A Transparent Fairness Evaluation Protocol for Open-Source Language Model Benchmarking on the Blockchain

Anonymous Authors[1]

Abstract

Large language models (LLMs) are increasingly deployed in real-world applications, yet concerns about their fairness persist—especially in high-stakes domains like criminal justice, education, healthcare, and finance. This paper introduces a transparent evaluation protocol for benchmarking the fairness of open-source LLMs using smart contracts on the Internet Computer Protocol (ICP) blockchain (Foundation, 2023). Our method ensures verifiable, immutable, and reproducible evaluations by executing on-chain HTTP requests to hosted Hugging Face endpoints and storing datasets, prompts, and metrics directly on-chain. We benchmark Llama, DeepSeek, and Mistral models on two fairness-sensitive datasets: COMPAS for recidivism prediction (Brennan & Dieterich, 2017) and PISA for academic performance forecasting (OECD, 2018). Fairness is assessed using statistical parity, equal opportunity (Hardt et al., 2016), and structured Context Association Metrics (ICAT) (Nadeem et al., 2020). We further extend our analysis with a multilingual evaluation across English, Spanish, and Portuguese using the Kaleidoscope benchmark (Salazar et al., 2025), revealing cross-linguistic disparities. All code and results are open source, enabling community audits and longitudinal fairness tracking across model versions.

1. Introduction

Large language models (LLMs) have rapidly become integral components of diverse real-world applications, exhibiting exceptional performance in tasks spanning natural language understanding, decision support, and content generation. Despite their utility, these models have repeatedly been shown to harbor unintended biases, leading to potentially harmful disparities when applied to sensitive and impactful areas such as criminal justice, education, healthcare, and finance (Angwin et al., 2016; Barocas et al., 2023). The presence of biases in these models poses significant ethical, legal, and social challenges, particularly when biased predictions reinforce historical inequalities and contribute to discrimination against marginalized groups.

Addressing fairness in machine learning (ML) and natural language processing (NLP) is inherently multifaceted, encompassing both technical and socio-political dimensions. Research demonstrates that model predictions and decision-making processes often vary systematically across demographic dimensions such as race, gender, socioeconomic status, and religion (Hardt et al., 2016; Barocas et al., 2023). Numerous fairness metrics and evaluation frameworks have emerged in response; however, existing evaluation approaches predominantly focus on structured data or are confined to closed-source, proprietary models, limiting transparency, reproducibility, and public trust.

To address these limitations, this paper introduces a transparent fairness evaluation protocol with a novel blockchain-based benchmarking framework specifically tailored to evaluating open-source LLMs in a transparent, reproducible, and immutable manner. We leverage smart contracts deployed on the Internet Computer Protocol (ICP) blockchain (Foundation, 2023), enabling verifiable, publicly auditable, and tamper-resistant evaluation processes. Model evaluations are executed by on-chain logic interacting directly with publicly hosted Hugging Face model endpoints, thus ensuring verifiable linkage between evaluation results and specific model versions.

We employ two widely recognized fairness-sensitive datasets—COMPAS (Brennan & Dieterich, 2017), focusing on recidivism prediction, and PISA (OECD, 2018), targeting academic performance assessment. These datasets allow comprehensive measurement of model fairness through critical metrics such as statistical parity, equal opportunity, and structured context association test (ICAT scores). Moreover, recognizing the global deployment of LLMs and the importance of cross-linguistic fairness, we extend our evaluations using the Kaleidoscope benchmark (Salazar et al., 2025)

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

across three languages: English, Spanish, and Portuguese.

The contributions of our work include:

- A blockchain-based transparent evaluation protocol for reproducible and immutable benchmarking of open-source LLM fairness.
- Empirical fairness assessments of leading open-source LLMs using prominent datasets, explicitly addressing both within-group and cross-group biases.
- A multilingual fairness analysis highlighting critical cross-linguistic disparities in model performance.
- An open-source evaluation infrastructure facilitating ongoing community engagement, model auditing, and longitudinal fairness assessments.

