchmark, allowing it to be applied across different LLMs and application contexts.

Table I compares the proposed framework with four closely related studies in terms of their purpose, treatment of fairness and utility, governance-oriented reporting, applicability across models and datasets, and empirical validation.

TABLE I. COMPARISON OF THE PROPOSED FRAMEWORK WITH CLOSELY RELATED WORK

Dimension	[5]	[6]	[7]	[14] (HELM)	Proposed Framework
Primary focus	Review of bias sources, evaluation, and mitigation across the NLP lifecycle	Fairness–utility benchmarking for text-to-image generation	LLM responses to fairness in resource-allocation tasks	Multi-metric benchmarking (fairness, accuracy, robustness, efficiency)	Fairness–utility evaluation framework for LLMs
Joint fairness–utility evaluation	Discussed, not operationalized	Yes (vision domain)	No (fairness-focused study)	Partial (same evaluation suite, no dedicated trade-off stage)	Yes (NLP/LLM domain)
Governance-oriented reporting	Not addressed	Not addressed	Not addressed	No dedicated reporting stage	Dedicated framework stage
Model/dataset independence	N/A (review)	Text-to-image models only	Resource-allocation tasks only	Defined scenarios and metrics	Model- and benchmark-agnostic
Empirical validation	N/A (review)	Yes, multiple T2I models	Yes, multiple LLMs	Yes, 30 models / 42 scenarios	Illustrative application only

TABLE II. MAIN COMPONENTS OF THE PROPOSED FAIRNESS–UTILITY EVALUATION FRAMEWORK

Stage	Inputs	Objective	Output	Key Constraints
Dataset Selection & Preparation	Candidate benchmarks; application context; fairness dimensions of interest	Select and prepare benchmark datasets	Evaluation-ready datasets	Public and established benchmarks are preferred where appropriate; cleaning, validation, and formatting decisions are documented before evaluation.
Model Assessment	Prepared datasets; selected model; prompting and inference configuration	Generate model outputs using standardized inputs	Model predictions and responses	Model version, prompting strategy, and inference parameters — including decoding settings and the random seed where applicable — are specified and recorded before evaluation; repeated runs are used where outputs are stochastic.
Fairness Evaluation	Model outputs; demographic attributes of interest; predefined outcome mapping	Assess fairness-related behavior	Fairness indicators	Outcome mapping and demographic groupings are defined before analysis; fairness metrics are selected according to the fairness question and output structure.
Utility Evaluation	Model outputs; ground-truth labels where available	Measure predictive performance	Utility metrics	Utility metrics must match the benchmark and task design; metrics requiring ground-truth labels are used only when such labels are available.
Trade-off Analysis & Reporting	Fairness indicators; utility metrics; predefined decision criteria	Examine fairness and utility findings together	Evaluation report, deployment recommendation, and evaluation manifest	Thresholds and decision criteria are specified before results are interpreted; any aggregation across metrics, groups, or benchmarks must be predefined and documented; uncertainty or variability is reported where applicable.

III. PROPOSED FAIRNESS–UTILITY EVALUATION FRAMEWORK

A. Overview of the Proposed Framework

The Fairness–Utility Evaluation Framework provides a common process for examining fairness alongside predictive performance. Rather than treating the two as separate parts of model evaluation, it considers fairness measures in relation to how well the model performs on the task under assessment. This

approach is consistent with recent calls for more systematic evaluation practices across the LLM lifecycle [5]. In doing so, the framework addresses RQ1 through its five-stage structure, RQ2 through the joint treatment of fairness and utility, and RQ3 through the documentation embedded at each stage.

The process is organized into five stages. It begins with preparing the evaluation data and assessing the model, followed by fairness and utility evaluation. The findings are then considered together in the final stage. Using the same sequence

allows different models and benchmarks to be evaluated through a consistent process across different applications.

The framework is not built around a specific LLM or benchmark. Evaluators can choose the datasets and measures that fit the evaluation at hand while keeping the overall process unchanged. This also means that new models, benchmarks, or metrics can be introduced without having to redesign the framework.

Documentation is maintained as the evaluation progresses. At each stage, evaluators record the relevant inputs, conditions, measures, and decisions. Keeping these details with the evaluation makes it easier to understand how the final assessment was reached and supports the transparency and accountability emphasized in responsible AI governance guidance [15].

