

MACE-XAI: A Multi-Agent Crowd Evaluation Framework for Robust Natural Language Explanations

Renato Okabayashi Miyaji

Escola Politécnica, Universidade de São Paulo
São Paulo, Brazil
re.miyaji@usp.br

Abstract

The rapid integration of Machine Learning in critical decision-making has shifted the focus of Explainable Artificial Intelligence (XAI) toward human-centric Natural Language Explanations (NLEs). While Large Language Models (LLMs) can synthesize narrative justifications from technical feature attributions, ensuring the faithfulness and robustness of these explanations remains an open challenge. Although multi-agent ensembles and Crowd-based Comparative Evaluation (CCE) are emerging standards for evaluating open-ended dialogues, they fail to capture domain-specific hallucinations and structural omissions inherent to technical explanations. To bridge this gap, we introduce MACE-XAI, a novel neuro-symbolic framework engineered specifically for the robust assessment of NLEs in the context of tabular feature-attribution XAI. Rather than proposing isolated evaluators, our key innovation lies in the architectural synergy of its components: we orchestrate a heterogeneous crowd of LLM agents to evaluate pragmatic text quality and fluency, while anchoring them with a deterministic symbolic verification layer that audits textual claims against raw technical ground truths. Experimental results demonstrate that while standalone LLM-as-judge setups are easily swayed by linguistic surface features, MACE-XAI achieves an agreement with human experts (Cohen’s $\kappa = 0.64$). While this represents a significant improvement over standalone neural judges, it also highlights the remaining alignment gap that must be addressed before fully autonomous evaluation in safety-critical deployments.

1 Introduction

The rapid evolution of Machine Learning (ML) has made the demand for transparency a cornerstone of responsible AI deployment (Paraschou et al., 2025). The problem of model explainability is widely discussed in the literature, particularly within the field

of Explainable Artificial Intelligence (XAI). Significant progress has been made through the development of technical feature attribution methods such as SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017), and LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016). However, while these techniques provide mathematical rigor for experts to interpret model outputs quantitatively, they often remain opaque to non-expert stakeholders (Martens et al., 2025).

Despite the effectiveness of numerical explainability, there is a persistent gap between mathematical evidence and human cognitive needs (Kroeger et al., 2023). Since the majority of end-users are not experts in ML or XAI, current technical methods lack the narrative structure required for effective communication in critical domains such as healthcare and finance (Martens et al., 2025). To bridge this barrier, recent literature proposes the use of language models (LMs) to synthesize Natural Language Explanations (NLEs) from XAI outputs (Burton et al., 2023). This transition toward "XAI stories" leverages the generative capabilities of Large Language Models (LLMs) to align AI reasoning with how humans naturally justify decisions (Martens et al., 2025).

While the application of LLMs for NLE generation via In-Context Learning (ICL) has been extensively explored (Kroeger et al., 2023), significant challenges remain regarding the evaluation of explanation quality. As highlighted by (Zytek et al., 2024), current evaluation paradigms often struggle with the faithfulness gap, where models generate linguistically fluent but factually unfaithful narratives. This introduces the risk of hallucinations, where explanations appear plausible but lack grounding in the underlying technical ground truth. Furthermore, existing metrics fail to account for the functional soundness and completeness necessary for safety-critical deployments (Zytek et al., 2024).

Research Statement and Contributions. Current LLM-as-judge paradigms, while effective for general open-ended dialogue, consistently struggle in the domain of Explainable AI due to the strict mathematical faithfulness required in feature attributions. To address these limitations, we propose the Multi-Agent Crowd Evaluation for XAI (MACE-XAI) framework. Rather than merely assembling existing evaluation tools, MACE-XAI introduces a novel neuro-symbolic architecture engineered specifically for the XAI domain. We synergize a heterogeneous ensemble of language models—adapting the Crowd-based Comparative Evaluation (CCE) (Zhang et al., 2025) to assess pragmatic quality and semantic intent—with a deterministic symbolic verification layer acting as a strict technical guardian. This architectural orchestration ensures that explanations are mathematically audited against raw attribution data, capturing latent technical omissions and hallucinations that purely neural evaluators inherently overlook.

In this work, we investigate two central research hypotheses: first, *that generic LLM evaluators are biased toward linguistic fluency over factual fidelity in XAI, requiring a domain-adapted ensemble to assess narrative quality*; and second, *that coupling this neural crowd with a deterministic Symbolic Evaluation Judge creates a neuro-symbolic synergy, successfully mitigating hallucinations and omissions*.

The main contributions of this work are threefold. First, we introduce MACE-XAI, a novel neuro-symbolic evaluation architecture that bridges the gap between semantic fluency and technical faithfulness by orchestrating specialized LLM agents under strict deterministic constraints. Second, we operationalize a domain-specific Symbolic Fidelity Score that integrates with crowd-based evaluation to penalize factual polarity errors and the exclusion of critical attributes. Finally, through empirical validation on the TEXEN dataset (Burton et al., 2023), we demonstrate the necessity of this architectural synergy, showing that our framework narrows the alignment gap between automated metrics and expert human judgment.

It is important to note that MACE-XAI is designed and validated specifically within the paradigm of tabular feature-attribution XAI. While the framework’s modular architecture allows for theoretical extensibility, its current empirical claims are strictly bounded to tabular classification tasks where explanations take the form of feature

importance scores.

2 Related Work

Based on foundational feature attribution methods (Lundberg and Lee, 2017; Ribeiro et al., 2016), research has increasingly prioritized the evolution of human-centered communication over local interpretability. While these quantitative frameworks established a necessary mathematical groundwork, subsequent studies have consistently identified a persistent gap where raw importance scores fail to satisfy the cognitive requirements of non-expert stakeholders (Martens et al., 2025; Kroeger et al., 2023). This limitation catalyzed a broader transition toward linguistic grounding in XAI, a shift underscored by the emergence of benchmarks like the TEXEN dataset (Burton et al., 2023), which formalizes the requirement for narrative structure in high-stakes decision-making domains.

