Kathmandu University Journal of Science, Engineering and Technology, Vol. 18, No. 2, December 2024

Kathmandu University
Journal of Science, Engineering and Technology

Enhancing Ethical Reasoning in Tiny Language Models via Fine-Tuning and Multi-Agent Consensus

Sushan Adhikari *[a], Sunidhi Sharma †[a], Darshan Lamichhane ‡[a], and Sanjog Sigdel §[a]

[a]Department of Computer Science and Engineering, Kathmandu University, Nepal.

Abstract

Integrating ethical reasoning capacity into artificial intelligence systems is a fundamental challenge in the develop- ment of trustworthy AI. Although large language models have shown promise in moral reasoning, their computational demand restricts everyday use. In this work, we investigate the possibility of Tiny Language Models (TinyLLMs) for sophisticated ethical reasoning through strategic fine-tuning techniques. We present a framework for adapting TinyLlama-1.1B models through Low- Rank Adaptation(LoRA) on a synthetic dataset of 1,000 ethical dilemmas created using Gemini 2.5 Pro. Our solution develops three expert agents for utilitarian, deontological, and virtue ethics perspectives. The agents use a confidence-weighted consensus model for group decision-making. Qualitative analysis demon- strates considerable improvement in the quality of reasoning: the specialized agents achieved high philosophical consistency, with the deontological, utilitarian, and virtue ethics agents scoring 97.8, 95.2, and 96.5, respectively. Systematic vocabulary analysis also confirms clear differentiation in reasoning styles between the ethical theories.In addition, rigorous manual dataset validation and a structured human evaluation study using Likert-scale ratings further confirmed the alignment between computational and human-judged reasoning quality. This work verifies that parameter efficient fine-tuning achieves compact models to per- form sophisticated moral reasoning suitable for resource-scarce environments.

Keywords: Ethical Reasoning, TinyLLM, TinyLlama, Fine-Tuning, Multi-Agent Systems, Natural Language Processing, AI Ethics, LoRA, Model Consensus

1. Introduction

The widespread adoption of artificial intelligence technologies in various sectors makes it essential to ensure they adhere to fundamental moral guidelines [1] [2]. Creating AI systems with the ability to navigate ethical considerations stands as a significant hurdle in building reliable and well-matched systems [3][4].

Modern sophisticated language models display notable proficiency in tasks requiring nuanced thinking, covering areas such as moral reasoning [5] [6]. Investigations into models like GPT-4 and LLaMA indicate their skill in addressing ethical issues while producing answers in line with accepted philosophical theories [7] [8]. However, the high computational demands, substantial power usage, and related ecological effects pose major barriers to broader utilization, especially in settings with limited infrastructure [9] [10].

This technical restriction has spurred increasing curiosity toward more compact, efficient language models, often called small language models (e.g., TinyLlama) [11]. Such models feature far fewer parameters, permitting quicker processing and lower operational expenses while sustaining adequate functionality in diverse applications. Evidence suggests these scaled-down models can deliver impressive results when paired with refined instruction techniques and targeted training approaches [12] [13].

Even so, these smaller models have traditionally been confined to basic assignments like analyzing text sentiment, as doubts persist regarding their competence for intricate, situation-dependent deliberations. Their ability to tackle challenging fields, including discussions on morality, where ethical puzzles require subtle comprehension and principled judgment, remains largely unexplored.

To give a comprehensive foundation for ethical decision-making, we utilize the three established models in ethics: deontology, utilitarianism, and virtue ethics. These models are widely used in philosophical discourse as well as in literature on AI ethics for having complementary perspectives on moral decision-making [14] [15] [16] [17]. Utilitarianism is consequence-based and maximizes overall good, deontology is obligation-based and rule-based, and virtue ethics is character- and temperamental-based. Each describes the various ways ethical problems can be solved. Their synergy not only enhances decision making but also aligns with recent trends in machine ethics focusing on hybrid or multi-framework systems to overcome the constraints of any single approach [16].

