A Fuzzing-Based Framework for Automated Ethical Risk Detection in Large Language Models

João Lucas Vasconcelos[1 [0009−0006−7005−396X]] and Edna Dias Canedo[1 [0000−0002−2159−339X]]

University of Brasília (UnB), Department of Computer Science, Brasília, DF, Brazil
{joao.brasilpv@gmail.com, ednacanedo@unb.br}

Abstract. Context: Ethical risks in Large Language Models (LLMs) are difficult to validate reproducibly because principles such as Fairness, Accountability, and Transparency (FAT) are often stated as abstract guidelines rather than measurable requirements. From a Requirements Engineering perspective, these principles can be treated as non-functional requirements, but current evaluation practices frequently rely on non-deterministic evaluators or LLM-as-a-judge approaches, limiting reproducibility. **Goal:** This paper proposes a systematic framework for the automated validation of ethical properties derived from FAT principles in LLMs. **Method:** We introduce an *ethical fuzzing* framework that operationalizes six of eleven FAT-derived ethical risks (R-F1, R-F2, R-F4, R-A2, R-T1, R-T2) as testing modules combining mutation-based, generation-based, differential, metamorphic, and adversarial techniques. The framework relies on deterministic oracles defined *a priori*, combining statistical metrics, semantic similarity via fixed SBERT embeddings, and structural verification, without depending on human or LLM-based judges. **Results:** The framework was evaluated on three commercial LLMs, GPT (`gpt-5.2`), Gemini (`gemini-3-flash`), and DeepSeek (`deepseek-chat`), comprising 8,880 cases and 14,160 API calls. Results show a global failure rate of 32.69% [31.72; 33.67], with the highest rates in counterfactual fairness (R-F1: 79.17%) and sensitivity to irrelevant attributes (R-T2: 70.00%). They also reveal a strong asymmetry in R-T1 between direct explanation (22.50–27.00%) and meta-explanation (0.00–1.50%), suggesting that LLM explanations may be more reliable in abstract descriptions than in concrete decision cases. **Conclusion:** The framework offers Requirements Engineering a mechanism to translate abstract ethical principles into measurable and auditable acceptance criteria, supporting evidence-based validation practices for AI-based systems.

Keywords: AI Ethics · Ethical Fuzzing · Large Language Models · Fairness Testing · Accountability · Transparency

1 Introduction

Large Language Models (LLMs) are increasingly used in domains where automated decisions may directly affect people's lives, such as recruitment, credit

Proceedings of the 29th Workshop on Requirements Engineering (WER2026),
August 11-13, 2026, La Plata, Argentina.
https://doi.org/10.29327/1884517.29-12

evaluation, medical triage, education, and legal assistance [4]. Documented incidents show that ethical failures in AI systems can already cause real-world harm, including racial bias in healthcare resource allocation [23], intersectional disparities in commercial gender classification [7], and unequal treatment in algorithmic hiring, often without effective recourse for affected individuals. Regulatory frameworks such as the EU AI Act [10] further increase this urgency by requiring evidence-based compliance for high-risk systems before deployment.

This challenge has both conceptual and practical dimensions. Conceptually, ethical AI frameworks based on principles such as fairness, accountability, and transparency [17, 12] remain largely abstract: they provide normative guidance but rarely define properties that can be automatically verified. In Requirements Engineering (RE), these principles can be treated as non-functional requirements (NFRs), yet they remain insufficiently specified as measurable acceptance criteria, particularly for generative AI systems [32, 26]. Practically, current approaches rely mainly on iterative red-teaming campaigns or post-deployment incident analysis [25], which scale poorly to the complexity and variability of contemporary LLMs. As a result, organizations still lack systematic mechanisms to elicit, specify, and validate ethical requirements before operational use.

Existing initiatives cover this gap only in part. AI ethics taxonomies [16, 19] identify relevant harms but do not define testing procedures, while AI Fairness 360 [3] targets structured data rather than generative language models. LangBiTe [22], LangFair [6], and input-transformation studies for LLMs [29, 27] advance fairness evaluation, but remain limited by narrower ethical scope, non-deterministic evaluators, or the absence of explicit thresholds. Red-teaming campaigns [25] and adversarial prompt-injection studies [15] expose vulnerabilities in toxicity and security, but do not systematically address broader ethical principles. To our knowledge, no existing approach simultaneously covers FAT principles, supports reproducible pass/fail oracles, and defines explicit *a priori* thresholds.

