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

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

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

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

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

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. Problem Statement

Ensuring ethical reasoning in AI-driven systems is paramount for their reliability and real-world acceptance. Large-scale language models exhibit superior ethical decision-making capabilities; however, their substantial computational cost limits accessibility, particularly in low-resource environments. This research aims to distill knowledge from larger models like LLaMA 3.2 8B into smaller, computationally efficient models such as LLaMA 3.2 3B and Phi models, maintaining ethical alignment while reducing resource constraints.

Formally, given a small language model M_s parameterized by θ_s, and a large language model M_l with parameters θ_l, where $|\theta_s| \ll |\theta_l|$, the objective is to optimize M_s such that it retains ethical reasoning capabilities while remaining computationally efficient. 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 optimization objective is formulated as:

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

where $\mathcal{L}$ measures alignment loss between M_s's predictions and ground truth labels, while D_{KL} represents the Kullback-Leibler divergence ensuring effective knowledge transfer from M_l to M_s. The primary challenge is achieving a balance between ethical reasoning retention and computational efficiency in M_s.

III. Methodology

This research follows a four-phase methodology:

A. Data Preparation

The ETHICS dataset, which encompasses multiple ethical paradigms—Virtue Ethics, Social Ethics, Deontology, Utilitarianism, and Commonsense Morality—is curated and preprocessed to ensure class balance and consistency. The dataset serves as the foundation for fine-tuning and evaluation.

B. Fine-tuning Small Language Models

We fine-tune smaller models, such as LLaMA 3.2 3B and Phi, using supervised learning with cross-entropy loss. Prompt-based learning techniques are leveraged to enhance the models' ethical reasoning capabilities. The fine-tuned models are evaluated iteratively to ensure stability and consistency in ethical decision-making.

C. Ethical Alignment Evaluation

To assess ethical alignment, we employ:

- Benchmark accuracy across different ethical paradigms
- Precision-recall metrics for classification tasks
- Perplexity scores to measure model coherence
- Qualitative human annotation evaluations for nuanced ethical reasoning

D. Knowledge Distillation for Optimization

Knowledge from larger models such as LLaMA 3.2 8B, Gemma 7B, and Mixtral 12B is transferred using logit-based and response-based distillation techniques. This process involves:

- Soft target optimization with temperature scaling
- Response-based distillation to align smaller models with ethical reasoning patterns observed in larger models
- Gradual refinement through iterative feedback loops

IV. Results

Table **??** summarizes the accuracy comparisons of various models. Our approach, utilizing LLaMA 3.2 3B and Phi model, achieves competitive performance.

TABLE I
Comparison of Accuracy, Precision, Recall, and F1-score across different models.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
SVM	69.4	65.0	68.0	66.5
Random Forest	78.6	82.9	66.9	74.0
Gradient Boosting	63.2	64.2	65.7	64.9
Logistic Regression	67.8	66.8	68.3	67.5
BERT	78.6	70.9	69.8	70.3
DistilBERT	78.2	74.5	73.1	73.8
Mistral 7B	42.4	44.3	46.2	45.2
Llama 2 7B	62.8	63.5	61.9	62.7
Ours (499A)	62.8	72.3	70.8	71.5
Phi3.5-mini	**87.8**	**75.3**	**76.8**	**72.5**
Llama 3.2 3B	**89.5**	**78.3**	**75.2**	**76.6**

V. Future work

For future work, we plan to investigate distilling knowledge from smaller models, such as LLaMA's lightweight variants and Phi models, into larger architectures to enhance their efficiency and performance. By leveraging the structured knowledge captured in these compact models, we aim to improve generalization and reduce training costs in larger-scale models. This reverse distillation strategy will be explored in our next update, focusing on optimizing both accuracy and computational efficiency.