The Efficacy of Finetuning Large Language Models for Interpersonal Conflict Resolution

Matthew Boraske
mb1020923@wcupa.edu

Faculty Advisor: Dr. Richard Burns
rburns@wcupa.edu

01

Overview

1. *Introduction to LLMs and Interpersonal Conflict Resolution*
2. *The Reddit AITA Datasets*
3. *Finetuning Methodology for Flan-T5-XXL and Llama-2-13B-Chat*
4. *Changes in Performance due to Finetuning*
5. *Recommended LLM Architecture and Finetuning Methodology*
6. *Conclusions and Future Work*

01

Introduction

What are Large Language Models (LLMs) and how are they being used?

What is interpersonal conflict resolution and why is it hard for LLMs?

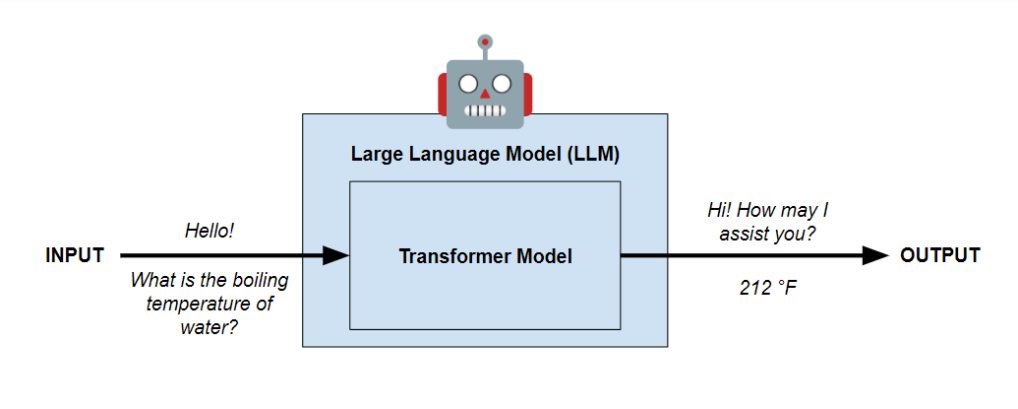
What are the contributions of this research?

What is a Large Language Model (LLM)?

- Artificial intelligence (AI) systems trained on extensive textual datasets to generate, understand, and interact with human language in a contextually appropriate manner.
- But how are LLMs implemented?

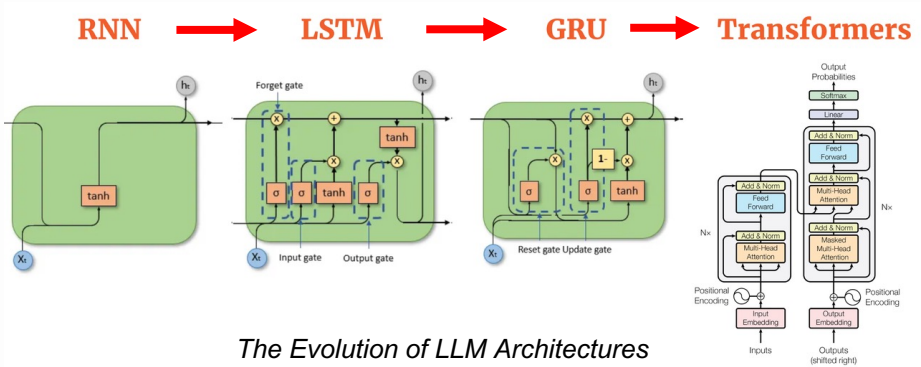
  … since 2017, via the transformer architecture!

