

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

Hugo Massaroli¹ Leonardo Iara¹ Emmanuel Iarussi² Viviana Siless^{2,1}

英寸

Abstract

大规模语言模型 (LLMs) 正越来越多地被部署在现实世界的应用中, 然而对于它们的公平性, 尤其是在刑事司法、教育、医疗和金融等高风险领域中, 仍然存在担忧。本文介绍了一种透明的评估协议, 该协议利用在互联网计算机协议 (ICP) 区块链上智能合约对开源 LLMs 的公平性进行基准测试 (Foundation, 2023)。我们的方法通过在链上执行到托管的 Hugging Face 端点的 HTTP 请求, 并直接在链上存储数据集、提示和指标, 确保可验证、不可变且可重复的评估。我们在 PISA 数据集上对 Llama、DeepSeek 和 Mistral 模型进行了学术表现预测的基准测试 (OECD, 2018), 该数据集适合使用统计均等和机会均等指标进行公平性评估 (Hardt et al., 2016)。我们还评估了从 StereoSet 数据集派生的结构化上下文关联指标 (Nadeem et al., 2020), 以衡量上下文关联中的社会偏见。我们进一步通过使用 Kaleidoscope 基准 (Salazar et al., 2025) 进行跨英语、西班牙语和葡萄牙语的多语言评估, 揭示了跨语言的差异。所有代码和结果均为开源, 支持社区审计和模型版本间的长期公平性跟踪。

1. 介绍

大型语言模型 (LLMs) 已迅速成为各种现实应用中的重要组成部分, 在跨越自然语言理解、决策支持和内容生成的任务中表现出卓越的性能。尽管它们非常有用, 但这些模型被多次证明存在无意的偏见, 当应用于刑事司法、教育、医疗保健和金融等敏感和有影响力的领域时可能导致潜在的有害差异。(Angwin et al., 2016; Barocas et al., 2023) 这些模型中存在的偏见提出了重大伦理、法律和社会挑战, 特别是当有偏见的预测加强了历史上的不平等, 并导致对边缘化群体的歧视时。

¹ FAI3, 布宜诺斯艾利斯, 阿根廷 ² 阿根廷迪特拉大学商学院. Correspondence to: Viviana Siless <viviana.siless@utdt.edu>.

在机器学习 (ML) 和自然语言处理 (NLP) 中解决公平性问题本质上是多方面的, 涵盖了技术和社会政治维度。研究表明, 模型预测和决策过程经常在种族、性别、社会经济地位和宗教等人口维度上系统性地变化 (Hardt et al., 2016; Barocas et al., 2023)。出现了许多公平性度量标准和评估框架以应对这一问题; 然而, 现有的评估方法主要集中于结构化数据或局限于闭源的专有模型, 这限制了透明性、可重复性和公众信任。

为了应对这些限制, 本文介绍了一种透明的公平性评估协议, 借助一种新颖的基于区块链的基准测试框架, 该框架专门用于以透明、可重复和不可更改的方式评估开源大规模语言模型 (LLMs)。我们利用部署在互联网计算机协议 (ICP) 区块链 (Foundation, 2023) 上的智能合约, 启用可验证、公开审核和防篡改的评估过程。模型评估由链上逻辑执行, 直接与公共托管的 Hugging Face 模型端点交互, 从而确保评估结果与特定模型版本之间具有可验证的链接。在链上托管评估和结果进一步加强了我们基准测试框架的可重复性、不可更改性和透明性。我们采用一个广为认可的关注公平性的数据集——PISA (OECD, 2018), 目标是学术表现评估。该数据集允许通过关键指标 (如统计平等和机会均等) 全面衡量模型的公平性。我们还使用 StereoSet (Nadeem et al., 2020) 数据集来衡量 LLMs 答案中的刻板偏见。此外, 鉴于 LLMs 的全球部署和跨语言公平性的重要性, 我们使用万花筒基准 (Salazar et al., 2025) 在三种语言 (英语、西班牙语和葡萄牙语) 中扩展了我们的评估。

我们的工作贡献包括:

- 一个基于区块链的透明评估协议, 用于开源 LLM 公平性的可重现和不可更改的基准测试。
- 利用著名数据集对领先的开源大型语言模型 (LLM) 进行实证公平性评估, 明确解决群体内和群体间偏见。
- 多语言公平性分析, 突出模型性能中关键的跨语言差异。
- 一个开源评估基础设施, 以促进持续的社区参与、模型审计和纵向公平性评估。