Within this landscape, recent frameworks advocate for NLEs that transcend mere parameter descriptions to embody communicative intent through contrastive and selective insights (Yu et al., 2022). Although LLMs facilitate the democratization of AI by translating algorithmic complexity into fluent prose (Paraschou et al., 2025; Biemans et al., 2025), their deployment remains hampered by inherent faithfulness gaps and the stochastic risk of hallucinations (Chang et al., 2023). Consequently, the research focus has moved beyond exploring generative feasibility toward the development of robust evaluation frameworks capable of rigorously verifying narrative grounding against the underlying technical ground truth (Zytek et al., 2024).

State-of-the-art approaches have recently begun to address these evaluative challenges through automated proxies. Techniques such as Crowd Comparative Reasoning (Zhang et al., 2025) and specialized judges like Prometheus 2 (Kim et al., 2024) aim to replicate human-aligned qualitative metrics at scale. However, these developments suggest that for an explanation to be truly human-centric, it must satisfy not only linguistic fluency but also functional soundness and completeness (Miyaji and Correa, 2025). As LLMs become the primary engines for both generating and auditing these explanations, ensuring their robustness becomes a vital frontier for maintaining safety-critical reliability (Zytek et al., 2024).

The synthesis of NLEs represents a broader shift toward Responsible AI, where model outputs are

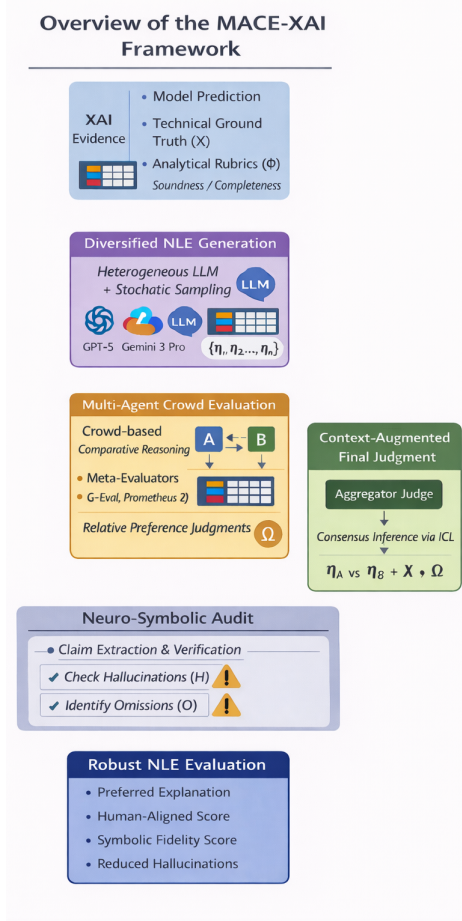


Figure 1: Overview of the MACE-XAI framework architecture.

dynamically tailored to the user’s expertise. By integrating a neuro-symbolic layer with neural generation, it is possible to achieve a high-fidelity interpretability framework that bridges the divide between mathematical evidence and human-readable narrative.

3 Methodology

3.1 Proposed Method

To address the fidelity-plausibility trade-off of NLEs, we propose the Multi-Agent Crowd Evaluation for XAI (MACE-XAI) framework, presented in Figure 1. This methodology shifts the NLE evaluation paradigm from a single-judge direct scoring system (Zytek et al., 2024) to a robust architecture. The core innovation of MACE-XAI lies in the utilization of a "crowd" of diverse language models (Zhang et al., 2025) to expose latent weaknesses in generated explanations that an isolated evaluator might overlook.

The evaluative core of our framework is grounded in the paradigm of automated judicial

arbitration (Chang et al., 2023). Rather than assigning isolated scores, we operationalize the assessment as a relative preference task between competing narratives. Formally, consider a pair of candidate NLEs, denoted as $\mathcal{E} = \{\eta_i, \eta_j\}$, both synthesized from a common context $\mathcal{X}$ representing the local feature importance derived from an XAI extractor. We define a meta-evaluator a , which is tasked with generating a qualitative rationale ω by scrutinizing the linguistic outputs against a set of predefined analytical rubrics Φ presented in Table 1. This judicial mapping is expressed in Equation 1.

$$\omega = a(\eta_i, \eta_j \mid \mathcal{X}, \Phi) \quad (1)$$

The goal of this formulation is to maximize the objective alignment between the machine-generated preference ω and the established expert consensus. To validate this framework, we measure the congruence of $\mathcal{A}$ by computing the pairwise accuracy of its decisions in relation to the human gold-standard annotations. This metric serves as a proxy for the system’s reliability in mirroring human-centric interpretability standards.

3.1.1 Diversified Crowd Generation

The first stage of the framework involves the creation of a pool of NLEs ($\mathcal{E}$) for each $\mathcal{X}$, with $\mathcal{E} = \{\eta_1, \eta_2, \dots, \eta_n\}$. Rather than relying on a single model architecture, we employ a heterogeneous ensemble of LLMs. This model diversity is critical for capturing different stylistic and structural approaches (Zhang et al., 2025) to explanation synthesis. Furthermore, we introduce stochastic variability through temperature scaling; for each model in the ensemble, explanations are generated at both a low temperature for precision and a high temperature for creative diversity.

3.1.2 Crowd-based Comparative Evaluation (CCE)

Departing from traditional absolute scoring methods, MACE-XAI utilizes the Crowd-based Comparative Evaluation (CCE) paradigm proposed by (Zhang et al., 2025). In this stage, to produce granular judicial assessments, we employ a group of meta-evaluators. For each individual narrative $\eta_i \in \mathcal{E}$ and each meta-evaluator $a_j \in \mathcal{A}$, a crowd judgment ω_{i,a_j} is generated by comparing η_i against the set of all other candidate narratives within the pool, as denoted in Equation 2.

Criterion	Definition	Symbolic Operationalization
Soundness	The NLE accurately reflects the feature polarities (positive/negative influence) present in $\mathcal{X}$ and contains no hallucinated features, fabricated class predictions, or false causal relationships absent from the raw XAI evidence.	Penalizes Hallucinations (H): any claim $c \in \mathcal{C}$ where the stated polarity contradicts $\mathcal{X}$.
Completeness	The NLE verbalizes the critical features identified in $\mathcal{X}$ —particularly the most heavily weighted attributes—without significant omissions that could mislead the end-user about the model’s decision logic.	Penalizes Omissions (O): any feature $f \in \mathcal{F}_{gt}$ absent from $\mathcal{F}_{claims}$.