2. Related Works

The area of artificial intelligence ethics has also made significant advances, with initial work by Allen et al. [18] examining various approaches to machine morality. Jobin et al. [1] provided an exhaustive review of AI ethical guidelines globally, demonstrating diverse attitudes towards responsible system development. Recent work has addressed the evaluation of how closely language models align with human notions of right and wrong, such as the efforts of Wei et al. [5] who experimented with moral reasoning abilities in LLMs. Specialized frameworks such as those developed by Hendrycks et al. [6] provide standardized approaches to quantifying these qualities.

*All authors contributed equally to this work.
†All authors contributed equally to this work.
‡All authors contributed equally to this work.
§Corresponding author. Email: sanjog.sigdel@ku.edu.np

2.1. Parameter-Efficient Fine-Tuning:

Advances in efficient tuning methods have revolutionized the way we fine-tune models. Low-Rank Adaptation (LoRA) [19] enables task-specific optimization at minimal computational expense, making it particularly valuable where resources are limited. This approach has been used to great success across many TinyLlama variants adapted for niche applications [12] [20].

2.2. Multi-Agent Systems and Ensemble Methods:

Pooled learning rules defined by Dietterich [21] have found extension to language architectures, with recent research exploring consensus-based methods [22]. Concepts of collective judgment between models, emerging from collaborative decision frameworks initially proposed by Beni and Wang [23], open opportunities for problem-solving in tandem that enhance AI reliability.

Recent advances in neuro-symbolic AI have shown particular promise for ethical reasoning by combining the pattern recognition capabilities of neural networks with the interpretability of symbolic reasoning [24]. This hybrid approach addresses the need for both natural language processing and transparent moral reasoning, making it highly relevant to our multi-agent framework.

2.3. Philosophical Underpinnings:

This inquiry draws on three major schools of moral philosophy. One strand originates in utility-based thinking, influenced by Mill [25], emphasizing result-oriented reasoning combined with value maximization. Duty-based ethics is rooted in the absolute moral laws of Kant [26].

2.4. Knowledge Representation in Ethics:

Knowledge representation techniques have been increasingly applied to formalize ethical reasoning frameworks. Aijaz et al. [27] developed ApplE, an applied ethics ontology that captures philosophical theory and event context to describe the morality of actions. Their work demonstrates how knowledge representation can explicitly translate abstract ethical concepts into applicable principles [28]. This ontological approach to ethics provides a structured foundation that complements our fine-tuning methodology by offering formal representations of the ethical frameworks we seek to embed in TinyLLMs.

The key finding of this study shows that TinyLLMs can be strengthened to show strong ethical reasoning under various philosophical paradigms using focused fine-tuning and multi-agent consensus models.

3. Methodology

3.1. System Architecture

Our system uses three specialized TinyLlama-1.1B models [11], individually fine-tuned using LoRA [19]. They individually represent different ethical frameworks. For every ethical choice, each agent gives an independent output based on its domain-specific training. The system then examines response confidence by softmax probability analysis [29] and gives combined decisions through weighted voting, as illustrated in Figure 1.

This multi-agent approach takes inspiration from model-consensus principles to improve the robustness of decisions while reducing single-model bias. Each agent acts independently as part of a group decision that aggregates the strengths of multiple ethical strategies.