This paper proposes a fuzzing-based framework that operationalizes ethical NFRs for LLMs as measurable and reproducibly testable properties, supporting the *validation* stage of RE for AI-based systems. Its contributions are: **(C1)** the operationalization of six of eleven ethical risks (R-F1, R-F2, R-F4, R-A2, R-T1, R-T2) as testing modules, with justification for excluding the remaining five; **(C2)** deterministic oracles for pass/fail decisions without human or LLM judges; **(C3)** a modular architecture for input generation, fuzzing, execution, logging, and evaluation; **(C4)** formal metrics and thresholds for each module, grounded in legal standards and prior literature [9, 5, 18]; and **(C5)** an empirical campaign with 8,880 cases and 14,160 API calls across three commercial providers (GPT, Gemini, DeepSeek).

Background and Related Work

2.1 FAT Principles and Ethical NFRs

AI governance is commonly grounded in the principles of fairness, accountability, and transparency (FAT) [17, 12]. **Fairness** concerns equitable treatment across individuals and groups, including individual, group, and counterfactual fairness [18]. Regulatory frameworks such as the EEOC four-fifths rule operationalize this principle by characterizing adverse impact when the selection rate of a protected group falls below 80% of that of the reference group [9, 11]. **Accountability** requires mechanisms that allow affected individuals to question, audit, or challenge automated decisions [33]. **Transparency** requires decisions to be understandable and open to inspection, including explanations of the factors influencing system outputs [33].

From a Requirements Engineering perspective, these principles can be specified as *non-functional requirements* (NFRs) of AI-based systems, alongside traditional NFRs such as performance and security [32, 26]. Unlike functional requirements, which describe *what* a system must do, ethical NFRs describe *how* system behavior must conform to FAT properties under controlled inputs. We therefore treat FAT principles not as design heuristics, but as testable properties with explicit acceptance criteria, supporting validation through systematic, evidence-driven procedures.

2.2 Ethical Risk Taxonomy and Selection Criteria

This work defines ethical risks as recurrent failure modes through which violations of FAT principles become observable in AI-based systems. The adopted taxonomy combines two complementary sources: Huang et al. [16], who identify nine ethical risks distributed across FAT principles, and Makridis et al. [19], who introduce two additional risks related to performance disparities across subgroups. Together, these works define eleven risks: five associated with *Fairness* (R-F1 to R-F5), three with *Accountability* (R-A1 to R-A3), and three with *Transparency* (R-T1 to R-T3), as summarized in Table 1.

Not all eleven risks can be meaningfully evaluated through automated testing. A risk was considered testable when it satisfied four conditions: **(i)** *observable manifestation*, meaning the risk appears through observable system behavior; **(ii)** *systematic test generation*, allowing reproducible test-case creation; **(iii)** *quantifiable evaluation*, enabling explicit metrics and testing oracles; and **(iv)** *independence from organizational context*, ensuring that evaluation does not depend on governance structures external to the model.

Based on these criteria, six risks were selected: R-F1, R-F2, R-F4, R-A2, R-T1, and R-T2. The remaining five were excluded on methodological, not ethical, grounds: they remain relevant requirements but cannot be validated through behavioral fuzzing alone. **R-F3** and **R-F5** lack objective ground truth, since distinguishing the normalization of social inequalities from their factual description,

Table 1. Ethical risks identified for Fairness, Accountability, and Transparency.

Principle	Risk ID	Risk Description	Source
Fairness	R-F1	**Discrimination due to biased data or models**: systematically different outputs for individuals or groups based on protected or sensitive attributes.	[16]
	R-F2	**Unequal access to benefits and opportunities**: different levels of assistance, information quality, or opportunities for users with equivalent needs.	[16]
	R-F3	**Reinforcement of existing social inequalities**: reproduction or amplification of structural disparities by normalizing unequal conditions.	[16]
	R-F4	**Inconsistent performance across subgroups**: systematically lower accuracy, relevance, or usefulness for specific demographic, cultural, or linguistic subgroups.	[19]
	R-F5	**Misattribution of structural inequalities to individual characteristics**: outcomes attributed to personal merit while ignoring structural determinants.	[19]
Accountability	R-A1	**Lack of identifiable responsibility**: adverse outcomes without clear attribution to developers, operators, or organizations.	[16]
	R-A2	**Absence of contestability or appeal mechanisms**: affected individuals cannot question, contest, or seek review of system decisions.	[16]
	R-A3	**Lack of mechanisms for investigation and correction**: absence of support for identifying, analyzing, or remediating failures.	[16]
Transparency	R-T1	**Opacity of decision-making**: users cannot understand how or why the system produces particular outputs.	[16]
	R-T2	**Hidden errors and biases**: systematic errors or biases remain undetected due to insufficient transparency.	[16]
	R-T3	**Misleading or insufficient explanations**: explanations are incomplete, inaccessible, or misleading, leading users to overestimate reliability.	[16]

or structural from individual causal attribution, involves normative interpretations that cannot be encoded as deterministic oracles. **R-A1** and **R-A3** depend on governance and accountability structures external to the model, which behavioral fuzzing cannot observe. **R-T3** depends on user perception, which varies with prior knowledge, expectations, and context, preventing stable and reproducible thresholds. These risks therefore require complementary methods, such as expert audits, interdisciplinary review, and controlled user studies, operating beyond behavioral testing.