Table 1: Analytical Rubrics (Φ) mapping the qualitative text evaluation to the deterministic symbolic audit.

$$\omega_{i,a_j} = a_j(\eta_i, \mathcal{E} \setminus \{\eta_i\} \mid \mathcal{X}, \Phi) \quad (2)$$

Formally, we collect a comprehensive set of judgments $\Omega = \{\omega_{i,a_j} \mid \eta_i \in \mathcal{E}, a_j \in \mathcal{A}\}$. While each individual judgment may not fully capture every technical detail of the candidate responses, they collectively provide a richer pool of evidence regarding how narratives differ in nuanced ways. This comparative reasoning forces the evaluator to identify subtle technical omissions and hallucinations that might appear plausible in isolation but become evident when contrasted with competing versions. Our implementation of CCE diverges from its original formulation by (Zhang et al., 2025), which was designed to benchmark open-ended conversational dialogues. While standard NLP evaluations focus on surface-level stylistic dimensions, MACE-XAI introduces a domain-specific adaptation engineered for XAI.

To mitigate vulnerabilities prevalent in LLM-as-a-judge frameworks, our evaluation protocol implements bias controls. First, to prevent identity-based coupling, all candidate narratives underwent identity masking. Second, to neutralize position bias, the presentation order of the narratives was systematically randomized across all pairwise comparisons. Finally, the framework inherently enforces cross-family judging, as the open-weight alignment judge evaluates proprietary models, ensuring that stylistic preferences are decoupled from the generator’s architectural family.

3.1.3 Specialized Judging Agents

The judicial process is conducted by two specialized agents with complementary roles. G-Eval (Liu et al., 2023) functions as the high-level semantic judge, focusing on complex reasoning and adherence to fidelity rubrics through Chain-of-Thought (CoT) prompting. Simultaneously, Prometheus

2 (Kim et al., 2024) acts as an alignment judge, specifically fine-tuned to follow rigorous human-like evaluation rubrics. By synergizing these distinct evaluative lenses, MACE-XAI achieves a balance between linguistic soundness and structural fidelity, mitigating the risk of plausible but incomplete explanations.

3.1.4 Context-augmented Judgment

The evaluative decision is derived through a stage of Context-augmented Inference. When comparing a pair of candidate explanations, η_A and η_B , the primary meta-evaluator a does not operate in isolation. Instead, it is conditioned on the input context $\mathcal{X}$, the analytical rubrics Φ , and the set of post-processed crowd judgments Ω generated in the previous phase. These prior judgments are integrated via In-Context Learning (ICL) into the model’s prompt, serving as multi-perspective evidence that guides the final reasoning process. Formally, the final judgment ω^* is defined in Equation 3.

$$\omega^* = a(\eta_A, \eta_B \mid \mathcal{X}, \Phi, \Omega) \quad (3)$$

By leveraging Ω as a rich pool of evidence, the meta-evaluator can synthesize a consensus that transcends the biases of any single model architecture. This mechanism ensures that the final preference is grounded in the collective judgments, leading to higher-quality CoT rationales.

3.1.5 Symbolic Verification and Hallucination Mitigation

In formalizing our verification layer, we draw a distinction between model-faithfulness (whether the XAI tool accurately reflects the underlying ML model’s internal logic) and tool-faithfulness (whether the NLE accurately verbalizes the numerical outputs of the XAI tool). MACE-XAI is explic-

itly designed to evaluate the latter. Given that feature attribution methods can be unreliable proxies for model behavior, our symbolic layer guarantees that the natural language layer does not introduce additional communicative distortions (hallucinations or omissions) on top of the provided XAI evidence ($\mathcal{X}$).

To ensure rigorous quality control and technical fidelity, MACE-XAI implements a secondary neuro-symbolic layer. NLEs are often susceptible to hallucinations—such as factual errors regarding class predictions, feature names, or the direction of feature influence (polarities)—as well as the omission of critical attributes (Zytek et al., 2024). To mitigate these risks, a specialized Symbolic Evaluation Judge is deployed to perform a structured audit of each narrative.

The process begins by extracting the set of textual claims $\mathcal{C}$ from the narrative η . To guarantee determinism and avoid the probabilistic variance, our generation pipeline enforces strict formatting constraints on the LLMs during inference. A deterministic parsing layer leveraging Regular Expressions (Regex) is then applied directly to the structured output. This Regex-based algorithm extracts the feature-polarity claims into a standardized format, such as $\{\text{"feature": "age", "polarity": "negative"}\}$, ensuring a robust pipeline for the subsequent symbolic verification.

These parsed claims are subsequently and cross-referenced with the raw feature attribution data $\mathcal{X}$. Because the extraction relies on deterministic rules, it allows the framework to verify whether the linguistic assertions are supported by the underlying numerical evidence. The result of this cross-referencing is the generation of a Symbolic Verification Score ($\mathcal{S}_{sym}$) for each explanation, providing a formal measure of factual consistency. $\mathcal{S}_{sym}$ is calculated as the ratio of valid technical assertions to the total number of relevant features, as presented in Equation 4.

$$\mathcal{S}_{sym} = \frac{|\mathcal{F}_{total}| - |\mathcal{H}| - |\mathcal{O}|}{|\mathcal{F}_{total}|} \quad (4)$$

In Equation 4, let $\mathcal{F}_{gt}$ denote the set of ground truth features present in the numerical attribution $\mathcal{X}$, and let $\mathcal{F}_{claims}$ represent the set of features extracted from the textual claims $\mathcal{C}$. We define the total relevant feature space as the union of these two sets: $\mathcal{F}_{total} = \mathcal{F}_{gt} \cup \mathcal{F}_{claims}$. The verification process identifies two primary error types. The first one are Hallucinations ($\mathcal{H}$), which are cases where

a feature $f \in \mathcal{F}_{claims}$ is mentioned in the narrative but is assigned an incorrect sign or polarity relative to $\mathcal{X}$. The second one are Omissions ($\mathcal{O}$), when a feature $f \in \mathcal{F}_{gt}$ is present in the ground truth but is entirely absent from the textual explanation ($f \notin \mathcal{F}_{claims}$).

