Ethical Alignment of Small Language Models Using the ETHICS Dataset and Knowledge Distillation from Large Language Models

1st Taskin Ahmed Prottoy
ID: 2031669642
ECE, North South University
Dhaka, Bangladesh
taskin.prottoy@northsouth.edu

2nd Amanullah Ahsan
ID: 2021769642
ECE, North South University
Dhaka, Bangladesh
amanullah.ahsan@northsouth.edu

3rd Atikur Rahman Shohag
ID: 2021746642
ECE, North South University
Dhaka, Bangladesh
atikur.shohag@northsouth.edu

4th Suraiya Akter
ID: 2022623642
ECE, North South University
Dhaka, Bangladesh
suraiya.akter@northsouth.edu

5th Afsana Ahmed Shammi
ID: 2021973642
ECE, North South University
Dhaka, Bangladesh
afsana.shammi@northsouth.edu

***Abstract*—The ethical alignment of language models is a critical challenge in the advancement of AI-driven decision-making systems. While large language models (LLMs) have demonstrated strong ethical reasoning capabilities, their high computational requirements make them inaccessible for many applications. This research aims to enhance the ethical reasoning of small language models (SLMs) by fine-tuning them on the ETHICS dataset and optimizing their performance using knowledge distillation from larger pre-trained models. Specifically, we focus on models such as Llama 3.2 1B, Llama 3.2 3B, Gemma 2 2B, and Phi 3.5-mini 4B, refining their decision-making process while ensuring computational efficiency. The methodology involves structured data preprocessing, fine-tuning with supervised learning, and rigorous evaluation using both quantitative and qualitative metrics. Additionally, knowledge distillation techniques, including logit-based and response-based distillation, are employed to transfer ethical reasoning capabilities from larger models to smaller ones. The expected outcomes include ethically aligned SLMs with improved reasoning efficiency, optimized models suitable for deployment in low-resource environments, and enhanced benchmarking methodologies for ethical AI evaluation. This research contributes to the broader discourse on AI ethics, fostering the development of accessible and responsible AI systems.**

***Index Terms*—LLM alignment, Knowledge Distillation, Small Langugae Model, AI ethics.**

I. Introduction

The increasing deployment of artificial intelligence (AI) in real-world applications has underscored the importance of ethical reasoning in decision-making systems. AI models are now integral to sectors such as healthcare, finance, legal systems, and automated customer service, where ethical considerations play a critical role in ensuring fairness, transparency, and accountability. However, large language models (LLMs) capable of sophisticated ethical reasoning, such as GPT-4, Llama-2, and Mistral, require extensive computational resources, making them impractical for many real-world deployments. Smaller language models (SLMs), on the other hand, offer a more efficient alternative but often lack the nuanced ethical reasoning capabilities of their larger counterparts. This imbalance presents a fundamental challenge: how to ensure that SLMs retain ethical alignment without sacrificing computational efficiency.

To address this challenge, this research investigates methods to enhance the ethical reasoning capabilities of SLMs through fine-tuning on the **ETHICS benchmark dataset** and **knowledge distillation** from larger pre-trained models. The ETHICS dataset provides structured ethical scenarios categorized into five primary ethical dimensions: *justice, deontology, virtue ethics, utilitarianism, and commonsense morality*. Prior work has demonstrated the feasibility of using supervised learning to align LLMs with ethical reasoning tasks, but the question remains whether these techniques can be effectively applied to smaller, resource-constrained models.

Knowledge distillation is a promising approach to bridge this gap. By transferring knowledge from large-scale models with strong ethical reasoning capabilities to smaller models, it is possible to retain performance while significantly reducing computational overhead. This study explores multiple knowledge distillation techniques, including **logit-based distillation** (which transfers probability distributions from larger models to smaller ones) and **response-based distillation** (which refines smaller models by aligning their ethical decision-making processes with those of larger models).

The key contributions of this research include:

- **Fine-tuning Small Language Models (SLMs) on the ETHICS Dataset**
  - Models such as *Llama 3.2 1B, Llama 3.2 3B, Gemma 2 2B, and Phi 3.5-mini 4B* are fine-tuned using su-

pervised learning to enhance their ethical reasoning capabilities.

- **Implementing Knowledge Distillation from Large Pre-Trained Models**
  - Ethical reasoning from models like *Llama 3.2 8B, Gemma 7B, and Mixtral 12B* is transferred to smaller models using advanced distillation techniques, improving their decision-making accuracy while maintaining efficiency.

The expected impact of this research is significant. By improving the ethical alignment of small language models, we can facilitate the development of AI systems that are not only computationally efficient but also responsible and fair in their decision-making processes. This work contributes to the broader discourse on AI ethics, providing a foundation for future advancements in the field of ethical AI model development.