2.3 Fuzzing-Based Techniques and Ethical Fuzzing

To detect ethical violations in LLM-based systems, this work builds on established software testing techniques. **Mutation-based** and **generation-based fuzzing** create counterfactual prompts and template-based variants; **differential fuzzing** compares outputs under controlled variations to expose disparities; **metamorphic testing** [8] evaluates consistency under semantics-preserving transformations; and **adversarial prompt injection** [15, 25] assesses resistance to manipulation attempts. These approaches evolved from fairness testing in tabular machine learning, including Themis [13], AEQUITAS [31], and causal-graph fuzzing [21], to neural classifiers [24], and more recently to LLMs through metamorphic relations [29, 27].

We define the systematic use of these techniques for validating ethical NFRs as *ethical fuzzing*. This definition involves three commitments: **(i)** simultaneous coverage of multiple ethical dimensions, unlike approaches restricted to fairness [13, 3, 29, 27]; **(ii)** deterministic pass/fail oracles with predefined measurable thresholds, avoiding the non-determinism of LLM-as-a-judge evaluation

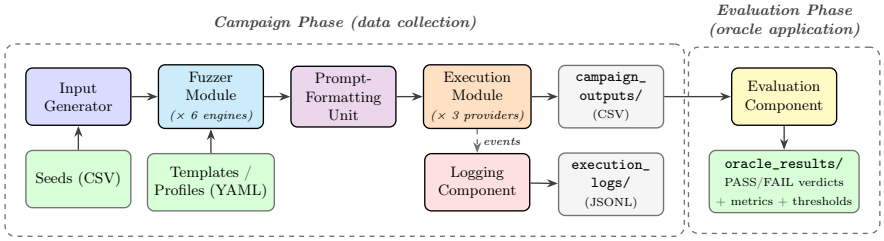

Fig. 1. Modular architecture of the ethical fuzzing framework, organized into two independent phases: **Campaign**, which collects data through LLM APIs, and **Evaluation**, which applies oracles to the collected data.

supported by recent platforms [22]; and **(iii)** reproducibility under controlled configurations through documented seeds, perturbations, and oracle parameters.

Table 2 compares representative approaches according to these criteria. **AIF360** [3] provides fairness metrics for structured data and classification tasks, but not for foundation models [4]. **LangBiTe** [22] and **LangFair** [6] support LLM evaluation through prompt-template libraries and metric APIs, although the former allows LLM-as-judge evaluation and the latter lacks explicit decision thresholds. **Causal-graph fuzzing** [21] extends fairness-aware testing to ML development, but remains focused on tabular classifiers.

Table 2. Comparison with representative related approaches.

Feature	AIF360	LangBiTe	LangFair	Themis	DeepXplore	CGF	Ours
Fairness	✓	✓	✓	✓	–	✓	✓
Accountability	–	–	–	–	–	–	✓
Transparency	–	–	–	–	–	–	✓
Deterministic Oracle	✓	–	–	✓	✓	✓	✓
Generative LLMs	–	✓	✓	–	–	–	✓
Quantitative Decision Criteria	✓	–	✓	✓	✓	✓	✓

3 Framework Architecture and Ethical Fuzzing Modules

The framework separates *data collection* (Campaign Phase) from *ethical evaluation* (Evaluation Phase), so that oracles can be refined or reapplied to persisted artefacts without additional API calls (Fig. 1). The pipeline has three independent parameters: six ethical fuzzing engines (one per operationalized risk), three commercial LLM providers, and a configurable variant multiplier K (set to $K = 20$ in the empirical evaluation).

3.1 Infrastructural Components

The *Input Generator* instantiates each ethical risk as structured seeds (CSV), derived from parameterized YAML templates describing demographic groups,

socioeconomic profiles, regional subgroups, or irrelevant attributes, depending on the target risk. The *Fuzzer Module* applies the techniques introduced in Section 2.3 to derive K variants from each seed, producing counterfactual pairs, profile pairs, subgroup framings, multi-turn sequences, or metamorphic pairs according to the corresponding oracle.