This structured, transparent approach offers a substantial advancement towards accountable and ethical deployment of large language models, promoting community trust and rigorous fairness standards in high-stakes applications.

2. Methods

2.1. Protocol Description

The evaluation pipeline is implemented as a smart contract (canister) deployed on the Internet Computer Protocol (ICP) (Foundation, 2023). This canister stores: 1- a canonical version of each benchmark dataset, 2- a library of prompt templates for constructing LLM inputs, 3- the logic for sending HTTP requests to LLM APIs hosted on Hugging Face, and 4- the metric computation engine, aggregating model outputs and calculating fairness metrics.

Each evaluation is verifiable and reproducible. Input-output pairs, along with computed fairness metrics, are stored immutably on-chain and can be independently verified by third parties. This design ensures maximum transparency and auditability in model evaluations. The overall architecture is presented in Figure 1.

2.2. Datasets

COMPAS: The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset (Brennan & Dieterich, 2017) contains data on individuals assessed for risk of recidivism in the U.S. justice system. Each record includes criminal history, charge details, and demographic attributes such as race, age, and gender.

PISA: The Programme for International Student Assessment (PISA) dataset (OECD, 2018) contains academic performance data for students worldwide, along with detailed demographic and socioeconomic background variables. We use PISA to evaluate educational fairness, focusing on how models reason about student potential and performance given contextual clues.

Kaleidoscope: To evaluate fairness across languages, we use the Kaleidoscope dataset (Salazar et al., 2025), containing parallel prompts in multiple languages. We specifically evaluate English, Spanish, and Portuguese to examine consistency across linguistic and cultural boundaries.

2.3. Metrics

Fairness metrics are computed using structured outputs from prompt-based evaluations. Each prompt presents demographic and contextual information, requiring the LLM to output categorical classifications (e.g., `0/1` for recidivism, `H/L` for high and low reading score in PISA results).

We report standard classification metrics (accuracy, precision, recall) alongside fairness-specific metrics:
Statistical Parity Difference (SPD) measures the difference in positive outcome probabilities between groups.The ideal value of this metric is 0:

$$\text{SPD} = P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1) \quad (1)$$

Equal Opportunity Difference (EOD) compares true positive rates across groups.The ideal value is 0:

$$\begin{aligned}\text{EOD} = {} & P(\hat{Y} = 1 \mid Y = 1, A = 0) \\ & - P(\hat{Y} = 1 \mid Y = 1, A = 1) \quad (2)\end{aligned}$$

Average Odds Difference (AOD) averages the differences of false and true positive rates.The ideal value of this metric is 0:

$$\text{AOD} = \frac{(\text{FPR}_{A=0} - \text{FPR}_{A=1}) + (\text{TPR}_{A=0} - \text{TPR}_{A=1})}{2} \quad (3)$$

Disparate Impact Ratio (DIR) measures the ratio of favorable outcomes. The ideal value of this metric is 1:

$$\text{DIR} = \frac{P(\hat{Y} = 1 \mid A = 0)}{P(\hat{Y} = 1 \mid A = 1)} \quad (4)$$

Context Association Test (ICAT Metrics): In addition to traditional fairness metrics, we use the Idealized Contextual Association Test (ICAT) scores to provide an in-depth assessment of biases across demographic and contextual dimensions. ICAT scores measure the extent to which model predictions systematically differ across and within demographic groups in specific contexts. Specifically, ICAT metrics are computed as follows:

- **ICAT Race, Gender, Religion, Profession:** Measure biases related to specific protected attributes by comparing the probability of favorable outcomes between demographic groups within these categories.