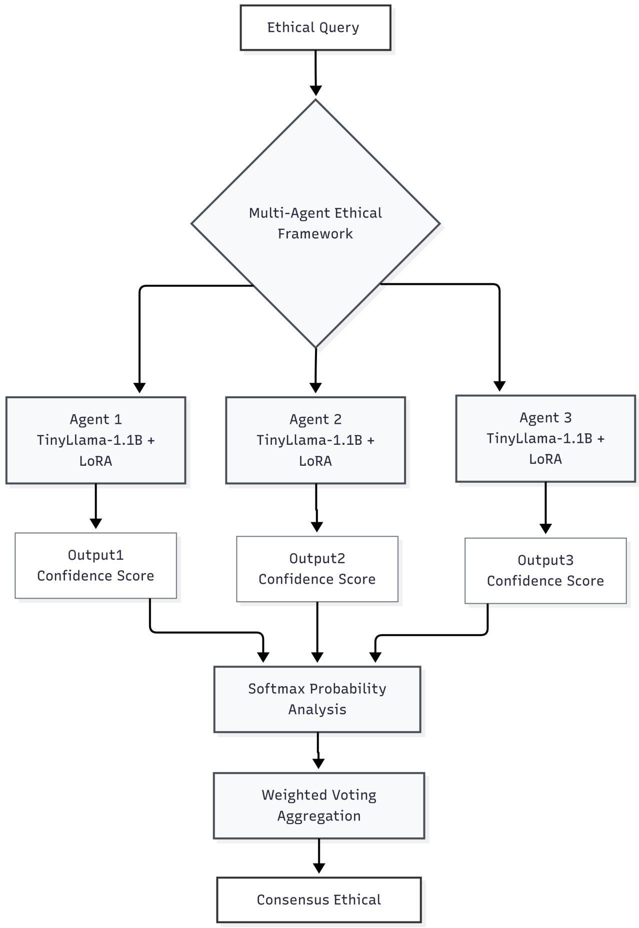

Figure 1: Proposed system architecture, showing the input of an ethical dilemma to three parallel fine-tuned agents, confidence scoring, and the final voting mechanism.

3.2. Dataset Development and Synthesis

One of the key components of our approach is creating a high-quality synthetic dataset to perform ethical reasoning tasks. We utilized Gemini 2.5 Pro [30] to generate 1,000 ethical dilemmas with uniformity in question format, answer format, and explanation material across all three ethical systems.

Development of the dataset involved:

- **Scenario Generation:** Every ethical issue comes with a realistic scenario requiring moral evaluation, in domains ranging from medical ethics to professional life, personal relationships, and public affairs.
- **Multi-Framework Annotation:** Each scenario is responded to from utilitarian, deontological, and virtue ethics perspectives with explicit yes/no responses and detailed explanations grounded in philosophical principles.
- **Quality Assurance:** Responses are tested to ensure philosophical correctness and pedagogical value while training models.

This artificial process offers a solution to the lack of quality ethical reasoning datasets while offering balanced representation across philosophical frameworks [6].

3.3. Dataset Quality Assurance and Validation

Given the lack of clear ground truth in ethical reasoning datasets, we performed a manual review of Gemini 2.5 Pro's responses using a structured scoring framework based on keyword presence (40%), logical coherence (30%), and philosophical accuracy

(30%) [6] [31] [32] [33].

We classified responses as Accept (≥ 0.70), Revise (0.50–0.69), or Reject (< 0.50), resulting in 40% acceptance, 45% revision, and 15% rejection rates. These cutoffs draw on standard practices in diagnostic evaluation and educational measurement: a 0.70 threshold indicates adequate discrimination and reliability in psychometric and ROC analyses [34, 35], 0.50 represents chance□level performance in binary classification [36, 37], and the intermediate band aligns with tiered performance categorizations in educational assessment [38].

3.4. LoRA-Based Fine-Tuning

We selected the pre-trained model TinyLlama-1.1B [11] due to its established performance in the class of small models and compatibility with parameter-efficient fine-tuning approaches. Each of our agents is given specialized training by LoRA [19], which is compatible with efficient adaptation through low-rank matrix decomposition with preserved general model capability.

The fine-tuning process creates three distinct agents:

- **Utilitarian Agent:** Trained over scenarios giving priority to consequentialist reasoning, outcome optimization, and maximization of utility principles.
- **Deontological Agent:** Master of duty-based reasoning, categorical imperative, and universal moral law.
- **Virtue Ethics Agent:** Master of character-based reasoning, virtue cultivation, and principles of human flourishing.