Natural Language Processing (NLP)



Attention: The Key to The Transformer

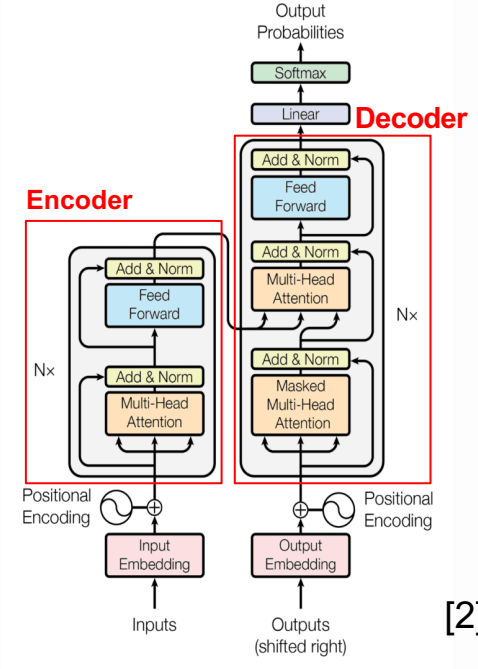
- Previously SOTA model for natural language processing (NLP) was Recurrent Neural Networks (RNNs) which learned relationships between elements in a sequential manner.
- RNNs are limited in learning complex relationship in larger texts due to the vanishing gradient problem.
- The attention mechanism in transformer models allows it to selectively focus on different parts of the input sequence when generating each output, enabling it to capture long-range dependencies and relationships in parallel.
- By attending to all positions in parallel, transformers directly learn relationships between any pair of tokens (pieces of text), regardless of their sequential distance



The Evolution of LLM Architectures [1]

The Two Types of Transformers

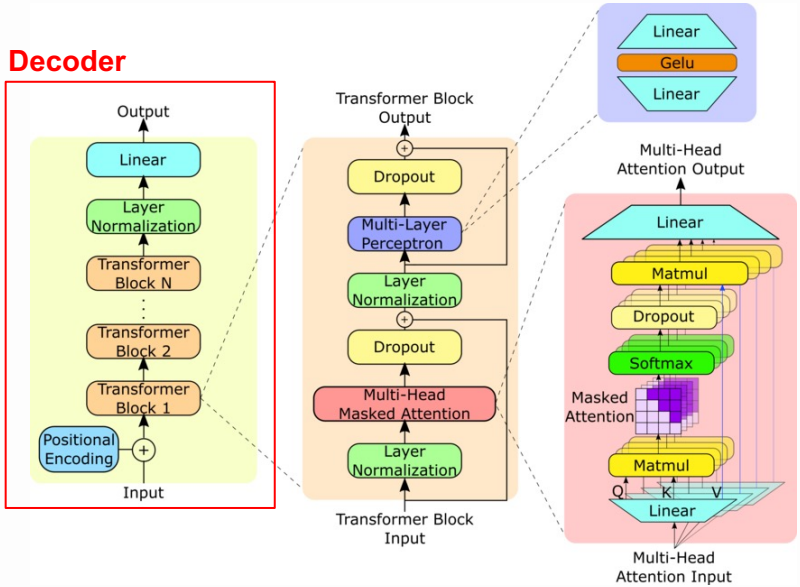
Encoder-Decoder (2017)



[2]

Original Transformer Architecture

Decoder-only (2019)



GPT Architecture

[3]