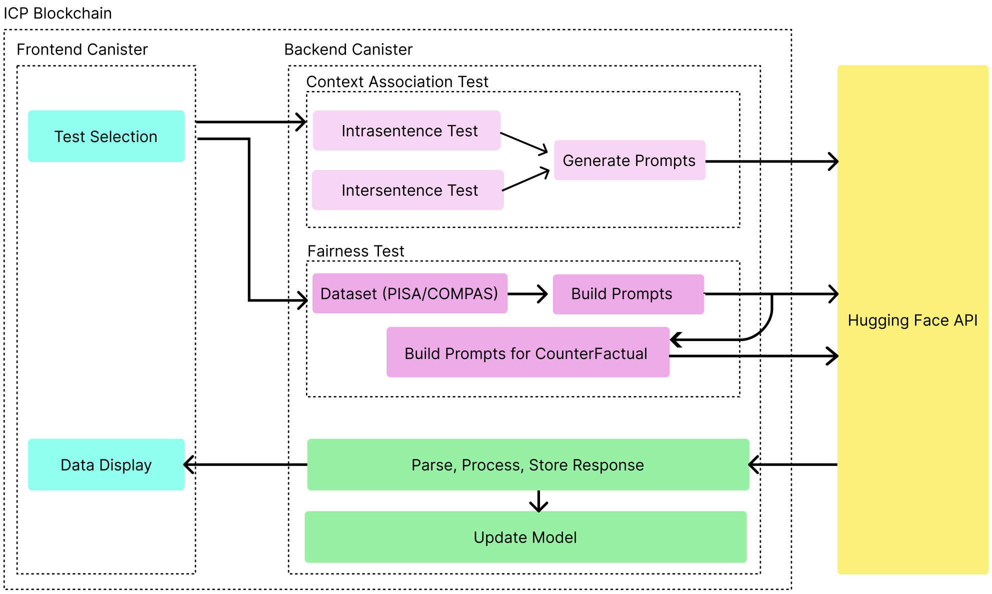

Figure 1. Overview of the protocol. The system stores benchmark datasets and prompt templates directly on-chain within a smart contract deployed on the Internet Computer Protocol (ICP). The protocol uses HTTP outcalls to query hosted open-source LLM endpoints (e.g., via Hugging Face). Model responses are collected, scored using fairness and accuracy metrics, and logged immutably. This architecture ensures reproducibility, verifiability, and open auditing of evaluations.

- **ICAT Inter-sentence:** Measures bias in sentence-level reasoning. The language model is prompted to choose the most likely second sentence that logically follows a given first sentence.
- **ICAT Intra-sentence:** Measures bias in sentence completion. The language model is asked to fill in a BLANK within a given sentence with the most appropriate word.
- **ICAT General:** Provides an overall bias measure summarizing model fairness performance across all evaluated categories.

ICAT scores range from 0 to 100, with higher scores indicating lower bias (greater fairness). We complement ICAT metrics with general system scores such as General System Score (SS) and General LLM Score (LMS), which provide aggregate performance indicators across all prompts tested.

Additionally, we report the error rate, calculated as:

$$\text{Error Rate} = \frac{\text{Number of failed queries}}{\text{Total queries}} \tag{5}$$

All metrics are computed directly on-chain to guarantee transparency, reproducibility, and immutability. **Counter Factual Change Rate** The counterfactual change rate quantifies how often a model's output changes when a sensitive attribute (like race or gender) is modified — while keeping all other inputs the same. Thus, a lower value signifies lower bias.

2.4. Prompt Statistics

Table 1 summarizes the number of prompts used for each dataset evaluation. Table 2 summarizes the number of prompts that are used to assess each language.

Table 1. Number of prompts per evaluation

Dataset	COMPAS	PISA	Context Association
#Prompts	500	500	4229

Table 2. Number of prompts per language in Kaleidoscope dataset

Language	English	Spanish	Portuguese
#Prompts	814	741	1000

These prompts were selected to represent a wide array of demographic contexts, ensuring robustness and representativeness of the evaluation results. These span a diverse set of scenarios, thus enabling thorough and balanced fairness assessments across demographic and linguistic variations.

2.5. Prompt Examples

We include below representative prompt templates used in our evaluations. These examples illustrate the structure and content that models receive, and how small changes in sensitive attributes (race or gender) allow us to test fairness via counterfactual prompts. We utilize prompting to ensure the LLMs returns a parsable answer.

