The Efficacy of Finetuning Large Language Models for Interpersonal Conflict Resolution

Matthew Boraske Advisor: Dr. Richard Burns

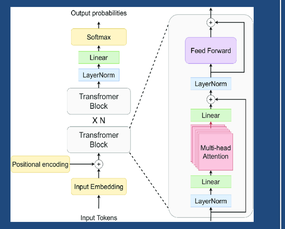
WCU WEST CHESTER UNIVERSITY

Large Language Models (LLMs) that leverage transformer architectures have become the predominant form of state-of-the-art artificial intelligence (AI).

The Two Types of Transformers

Encoder-decoder

- Designed for sequence-to-sequence conversions tasks such as summarization and translation
- Purpose of the encoder is to give the LLM an intricate understanding of the input context



Decoder-only

- Designed for open-ended generation tasks
- Input is directly fed to decoder without any intermediate processing.
- Reduced training requirements and improved generation speed by eliminating the encoder.

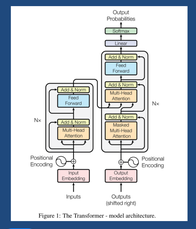


Figure 1: The Transformer - model architecture.

Google FLAN-T5

11 Billion Parameters

State-of-the-art and open-source models that utilize each architecture

13 Billion Parameters

Meta Llama-2

Interpersonal conflict resolution is a challenging task for LLMs

Solutions are inherently ambiguous

Tendency for humans to resort to toxic language

Task is two-fold: classification and justification

Which LLM is superior at interpersonal conflict resolution when finetuned on these Reddit AITA datasets?

The Reddit "Am I the A**hole" (AITA) Subreddit

- An online forum with over fifteen million members where interpersonal conflicts are shared for judgement, which consists of choosing one of five AITA classifications and then writing a justification.
- We created two datasets consisting of subreddit submissions and the classifications and justifications for the top ten comments by community score
  - Multiclass dataset: Contains all five possible AITA classifiations.
  - Binary dataset: Only includes the extreme classifications of NTA and YTA.

AITA Classifications

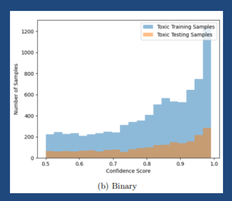
Classification	Abbreviation Meaning	Description
YTA	You're the a**hole	The writer is causing the conflict.
NTA	Not the a**hole	The writer is not causing the conflict.
NAH	No a**holes here	No one is causing a conflict.
ESH	Everyone sucks here	Everyone is causing the conflict.
INFO	More Information Needed	The conflict lacks context for fair judgment.

By finetuning Flan-T5-XXL and Llama-2-13B-Chat on these datasets, we evaluated their ability to learn to solve real-world interpersonal conflicts while also assessing their robustness against adopting the generation of toxic language.

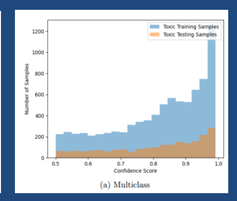
Dataset	Total Samples	YTA	NTA	ESH	NAH	INFO
Multiclass	50000	4465	32431	1071	1509	524
Binary	36896	4465	32431	0	0	0

Table 3.2: Reddit AITA Dataset Classifications

Toxicity Rates of Top Comments using ConfliBERT Finetuned on Toxigen dataset



(b) Binary



(a) Multiclass

Dataset	Train Partition	Test Partition
Multiclass	0.219	0.224
Binary	0.225	0.231

Table 3.4: Top Comment Toxicity Rates in Reddit AITA Datasets

Comment Agreement Analysis Using Krippendorff's Alpha

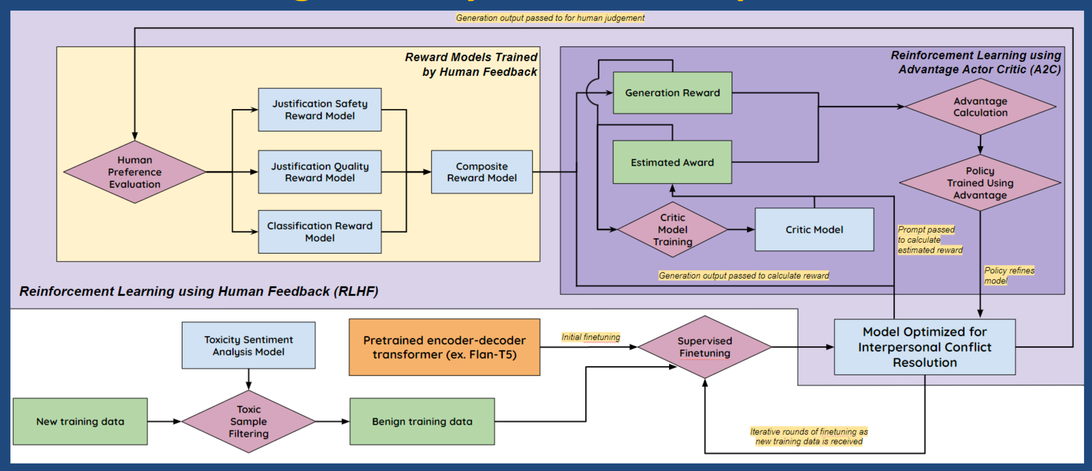
Dataset	Train Partition	Test Partition
Multiclass	0.731	0.737
Binary	0.752	0.759