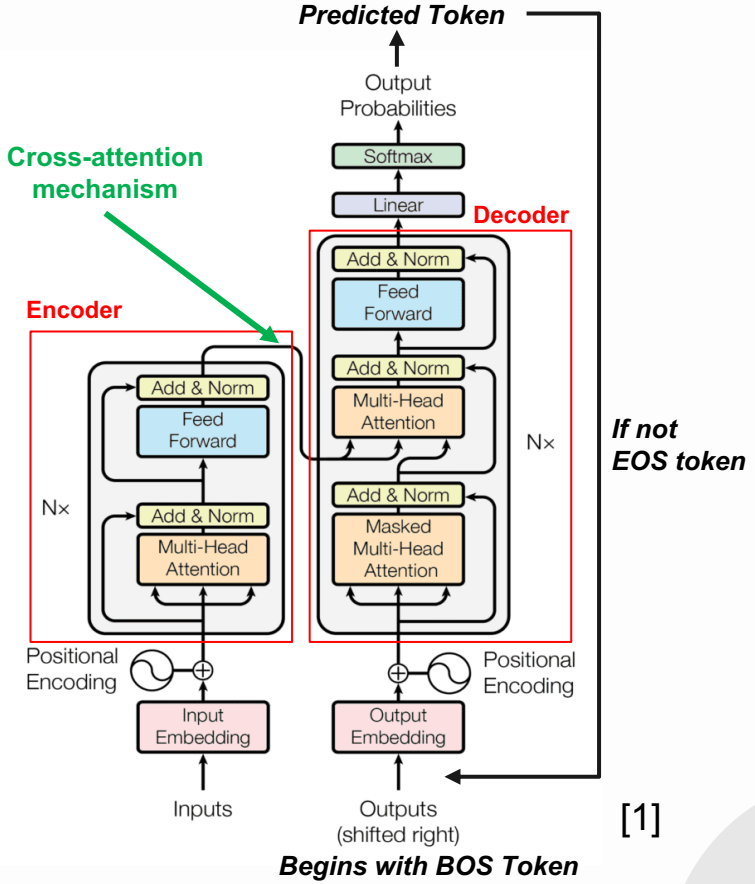
Encoder-Decoder

- Introduced by Vaswani et al. in 2017.

- Based on ***sequence-to-sequence translation***, where the encoder converts the input sequence into a fixed-length context vector, which is used by the decoder via a **cross-attention mechanism**

- Computes **multiple self-attention operations in parallel**, allowing the model to attend to different aspects of the input sequence simultaneously.

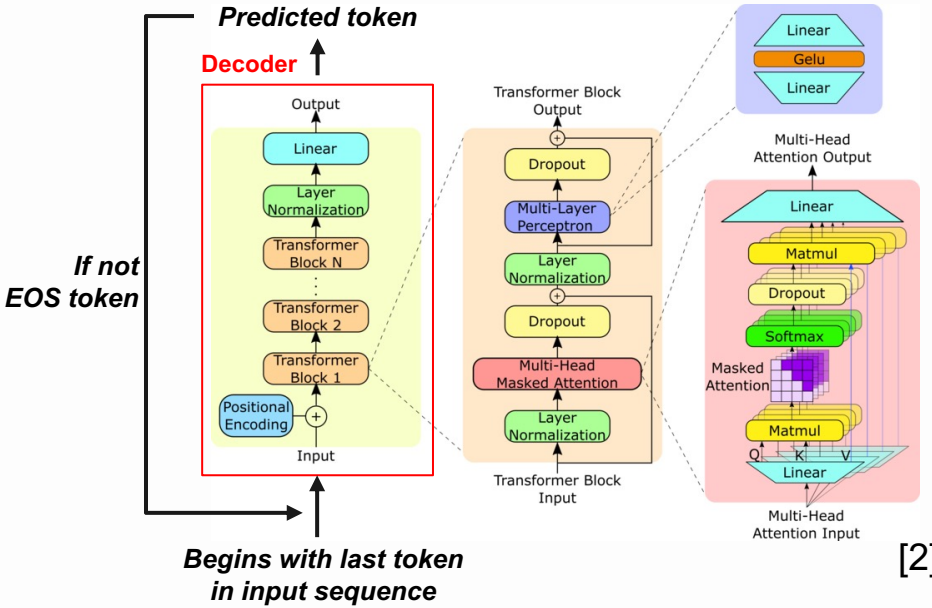
- Encoder-decoder models have access to the entire input sequence. The encoder self-attention mechanism allows each token to attend to **all other tokens** in the sequence, regardless of their position.