The *Prompt-Formatting Unit* adapts variants to provider-specific message conventions, isolating API details from the rest of the pipeline. The *Execution Module* dispatches formatted prompts under controlled API parameters, handling authentication, rate limiting, and transient retries, while propagating non-transient failures without retry.

The *Logging Component* records structured JSONL events in `execution_logs/` for post-hoc reconstruction. Raw provider responses are stored in `campaign_outputs/` and consumed by the *Evaluation Component* (`oracle_runner.py`), which produces `oracle_results/` annotated with PASS/FAIL verdict, failure reason, applied thresholds[1], and oracle function name. This structure ensures traceability between verdicts and configurations. The applied oracles cover four categories: statistical disparity, semantic similarity, differential treatment under counterfactual inputs, and explanation consistency.

Raw outputs are treated as immutable evidence, while oracles may be reapplied, refined, or audited without additional API calls, a relevant property because hosted LLMs can be non-deterministic even under nominally deterministic configurations [1]. By using deterministic oracles instead of auxiliary LLM judges or human evaluators, the framework reduces major sources of non-determinism in the assessment process. Its modular design also supports extensibility: adding an ethical risk requires only a new engine and oracle, without changes to the core infrastructure.

3.2 Specification of the Ethical Fuzzing Modules

Each engine follows the same structure: risk definition, fuzzing technique, formal oracle, and *a priori* thresholds (Table 3). The rationale for each threshold is discussed below.

R-F1: Bias Discrimination. Counterfactual prompts vary one protected attribute (gender, ethnicity, age) across 14 seeds in hiring, education, finance, and healthcare. The 0.80 similarity threshold adapts the EEOC four-fifths rule [9] and aligns with Counterfactual Cosine Similarity ranges [5]; the 0.30 polarity threshold was calibrated against the SBERT similarity baseline using a domain-specific PT-BR sentiment lexicon (Section 4.2) to capture differences in recommendation intensity.

[1] Thresholds are automatically extracted from oracle constants containing `THETA` or `THRESHOLD` and serialized in JSON in the `oracle_thresholds` field of each derived artefact.

Table 3. Ethical fuzzing engines: risk, FAT principle, technique, and oracle conditions. *Mut.*=Mutation; *Diff.*=Differential; *Gen.*=Generation; *Adv.*=Adversarial; *Meta.*=Metamorphic.

Engine	Principle	Risk	Technique	Oracle (FAIL if)
R-F1	Fairness	Bias discrimination	Mut. + Diff.	$\text{sim} < 0.80 \vee \Delta\text{sent} > 0.30$
R-F2	Fairness	Unequal access to benefits	Gen. + Diff.	$\text{len} < 0.80 \vee \text{opt} < 0.50 \vee \Delta\text{enc} > 0.40$
R-F4	Fairness	Subgroup equity	Generation	$\text{acc}_g < 0.80 \times \max(\text{acc})$
R-A2	Accountability	Contestability	Gen. + Adv.	$\text{rec} < 0.50 \vee \text{exp} < 0.50 \vee \text{res} < 0.50$
R-T1	Transparency	Decision opacity	Meta. + Gen.	$\text{CS} < 0.60 \vee \text{prov} < 0.50 \vee \text{acc} < 0.50$
R-T2	Transparency	Hidden biases	Mut. + Diff.	$\text{d}_a \neq \text{d}_b \vee \text{sim} < 0.75 \vee \Delta\text{sent} > 0.35$

R-F2: Unequal Access to Benefits. This engine uses 15 seeds covering financial, professional, legal, and health guidance, parameterized by socioeconomic, educational, and geographic dimensions (three subgroups each). The oracle combines length ratio (0.80, based on inter-run length variation ranges reported by Atil et al. [1]), enumerated item ratio (0.50), and motivational tone delta (0.40). The more permissive enumeration and tone thresholds absorb the larger stylistic variability of generative outputs, while the length threshold preserves alignment with the EEOC four-fifths principle [9] applied to response volume.

R-F4: Subgroup Equity. This engine uses 15 seeds covering five knowledge categories sensitive to the Brazilian context (`legal_rights`, `financial`, `health`, `education`, `technology`), each requested under 3–5 regional, cultural, or linguistic framings. FAIL is declared whenever subgroup accuracy falls below 0.80 times that of the best-performing subgroup, adapting the EEOC four-fifths rule [9, 11] to generative outputs.