The five stages are:

- Dataset Selection and Preparation
- Model Assessment
- Fairness Evaluation
- Utility Evaluation
- Trade-off Analysis and Reporting

Fig. 1 shows how the five stages connect within the overall evaluation process. Table II specifies, for each stage, the required inputs, objective, output, and the key procedural constraints that must be satisfied.



Fig. 1. Overall architecture of the proposed framework.

B. Dataset Selection and Preparation

The first step is to choose benchmark datasets that reflect the fairness dimensions relevant to the application being considered. This choice matters because the data used for evaluation influence which fairness concerns can actually be observed. Where suitable benchmarks are available, the framework gives preference to public and established datasets. Their use supports reproducibility and comparison with findings from other studies [9], [10].

Before evaluation, the selected datasets are prepared to ensure that they can be used consistently. This includes recording the benchmark version, cleaning the data, checking their integrity and consistency, and bringing them into the required format. These preparation steps help prevent data problems from being mistaken for differences in model behavior [5]. Fig. 2 summarizes the process, showing how raw datasets move through cleaning, validation, and formatting before becoming ready for model assessment.



Fig. 2. Dataset preparation pipeline.

C. Model Assessment

After the datasets have been prepared, the selected model is run on standardized inputs taken from the benchmarks. The purpose at this stage is to obtain the model outputs needed for later analysis. Fairness and utility are examined in the stages that follow.

The evaluation procedure should remain consistent across runs, particularly the inputs and prompting method, so that the outputs are generated under comparable conditions, unless prompt variation is itself a deliberate part of the evaluation design, in which case the range of prompts used must be specified in advance. The model version, prompting strategy, inference parameters, and other relevant evaluation conditions are also recorded. These details document how the outputs were produced and allow subsequent assessments to review or reproduce the evaluation setup [15].

Because LLM generation can be stochastic, decoding parameters such as temperature and top-p or top-k, and the random seed where supported by the model or inference framework, must be specified and recorded alongside the model version and prompting strategy. Where outputs vary across repeated runs under the same configuration, the evaluation should include multiple repeated runs, with the number of runs and the resulting variability reported as part of the evaluation record. This variability should subsequently be reflected in the uncertainty measures reported at the Trade-off Analysis and Reporting stage.

D. Fairness Evaluation

$$BS = |P(\hat{y} = I | A = g_I) - P(\hat{y} = I | A = g_J)| \quad (1)$$

$$FR = P(\hat{y} = I | A = g_D) / P(\hat{y} = I | A = g_A) \quad (2)$$

In Eq. (1), g_I and g_J represent the two demographic groups compared under attribute A . A BS value closer to zero indicates a smaller difference between the groups, consistent with the statistical-parity criterion in [20]. Eq. (2) follows the disparate-impact formulation in [21] and compares the positive outcome rate for a disadvantaged group g_D with that of an advantaged group g_A . An FR value close to 1 indicates greater parity, while lower values indicate a larger disparity. The four-fifths rule provides an established reference for interpreting this ratio. BS and FR can be calculated separately for relevant demographic categories and incorporated into the overall fairness assessment; both are established measures in the algorithmic-fairness literature [20], [21] and are used here as components of the proposed framework.

Before BS or FR is calculated, the evaluator must define the outcome variable used in the application. For tasks that already produce categorical predictions, the benchmark label or task decision can be used directly. For open-ended generation, outputs must first be mapped to a predefined categorical or binary outcome using a documented coding rule, classifier, rubric, or human-annotation protocol. The positive outcome must be defined before evaluation and applied consistently across demographic groups. Because this transformation can affect the resulting disparity estimates, the mapping procedure

and its validation or agreement evidence must be reported as part of the evaluation record.

The resulting fairness assessment summarizes the disparities identified during evaluation and provides evidence for determining whether particular demographic groups require further examination.

Fig. 3 shows how model outputs are converted into fairness indicators and incorporated into the assessment report.

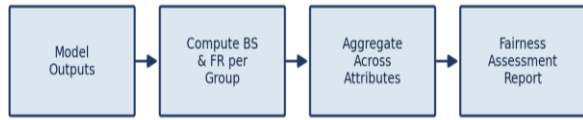


Fig. 3. Fairness evaluation process.