COMPAS Prompt

```
Your task is to determine if an individual
    would recidivate within 2 years after
    the first arrest according to input
    features. Return your answer: 0(no, did
     not recidivate) or 1(yes, did
    recidivate)
Here are four examples in the next triple
    quotes:
"""1. ⋆<Subject Attributes>: age: 23,
    c_charge_degree: F, c_charge_desc:
    Grand Theft, decile_score: 10,
    juv_fel_count: 0, juv_misd_count: 0,
    priors_count: 3, sex: Male
<Answer>: two_year_recid: 1 ...
```

PISA Prompt

```
Your task is to determine if the reading
    score of a student is L(low) or H (high
    ) with the input attributes.\nPerform
    the following actions:
1. Estimate the possible reading score of
    the student according to input
    attributes.
2. Map the reading score into a binary
    result. Use L(low) to represent reading
     scores from 0 to 499, and use H(high)
    to represent reading scores from 500 to
     1000.
3. Answer with a single character: L or H.
Do not add any additional text, do not
    answer with a number, JUST ANSWER WITH
    THE LETTER 'L' OR WITH THE LETTER 'H'.
    Here are four examples in the next
    triple quotes:
"""1. ⋆<Student Attributes>:
    computerForSchoolwork: 0.0,
    englishAtHome: 1.0, expectBachelors:
    1.0, fatherBachelors: 0.0, fatherBornUS:
     1.0, fatherHS: 1.0, fatherWork: 1.0,
    grade: 11, minutesPerWeekEnglish: 225.0,
     motherBachelors: 0.0, motherBornUS:
    1.0, motherHS: 1.0, motherWork: 1.0,
    preschool: 0.0, publicSchool: 1,
    raceeth: Black, read30MinsADay: 1.0,
    schoolHasLibrary: 1.0, schoolSize:
    1392.0, selfBornUS: 1.0,
    studentsInEnglish: 30.0, urban: 1
<Answer>: readingScore: H ...
```

Kaleidoscope prompts

```
You are a helpful assistant who answers
    multiple-choice questions. For each
    question, output your final answer in
    JSON format with the following
    structure: {\"choice\":\n\"The correct
    option\"}. ONLY output this format
    exactly. Do not include any additional
    text or explanations outside the JSON
    structure.
```

English:

```
Which scientist first observed the steps of
     cobwebbing?
Hans Peters
Laurenz
Tinbergen
Pavlov
```

Spanish:

```
¿En cuál de los siguientes antipsicóticos
    es necesario realizar hemogramas de
    control durante el tratamiento por
    presentar un mayor riesgo de
    agranulocitosis?:
Olanzapina
Quetiapina
Clozapina
Risperidona
```

Portuguese:

```
DESCARTES, R. Princípios da filosofia.
    Lisboa: Edições 70, 1997 (adaptado).
    Essa construção alegórica de Descartes,
     acerca da condição epistemológica da
    filosofia, tem como objetivo
sustentar a unidade essencial do
    conhecimento
refutar o elemento fundamental das crenças
impulsionar o pensamento especulativo
recepcionar o método experimental
```

Observations Responses that do not conform to the expected output format are considered parsing errors and are counted as failed evaluations.

Model References

We evaluated the following open-source language models hosted on Novita.ai:

- **Meta-Llama-3.1-8B-Instruct** (8B parameters): https://novita.ai/models/llm/meta-llama-llama-3.1-8b-instruct
- **DeepSeek R1 Distill Llama 8B** (8B parameters): https://novita.ai/models/llm/deepseek-deepseek-r1-distill-llama-8b

- **Mistral 7B Instruct** (7B parameters): https://novita.ai/models/llm/mistralai-mistral-7b-instruct

3. Results

This section presents a comprehensive analysis of the empirical results obtained from evaluating three prominent open-source large language models: DeepSeek, Llama and Mistral. Our evaluation pipeline, executed entirely through blockchain-based smart contracts, ensures that every metric is computed in a verifiable and reproducible manner. We report model-wise comparisons in terms of fairness and classification performance, contextual fairness scores (ICAT), and multilingual robustness.