Although the formulation of the Symbolic Verification Score relies on a precision-recall mechanism, this is a deliberate architectural choice. A core tenet of Explainable AI is that the evaluation framework itself must be transparent, deterministic, and auditable. While the heterogeneous LLM agents in our crowd evaluate the pragmatics and linguistic fluency of the explanations, the symbolic layer is designed to act as a deterministic anchor. This dual-layer approach ensures that the final outputs are not only linguistically persuasive but also mathematically sound and complete.

A critical assumption of MACE-XAI is that the symbolic layer acts as a reliable deterministic anchor. To validate the regex-based claim extractor, we conducted a manual audit of 50 randomly sampled instances. Two independent experts manually annotated all feature-polarity claims (inter-annotator agreement $\kappa = 0.95$) to establish a gold-standard extraction set. Against this gold standard, the automated regex extractor achieved a precision of 98.9% and a recall of 97.1%. The minor recall shortfall (2.9%) was heavily concentrated in implicit polarity statements involving hedging language. When recomputing $\mathcal{S}_{sym}$ with the gold-standard manual extractions, the mean absolute difference was negligible (0.018 ± 0.012 , Spearman’s $\rho = 0.97$). Most importantly, in no case did an extraction error alter the ordinal ranking between a pair of competing NLEs. This confirms that the deterministic extraction is highly reliable in practice, contingent on the generation prompt enforcing canonical feature naming conventions.

3.2 Experiments

3.2.1 Dataset

To evaluate the proposed MACE-XAI framework, we utilize the TEXEN (Textual Explanations for Machine Learning) dataset (Burton et al., 2023), a human-curated benchmark for structured data-to-text generation in XAI. TEXEN comprises 469 expert-validated instances in which local numerical explanation outputs — consisting of class probabilities and feature importance scores produced by post-hoc XAI techniques such as LIME, SHAP,

Integrated Gradients, and Layer-wise Relevance Propagation — are paired with expert-written natural language narratives that verbalize why a classifier made a particular decision.

The numerical explanation outputs in TEXEN (Burton et al., 2023) originate from multiple binary classification tasks. These predictions are produced by commonly used machine learning classifiers, including Logistic Regression, Support Vector Machines, Random Forests, and Deep Neural Networks, whose decisions are subsequently explained using post-hoc XAI methods (Burton et al., 2023). This dataset was selected for our case study because it provides expert-authored ground truth (GT) explanations that directly verbalize feature-attribution outputs, enabling precise evaluation of explanation faithfulness. Moreover, its use of heterogeneous tasks, standard classifiers, and multiple XAI techniques makes TEXEN well-suited for assessing the robustness of the proposed method.

3.2.2 Evaluation

To ensure a robust validation of the MACE-XAI framework, we employ a multi-dimensional evaluation strategy that encompasses linguistic fluency, semantic fidelity, and functional accuracy. This approach allows us to triangulate the performance of our multi-agent system by comparing its outputs against the expert-authored TEXEN ground truth. The evaluation is structured into three pillars: automatic linguistic metrics, human-aligned comparative judgment, and symbolic error analysis.

The quality of the generated NLEs is first assessed using standard automatic metrics, specifically METEOR, BLEU, and BLEURT. We employ BLEU and METEOR to measure the surface-level similarity between GTs and NLEs. These metrics focus on overlap and linguistic precision, providing a baseline for the model’s fluency. However, to capture the deeper logic of the explanation, we prioritize BLEURT, a semantic equivalence-based metric. BLEURT indicates how effectively the machine-produced text communicates the underlying meaning and intent of the reference.

To ensure that the automated evaluation effectively mirrors expert human intuition, we quantify Human Alignment Preferences within the MACE-XAI comparative framework. In this setup, the judging agents perform pairwise comparisons over candidate explanations $\{\eta_i, \eta_j\}$, which are benchmarked against the GT. This evaluation is governed by two core qualitative criteria: Soundness, which

assesses the technical correctness of the information included in the narrative, and Completeness, which measures the sufficiency of the feature coverage relative to the original XAI evidence (Zytek et al., 2024). To validate the framework’s reliability, we measure the congruence between the preferences indicated by MACE-XAI and those of human annotators using Win Rates and Cohen’s Kappa.

Finally, to provide an assessment of technical integrity, we conduct a symbolic error analysis centered on the quantification of Hallucinations ($\mathcal{H}$) and Omissions ($\mathcal{O}$). Moving beyond general similarity, this stage utilizes $\mathcal{S}_{sym}$ to perform an audit of the data-to-text transformation. By cross-referencing the claims extracted from the NLEs with the TEXEN ground truth, we measure the framework’s capacity to accurately preserve class predictions, feature polarities, and relative importance rankings.

3.2.3 Ablation Studies

To provide a validation of the MACE-XAI framework, we conducted targeted ablation studies to isolate the contributions of multi-agent synergy and the symbolic verification layer.

A1: Synergy of Diverse Meta-Evaluators. We investigated the necessity of a heterogeneous "crowd" by comparing the performance of individual meta-evaluators—G-Eval (Liu et al., 2023) and Prometheus 2 (Kim et al., 2024)—against the combined MACE-XAI ensemble. Each configuration was benchmarked against human gold-standard annotations using Win Rate and Cohen’s Kappa.

A2: Impact of Symbolic Verification. To isolate the contribution of the neuro-symbolic bridge, we evaluated the framework in two scenarios: the MACE base architecture without symbolic grounding and the full framework including the Symbolic Verification layer. We measured the delta in Win Rate and Human Alignment to assess how the verification of textual claims against raw data affects the final judgment.

3.2.4 Experimental Details

For the initial generation stage of the framework, we utilized a heterogeneous ensemble consisting of two LLMs: OpenAI GPT-4.1 and Gemini 3 Pro. To capture the trade-off between technical precision and linguistic diversity, each model synthesized a pool of $n = 5$ candidate explanations per data instance (X) for each of two distinct temperature settings: a low-variability configuration (0.2) and a

high-diversity configuration (0.8), yielding a comprehensive pool ($\mathcal{E}$) of 20 candidate NLEs per instance. In the evaluative phase, specialized model variants were assigned to specific judicial roles to optimize reasoning depth. To evaluate the framework across different scales, we benchmarked both proprietary and open-weights judges: we compared frontier models (OpenAI GPT-5.1 and GPT-5.2) against an open-weights baseline using Meta’s Llama-3.1-8B in identical judicial roles. Specifically, the assigned model (either the GPT variants or Llama-3.1-8B) was deployed for the G-Eval semantic assessment and the Context-augmented Final Judgment stage to evaluate capabilities in synthesizing multi-perspective evidence via In-Context Learning (ICL). The aggregator judge receives the complete set of crowd judgments (Ω) injected directly into a specifically structured meta-prompt to synthesize the final consensus.