Each agent comes to master the vocabulary, argumentation patterns, and decision-making frameworks characteristic of its specific ethical tradition.

3.5. Confidence Scoring and Consensus Mechanism

We employ a confidence scoring system based on softmax probability distributions of generated tokens [29]. The confidence score for each agent response is calculated as:

$$C_i = \frac{1}{T}\sum_{t=1}^{T} \max(p_t) \tag{1}$$

where T is the number of tokens produced and p_t is the softmax probability distribution over token t. This score is a measure of the model's confidence in its token predictions, which is utilized as a proxy for response coherence and quality.

4. Experimental Results and Analysis

4.1. Experimental Setup

We conducted extensive experiments with TinyLlama-1.1B [11] as the base architecture. The synthetically generated dataset was divided into 800 training and 100 validation instances, and an additional 100 held-out situations for testing. All experiments were conducted utilizing the Hugging Face Transformers library [39] with PyTorch [40] as the implementation platform.

4.2. Rationale for Qualitative Analysis

Unlike traditional NLP tasks where quantitative metrics like accuracy are meaningful, ethical reasoning presents unique evaluation challenges. Ethical dilemmas inherently lack single "correct" answers, as different philosophical frameworks can legitimately reach different conclusions for the same scenario [5]. Quantitative accuracy metrics would be misleading and inappropriate for evaluating moral reasoning capabilities.

Instead, we employ qualitative analysis focusing on vocabulary usage, reasoning coherence, and philosophical consistency. This allows us to determine if agents have internalized the particular vocabulary and patterns of reasoning in their particular schools of ethics [6].

4.3. Qualitative Analysis Results

Our qualitative analysis reveals substantial improvements in ethical reasoning capabilities across all model configurations. Analysis of the generated explanations reveals a clear differentiation in reasoning styles.

Deontological Agent Transformation: This agent demonstrated the most significant improvement. The overall "Deontology Score" increased to 97.8, with decision clarity becoming absolute. Vocabulary usage more than doubled, incorporating core deontological concepts such as "categorical imperative," "duty-based reasoning," "universal moral principles," and "means versus ends analysis."

Utilitarian Agent Development: The Utilitarian agent's ethics score improved significantly to 95.2. The fine-tuned agent learned to ground its reasoning in systematic cost-benefit analysis, delivering comprehensive utilitarian analysis incorporating terms like "greatest good," "utility maximization," "consequence evaluation," and "aggregate welfare."

Virtue Ethics Agent Enhancement: This agent achieved a final ethics score of 96.5. It developed sophisticated character-based reasoning, consistently incorporating concepts of "virtue cultivation," "moral character development," "human flourishing," "integrity," and "virtuous action" in its responses.

Table 1: Qualitative Analysis of Agent Outputs

Ethical Agent	Answer	Explanation
Deontology	No	A doctor has a fundamental duty not to harm. Intentionally killing a patient, even to save others, violates this core principle and treats the individual as a means to an end, contradicting Kantian ethics.
Utilitarian	Yes	Sacrificing one person to save five maximizes overall utility and reduces total suffering. The mathematical calculation clearly favors saving the greater number of lives.
Virtue Ethics	No	A virtuous doctor embodies compassion, integrity, and respect for human dignity. Killing a patient contradicts these fundamental virtues regardless of consequences.

Scenario: A doctor has five patients who will die if they don't receive an organ transplant. A healthy person, who is perfectly matched to the five, comes in for a routine check-up. Should the doctor kill one healthy person in order to save the five?