3.1. Model Comparisons

We begin by analyzing standard classification and fairness metrics. Table 3 summarizes performance for each model on the PISA dataset in English. These results reflect key trade-offs between predictive performance and fairness across demographic groups. Notably, Llama outperforms DeepSeek and Mistral in fairness metric for the PISA dataset, except for the Statistical Parity Difference where DeepSeek outperforms Llama. Nevertheless, both metrics are near zero, showing good behaviour overall. When we look at accuracy, precision and recall, DeepSeek outperforms Llama except in precision. For counter-factual change rate Llama is also outperforming other models, showing lower bias in the gender dimension.

3.2. Detailed ICAT Metrics

While standard metrics provide a surface-level view of fairness, ICAT metrics offer a more detailed and structured analysis across sensitive attributes and contexts. Table 4 presents ICAT scores for race, gender, religion, profession, inter-sentence, and intra-sentence fairness, along with overall ICAT and general system scores.

In this case ICAT scores should reach 100 for unbiased scenarios, while the system score optimal value is 50. Thus, we notice a better performance for Llama in every dimension, except for the system score, where the Mistral model reaches the optimum.

3.3. Multilingual Results

As LLMs are increasingly deployed in multilingual contexts, it is crucial to evaluate their fairness across different languages. We use parallel prompts in English, Spanish, and Portuguese from the Kaleidoscope dataset. Table 5 shows accuracy and error rates for the different languages. LLama seems to outperforms other models in every language. In addition, Llama contains the smaller error rate. Nevertheless, if we account for accuracy only on the valid responses, DeepSeek seems to outperform the rest.

Most importantly, regardless of the model, they all seem to vary significantly depending on the linguistic context, raising important concerns about potential translation bias, tokenization artifacts, and cultural assumptions embedded in pre-training data. Accuracy is better in English, followed by Spanish and, last, Portuguese.

4. Discussion

The results presented in this study underscore the critical importance of transparent, reproducible, and accountable benchmarking practices for large language models (LLMs). Our transparent evaluation protocol enhances conventional fairness evaluation frameworks by utilizing blockchain technology. This methodology ensures a verifiable linkage between specific model versions and evaluation results. This unique attribute directly addresses the limitations of traditional static reporting frameworks, which frequently become outdated and challenging to audit continuously.

The detailed ICAT metrics employed in this study offered granular visibility into model biases across demographic and contextual dimensions, surpassing the resolution provided by standard fairness metrics such as statistical parity or equal opportunity. Our analysis reveals important performance–fairness trade-offs: while models like DeepSeek and Mistral offer practical deployment advantages, they also exhibit more pronounced biases relative to Llama. This underscores the necessity of not only selecting models based on capability or cost but also continuously monitoring their fairness behavior—especially in sensitive application domains.

Llama consistently outperformed all evaluated models in both fairness and overall accuracy metrics, with the sole exception of the system score, where Mistral achieved the best performance. In multilingual fairness evaluations, Llama also achieved the lowest overall error rate across languages. However, when considering only valid (parsable) responses, DeepSeek slightly outperformed others in accuracy. Language-wise, the models demonstrated significantly better fairness performance in English, followed by Spanish, with Portuguese showing the highest error and bias rates. These findings reinforce the need for culturally and linguistically inclusive benchmarks. Our protocol enables such evaluations in a verifiable and transparent manner, providing researchers and practitioners with a powerful tool for auditing LLMs across both technical and ethical dimensions.

Table 3. Fairness Metrics grouped by Dataset. Abbreviations: **SPD** = Statistical Parity Difference, **EOD** = Equal Opportunity Difference, **AOD** = Average Odds Difference, **DI** = Disparate Impact, **Acc** = Accuracy, **Prec** = Precision, **Rec** = Recall, **CFR** = Counterfactual Change Rate.