这种结构化、透明的方法为大规模语言模型的负责和伦理部署提供了实质性的进展, 促进了社区信任和高风险应用中的严格公平标准。

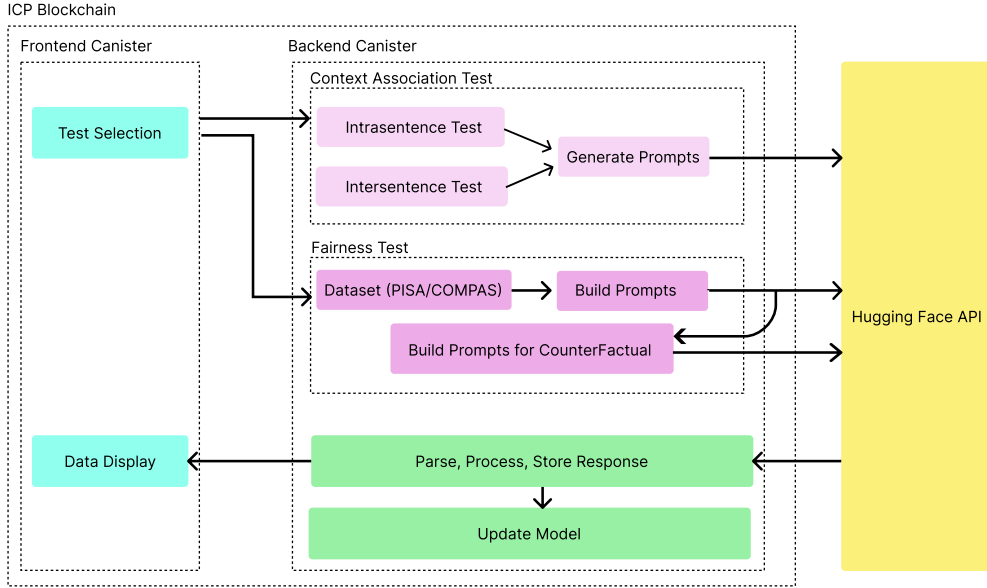


Figure 1. 协议概述。系统直接在部署在 Internet 计算机协议 (ICP) 上的智能合约中存储基准数据集和提示模板。该协议使用 HTTP 外调查询托管的开源 LLM 端点 (例如, 通过 Hugging Face)。收集模型响应, 使用公平性和准确性指标进行评分, 并进行不可变的记录。该架构确保评估的可重现性、可验证性和开放审核。

2. 方法

2.1. 协议描述

评估流程作为一个智能合约 (canister) 实现, 并部署在互联网计算机协议 (ICP) (Foundation, 2023) 上。这个 canister 存储了: 1- 每个基准数据集的权威版本, 2- 用于构建 LLM 输入的提示模板库, 3- 向托管在 Hugging Face 上的 LLM API 发送 HTTP 请求的逻辑, 以及 4- 度量计算引擎, 用于聚合模型输出并计算公平性度量。

每个评估都是可验证和可再现的。输入-输出对以及计算的公平性指标不可变地存储在区块链上, 可以由第三方独立验证。该设计确保了模型评估的最大透明度和可审计性。整体架构如图 1 所示。

2.2. 数据集

PISA: 国际学生评估项目 (PISA) 数据集 (OECD, 2018) 包含来自全球学生的学业表现数据, 以及详细的人口统计和社会经济背景变量。我们使用 PISA 来评估教育公平性, 重点关注在给定上下文线索的情况下, 模型如何推理学生的潜力和表现。StereoSet: StereoSet 数据集 (Nadeem et al., 2020) 由多项选择题组成, 旨在评估两种任务类型。第一种任务涉及预测句子中缺失的单词 (句内任务), 而第二种任务要求在给定前一句的情况下选择最合适的下一句 (句间任务)。该数据集能够通过为每个提示提供刻板印象和反刻板印象的答案选项来评估刻板印象偏见, 并通过包含无意义的 (干扰) 选项评估语言建模性能。评估涵盖四个领域: 职业、性别、宗教和种族。

万花筒: 为了评估不同语言之间的公平性, 我们使用了万花筒数据集 (Salazar et al., 2025), 其中包含多种语言的平行提示。我们特别评估了英语、西班牙语和葡萄牙语, 以检查跨语言和文化界限的一致性。

2.3. 指标

公平性指标是使用基于提示的评估中的结构化输出来计算的。每个提示都呈现人口统计和背景信息, 要求大型语言模型输出分类结果 (例如, H / L 用于 PISA 结果中的高低阅读分数)。

我们报告了标准分类指标 (准确性、精密度、召回率) 以及公平性特定指标:

统计平等差异 (SPD) 测量组间正向结果概率的差异。该指标的理想值是 0:

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

机会平等差异 (EOD) 比较各组之间的真正例率。理想值为 0:

$$EOD = P(\hat{Y} = 1 | Y = 1, A = 0) - P(\hat{Y} = 1 | Y = 1, A = 1) \quad (2)$$