Table 3.5: Krippendorff's Alpha for Reddit AITA Datasets

*A Krippendorff's alpha of less than 0.8 indicates **statistically significant disagreement** between the AITA classifications by commenters [1]*

Metric	Description
ROUGE Lsum	A summarization evaluation metric that measures the longest common subsequence between a system-generated summary and a reference summary.
Average COMET	Evaluates the performance of machine translation systems by predicting how much a human would understand and appreciate the translation.
Toxicity Rate	The proportion of generated texts that contained toxic language
Precision	The proportion of true positive results out of all the positive results
Recall	The proportion of true positive cases that were correctly identified.
F1 Score	The harmonic mean of precision and recall.
Matthew's Correlation Coefficient (MCC)	Accounts for both true and false positives and negatives and is regarded as a balanced measure that is particularly useful when the classes are of very different sizes.

A Proposed LLM Architecture and Training Process for Learning to Safely Resolve interpersonal Conflicts



Key Components

- **Usage of an encoder-decoder transformer** like in Flan-T5 XXL, as the finetuned binary model achieved the greatest classification performance and justification quality.
- Supervised finetuning only on samples validated to **not contain toxic language**.
- Iteratively improving model alignment by implementing a **Reinforcement Learning with Human Feedback** loop that utilizes rewards models for safety, justification quality, and classification accuracy. This reinforces the idea that any AI tool that will be successfully used in sensitive contexts such as therapy will **require close supervision by humans** to ensure it doesn't deviate towards dangerous behavior.

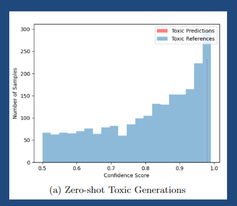
By using a parameter efficient finetuning technique called **QLoRA,** all models were finetuned in **less than 48 hours** on a single, high-end Nvidia L40 GPU with 48 GB of VRAM

AITA Multiclass Results

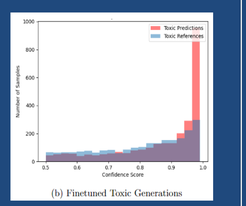
Flan-T5 XXL

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.025	0.314	0.063	0.69	0.35	0.40	0.032
Finetuned	0.161	0.515	0.268	0.75	0.81	0.78	0.314

Table 4.1: Performance of Flan-T5 XXL on Reddit AITA Multiclass Dataset



(a) Zero-shot Toxic Generations

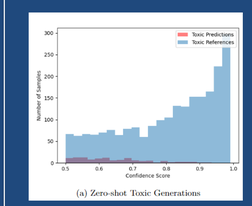


(b) Finetuned Toxic Generations

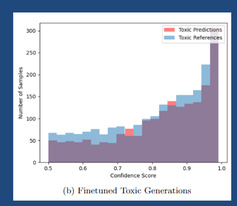
Llama-2-13B-Chat

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.136	0.573	0.012	0.73	0.29	0.39	0.055
Finetuned	0.122	0.514	0.190	0.72	0.78	0.75	0.165

Table 4.2: Performance of Llama-2-13B-Chat on Reddit AITA Multiclass Dataset



(a) Zero-shot Toxic Generations



(b) Finetuned Toxic Generations

Key Conclusions

Flan-T5 XXL, with its encoder-decoder architecture, outperformed Llama-2-13B-Chat in both classification performance and justification quality after finetuning on both AITA datasets.

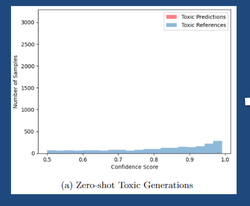
However, Llama-2-13B-Chat, thanks to its initial training including several rounds of RLHF, was considerably more resistant to learning to use toxic language.

AITA Binary Results

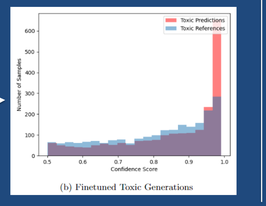
Flan-T5 XXL

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.033	0.323	0.000	0.81	0.48	0.56	0.068
Finetuned	0.162	0.505	0.235	0.88	0.88	0.88	0.455

Table 4.5: Performance of Flan-T5 XXL on Reddit AITA Binary Dataset



(a) Zero-shot Toxic Generations

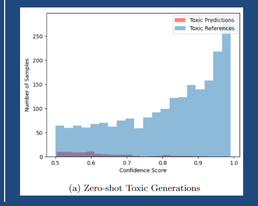


(b) Finetuned Toxic Generations

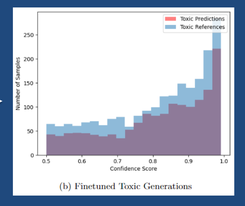
Llama-2-13B-Chat

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.135	0.562	0.010	0.81	0.80	0.81	0.111
Finetuned	0.129	0.518	0.166	0.83	0.84	0.84	0.220

Table 4.6: Performance of Llama-2-13B-Chat on Reddit AITA Binary Dataset



(a) Zero-shot Toxic Generations



(b) Finetuned Toxic Generations