[1]

Decoder-only

- Introduced by Radford et al. in 2019 via the Generative Pretrained Transformer (GPT)
- Based on ***casual language modeling***, where the goal is to predict the next token in a sequence given all the previous tokens.
- When generating text using a decoder-only LLM, only the portion of the sequence that the model has seen so far is considered as context **(it can't attend to future tokens)**
- This can be problematic if the optimal response requires knowledge contained in future context.



[2]

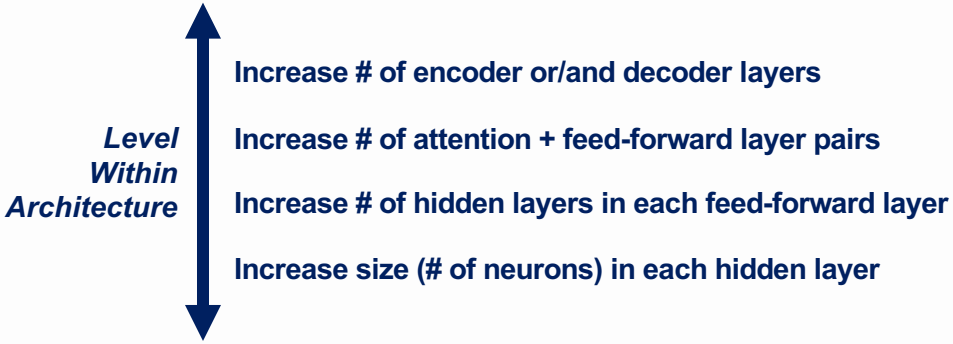
Why are Transformers Ideal for LLMs?

They are very scalable!

Why do we want to scale up LLMs?

1. Improved understanding in learning complex relationships
2. Improved performance on downstream tasks
3. Better at learning from less data
4. Able to handle larger vocabulary sizes and context window, which is the amount of information that is considered when generating the next token

Transformers can scale in 4 ways



Breaking down Interpersonal Conflict Resolution

What is it?

Relationship problems!

Why is this challenging for AI?

1. Ambiguity
2. Toxicity
3. Human Biases

What does interpersonal conflict resolution look like in practice?

1. A classification of behavior
2. A justification for this classification

Two Pressing Questions to Address

- Adapting a pre-trained LLM to a downstream task by training it on a smaller ***task-specific*** dataset
- A finetuned model ***combines*** pre-trained knowledge with task-specific knowledge

What transformer architecture should LLMs utilize for interpersonal conflict resolution?

What finetuning techniques should we apply to create an optimal AI agent?

Experimentation on a Real-Life Data

Encoder-Decoder LLM

Decoder-only LLM

Example Input

"I (f34) bought my boyfriend (m32) a 2018 MacBook Air … Two days ago, he got annoyed at me that I asked him to come to bed (it was 2 am), slammed the laptop shut and came. Then, he discovered the next day that his slamming the laptop shut broke the screen. He started demanding that I pay for it because "I provoked him" … AITA?"

Instruction Finetuning on Interpersonal Conflict Dataset

*r/amithea**hole*
(AITA)

Finetuned Model

Example Output

NTA. That is classic manipulative behavior. I'd watch out that he doesn't continue to blame you for his actions.

Classification
Justification

02

The Reddit AITA Datasets

How was the data collected, prepared, and analyzed for instruction finetuning?

Dataset Creation

Filtering for submissions that have…

1. a high community score
2. significant engagement.