II. Previous Work

The ethical alignment of AI systems has been a growing area of research, particularly as artificial intelligence (AI) is increasingly deployed in sensitive domains such as healthcare, finance, and law. One of the significant contributions in this field is the ETHICS benchmark, which provides a structured framework for evaluating AI decision-making in ethically charged scenarios. The benchmark categorizes ethical principles into five key dimensions: justice, deontology, virtue ethics, utilitarianism, and commonsense morality.

Prior research, including the seminal paper Aligning AI with Shared Human Values, has explored the fine-tuning of large language models (LLMs) such as BERT, RoBERTa, and ALBERT using the ETHICS dataset. These models were evaluated based on their ability to classify ethical scenarios into predefined categories, with results demonstrating varying degrees of alignment with human ethical expectations. Despite these advancements, challenges remain in ensuring that AI systems generalize effectively across diverse ethical dilemmas and maintain computational efficiency.

Replication studies have played a crucial role in validating the robustness of ethical AI benchmarks. Previous work has focused on reproducing ETHICS benchmark experiments to verify the reliability of its methodologies and results. Studies have confirmed that while large models achieve reasonable accuracy in ethical classification tasks, there is a growing need for smaller, resource-efficient models capable of ethical reasoning in real-world applications.

Furthermore, recent work has examined the potential of fine-tuning modern LLMs, such as Llama-2 and Mistral, for ethical alignment. While these models offer improved language understanding, their ethical decision-making capabilities require further refinement. Additionally, researchers have emphasized the need for expanding the ETHICS benchmark to cover new ethical dilemmas beyond the initial predefined categories, ensuring broader applicability.

This prior research serves as a foundation for the proposed work, which aims to extend the ethical alignment of AI by leveraging state-of-the-art models and exploring more computationally efficient solutions. By addressing existing gaps in model generalization and resource efficiency, the project will contribute to the ongoing efforts to develop ethically responsible AI systems.

III. Problem Statement

Ensuring that AI-driven systems adhere to ethical reasoning principles is crucial for their reliability and acceptance in real-world applications. Large-scale language models exhibit superior ethical decision-making capabilities; however, their high computational cost restricts their use in low-resource settings. Given a small language model M_s parameterized by $heta_s$, and a large language model M_l with parameters $heta_l$, where $|heta_s| \ll |heta_l|$, the objective is to align M_s with ethical reasoning while maintaining computational efficiency. Formally, for a dataset $D = \{(x_i, y_i)\}_{i=1}^N$, where x_i represents an ethical prompt and y_i the corresponding ethically aligned response, the goal is to optimize $heta_s$ such that:

$$\theta_s^* = \arg\min_{\theta_s} \mathcal{L}(M_s(x_i;\theta_s), y_i) + \lambda D_{KL}(M_s(x_i;\theta_s)||M_l(x_i;\theta_l)),$$

where $\mathcal{L}$ is the loss function measuring the alignment between M_s's predictions and ground truth labels, and D_{KL} is the Kullback-Leibler divergence ensuring knowledge transfer from M_l to M_s. The challenge lies in training M_s to retain ethical reasoning capabilities comparable to M_l while being resource-efficient.

A. Methodology

The research follows a four-phase methodology: data preparation, fine-tuning of small language models, evaluation of ethical alignment, and knowledge distillation for optimization. The ETHICS dataset, comprising multiple ethical frameworks—Virtue Ethics, Social Ethics, Deontology, Utilitarianism, and Commonsense Morality—is preprocessed to ensure balance and consistency. The selected SLMs undergo supervised fine-tuning using cross-entropy loss, leveraging prompt-based learning for enhanced reasoning. Ethical alignment is assessed through benchmark accuracy, precision-recall metrics, perplexity scores, and qualitative evaluations involving human annotation. Finally, knowledge distillation from larger models such as Llama 3.2 8B, Gemma 7B, or Mixtral 12B is implemented using logit-based and response-based approaches, optimizing SLMs through soft targets and temperature scaling to improve ethical reasoning performance.

IV. Expected Outcome

This research anticipates developing ethically aligned SLMs that can reason effectively across various ethical frameworks while remaining computationally efficient. The optimized models will be deployable on low-resource devices, ensuring accessibility without sacrificing ethical integrity. Furthermore, new evaluation methodologies for ethical AI alignment will be established, contributing to AI fairness and transparency research. The findings will inform best practices for deploying responsible AI systems, bridging the gap between large-scale and resource-efficient models in ethical decision-making contexts.

V. Conclusion

The increasing reliance on AI-driven systems necessitates the development of models that align with ethical principles while being computationally efficient. By fine-tuning small language models on the ETHICS dataset and employing knowledge distillation from larger models, this research seeks to create scalable, responsible AI solutions. The proposed approach ensures that ethical reasoning capabilities are maintained in low-resource environments, paving the way for more inclusive and accessible AI applications while upholding fairness and human values.