To establish a robust ground truth for evaluating the MACE-XAI framework, we extracted a random sample of 110 instances from the TEXEN dataset, resulting in 550 pairwise NLE comparisons. These pairs were independently evaluated by four domain experts with extensive backgrounds in ML and XAI. To ensure consistency and minimize subjective variance, all annotators underwent a dedicated training phase centered on the identical analytical rubrics (Φ) provided to the LLM agents, followed by a formal calibration session to synchronize interpretative standards.

Prior to any harmonization, the raw expert-expert inter-annotator agreement yielded a Cohen’s Kappa of 0.82, indicating substantial initial consensus. To produce the final gold-standard labels, all diverging judgments were subjected to an adjudication process, where conflicts were resolved through open consensus discussion. This ensures the validity of the preference labels and serves as a benchmark for calculating Win Rates and our evaluation metrics.

4 Results and Discussions

4.1 Linguistic Metrics

The performance of the candidate LLMs—GPT-4.1 and Gemini 3 Pro—was first evaluated using standard automated linguistic metrics to establish a baseline for fluency and semantic alignment with the expert-authored ground truth (GT). As presented in Table 2, the two models exhibit distinct strengths depending on whether surface-level lexi-

cal overlap or deeper semantic equivalence is prioritized.

Table 2: Quantitative comparison of automated linguistic metrics for GPT-4.1 and Gemini 3 Pro narratives with the expert-authored ground truth (GT).

Model	BLEU	METEOR	BLEURT
GPT-4.1	1.608	0.297	-0.718
Gemini 3 Pro	1.324	0.290	-0.604

GPT-4.1 achieved higher scores in both BLEU and METEOR, indicating a superior level of surface-level similarity and linguistic precision relative to the reference explanations. These results suggest that the narratives generated by GPT-4.1 align more closely with the specific vocabulary and structural patterns utilized by human experts in the TEXEN dataset. In contrast, Gemini 3 Pro outperformed GPT-4.1 in terms of BLEURT. Since BLEURT is a semantic equivalence-based metric, this suggests that Gemini 3 Pro is more effective at communicating the underlying meaning and technical intent of the original XAI evidence, even when its linguistic choices diverge from the GT. This semantic robustness is critical for explainability, as it implies that Gemini 3 Pro may capture the logical essence of the decision-making process more accurately.

A notable observation in our results is the limited correlation between surface-level linguistic metrics and expert human preference. Although GPT-4.1 exhibited superior BLEU and METEOR scores, human evaluators overwhelmingly favored Gemini 3 Pro with a benchmark win rate of 81.8%. This sharp contrast indicates that while n-gram-based metrics capture lexical precision and mimicry of the GT’s style, they fail to account for the technical soundness and narrative logic that experts prioritize in XAI. Gemini’s higher BLEURT score, which focuses on semantic equivalence, aligns more closely with human intuition, yet it still fails to fully capture the nuances of expert judgment. This gap confirms that linguistic fluency is an insufficient proxy for evaluating natural language explanations.

4.2 Human Alignment and Win Rate Analysis

To evaluate the effectiveness of the MACE-XAI framework in replicating expert judgment, we compared the Win Rates of the candidate models and the inter-rater agreement (Cohen’s Kappa) between various automated judging configurations and human experts. Human experts exhibited a strong

preference for Gemini 3 Pro, yielding a benchmark Win Rate of 81.8%. This high preference serves as the "gold standard" for our comparative analysis.

Table 3: Comparison of Gemini 3 Pro Win Rates and Cohen’s Kappa relative to Human Expert. All metrics are reported with 95% Confidence Intervals (CI).

Configuration	Win Rate (95% CI)	Cohen’s Kappa ($\kappa \pm 95\%$ CI)
Human Expert (Baseline)	81.8% $\pm$ 2.1%	-
G-Eval	63.6% $\pm$ 3.2%	0.42 $\pm$ 0.02
Prometheus 2	72.7% $\pm$ 2.8%	0.56 $\pm$ 0.01
MACE-XAI (Base, w/o Symbolic)	66.5% $\pm$ 3.0%	0.58 $\pm$ 0.01
MACE-XAI (Full, w/ Symbolic)	75.1% $\pm$ 2.4%	0.64 $\pm$ 0.01

The results in Table 3 demonstrate that independent judging agents significantly underperform in mirroring human intuition. G-Eval, while capable of semantic reasoning, achieved the lowest alignment, likely due to its susceptibility to plausibility bias where fluent but factually incorrect narratives are unfairly rewarded. Although Prometheus 2 showed a higher independent Win Rate its agreement with human experts remained moderate.

The integration of the symbolic verification layer proved to be the most critical factor for alignment. The transition from the base MACE-XAI configuration to the full framework—incorporating $\mathcal{S}_{sym}$ —resulted in a significant increase in both Win Rate and alignment. This improvement indicates that human experts heavily penalize technical hallucinations and omissions—factors that the symbolic verifier explicitly mathematically identifies. By grounding the neural preference in symbolic truth, MACE-XAI successfully narrows the gap between automated proxy evaluation and expert human review, providing a more reliable metric for high-stakes XAI deployments.

MACE-XAI achieves a Cohen’s κ of 0.64 when aligned with human expert preferences. Following the standard interpretation guidelines by Landis and Koch (1977), a score of 0.64 falls within the "substantial agreement" range. Given the inherent multi-dimensionality and subjectivity of evaluating NLEs, this baseline represents a major leap over traditional surface-level proxies. Nonetheless, a κ of 0.64 indicates that a non-trivial alignment gap remains.