Score = Upvotes - Downvotes

*r/amithea**hole*
2019-2022

Pushshift API

1,688,066 Submissions

72,119,169 Comments

50+ score

10+ score

High Quality Submissions and Their Top 10 Comments

4 Unique AITA Datasets

Reddit AITA Multiclass Top 2K
NTA | NAH | ESH | INFO | YTA

Filtering by score
Stratified to contain equal classes

Reddit AITA Multiclass
NTA | NAH | ESH | INFO | YTA

Reddit AITA Binary
NTA | YTA

Filtering by score
Stratified to contain equal classes

Reddit AITA Binary Top 2K
NTA | YTA

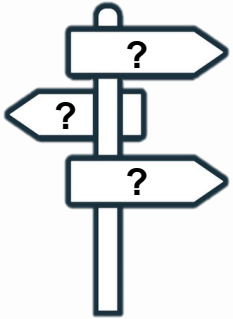
Datasets publicly released on HuggingFace 

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.

Dataset	Total Samples	YTA	NTA	ESH	NAH	INFO
Multiclass	50000	5576	40549	1331	1887	657
Multiclass-Top-2K	2000	400	400	400	400	400
Binary	46125	5576	40549	0	0	0
Binary-Top-2K	2000	1000	1000	0	0	0

Dataset Analysis

Ambiguity



Commenter Disagreement

Toxicity

Derogatory, vulgar, and racist, and bigoted language

Ambiguity Analysis

Disagreement between commenter AITA classifications can be evaluated on **three** different levels

Holistic

All ten commenters across entire dataset

↓

Krippendorff's Alpha

Pair-wise

Pairs of commenters across entire dataset

↓

Cohen's Kappa

Intra-sample

Classifications the ten commenters within each sample

↓

Custom Ambiguity Score

Holistic Commenter Agreement (Krippendorff's Alpha)

- Measures inter-rater reliability, or how much homogeneity, or consensus, there is in given ratings.
- Can be used for any number of raters >= 2

Scores below 0.8 signal moderate disagreement.

Dataset	Train Partition	Test Partition
Multiclass	0.731	0.737
Binary	0.752	0.759
Multiclass Top 2K	0.646	0.650
Binary Top 2K	0.844	0.832

$$\alpha = \frac{p_a - p_e}{1 - p_e}$$ [4]

- *α (Alpha): Krippendorff's alpha.*
- *p_0: The observed agreement among raters, aka the proportion of times that the raters agree, relative to what would be expected by chance.*
- *p_e: The expected agreement by random chance*

Pairwise Commenter Agreement (*Cohen's Kappa*)

- Similar to Krippendorrf's alpha, but measures agreement between two raters only.

Scores below 0.75 indicate moderate disagreement.

- *K: Cohen's Kappa*
- *P_0 : The observed agreement between the two raters.*
- *P_e : The expected agreement by random chance.*

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \quad [5]$$

	TC1	TC2	TC3	TC4	TC5	TC6	TC7	TC8	TC9	TC10
TC1	-	0.72	0.67	0.64	0.60	0.59	0.58	0.58	0.57	0.57
TC2	-	-	0.63	0.62	0.59	0.58	0.57	0.57	0.57	0.57
TC3	-	-	-	0.60	0.59	0.57	0.57	0.56	0.57	0.56
TC4	-	-	-	-	0.56	0.57	0.56	0.56	0.57	0.56
TC5	-	-	-	-	-	0.55	0.56	0.56	0.57	0.56
TC6	-	-	-	-	-	-	0.54	0.54	0.56	0.56
TC7	-	-	-	-	-	-	-	0.53	0.55	0.55
TC8	-	-	-	-	-	-	-	-	0.53	0.54
TC9	-	-	-	-	-	-	-	-	-	0.53
TC10	-	-	-	-	-	-	-	-	-	-

Reddit AITA Multiclass

	TC1	TC2	TC3	TC4	TC5	TC6	TC7	TC8	TC9	TC10
TC1	-	0.72	0.67	0.64	0.60	0.59	0.58	0.58	0.57	0.57
TC2	-	-	0.63	0.62	0.59	0.58	0.57	0.57	0.57	0.57
TC3	-	-	-	0.60	0.59	0.57	0.57	0.56	0.57	0.56
TC4	-	-	-	-	0.56	0.57	0.56	0.56	0.57	0.56
TC5	-	-	-	-	-	0.55	0.56	0.56	0.57	0.56
TC6	-	-	-	-	-	-	0.54	0.54	0.56	0.56
TC7	-	-	-	-	-	-	-	0.53	0.55	0.55
TC8	-	-	-	-	-	-	-	-	0.53	0.54
TC9	-	-	-	-	-	-	-	-	-	0.53
TC10	-	-	-	-	-	-	-	-	-	-