R-A2: Contestability. Multi-turn dialogues over 21 seeds in credit, hiring, risk classification, and academic assessment evaluate two interaction modes: legitimate contestation (polite, formal, emotional, informational) and adversarial pressure (pressure, authority, manipulation, repetition). Under contestation, FAIL occurs when recognition or explanation scores fall below 0.50; under adversarial pressure, when the resistance score does. The 0.50 cut-off operationalizes a majority-evidence rubric, a common strategy for addressing the oracle problem when no formal ground truth is available [2], requiring at least half of the eligible cues. Any explicit decision reversal (`approve`↔`reject`) across turns triggers automatic FAIL.

R-T1: Decision Opacity. This engine uses 20 seeds combining two scenarios. In the *metamorphic* scenario, two equivalent cases are independently submitted and compared using factor overlap (Jaccard) and structural similarity (sections, sentences, lists), with a composite threshold of 0.60, a structural counterpart to the 0.80 semantic threshold of R-F1 that accommodates the larger structural variation across independent generations [8]. In the *explanation* scenario, a decision is followed by an explanation request at three depth levels (*basic*, *detailed*,

challenge), evaluated through provision (substantive explanation, ≥ 0.50) and accessibility (jargon density, sentence length, simplification markers, ≥ 0.50), following the same majority-evidence rubric [2].

R-T2: Hidden Biases. This engine is structurally similar to R-F1, but perturbs attributes that should be irrelevant (hobbies, food, music, pets, transportation, weekend activities) across 18 seeds in credit, hiring, academic, insurance, and service scenarios, a property close to counterfactual fairness with respect to irrelevant variables [18]. FAIL is triggered by any of three conditions: decision change (via lexical patterns), semantic similarity below 0.75, or polarity delta above 0.35. Both thresholds are slightly more permissive than in R-F1 to absorb minor lexical variation from irrelevant perturbations, while remaining aligned with Counterfactual Cosine Similarity ranges [5] and the SBERT-calibrated polarity baseline (Section 4.2).

4 Empirical Evaluation

We evaluate the proposed framework on three commercial LLMs: GPT (`gpt-5.2`), Gemini (`gemini-3-flash`), and DeepSeek (`deepseek-chat`), addressing three research questions: *RQ1: How frequently do commercial LLMs violate operationalized FAT-derived requirements under ethical fuzzing? RQ2: How are ethical failure rates distributed across providers and FAT principles? RQ3: Do observed failure patterns indicate structural, principle-specific limitations or incidental provider-specific behavior?*

4.1 Experimental Setup

The campaign comprises 8,880 evaluated cases and 14,160 API calls distributed across the six engines. Each case corresponds to one oracle judgment: a counterfactual pair (R-F1, R-T2), a profile pair (R-F2), a multi-turn sequence (R-A2, R-T1), or a subgroup row (R-F4). R-F4 yields 3,600 cases with one API call each; the remaining engines require two calls per pair. Confidence intervals were estimated with the Wilson method [34] using $z = 1.96$ (95% confidence level).

4.2 Calibration of the Primary Similarity Metric

Responses for semantically equivalent counterfactual pairs frequently reach thousands of characters (medians of 3,254 in R-F1 and 2,846 in R-T2). For such long outputs, TF-cosine similarity systematically falls below 0.80 even without differential treatment, reflecting the cosine distortion known for long documents [30]. Sentence embeddings mitigate this effect through dense semantic representations [28]: TF-cosine medians are 0.6198 (R-F1) and 0.6100 (R-T2), whereas SBERT reaches 0.8039 and 0.8305, respectively. We therefore adopt SBERT as the primary similarity metric and retain TF-cosine for cross-audit, preserving the oracle's discriminative capacity.

4.3 Aggregate Results

Table 4 reports failure rates per engine and provider; Wilson 95% confidence intervals are reported throughout this section. For *RQ1*, the overall failure rate is 32.69% [31.72; 33.67], with module-specific rates ranging from 12.28% to 79.17%. Aggregated by FAT principle, *Transparency* has the highest rate at 39.87% [37.88; 41.89], followed by *Accountability* at 30.79% [28.31; 33.40] and *Fairness* at 30.07% [28.86; 31.32]. By module, R-F1 has the highest failure rate (79.17% [76.29; 81.78]), followed by R-T2 (70.00% [67.20; 72.66]), R-F2 (55.44% [52.18; 58.66]), R-A2 (30.79% [28.31; 33.40]), R-T1 (12.75% [10.98; 14.76]), and R-F4 (12.28% [11.25; 13.39]).

Table 4. Failure rate per engine and provider (%). Wilson 95% confidence intervals are reported in the text. The last column shows the engine total.