4.3 Symbolic Error Analysis and Technical Fidelity

To move beyond the limitations of surface-level linguistic evaluation, we conducted a granular Symbolic Error Analysis by cross-referencing the textual claims $\mathcal{C}$ with the raw feature attribution data $\mathcal{X}$. As illustrated in Figure 2, there is a significant

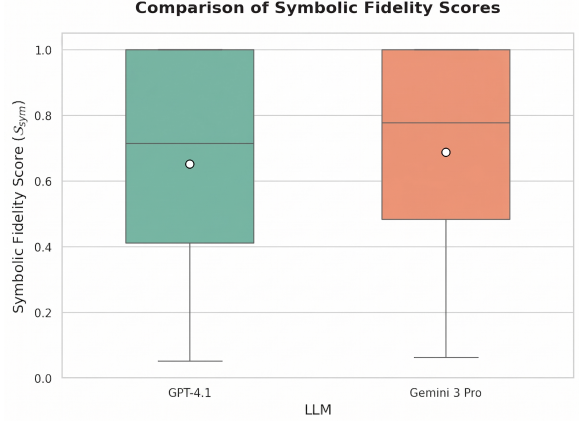


Figure 2: Empirical distribution of the Symbolic Fidelity Score ($\mathcal{S}_{sym}$) across for GPT-4.1 and Gemini 3 Pro. The boxplots depict the factual consistency of generated narratives by cross-referencing textual claims with raw feature attribution data. The central horizontal lines denote the median, the white circles represent the mean fidelity, and the boxes indicate the interquartile range (IQR).

discrepancy between mere narrative plausibility and technical truth. While both models were capable of generating fluent prose, the symbolic audit revealed that Gemini 3 Pro consistently achieved higher fidelity across the test set, maintaining a superior median score and a more concentrated distribution of high-accuracy explanations.

Table 4: Functional Fidelity and Symbolic Audit Results.

Functional Metric	GPT-4.1	Gemini 3 Pro
Class Prediction Accuracy	100.0%	100.0%
Feature Polarity Accuracy	95.54%	98.65%
Omissions Ratio	56.46%	51.52%

As presented in Table 4, a deeper investigation into the error profiles reveals that Omissions ($\mathcal{O}$) represent the primary driver of fidelity loss for both architectures. Both models occasionally failed to incorporate the full set of ground-truth features $\mathcal{F}_{gt}$, particularly in complex samples with high feature dimensionality. However, GPT-4.1 exhibited a higher propensity for substantial omissions in samples where the ground truth exceeded ten features, often resulting in narratives that provided a localized but incomplete view of the model’s decision-making process. Furthermore, the audit identified critical Hallucinations ($\mathcal{H}$) regarding feature polarities. While these errors were less frequent than omissions, they carry a significantly higher risk in high-stakes XAI applications as they provide

false causal narratives. Gemini 3 Pro demonstrated greater robustness in preserving the correct signs of feature influences, whereas GPT-4.1’s outliers were often linked to these polarity inversions.

The high omission ratios represent an indictment of current LLM capabilities, demonstrating that frontier models fail to report roughly half of the ground-truth technical features. This omission gap directly jeopardizes the practical utility of autonomous NLEs in high-stakes decision-making, leaving end-users blind to critical decision drivers. This finding highlights the core necessity of MACE-XAI: while standard semantic metrics fail to penalize these incomplete narratives due to their high linguistic fluency, our neuro-symbolic verification layer successfully exposes this critical systemic failure.

Although GPT-4.1 excels in linguistic mimicry—as evidenced by its higher BLEU and METEOR scores—it falters in technical precision. These findings suggest that Gemini 3 Pro is not only more effective at capturing the correct direction of feature influence but also more comprehensive in its coverage of relevant attributes. This technical superiority in symbolic grounding directly correlates with the substantial agreement found in the Cohen’s Kappa analysis, confirming that experts prioritize factual completeness and sign correctness over stylistic alignment with the reference text.

Conclusions

This study demonstrates that the MACE-XAI framework effectively resolves the fidelity-plausibility trade-off inherent in LLM generated explanations. Our empirical findings using the TEXEN dataset reveal that while state-of-the-art models excel in linguistic mimicry, the integration of a multi-agent comparative paradigm and a symbolic verification layer is essential for achieving high quality NLEs. The superior performance of the full framework highlights that human experts prioritize technical soundness and the correct representation of feature polarities over mere surface-level fluency. Consequently, the Symbolic Fidelity Score serves as a critical audit mechanism, successfully penalizing the omissions and hallucinations that typically undermine user trust in critical domains.

Looking forward, the shift toward neuro-symbolic interpretability frameworks represents a

significant step in the democratization of ethical AI. By ensuring that natural language narratives are mathematically grounded in verifiable ground truths, MACE-XAI allows users to transition from passive recipients of AI predictions to informed participants in the decision-making loop. Future research should explore the scalability of multi-agent approaches to high-dimensional multimodal datasets and real-time interactive XAI systems. Ultimately, this work establishes a new benchmark for robust NLE evaluation, providing a foundation for deploying safer and more transparent AI engines.

Limitations

Single-Dataset Scope and Modality Constraints.

The most primary limitation of this work is its exclusive empirical validation on the TEXEN dataset. Consequently, MACE-XAI’s current claims are strictly bounded to binary tabular classification evaluated via post-hoc feature attribution methods. We provide no empirical evidence that the analytical rubrics, the regex claim extractor, or the crowd composition successfully transfer to other explanation paradigms, such as regression tasks, graph neural network explanations, text-based saliency maps, or concept-based explanations. Extending MACE-XAI to these modalities would require redesigning the extraction module ($\mathcal{F}_{claims}$) and remains an open area for future work.

Tool-Faithfulness Boundary. As formalized in Section 3.1.5, MACE-XAI audits tool-faithfulness, not model-faithfulness. A high Symbolic Fidelity Score ($\mathcal{S}_{sym}$) certifies that the NLE accurately verbalizes the provided XAI evidence, but it does not guarantee that the upstream XAI tool itself is a reliable proxy for the underlying model. A complete end-to-end faithfulness guarantee would require coupling MACE-XAI with intervention-based model verification routines.

Deployment Readiness. While MACE-XAI yields a substantial agreement ($\kappa = 0.64$) with human experts, representing a significant improvement over generic LLM-as-a-judge baselines, this alignment gap is arguably insufficient for full autonomy in risk-sensitive environments. MACE-XAI should not be deployed as an unmonitored judge in high-stakes fields. Instead, it is intended as a human-in-the-loop assistive tool to pre-filter explanations before final expert review.