Reddit AITA Binary

	TC1	TC2	TC3	TC4	TC5	TC6	TC7	TC8	TC9	TC10
TC1	-	0.49	0.46	0.42	0.42	0.41	0.42	0.37	0.37	0.35
TC2	-	-	0.56	0.54	0.55	0.54	0.52	0.50	0.47	0.44
TC3	-	-	-	0.50	0.54	0.52	0.51	0.50	0.47	0.46
TC4	-	-	-	-	0.50	0.49	0.49	0.47	0.49	0.37
TC5	-	-	-	-	-	0.49	0.53	0.48	0.48	0.49
TC6	-	-	-	-	-	-	0.50	0.48	0.49	0.45
TC7	-	-	-	-	-	-	-	0.47	0.49	0.45
TC8	-	-	-	-	-	-	-	-	0.43	0.40
TC9	-	-	-	-	-	-	-	-	-	0.42
TC10	-	-	-	-	-	-	-	-	-	-

Reddit AITA Multiclass Top 2K
Cohen's Kappa Scores

Scores below 0.75 indicate moderate disagreement.

	TC1	TC2	TC3	TC4	TC5	TC6	TC7	TC8	TC9	TC10
TC1	-	0.85	0.84	0.80	0.80	0.78	0.77	0.71	0.71	0.68
TC2	-	-	0.79	0.77	0.76	0.75	0.74	0.70	0.70	0.66
TC3	-	-	-	0.76	0.77	0.76	0.74	0.72	0.70	0.67
TC4	-	-	-	-	0.72	0.73	0.74	0.69	0.69	0.56
TC5	-	-	-	-	-	0.73	0.74	0.69	0.71	0.66
TC6	-	-	-	-	-	-	0.72	0.71	0.68	0.65
TC7	-	-	-	-	-	-	-	0.67	0.70	0.64
TC8	-	-	-	-	-	-	-	-	0.66	0.63
TC9	-	-	-	-	-	-	-	-	-	0.63
TC10	-	-	-	-	-	-	-	-	-	-

Reddit AITA Binary Top 2K
Cohen's Kappa Scores

Scores below 0.40 indicate major disagreement.

Intra-Sample Commenter Agreement *(Ambiguity Score)*

- Custom metric that calculates the agreement between the raters within a sample
- Designed to account for the fact that some classifications are more similar than others.
- Normalized so that zero represents no ambiguity and one represents complete ambiguity