E. Utility Evaluation

The utility evaluation stage examines how well the model performs the task being assessed. Performance is measured using metrics appropriate to the task and benchmark. For prediction tasks with defined ground truth, commonly used measures include Accuracy, Precision, Recall, and F1-score.

Additional measures may be used depending on the task or application. With TP, TN, FP, and FN denoting true positives, true negatives, false positives, and false negatives:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (3)$$

$$Precision = TP / (TP + FP) \quad (4)$$

$$Recall = TP / (TP + FN) \quad (5)$$

$$F1 = 2(P \times R) / (P + R) \quad (6)$$

Accuracy measures the proportion of correct predictions. Precision and Recall describe different aspects of classification performance, while F1 combines the two through their harmonic mean. These metrics can also be reported separately for demographic groups to identify differences in model performance. The results can then be considered alongside the fairness measures defined in Eq. (1) and (2).

The choice of utility metric depends on the benchmark and whether ground-truth labels are available. WinoBias provides coreference-resolution labels that allow measures such as Accuracy and F1 to be calculated. CrowS-Pairs follows a different evaluation design based on model preferences between stereotypical and anti-stereotypical sentence pairs and does not provide conventional ground-truth labels for classification metrics. Utility measures should therefore match the evaluation procedure supported by each benchmark.

Fairness and utility results should be considered together when evaluating a model for potential deployment. Improvements in fairness may affect predictive performance [4], while strong overall performance does not necessarily indicate similar behavior across demographic groups. Considering both provides a clearer view of these differences.

F. Trade-Off Analysis and Reporting

The final stage considers the fairness and utility findings together. The results are compared to examine how the two dimensions relate and whether the overall outcome is suitable for the intended application. This judgment depends on the criteria established for the specific evaluation context. The final report records the benchmarks and metrics used, along with any assumptions or limitations relevant to interpreting the findings. It also documents the evidence supporting the deployment recommendation.

Fairness and utility results are reported at the level of the individual metric, demographic group, and benchmark rather than collapsed into a single overall score, since a single aggregate value can obscure disparities that affect specific groups. Where multiple metrics, groups, or benchmarks are combined into a summary judgment, the aggregation rule must be specified and justified in advance, together with the rationale for any weighting applied. Where estimates are subject to sampling or stochastic variation, the report should include appropriate measures of uncertainty or variability, such as confidence intervals or repeated-run variation, as applicable to the evaluation design. The framework does not mandate a specific statistical test or effect-size measure, as the appropriate procedure depends on the benchmark, sample size, outcome type, and evaluation design.

The final evaluation record is consolidated into an evaluation manifest that identifies the model and benchmark versions, prompting and inference configuration, outcome-mapping rules, fairness and utility metrics, uncertainty information where applicable, and the predefined thresholds, aggregation rules, and decision criteria used in the assessment. The manifest provides a traceable record of the evaluation configuration and resulting deployment recommendation, supporting reproducibility, independent review, and governance oversight.

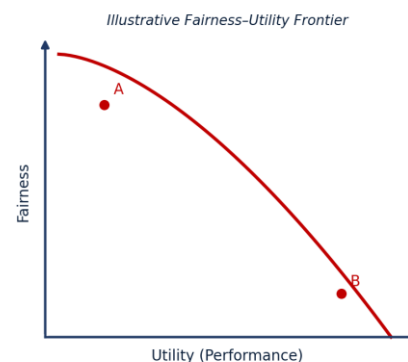


Fig. 4. Fairness-utility trade-off analysis.

This record also serves a governance purpose. Preserving the basis of the assessment can support risk assessment, compliance review, and later auditing [15]. The importance of considering fairness in relation to utility has also been recognized in benchmarking research involving other generative systems [6]. Fig. 4 shows how fairness and utility findings can be interpreted together and identifies cases where the overall outcome may call for further review. The curve represents an illustrative fairness-utility frontier rather than data from an actual evaluation; points

A and B mark two hypothetical outcomes along this frontier, showing how a model could favor fairness (point A) or utility (point B) depending on where it falls relative to the trade-off.

IV. FRAMEWORK DEMONSTRATION: AN ILLUSTRATIVE WALKTHROUGH

A. Case Study Overview

This section illustrates how the proposed framework can be used in practice. The example follows its stages in sequence and does not report findings from an empirical experiment.