Engine	DeepSeek	Gemini	GPT	Total
R-F1	79.29	81.07	77.14	**79.17**
R-F2	72.00	40.33	54.00	**55.44**
R-F4	12.58	15.92	8.33	**12.28**
R-A2	32.86	27.38	32.14	**30.79**
R-T1	14.25	11.25	12.75	**12.75**
R-T2	70.56	74.72	64.72	**70.00**
Global	**35.07**	**32.70**	**30.30**	**32.69**

For *RQ2*, provider-level differences are relatively small (GPT 30.30%, Gemini 32.70%, DeepSeek 35.07%), although variability across modules indicates strong context dependence. R-F2 spans more than 30 percentage points (Gemini 40.33% versus DeepSeek 72.00%), whereas R-F1 converges around 79% across providers (77.14–81.07%), suggesting a shared limitation in handling counterfactual prompts involving protected attributes (*RQ3*). R-T2 is also consistently high across all providers (64.72–74.72%).

The global R-T1 rate of 12.75% masks a strong internal asymmetry. In the *expl* scenario (direct explanation of a decision), failure rates range from 22.50% (Gemini) to 27.00% (DeepSeek); in the *meta* scenario (equivalent cases evaluated independently), rates drop to 0.00–1.50%. Section 5 discusses the implications.

4.4 Failure Categorization

The most frequent failure categories are `decision_changed` ($n = 478$, 16.5%, R-T2), `sim_below_threshold` ($n = 400$, 13.8%, R-F1), `enc_delta_exceeded` ($n = 276$, 9.5%, R-F2), `sent_delta_exceeded` ($n = 265$, 9.1%, R-F1), `four_fifths_violated` ($n = 237$, 8.2%, R-F4), and `decision_reversal` ($n = 231$, 8.0%, R-A2). These categories capture decision changes under irrelevant perturbations, semantic divergence, profile-based asymmetries, subgroup accuracy violations, and adversarially induced reversals.

The predominance of *structural* causes, namely decision changes and similarity violations, over *marginal* causes, such as polarity variations near the threshold, suggests that observed failures are not artefacts of threshold calibration

but reflect consistent behavioral patterns. Combined with cross-provider convergence in R-F1 (77.14–81.07%) and R-T2 (64.72–74.72%), this indicates that the highest-rate failures are structural and principle-specific rather than incidental, an issue revisited in the threats-to-validity analysis (Section 6).

5 Discussion

From Performance-Centric to Ethics-Aware Evaluation. The reported failure rates show that performance-oriented evaluation is insufficient to capture ethical dimensions in LLMs. Models produced failures even when no explicit task error was present, especially under counterfactual perturbations of protected attributes and variations of attributes that should be irrelevant. The cross-provider convergence in R-F1 and R-T2 suggests that these limitations are not provider-specific, but reflect broader patterns in how current LLMs process ethically sensitive variations.

Operationalizing Ethical NFRs. From a Requirements Engineering perspective, the results support treating ethical properties as first-class non-functional requirements validated through evidence-driven procedures [32, 26]. By translating FAT principles into measurable thresholds grounded in regulatory standards (EEOC 4/5 rule) and empirical literature (Counterfactual Cosine Similarity ranges, inter-run variability), the framework makes ethical validation auditable. Requirements are elicited as testable properties, validated through fuzzing campaigns before deployment, and monitored through oracle reapplication over persisted outputs, without relying on non-deterministic LLM-based or human judges.

Direct versus Meta-Explanation. The R-T1 asymmetry reported in Section 4.3 highlights a central finding: models produce coherent abstract narratives describing how decisions *would* be made, but fail more often when explaining decisions actually made in concrete cases. This creates an *explanation illusion*, in which structured meta-explanations suggest transparency that does not extend to the operational level where it is required. This finding is especially relevant because regulatory frameworks such as the EU AI Act [10] require transparency at the level of individual decisions, where the evaluated models failed more often. Therefore, verifying transparency requirements cannot rely only on abstract explanatory text; it must include controlled evaluation of decision-specific explanations, as operationalized by R-T1.

6 Threats to Validity

We organize threats to validity into three dimensions commonly used in empirical Software Engineering [35]: construct, external, and conclusion validity.

Construct Validity. The R-F1 failure rate (79.17%) can be interpreted in two ways: substantial differential behavior under counterfactual prompts involving protected attributes, or excessive sensitivity of the 0.80 threshold, derived from the EEOC four-fifths rule, to the natural variability of generative outputs.

The distribution of failure causes helps distinguish these interpretations: `sim_below_threshold` accounts for 60.2% of R-F1 failures and is SBERT-based, making it less sensitive to surface variation. This supports the former interpretation, although it does not prove it conclusively. Thresholds were defined as explicit methodological choices grounded in technical and regulatory literature; more definitive calibration would require longitudinal studies or expert evaluation, both beyond the scope of this work. The internally developed sentiment lexicons (R-F1, R-F2, R-T2) also depend on manual curation, a limitation mitigated through full documentation in the source code.