AITA Classification	*Ambiguity Score*
YTA	1
ESH	2
INFO	3
NAH	4
NTA	5

Similarity

Each of the ten classification is encoded into its integer score

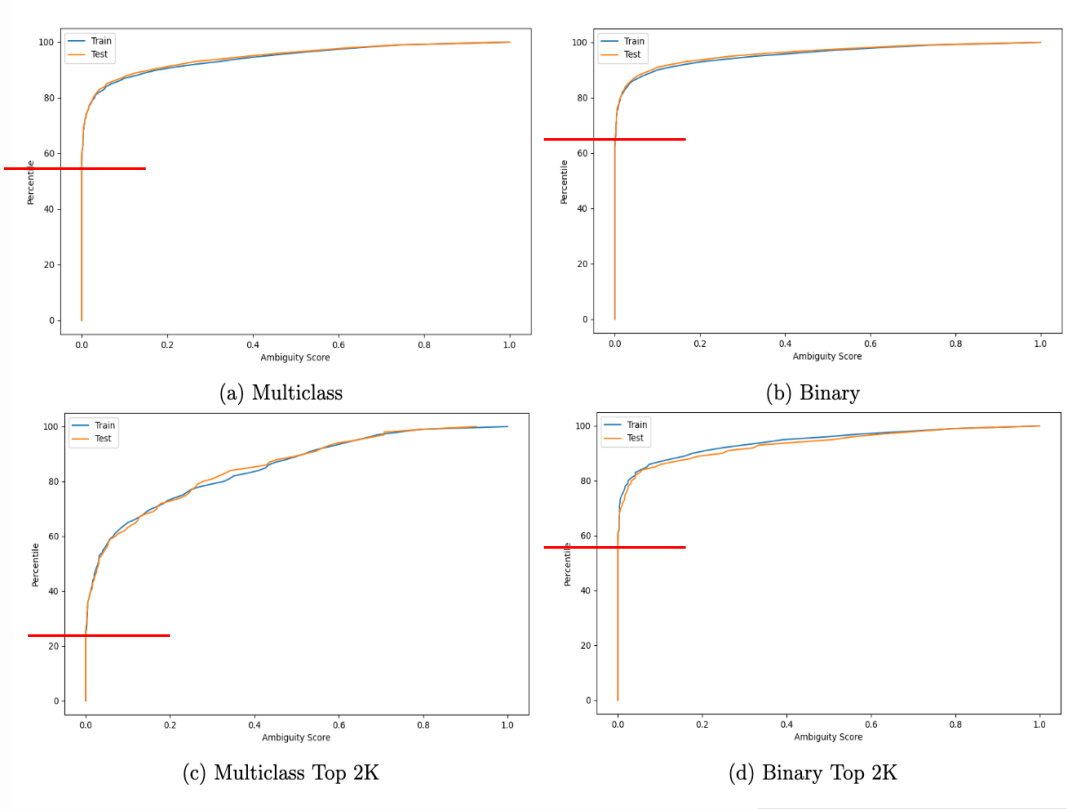
Dataset	Zero Ambiguity Rate
Multiclass	0.576
Binary	0.623
Multiclass Top 2K	0.242
Binary Top 2K	0.588