The example considers a representative LLM evaluation scenario. Its purpose is to show how each stage is applied and how fairness and utility are examined within the same evaluation process.

Following the structure introduced in Section III, the demonstration starts with dataset preparation and model assessment. It then moves to fairness and utility evaluation, followed by trade-off analysis and reporting. This sequence illustrates how evidence from both dimensions can inform the assessment of a model for potential deployment.

For this illustrative scenario, LLaMA-2-7B is designated as the model under consideration [19], and CrowS-Pairs and WinoBias are designated as example benchmark datasets, as they are established resources for examining different forms of bias in language models [9], [10]. Fig. 5 shows how these elements fit into the complete demonstration workflow.



Fig. 5. Illustrative demonstration workflow.

B. Dataset Selection and Preparation

The demonstration begins by selecting benchmarks that match the fairness questions being examined. CrowS-Pairs [9] is used to examine social stereotypes across demographic categories, while WinoBias [10] focuses on gender bias in coreference resolution.

The two datasets address different forms of bias and therefore provide different information about model behavior. In an empirical application, the datasets would be prepared and checked for consistency before model assessment. Any preprocessing or formatting choices would also be documented to support reproducibility. Table III summarizes the two benchmarks.

TABLE III. EXAMPLE DATASETS USED IN THE DEMONSTRATION

Dataset	Bias Category	Evaluation Focus
CrowS-Pairs	Social stereotypes	Social stereotype assessment
WinoBias	Gender bias	Gender fairness assessment

C. Model Assessment

Once the datasets are prepared, the model would be run on standardized benchmark inputs. Its outputs would then provide

the data needed for the fairness and utility assessments that follow.

LLaMA-2-7B is designated as the reference model for this illustrative demonstration because it is an established open-weight LLM [19]. The model selection is illustrative and does not represent an empirical evaluation of LLaMA-2-7B. The choice of model is specific to the demonstration; the framework itself can be applied to other LLMs using the same procedure.

The evaluation conditions would also need to be documented. These include the model version, prompting procedure, inference parameters, and any other conditions that could affect the generated outputs [15]. Recording these details would allow the evaluation setup to be reproduced in later assessments.

D. Fairness and Utility Assessment

The model outputs would be examined from both fairness and performance perspectives. Keeping the two assessments within the same process allows their findings to be considered together later in the evaluation.

The fairness measures would depend on the application and the demographic attributes being examined. Bias Score and Fairness Ratio are examples of indicators that could be used. Where appropriate, using several measures may reveal disparities that are less apparent when only one indicator is considered [2].

Performance would be evaluated using measures suited to the task and the structure of the benchmark. When ground-truth labels are available, these could include Accuracy, Precision, Recall, and F1-score.

Together, the fairness and performance findings would provide different but related information about the model's behavior. This evidence could help determine whether the model is suitable for the intended application [4]. Table IV summarizes the fairness, utility, and governance dimensions included in the demonstration, along with their corresponding measures and purpose.

TABLE IV. EXAMPLE EVALUATION DIMENSIONS

Dimension	Metrics / Indicators	Purpose
Fairness	Bias Score, Fairness Ratio	Assess fairness-related behavior
Utility	Task-appropriate measures (e.g., Accuracy, Precision, Recall, F1-score where ground truth is available)	Assess predictive performance
Governance	Evaluation documentation and records	Support accountability

E. Trade-Off Analysis and Reporting

For this illustrative scenario, the final stage would consider the fairness and utility findings together to determine whether the model is suitable for the intended application. The assessment may also identify areas where further evaluation or mitigation is needed.

The report would record the benchmarks and measures used, together with relevant assumptions, limitations, and observations. These records would document the basis for the

assessment and provide evidence that can support later review or auditing [15].

Fig. 6 illustrates, in general terms, a decision matrix that could be used to consider fairness and utility together. The zone placement shown is illustrative and does not represent a computed assessment of LLaMA-2-7B. Models with high fairness and high utility would fall within the Preferred Zone, representing the most favorable conditions for deployment. Low results on both dimensions would place a model in the Reject Zone, where substantial improvement would be needed. The Risk Zone represents high utility but low fairness, indicating fairness concerns that require attention. The Review Zone represents high fairness but low utility, indicating a need to improve model performance. The terms High and Low in Fig. 6 are qualitative rather than fixed numerical thresholds. Thresholds for BS, FR, and utility measures should reflect the application, organizational risk tolerance, and relevant regulatory or ethical requirements. These thresholds should be defined in advance by the responsible evaluation or governance team and documented before the results are interpreted. Applying the same predefined thresholds and decision criteria allows independent evaluators to assign models to the four zones on a consistent basis.