平均赔率差异 (AOD) 则是对错误和真正例率的差异进行平均。该指标的理想值为 0:

$$AOD = \frac{(FPR_{A=0} - FPR_{A=1}) + (TPR_{A=0} - TPR_{A=1})}{2} \quad (3)$$

差异影响比 (DIR) 衡量有利结果的比率。该指标的理想值为 1:

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

上下文关联测试 (ICAT 指标): 除了传统的公平性指标之外, 我们还使用理想化上下文关联测试 (ICAT) 得分来提供对跨人口和上下文维度偏见的深入评估。ICAT 得分衡量的是在特定上下文中, 模型预测在跨人口群体和群体内部的系统性差异程度。具体而言, ICAT 指标的计算如下:

- ICAT 种族、性别、宗教、职业: 通过比较这些类别中人口群体之间的有利结果概率, 衡量与特定受保护属性相关的偏见和语言建模性能。
- ICAT 句间: 衡量句子层次推理中的偏差。提示语言模型选择在逻辑上最有可能跟随给定第一句的第二句。
- ICAT 句内: 用于衡量句子完成中的偏差。要求语言模型在给定句子中的空白处填入最合适的单词。
- ICAT 通用: 提供一种总体偏差衡量, 用于总结模型公平性和语言建模性能在所有评估类别中的表现。

ICAT 评分范围从 0 到 100, 分数越高表明偏见越低 (公平性越高), 有意义的答案率越高。我们用一般的系统评分如一般刻板印象评分 (SS) 和一般语言建模评分 (LMS) 来补充 ICAT 指标, 这些为所有测试的提示提供了综合的性能指标。

此外, 我们报告了 Kaleidoscope 的错误率, 即文本 LLM 响应不匹配任何可能的答案或所需的 JSON 格式。其计算公式为:

$$\text{Format Error Rate} = \frac{\text{Number of invalid answers}}{\text{Total answers}} \quad (5)$$

所有指标均直接在链上计算, 以确保透明性、可重复性和不可变性。

反事实变化率反事实变化率量化了当敏感属性 (如种族或性别) 被修改时, 模型输出改变的频率——同时保持所有其他输入不变。因此, 较低的值表示较低的偏差。

Table 1. 每次评估的提示数量

Dataset	PISA	Context Association
# Prompts	500	4229

Table 2. Kaleidoscope 数据集中每种语言的提示数量

Language	English	Spanish	Portuguese
# Prompts	814	741	1000

2.4. 提示统计

表 1 总结了用于每个数据集评估的提示数量。表 2 总结了用于评估每种语言的提示数量。

这些提示被选中以代表广泛的人口统计背景, 确保评估结果的稳健性和代表性。这些情境跨越多样化的场景, 从而使跨人口统计和语言差异的公平性评估能够更加全面和平衡。

2.5. 提示示例

我们在下面包含了评估中使用的具有代表性的提示模板。这些例子展示了模型接收到的结构和内容, 以及通过在敏感属性 (种族或性别) 上进行小的改变, 如何利用反事实提示来测试公平性。我们利用提示来确保 LLMs 返回可解析的答案。

2.5.1.

PISA 提示

Your task is to determine if the reading score of a student is L(low) or H (high) with the input attributes.
Perform 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 ...

Table 3. 按数据集分组的公平性指标。缩写: SPD = 统计平等差异, EOD = 机会平等差异, AOD = 平均差异, DI = 差异影响, Acc = 准确率, Prec = 精确率, Rec = 召回率, CFR = 反事实变化率。

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

2.5.2.

万花筒提示

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

英语:

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

西班牙语:

¿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

葡萄牙语:

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

2.6.

模型参考 我们评估了由 Novita.ai 托管的以下开源语言模型:

不符合预期输出格式的回复被视为解析错误, 并被计入失败评估。

本节对评估三种著名开源大型语言模型: DeepSeek、Llama 和 Mistral 的实证结果进行了全面分析。我们的

评估流程完全通过基于区块链的智能合约进行, 确保每个指标均以可验证且可重复的方式计算。我们报告了模型在公平性和分类性能、背景公平性得分 (ICAT) 以及多语言鲁棒性方面的比较。

2.7. 模型比较

我们首先分析标准分类和公平性指标。表格 3 总结了每个模型在 PISA 数据集上的性能 (英文)。这些结果反映了在不同人口群体之间预测性能和公平性之间的关键权衡。值得注意的是, 除了统计平等差异指标以外, Llama 在 PISA 数据集的公平性指标上优于 DeepSeek 和 Mistral, 在这一指标上 DeepSeek 优于 Llama。然而, 这两个指标都接近零, 整体表现良好。再看准确率、精确率和召回率, 除了精确率外, DeepSeek 均优于 Llama。在反事实变化率方面, Llama 也优于其他模型, 在性别维度上表现出更低的偏见。