External Validity. Three factors limit generalization. First, the campaign was conducted in Brazilian Portuguese using seeds and lexicons grounded in the Brazilian context (SUS, INSS, Pronampe, LGPD), which may require adaptation to other environments. Second, the evaluated providers (GPT, Gemini, and DeepSeek) represent high-capability commercial systems; smaller or open-source models may show different failure patterns. Third, hosted LLMs exhibit non-determinism even under nominally deterministic configurations [1], so exact rates may vary across re-runs. This limitation is partially mitigated by separating Campaign and Evaluation Phases, which enables oracle reapplication over persisted outputs without additional API calls.

Conclusion Validity. Failure rates were estimated with the Wilson method using $z = 1.96$ (95% confidence intervals), limiting claims driven by sample noise. Deterministic *a priori* oracles eliminate variance from evaluators and support reproducibility. A remaining limitation is the absence of human-in-the-loop validation of oracle verdicts, identified as future work. This does not compromise the internal consistency of the reported comparisons, since all providers were evaluated under identical seeds, perturbations, oracles, and thresholds.

7 Conclusion

This paper presented a fuzzing-based framework for automated ethical risk detection in Large Language Models, formalizing *ethical fuzzing* through multi-dimensional coverage, deterministic *a priori* oracles, and controlled reproducibility. Six FAT-derived risks were operationalized as testing modules within a two-phase architecture that separates data collection from oracle application and treats raw outputs as immutable evidence. The empirical evaluation across three commercial LLM providers (8,880 cases, 14,160 API calls) showed that ethical failures are frequent and recurrent across providers, especially in counterfactual fairness (R-F1) and sensitivity to irrelevant attributes (R-T2). It also exposed an *explanation illusion* in R-T1, with models explaining hypothetical decisions more reliably than concrete ones. Beyond these findings, the framework offers Requirements Engineering a mechanism for translating abstract ethical principles into measurable acceptance criteria and treating ethical validation as an auditable activity across the lifecycle of AI-based systems.

Future work includes extending coverage with additional ethical-risk engines, evaluating smaller-scale and open-source models, adapting the framework to

other languages and institutional contexts, and refining thresholds and lexicons through expert-driven validation. We also plan to integrate the framework with CI/CD pipelines and align it with governance artefacts such as model cards [20] and datasheets [14], supporting regulatory demands such as the EU AI Act [10] for systematic ethical-compliance verification. All source code, configuration artefacts, and oracle implementations are available at `https://github.com/VasconcelosJoao/ethical-fuzzing-llms`.

Acknowledgements

We thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Grant Nº 300883/2025-0 and 406266/2025-5.

References

1. Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R.J., et al.: Non-determinism of "deterministic" LLM system settings in hosted environments. In: Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems. pp. 135–148. Association for Computational Linguistics (Dec 2025). https://doi.org/10.18653/v1/2025.eval4nlp-1.12
2. Barr, E.T., Harman, M., McMinn, P., Shahbaz, M., Yoo, S.: The oracle problem in software testing: A survey. IEEE Transactions on Software Engineering **41**(5), 507–525 (2015). https://doi.org/10.1109/TSE.2014.2372785
3. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., et al.: Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development **63**, 4:1–4:15 (2019). https://doi.org/10.1147/JRD.2019.2942287
4. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., et al.: On the opportunities and risks of foundation models (2022), `https://arxiv.org/abs/2108.07258`
5. Bouchard, D.: An actionable framework for assessing bias and fairness in large language model use cases (2026), `https://arxiv.org/abs/2407.10853v4`, v4
6. Bouchard, D., Chauhan, M.S., Skarbrevik, D., Bajaj, V., Ahmad, Z.: Langfair: A python package for assessing bias and fairness in large language model use cases. Journal of Open Source Software **10**(105), 7570 (2025). https://doi.org/10.21105/joss.07570
7. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 77–91. PMLR (2018)
8. Chen, T.Y., Kuo, F.C., Liu, H., Poon, P.L., Towey, D., et al.: Metamorphic testing: A review of challenges and opportunities. ACM Computing Surveys **51**(1), 1–27 (2018). https://doi.org/10.1145/3143561
9. Commission, E.E.O.: Uniform guidelines on employee selection procedures (1978), `https://www.ecfr.gov/current/title-29/subtitle-B/chapter-XIV/part-1607`