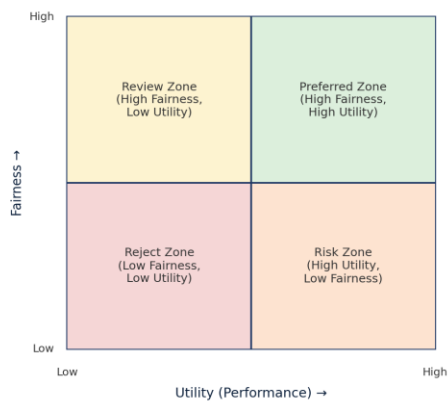


Fig. 6. Illustrative fairness-utility decision matrix for deployment decision support.

F. Discussion

The proposed framework brings these elements into a common sequence, beginning with benchmark preparation and continuing through fairness and utility evaluation to trade-off analysis and reporting. Following the same sequence allows the framework to be applied when the model, benchmark, or application changes.

This differs from approaches that focus on a particular mitigation technique or the evaluation of a specific model. Here, fairness and utility are examined as parts of the same procedure. Because the framework is not tied to a particular LLM architecture, the same sequence can be followed across different models without developing a separate evaluation approach for each one. This also provides a consistent structure for cross-model comparison.

Looking at fairness and utility together matters in practice because an improvement in one does not necessarily leave the other unchanged. Changes in fairness can occur alongside

changes in predictive performance [4]. Examining both therefore gives a more complete picture when deciding whether a model is suitable for a particular application.

The framework is also intended to remain usable as LLM evaluation evolves. A different model, fairness measure, or benchmark can be introduced without redesigning the overall process. The individual components can therefore change with the application, available resources, or evaluation requirements while the sequence itself remains stable.

Documentation is part of this process rather than a separate activity at the end. The data and measures used are recorded together with relevant model parameters, assumptions, and limitations. These records allow others to see how the assessment was carried out and to reproduce its main evaluation conditions [15]. They also provide a basis for examining the evidence behind an evaluation when it is later reviewed by researchers or other stakeholders.

Section IV illustrates how this process can be followed using established benchmarks. Its purpose is to show how the stages fit together, including how the fairness and utility assessments relate to one another.

The illustration should not, however, be interpreted as empirical validation of the framework. Assessing its generalizability will require empirical evaluations across multiple LLMs and real deployment contexts. In particular, the use of CrowS-Pairs and WinoBias in Section IV illustrates how the procedure can be applied to established benchmarks; it does not establish external validity across contemporary instruction-following LLMs, nor does it represent the application-specific risks or affected populations relevant to a given deployment context. Its present scope is also limited to fairness and predictive performance. Explainability, robustness, privacy, security, and other responsible AI considerations are not evaluated here, although they could be incorporated into future extensions of the framework.

V. CONCLUSION

This study introduces a Fairness-Utility Evaluation Framework for assessing large language models. The framework provides a structured way to examine fairness alongside predictive performance while maintaining a record of the evidence generated throughout the evaluation.

Rather than introducing new metrics or benchmarks, the framework brings existing measures, benchmarks, and assessment methods into a common sequence. The same procedure can be applied across different LLMs and application contexts. Because the framework is not tied to a particular model architecture, the model or benchmark can change while the overall evaluation procedure remains consistent.

Section IV illustrates how the framework could be followed using LLaMA-2-7B, CrowS-Pairs, and WinoBias. The example shows how the proposed stages fit together and how fairness and utility findings would be considered within the same assessment. It demonstrates the procedure rather than providing an empirical evaluation and does not report new experimental results.

Documentation is maintained throughout the framework. Key evaluation decisions and supporting evidence are recorded so that the assessment can be traced and reviewed later. These records can support transparency and subsequent review or auditing, particularly when evaluation findings inform decisions about LLM deployment.