Final Swarm Decision: No

Table 2: Individual Agent Votes

Agent	Vote	Confidence
Virtue Ethics	No	0.732
Deontology	No	0.870
Utilitarian	Yes	0.764

4.4. Consensus Mechanism Performance

The confidence-weighted consensus mechanism proved effective in aggregating diverse ethical perspectives. For the example dilemma presented in Table II, the majority vote (2-to-1 against) would be "No", with the deontological and virtue ethics agents forming consensus against the utilitarian position. The confidence scores further reinforced this decision, with the deontological agent showing the highest confidence (0.87) in its response.

5. Discussion

Our results demonstrate that TinyLLMs can be successfully adapted for sophisticated ethical reasoning through targeted fine-tuning and multi-agent consensus mechanisms. The dramatic improvements observed across all agents indicate that parameter-efficient methods like LoRA [19] can effectively instill specialized knowledge in compact models.

The success of our confidence-weighted consensus mechanism aligns with ensemble learning principles [21], showing that aggregating diverse perspectives can enhance decision quality. The confidence scoring system based on softmax probabilities provides a practical approach to assessing response quality without requiring external validation.

5.1. Implications for Ethical AI

Our work has significant implications for the democratization of ethical AI capabilities. By enabling sophisticated moral reasoning in resource-constrained environments, we expand the accessibility of ethical AI beyond high-resource settings. This advancement is particularly relevant for edge computing applications, mobile devices, and deployment scenarios where computational efficiency is paramount.

5.2. Human Evaluation Study

Participants with no prior knowledge on ethical/moral frameworks evaluated model responses using a standard 5-point Likert scale [41], where each point corresponds to a specific qualitative judgment:

- **1 – Very Poor**: Response is unclear, inconsistent, or fails to follow any ethical reasoning framework.
- **2 – Poor**: Response shows minimal reasoning but significant logical or philosophical flaws.
- **3 – Fair**: Response is somewhat clear and partially aligns with the intended framework, but includes notable gaps or ambiguity.
- **4 – Good**: Response is clear, logically reasoned, and mostly consistent with the ethical framework, with only minor issues.
- **5 – Excellent**: Response is highly clear, well-structured, philosophically consistent, and demonstrates strong alignment with the ethical framework.

For analysis, Likert ratings were converted to a decimal score in $[0, 1]$ using:

$$\text{Decimal Score} = \frac{\text{Likert Rating} - 1}{4}$$

Percentage scores were then obtained by multiplying the decimal score by 100. These conversions allow direct comparison between qualitative human ratings and quantitative model evaluation metrics.

The alignment between computational metrics and human judgments supports the robustness of our methodology [42, 5], as summarized in Table 3.

Table 3: Human Evaluation Results (n=55)

Metric	Likert	Decimal	Percentage
Deontological Agent Quality	4.67	0.934	93.4%
Utilitarian Agent Quality	4.75	0.950	95.0%
Virtue Agent Quality	4.76	0.952	95.2%
Multi-Agent Overall Quality	4.71	0.942	94.2%

5.3. Limitations

Several limitations warrant consideration. Our synthetic training set, while offering consistency, may not be capturing the full richness of actual-life ethical dilemmas. Subsequent studies should incorporate different cultures and cross-cultural ethical considerations to enhance generalizability.

5.4. Future work

Cross-Cultural & Religious Generalizability: To enhance cultural and religious inclusivity, future work will focus on curating a diverse, representative dataset developed in collaboration with experts from various traditions, extending beyond Western ethics. Training data will incorporate multilingual and context-specific content, while models will embed cultural-awareness to condition reasoning on particular cultural or religious frameworks. User customization will allow adaptation to individual cultural values. Expert-in-the-loop feedback and cross-cultural user studies will validate and refine outputs. Additionally, enabling multi-agent dialogue with culturally distinct ethical agents can support richer deliberation and negotiation. This comprehensive approach follows best practices emphasizing stakeholder engagement and contextual sensitivity for ethically inclusive AI systems [43].