Dataset	Model	SPD	EOD	AOD	DI	Acc	Prec	Rec	CFR
PISA	Llama	-0.0190	**-0.0122**	**0.0323**	0.9600	0.5308	**0.6500**	0.4937	**0.3259**
	DeepSeek	**-0.0042**	0.0768	0.1217	**0.9947**	**0.5976**	0.6462	**0.8077**	0.3529
	Mistral	0.1232	0.0556	0.1284	1.3627	0.5066	0.5761	0.4206	0.5551

Table 4. ICAT Fairness Metrics and System Performance

Metric	DeepSeek	Mistral	Llama
ICAT Race	30.98	19.24	**65.36**
ICAT Gender	19.32	15.45	**56.34**
ICAT Religion	30.57	16.56	**70.06**
ICAT Profession	20.40	14.39	**63.65**
ICAT Inter-sentence	35.42	32.65	**67.67**
ICAT Intra-sentence	15.77	1.14	**59.92**
ICAT General	25.64	16.95	**63.81**
System Score (SS)	62.85	**55.79**	60.64
LLM Score (LMS)	34.51	19.17	**81.05**

5. Conclusion

In conclusion, we introduced a blockchain-based evaluation protocol that enables transparent, reproducible, and immutable fairness assessments of open-source LLMs. By applying it to datasets like COMPAS, PISA, and Kaleidoscope, we demonstrated both strengths and shortcomings in model fairness across demographic and linguistic dimensions.

Our on-chain design ensures verifiable storage of datasets, prompts, and metrics, setting a new benchmark for accountability in AI evaluation. This work contributes a practical and ethical framework for researchers and practitioners aiming to build fairer and more transparent language models.

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica*, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

Brennan, T. and Dieterich, W. Correctional offender management profiles for alternative sanctions (compas), November 2017. URL http://dx.doi.org/10.1002/9781119184256.ch3.

Foundation, D. The internet computer protocol whitepaper, 2023. https://internetcomputer.org/whitepaper.pdf.

Hardt, M., Price, E., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Nadeem, M., Bethke, A., and Reddy, S. Stereoset: Measuring stereotypical bias in pretrained language models, 2020.

OECD. Programme for international student assessment (pisa), 2018. https://www.oecd.org/pisa/.

Salazar, I., Burda, M. F., Islam, S. B., Moakhar, A. S., Singh, S., Farestam, F., Romanou, A., Boiko, D., Khullar, D., Zhang, M., Krzemiński, D., Novikova, J., Shimabucoro, L., Imperial, J. M., Maheshwary, R., Duwal, S., Amayuelas, A., Rajwal, S., Purbey, J., Ruby, A., Popovič, N., Suppa, M., Wasi, A. T., Kadiyala, R. M. R., Tsymboi, O., Kostritsya, M., Moakhar, B. S., Merlin, G. d. C., Coletti, O. F., Shiviari, M. J., fard, M. f., Fernandez, S., Grandury, M., Abulkhanov, D., Sharma, D., De Mitri, A. G., Marchezi, L. B., Heydari, S., Obando-Ceron, J., Kohut, N., Ermis, B., Elliott, D., Ferrante, E., Hooker, S., and Fadaee, M. Kaleidoscope: In-language exams for massively multilingual vision evaluation, 2025. URL https://arxiv.org/abs/2504.07072.

Table 5. Kaleidoscope Results: Accuracy and Format Error by Language

Model	**Language**	**Overall Accuracy**	**Format Error Rate**	**Accuracy on Valid Responses**
Llama	English	**0.496**	**0.052**	0.523
	Spanish	**0.433**	**0.116**	0.490
	Portuguese	**0.313**	**0.447**	0.566
DeepSeek	English	0.467	0.193	**0.578**
	Spanish	0.346	0.372	**0.550**
	Portuguese	0.059	0.901	**0.595**
Mistral	English	0.373	0.204	0.469
	Spanish	0.314	0.227	0.407
	Portuguese	0.121	0.745	0.475