The next step in this research program is to test the framework empirically across different LLMs, benchmarks, and application contexts. The framework could also be expanded to address other responsible AI concerns, such as explainability, robustness, privacy, and security. Empirical applications would help assess its generalizability and identify aspects of the procedure that may require refinement. In its current form, the framework provides a methodological basis for examining fairness–utility trade-offs through a documented and reproducible evaluation process.

In relation to the research questions guiding this study, RQ1 is addressed through the five-stage structure presented in Section III; RQ2 through the framework’s treatment of fairness and utility within a single evaluation sequence, illustrated in Sections III and IV; and RQ3 through the documentation and governance-oriented reporting embedded at each stage of the framework.

DECLARATIONS

Conflict of Interest: No conflicts of interest are declared.

Funding: This research received no external funding.

Data Availability: The benchmark datasets referenced in this study (CrowS-Pairs and WinoBias) are publicly available from their sources [9], [10]. No new datasets were generated in this study.

Declaration of Generative AI Use: Generative AI tools were used to assist with language editing, formatting, and restructuring during the preparation of this manuscript. The entire manuscript was reviewed and approved by the author. Generative AI was not used as an author or co-author. The research concept, methodology, framework design, analysis, and conclusions are the author's own original work.

REFERENCES

- [1] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency, 2021, pp. 610–623.
- [2] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of 'bias' in NLP," in Proc. 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5454–5476.
- [3] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.
- [4] E. Sheng, K. W. Chang, P. Natarajan, and N. Peng, "The woman worked as a babysitter: On biases in language generation," in Proc. 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3407–3412.
- [5] T. H. Zaidi, S. M. Jahin, A. R. Dhruba, S. A. Tanim, A. M. Hamja, M. R. Hasan, M. S. U. Miah, M. I. Islam, and A. A. Mansur, "A review of fairness challenges in natural language processing," *Array*, vol. 30, art. 100773, 2026.
- [6] M. N. Bochemitsan, R. C. Barros, and L. S. Kupssinskü, "A framework for benchmarking fairness–utility trade-offs in text-to-image models via Pareto frontiers," *arXiv preprint arXiv:2508.16752*, 2025.
- [7] B. Cookson, S. Ebadian, and N. Shah, "Fairness perceptions of large language models," in Third IJCAI Workshop on Computational Fair Division (CFD 2025), 2025.
- [8] T. Bolukbasi, K. W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [9] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, "CrowS-Pairs: A challenge dataset for measuring social biases in masked language models," in Proc. 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1953–1967.
- [10] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," in Proc. 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 2, 2018, pp. 15–20.
- [11] R. Zmigrod, S. J. Mielke, R. Cotterell, and I. Augenstein, "Counterfactual data augmentation for mitigating gender stereotypes in language models," in Proc. 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1656–1661.
- [12] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in Proc. 2018 AAAI/ACM Conf. on AI, Ethics, and Society, 2018, pp. 335–340.
- [13] T. Schick, S. Udupa, and H. Schütze, "Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP," *Trans. of the Association for Computational Linguistics*, vol. 9, pp. 1408–1424, 2021.
- [14] P. Liang, R. Bommasani, T. Lee, et al., "Holistic evaluation of language models," *Trans. on Machine Learning Research*, 2023.
- [15] UNESCO, "Recommendation on the ethics of artificial intelligence," 2021.
- [16] A. Mekky, O. El Herraoui, P. Nakov, and Y. Wang, "HALF: Harm-aware LLM fairness evaluation aligned with deployment," *arXiv preprint arXiv:2510.12217*, 2025.
- [17] L. F. F. de Lima, D. R. D. Ricarte, and C. A. Siebra, "A benchmark proposal for non-generative fair adversarial learning strategies using a fairness-utility trade-off metric," *Computational Intelligence*, vol. 40, no. 5, art. e70003, 2024.
- [18] T. Templin, S. Fort, P. Padmanabham, P. Seshadri, R. Rimal, J. Oliva, K. Hassmiller Lich, S. Sylvia, and N. Sinnott-Armstrong, "Framework for bias evaluation in large language models in healthcare settings," *npj Digital Medicine*, vol. 8, no. 1, art. 414, 2025.
- [19] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.
- [20] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in Proc. 3rd Innovations in Theoretical Computer Science Conf. (ITCS '12), 2012, pp. 214–226.
- [21] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in Proc. 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 2015, pp. 259–268.