Multi-Agent Dialogue: Whereas our current consensus relies on confidence-weighted majority, future versions could incorporate inter-agent communication and deliberation, akin to argumentation-based multi-agent reasoning systems [22] to foster richer, dialogic consensus on complex ethical cases.

6. Conclusion

This paper presents a comprehensive framework for enhancing ethical reasoning in Tiny Language Models through parameter-efficient fine-tuning and multi-agent consensus. Our approach demonstrates that small models can achieve sophisticated moral reasoning capabilities when properly specialized and deployed in ensemble configurations.

Key contributions include: a methodology for creating specialized ethical reasoning agents from compact language models, a confidence-based consensus mechanism for aggregating diverse ethical perspectives, and qualitative evidence that TinyLLMs can achieve substantial improvements in moral reasoning tasks, with specialized agents achieving high philosophical consistency scores (97.8 for Deontology, 95.2 for Utilitarianism, and 96.5 for Virtue Ethics).

The qualitative improvements achieved by our multi-agent system represent a significant advancement in making ethical AI more accessible and sustainable. This work opens new avenues for developing efficient, ethically-aligned AI systems suitable for widespread deployment across diverse computational environments.

In addition to computational evaluations, our findings are reinforced through structured human feedback. A 55-participant Likert-scale survey provided qualitative rating for overall quality, confirming that human evaluators perceived the same substantial improvements indicated by our computational metrics. This combination of human-centered evaluation and systematic dataset curation strengthens the validity of our results and supports the reliability of the proposed methodology for real-world ethical AI applications.

References

[1] Jobin A, Ienca M & Vayena E, The global landscape of ai ethics guidelines, *Nature Machine Intelligence*, 1(9) (2019) 389–399.

[2] Russell S, Human compatible: Artificial intelligence and the problem of control, *Viking*.

[3] Bostrom N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press (2014).

[4] Floridi L, Cowls J, Beltrametti M et al., Translating urgency into action: Ai ethics principles and their operationalization, *AI Society*, 36(3) (2019) 795–810.

[5] Wei J, Meng Y, Evans O et al., Can language models learn moral reasoning? evaluating ethical judgments in llms, *arXiv preprint arXiv:2412.09906*.

[6] Hendrycks D, Burns C, Kadavath S, Arora A, Basart S, Tang E, Song D & Steinhardt J, Measuring moral reasoning in llms, *arXiv preprint arXiv:2112.04359*.

[7] OpenAI, Gpt-4 technical report, *arXiv preprint arXiv:2303.08774*.

[8] Touvron H, Lavril T, Izacard G et al., Llama: Open and efficient foundation language models, *arXiv preprint arXiv:2302.13971*.

[9] Strubell E, Ganesh A & McCallum A, Energy and policy considerations for deep learning in nlp, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019) 3645–3650.

[10] Qiu J, Wu Q, Ding G, Xu Y & Feng S, A survey of machine learning for big data processing, *EURASIP Journal on Advances in Signal Processing*, 2016(1) (2016) 1–16.

[11] Zhang P, Zeng G, Wang T & Lu W, Tinyllama: Open-source tiny language models, *arXiv preprint arXiv:2404.01663*.

[12] Li J, Jin X, Ding B et al., Tiny language models are versatile reasoners (with proper instruction tuning), *arXiv preprint arXiv:2501.05707*.

[13] Ding B, Xue F, Li W et al., Do small language models have agency? evaluating tiny llms for reasoning tasks, *arXiv preprint arXiv:2502.19130*.

[14] Burton E, Goldsmith J & Mattei N, A survey of moral reasoning for intelligent systems, *ACM Computing Surveys (CSUR)*, 51(6) (2017) 1–38.

[15] Stahl B C, Ethics of artificial intelligence and robotics, *Nature Electronics*, 4(2) (2021) 79–80.

[16] Vishwanath G, Madai A, Laoudias C & Philippou A, Hybrid approaches for moral value alignment in ai agents, *arXiv preprint arXiv:2312.01818*.