10. European Union: Artificial intelligence act: Regulation (EU) 2024/1689 of the european parliament and of the council. Official Journal of the European Union (Jul 2024), https://eur-lex.europa.eu/eli/reg/2024/1689
11. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 259–268. KDD '15, Association for Computing Machinery (2015). https://doi.org/10.1145/2783258.2783311
12. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., et al.: Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. Minds and Machines **28**(4), 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
13. Galhotra, S., Brun, Y., Meliou, A.: Fairness testing: testing software for discrimination. In: Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering. pp. 498–510. ESEC/FSE'17, ACM (2017). https://doi.org/10.1145/3106237.3106277
14. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., et al.: Datasheets for datasets. Communications of the ACM **64**(12), 86–92 (2021). https://doi.org/10.1145/3458723
15. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M.: Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. pp. 79–90. AISec '23, Association for Computing Machinery (2023). https://doi.org/10.1145/3605764.3623985
16. Huang, C., Zhang, Z., Mao, B., Yao, X.: An overview of artificial intelligence ethics. IEEE Transactions on Artificial Intelligence **4**(4), 799–819 (2023). https://doi.org/10.1109/TAI.2022.3194503
17. Jobin, A., Ienca, M., Vayena, E.: The global landscape of ai ethics guidelines. Nature Machine Intelligence **1**(9), 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
18. Kusner, M., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 4069–4079. NIPS'17, Curran Associates Inc. (2017)
19. Makridis, C.A., Mueller, J., Tiffany, T., Borkowski, A.A., Zachary, J., Alterovitz, G.: From theory to practice: Harmonizing taxonomies of trustworthy ai. Health Policy OPEN **7**, 100128 (2024). https://doi.org/10.1016/j.hpopen.2024.100128
20. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., et al.: Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 220–229. ACM (2019). https://doi.org/10.1145/3287560.3287596
21. Monjezi, V., Kumar, A., Tan, G., Trivedi, A., Tizpaz-Niari, S.: Causal graph fuzzing for fair ml software development. In: Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings. pp. 402–403. ICSE-Companion '24, ACM (2024). https://doi.org/10.1145/3639478.3643530
22. Morales, S., Clarisó, R., Cabot, J.: Langbite: An open-source platform to automate bias testing of large language models. SoftwareX **31**, 102248 (2025). https://doi.org/10.1016/j.softx.2025.102248
23. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019). https://doi.org/10.1126/science.aax2342

24. Pei, K., Cao, Y., Yang, J., Jana, S.: Deepxplore: Automated whitebox testing of deep learning systems. Commun. ACM **62**(11), 137–145 (2019). https://doi.org/10.1145/3361566
25. Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., et al.: Red teaming language models with language models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3419–3448. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.emnlp-main.225
26. Porto, D., Prado, R., Marques, G., Serrano, A., Mendonça, F., Canedo, E.: Ethical requirements in the age of artificial intelligence: A systematic literature review. In: Anais do XXI Simpósio Brasileiro de Sistemas de Informação. pp. 663–672. SBC (2025). https://doi.org/10.5753/sbsi.2025.246613
27. Reddy, H., Srinivasan, M., Kanewala, U.: Metamorphic testing for fairness evaluation in large language models: Identifying intersectional bias in LLaMA and GPT. In: 2025 IEEE/ACIS 23rd International Conference on Software Engineering Research, Management and Applications (SERA). pp. 239–246. IEEE Computer Society (May 2025). https://doi.org/10.1109/SERA65747.2025.11154488
28. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. pp. 3980–3990. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1410
29. Salimian, S., Uddin, G., Biswas, S., Leung, H.: Bias testing and mitigation in black box llms using metamorphic relations (2025), `https://arxiv.org/abs/2512.00556`
30. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 21–29. ACM (1996). https://doi.org/10.1145/243199.243206
31. Udeshi, S., Arora, P., Chattopadhyay, S.: Automated directed fairness testing. In: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering. pp. 98–108. ACM (2018). https://doi.org/10.1145/3238147.3238165
32. Vicenzi, A.A.F.C., de Cerqueira, J.S., Abrahamsson, P., Canedo, E.D.: Specifying fairness and transparency requirements for public benefit allocation. In: Requirements Engineering: Foundation for Software Quality: 32nd International Working Conference, REFSQ 2026, Poznań, Poland, March 23-26, 2026, Proceedings. pp. 245–253. Springer-Verlag (2026). https://doi.org/10.1007/978-3-032-21423-2_17
33. Wachter, S., Mittelstadt, B., Floridi, L.: Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law **7**(2), 76–99 (05 2017). https://doi.org/10.1093/idpl/ipx005
34. Wilson, E.B.: Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association **22**(158), 209–212 (1927)
35. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in Software Engineering, Second Edition. Springer (2024). https://doi.org/10.1007/978-3-662-69306-3