Computational Overhead. Finally, MACE-XAI’s multi-stage pipeline introduces significant

computational and financial overhead compared to single-judge proxies. Evaluating datasets at scale requires candidate generation across multiple temperatures, pairwise meta-evaluations, and context-augmented aggregation.

References

- B. Biemans, P. Troubil, I. Grau, and W. P. M. Nuijten. 2025. Explainable optimization: Leveraging large language models for user-friendly explanations. In R. Guidotti, U. Schmid, and L. Longo, editors, *Communications in Computer and Information Science*, volume 2578. Springer, Cham.
- J. Burton, N. Al Moubayed, and A. Enshaei. 2023. Natural language explanations for machine learning classification decisions. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.
- Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, and 1 others. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3).
- S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA.
- N. Kroeger, D. Ley, S. Krishna, C. Agarwal, and H. Lakkaraju. 2023. Are large language models post hoc explainers? In *Proceedings of the NeurIPS Workshop on Robustness of Zero/Few-Shot Learning in Foundation Models (R0-FoMo)*.
- J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore.
- S. M. Lundberg and S. Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*.
- D. Martens, J. Hinns, C. Dams, M. Vergouwen, and T. Evgeniou. 2025. Tell me a story! narrative-driven xai with large language models. *Decision Support Systems*, 191:114402.
- R. O. Miyaji and P. L. P. Correa. 2025. Leveraging large language models for generating and evaluating natural language explanations in xai: A comparative study. In *Proceedings of EXPLAINS 2025 – International Conference on Explainable AI, part of the 17th International Joint Conference on Computational Intelligence (IJCCI 2025)*.
- E. Paraschou, I. Arapakis, S. Yfantidou, S. Macaluso, and A. Vakali. 2025. Mind the xai gap: A human-centered llm framework for democratizing explainable ai. In R. Guidotti, U. Schmid, and L. Longo, editors, *Explainable Artificial Intelligence, Part I*. Springer, Cham.
- M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.
- J. Yu, A. I. Cristea, A. Harit, Z. Sun, O. T. Aduragba, L. Shi, and N. Al Moubayed. 2022. Interaction: A generative xai framework for natural language inference explanations. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.
- Q. Zhang, Y. Wang, Y. Jiang, L. Li, C. Wu, Y. Wang, X. Jiang, L. Shang, R. Tang, F. Lyu, and C. Ma. 2025. Crowd comparative reasoning: Unlocking comprehensive evaluations for llm-as-a-judge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria.
- A. Zytek, S. Pidò, and K. Veeramachaneni. 2024. Llms for xai: Future directions for explaining explanations. In *Proceedings of the 2024 ACM CHI Workshop on Human-Centered Explainable AI (HCXAI)*.

Appendix A: Reproducibility Details and Evaluation Protocol

To ensure reproducibility and address the inherent variance in LLM-as-a-judge settings, this appendix details the decoding configurations, the analytical rubrics (Φ), and the prompt templates utilized across generation and evaluation phases of the MACE-XAI framework.

A.1 Decoding Settings and Hyperparameters

To isolate the performance gains derived from our multi-agent architecture from hyperparameter tuning, we standardized the decoding settings as follows:

- **Generation Phase:** To capture the trade-off between technical precision and linguistic diversity, models generated explanations at two temperature settings: $T = 0.2$ (low-variability) and $T = 0.8$ (high-diversity). Top- p was set to 0.95.
- **Evaluation Phase:** All judging agents and the final aggregator were strictly set to a greedy decoding strategy (Temperature $T = 0.0$, Top- $p = 1.0$) to guarantee deterministic and reproducible evaluations.

A.2 Generation Prompt Template

The prompt utilized to synthesize the candidate Natural Language Explanations (NLEs) from the raw feature attributions ($\mathcal{X}$) is as follows:

System: You are an expert AI explainer. Given a set of raw feature attributions (Technical Ground Truth) for a model's prediction, your task is to generate a clear, accessible Natural Language Explanation for an end-user.

Technical Ground Truth:
{technical_context_X}

Instruction: Write a short explanation (max 5 sentences). Focus strictly on the most influential features. When mentioning features from the data, you MUST refer to them using their exact names as presented in the Technical Ground Truth to ensure technical accountability. Do not hallucinate features or causal links not present in the data.

A.3 Analytical Rubrics (Φ)

The Crowd-based Comparative Evaluation (CCE) evaluates the candidate NLEs based on two criteria:

- **Soundness:** Evaluates the technical factuality of the narrative. An explanation is deemed sound if it accurately reflects the feature polarities (positive/negative influence) and strictly avoids hallucinated features, incorrect class predictions, or false causal relationships not present in the raw XAI evidence ($\mathcal{X}$).
- **Completeness:** Measures the sufficiency of feature coverage. An explanation is complete if it successfully verbalizes the critical features (e.g., the most heavily weighted attributes) identified in the ground truth, avoiding significant technical omissions.

A.4 Meta-Evaluator Prompts & In-Context Examples

The framework utilizes specialized prompt architectures. Below we detail the prompt for the Semantic Judge (G-Eval), the Alignment Judge (Prometheus 2), and crucially, the Context-Augmented Final Aggregator, which employs In-Context Learning (ICL).

G-Eval (Semantic Judge) Prompt Template

G-Eval is explicitly instructed to use Chain-of-Thought (CoT) reasoning to prioritize technical fidelity over linguistic fluency.

Task Description:

You will be given a Technical Ground Truth containing feature attributions for a prediction, and two candidate Natural Language Explanations (Explanation A and Explanation B).

Your task is to evaluate and compare both explanations based on the evaluation criteria below, and determine which one is better.

Evaluation Criteria:

- Soundness: The explanation must accurately reflect feature polarities and avoid false causal claims or hallucinations not present in the Technical Ground Truth.
- Completeness: The explanation must verbalize the critical features identified

in the Technical Ground Truth without significant omissions that could mislead the user.

Technical Ground Truth:
{technical_context_X}

Explanation A:
{candidate_eta_A}

Explanation B:
{candidate_eta_B}

Evaluation Steps:

1. Analyze Explanation A against the Technical Ground Truth for Soundness and Completeness.
2. Analyze Explanation B against the Technical Ground Truth for Soundness and Completeness.
3. Compare the two analyses to determine which explanation is more faithful and comprehensive.
4. Output your detailed Chain-of-Thought reasoning based on steps 1-3.
5. On a new line at the very end, output your final verdict using exactly one of the following formats: "Verdict: Winner A", "Verdict: Winner B", or "Verdict: Tie".