2.8. 详细的 ICAT 指标

虽然标准指标提供了一个关于公平性的浅层视角, ICAT 指标则在敏感属性和上下文之间提供了更详细和结构化的分析。表 4 展示了种族、性别、宗教、职业、句间和句内公平的 ICAT 分数, 以及整体 ICAT 和一般刻板印象及语言建模分数。

在这种情况下, 对于无偏见场景, ICAT 得分应达到 100, 而刻板印象得分的最佳值是 50。因此, 除了刻板印象得分外, 我们注意到 Llama 在每个维度上表现更好, 其中 Mistral 模型达到了最佳值。

Table 4. ICAT 公平性指标和系统性能

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
Stereotype Score (SS)	62.85	55.79	60.64
LLM Score (LMS)	34.51	19.17	81.05

2.9. 多语言结果

随着大型语言模型（LLM）的日益广泛应用于多语言环境，评估其不同语言间的公平性至关重要。我们使用来自万花筒数据集的英语、西班牙语和葡萄牙语的平行提示。表格 5 显示了不同语言的准确性和格式错误率。Llama 似乎在每种语言中都优于其他模型。此外，Llama 的错误率是最小的。尽管 DeepSeek 在生成提示中规定的 JSON 格式时遇到困难——经常包含不必要的文本在预期结构之外，从而增加了其错误率——但在仅考虑有效响应的情况下，其准确性优于其他模型。

最重要的是，无论哪个模型，它们似乎都严重依赖于语言环境而变化，这引发了对潜在翻译偏差、分词失误以及嵌入在预训练数据中的文化假设的重要担忧。准确性在英语中最好，其次是西班牙语，最后是葡萄牙语。

3. 讨论

本研究中呈现的结果强调了对大型语言模型（LLMs）的透明、可重复和负责任的基准测试实践的重要性。我们透明的评估协议通过利用区块链技术增强了传统的公平性评估框架。这种方法确保了特定模型版本与评估结果之间的可验证关联。这一独特属性直接解决了传统静态报告框架的局限性，这些框架经常变得过时并难以持续审计。

本研究中使用的详细 ICAT 指标提供了对模型在人口和背景维度上的偏差的细致可见性，超越了诸如统计平等或机会平等这样的标准公平性指标所提供的分辨率。我们的分析揭示了重要的性能与公平性权衡：尽管像 DeepSeek 和 Mistral 这样的模型在实际部署中有优势，它们相对于 Llama 也表现出更明显的偏差。这强调了不仅要基于能力或成本选择模型，还要持续监控它们的公平性行为——尤其是在敏感的应用领域中。

在公平性和整体准确性指标上，Llama 一贯表现优于所有评估的模型，唯一的例外是系统得分，在这一方面 Mistral 获得了最佳表现。在多语言公平性评估中，Llama 在多种语言中也实现了最低的整体错误率。然而，在仅考虑有效（可解析）响应时，DeepSeek 的准确性略优于其他模型。从语言角度看，各模型在英语中的公平性表现显著更好，随后是西班牙语，而葡萄牙语则显示出最高的错误率和偏差率。这些发现加强了文化和语言包容性基准的必要性。我们的协议允许以可验证和透明的方式进行此类评估，为研究人员和实践者提供了一个强大的工具，以便在技术和伦理维度上对 LLMs 进行审查。

4. 结论

总之，我们介绍了一种基于区块链的评估协议，该协议能够对开源 LLM 进行透明、可重复和不可更改的公平性评估。通过将其应用于 PISA、StereoSet 和

Kaleidoscope 等数据集，我们展示了模型在跨人口统计和语言维度的公平性上的优势和不足之处。

我们的链上设计确保了数据集、提示和指标的可验证存储，为 AI 评估中的责任设立了新的基准。这项工作为研究人员和从业者提供了一个实用且符合伦理的框架，旨在构建更公平和更透明的语言模型。

5.

代码可用性 本研究中使用的所有代码和数据集可在以下网址获取：https://github.com/FAI3network/ICP_MVP。

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.
- 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., Kostitsya, 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,

Table 5. 万花筒结果：按语言划分的准确性和格式错误

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

* A response is deemed valid when the LLM’s answer corresponds to one of the available multiple-choice options.

S., and Fadaee, M. Kaleidoscope: In-language exams for massively multilingual vision evaluation, 2025. URL <https://arxiv.org/abs/2504.07072>.