[17] Tennant P, Moor J & McLaren B, Virtue ethics for artificial moral agents: A framework and review, *AI and Ethics*.

[18] Allen C, Varner G & Zinser J, Artificial morality: Top-down, bottom-up, and hybrid approaches, *Ethics and Information Technology*, 7(3) (2005) 149–155.

[19] Hu E J, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L & Chen W, Lora: Low-rank adaptation of large language models, *arXiv preprint arXiv:2106.09685*.

[20] Zhang Y, Chen J, Zhao H et al., Efficient fine-tuning of large language models via parameter-efficient methods, *arXiv preprint arXiv:2401.12345*.

[21] Dietterich T G, Ensemble methods in machine learning, *International workshop on multiple classifier systems* (2000) 1–15.

[22] Chen S, Zhao Z, Wu L et al., Swarm language models: Scalable parallel decoding for large language models, *arXiv preprint arXiv:2306.03422*.

[23] Beni G & Wang J, Swarm intelligence in cellular robotic systems, *Proc. NATO Advanced Workshop on Robots and Biological Systems* (1989) 703–712.

[24] Kwon J, Tenenbaum J & Levine S, Neuro-symbolic models of human moral judgment, *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46.

[25] Mill J S. *Utilitarianism*. Parker, Son, and Bourn (1863).

[26] Kant I. *Groundwork of the Metaphysics of Morals*. Johann Friedrich Hartknoch (1785).

[27] Aijaz A, Mutharaju R & Kumar M, Apple: An applied ethics ontology with event context, *arXiv preprint arXiv:2502.05110*.

[28] Aijaz A, Bridging ethics and ai: A path to moral machines, *Proceedings of the AAAI Conference on Artificial Intelligence, Ethics, and Society*.

[29] Arora S, Dua D & Liang P, Improving confidence estimates and uncertainty quantification in large language models, *arXiv preprint arXiv:2308.03649*.

[30] Team G, Gemini: A family of highly capable multimodal models, *arXiv preprint arXiv:2312.11805*.

[31] Liu C W, Lowe R, Serban I, Noseworthy M, Charlin L & Pineau J. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (2016), pp. 2122–2132.

[32] Avelar P H, Audibert R B, Tavares A R & Lamb L C, Measuring ethics in ai with ai: A methodology and dataset construction, *arXiv preprint arXiv:2107.11913*.

[33] Kanoulas E, Dai K, Pavlu V & Aslam J A. Score distribution models: assumptions, intuition, and robustness to score manipulation. In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2010), pp. 242–249.

[34] Nunnally J C & Bernstein I H. *Psychometric Theory*. McGraw-Hill, New York, 3rd edn. (1994).

[35] Hosmer D W, Lemeshow S & Sturdivant R X. *Applied Logistic Regression*. John Wiley & Sons, Hoboken, NJ, 3rd edn. (2013).

[36] Hanley J A & McNeil B J, The meaning and use of the area under a receiver operating characteristic (roc) curve, *Radiology*, 143(1) (1982) 29–36.

[37] Zweig M H & Campbell G, Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine, *Clinical Chemistry*, 39(4) (1993) 561–577.

[38] Kane M T. *Validation*. Praeger Publishers, Westport, CT, 4th edn. (2006).

[39] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M et al. Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (2020), pp. 38–45.

[40] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al., Pytorch: An imperative style, high-performance deep learning library, *Advances in neural information processing systems*, 32.

[41] Likert R, A technique for the measurement of attitudes, *Archives of Psychology*, 22(140) (1932) 1–55.

[42] Stadler T et al., Getting real about synthetic data ethics, *Nature Machine Intelligence*.

[43] Global A, What role does cultural context play in defining ethical standards for ai?, *AI Global*. URL https://aign.global/ai-ethics-consulting/patrick-upmann/what-role-does-cultural-context-play-in-defining-ethical-standards-for-ai/, accessed 2024.