Prometheus 2 (Alignment Judge) Prompt Template

Prometheus 2 utilizes a direct scoring approach fine-tuned for rigorous alignment with human-like analytical rubrics.

###Task Description:

An AI model has made a prediction based on specific feature attributions. Evaluate which explanation best communicates the ground truth to a user without misleading them.

###Ground Truth Evidence:
{technical_context_X}

###Rubric:

Evaluate based on [Soundness] (no hallucinations) and [Completeness] (no critical omissions). A perfect explanation covers the ground truth

accurately and fully.

###Explanation A:
{candidate_eta_A}

###Explanation B:
{candidate_eta_B}

###Instruction:

Provide a critique analyzing the functional accuracy of both explanations. Conclude with your final preference strictly based on the rubric.

Context-Augmented Aggregator Prompt

To synthesize the multi-perspective evidence, the aggregator uses few-shot In-Context Learning (ICL). It receives the set of crowd judgments (Ω) alongside few-shot demonstration examples.

Task Description:

You are the lead meta-evaluator. You will receive a Technical Ground Truth, two candidate explanations (A and B), and a set of evaluations from a crowd of AI judges. Your task is to aggregate these judgments and output the final, most robust verdict.

In-Context Examples

Example 1:

[Ground Truth]: {example_1_ground_truth}
[Expl. A]: {example_1_A}
[Expl. B]: {example_1_B}
[Crowd Judgments]: 3 votes for A,
1 vote for B. Judge 2 noted B
hallucinated a feature.
[Aggregator Rationale]:
{example_1_rationale}
[Final Verdict]: Winner A

Example 2:

[Ground Truth]: {example_2_ground_truth}
[Expl. A]: {example_2_A}
[Expl. B]: {example_2_B}
[Crowd Judgments]: 2 votes A, 2 votes B.
[Aggregator Rationale]:
{example_2_rationale}
[Final Verdict]: Tie
End of Examples

Now, evaluate the following real instance:

Technical Ground Truth:

{technical_context_X}

Explanation A:
{candidate_eta_A}

Explanation B:
{candidate_eta_B}

Crowd Judgments (Ω):
{crowd_judgments_omega}

Instruction:
Provide your Aggregator Rationale and your
Final Verdict (Winner A, Winner B, or Tie).

Appendix B: Performance of the Open-Weights Baseline (Llama-3.1-8B)

To evaluate the MACE-XAI framework across different scales and accessibility paradigms, we benchmarked an open-weights baseline against the proprietary frontier models utilized in our primary experiments. Specifically, we deployed Meta’s Llama-3.1-8B as a substitute for the proprietary agents (e.g., GPT-5.1, GPT-5.2) within identical judicial roles, including the semantic assessment and the Context-augmented Final Judgment stages.

For this ablation, Llama-3.1-8B was evaluated using the exact same prompt templates and analytical rubrics (Φ) detailed in Appendix A.

Configuration	Win Rate (95% CI)	Cohen’s κ (95% CI)
Human Expert (Baseline)	81.8% $\pm$ 2.1%	-
MACE-XAI (Proprietary Full)	75.1% $\pm$ 2.4%	0.64 $\pm$ 0.01
MACE-XAI (Llama-3.1-8B Full)	73.8% $\pm$ 2.1%	0.62 $\pm$ 0.02
G-Eval (Llama-3.1-8B)	62.1% $\pm$ 2.5%	0.42 $\pm$ 0.02

Table 5: Comparison of Human Alignment Metrics for the open-weights baseline (Llama-3.1-8B) versus the proprietary MACE-XAI setup.

Appendix C: Algorithm for MACE-XAI Evaluation Pipeline

Algorithm 1 MACE-XAI Evaluation Pipeline

Require: Raw XAI dataset $\mathcal{X}$, Generator LLMs $\mathcal{G}$, Evaluator Crowd $\mathcal{J}$, Aggregator Judge $\mathcal{A}$, Analytical Rubrics Φ , Temperatures $\mathcal{T} = \{0.2, 0.8\}$

Ensure: Final Aggregated Verdicts $\mathcal{V}$

```

1:  $\mathcal{V} \leftarrow \emptyset$ 
2: for each instance  $x \in \mathcal{X}$  do
3:    $\triangleright$  1. Candidate Generation
4:    $\mathcal{E}_x \leftarrow \emptyset$ 
5:   for each model  $g \in \mathcal{G}$  and temp  $\tau \in \mathcal{T}$  do
6:     Generate  $n$  NLEs:  $\eta \sim g(x, \tau)$ 
7:      $\mathcal{E}_x \leftarrow \mathcal{E}_x \cup \{\eta\}$ 
8:   end for
9:    $\triangleright$  2. Pair Sampling & Masking
10:  Sample pairs  $(A, B)$  from  $\mathcal{E}_x$ 
11:  Apply Identity Masking and Order Randomization to  $(A, B)$ 
12:  for each pair  $(A, B)$  do
13:     $\triangleright$  3. Neuro-Symbolic Verification
14:    Extract claims  $\mathcal{C}_A, \mathcal{C}_B$  from texts  $A, B$ 
15:    Compute symbolic fidelity scores  $S_A, S_B$  against  $x$ 
16:     $\triangleright$  4. Crowd-based Judge Comparisons
17:     $\Omega \leftarrow \emptyset$   $\triangleright$  Initialize crowd judgments
18:    for each judge  $j \in \mathcal{J}$  do
19:      Evaluate  $(A, B)$  using  $j$ , rubrics  $\Phi$ , and scores  $S_A, S_B$ 
20:       $\Omega \leftarrow \Omega \cup \{\text{Verdict}_j\}$ 
21:    end for
22:     $\triangleright$  5. Context-Augmented Aggregation
23:    Output final verdict  $v \leftarrow \mathcal{A}(A, B, x, \Omega)$  via ICL
24:     $\mathcal{V} \leftarrow \mathcal{V} \cup \{v\}$ 
25:  end for
26: end for
27: return  $\mathcal{V}$ 

```