$$\text{Ambiguity Score} = \sigma \times (2 - |3 - \mu|)^2$$

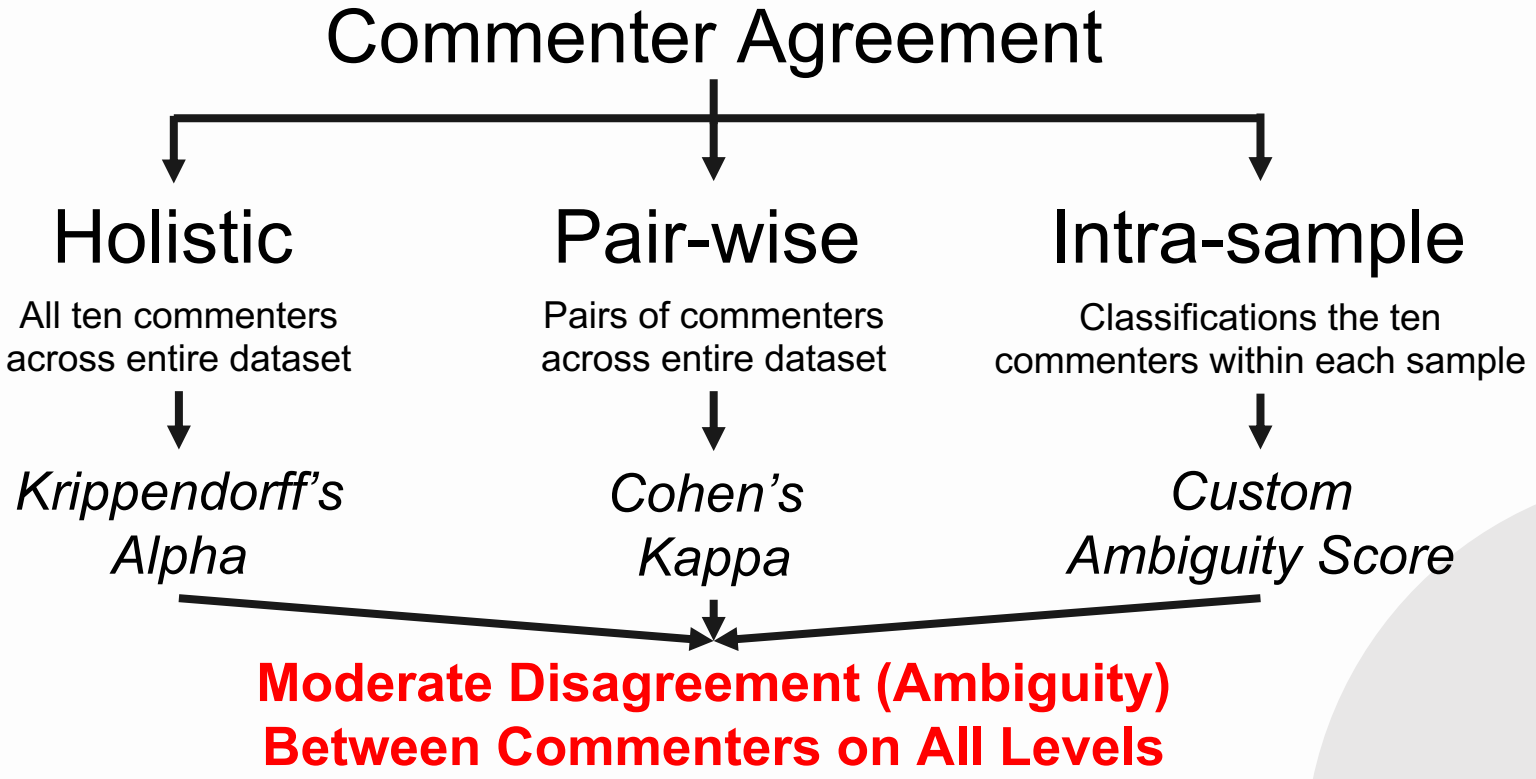
- **σ**: Standard deviation
- **μ**: Mean

At least some degree of disagreement between commenter classifications within samples for …

- **42.4%** of Multiclass dataset
- **37.3%** of Binary dataset
- **75.8%** of Multiclass Top 2K
- **41.2%** of Binary Top 2K

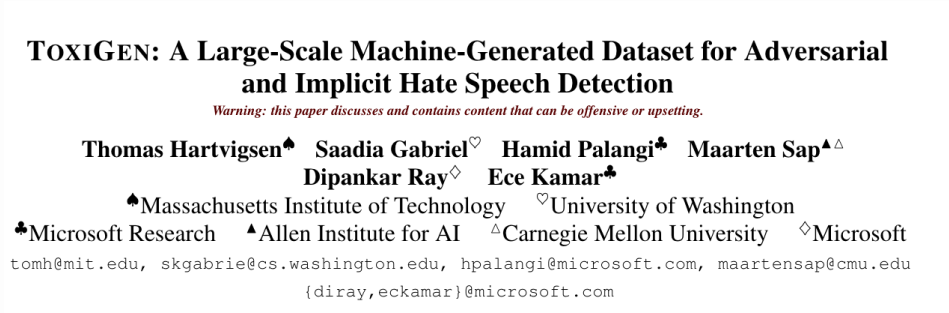


Bringing it all together…



Toxicity Analysis

- Utilized a RoBERTa model finetuned on the Toxigen dataset, which contains 274k samples of toxic and benign sentences mentioning thirteen minority groups



TOXIGEN: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection

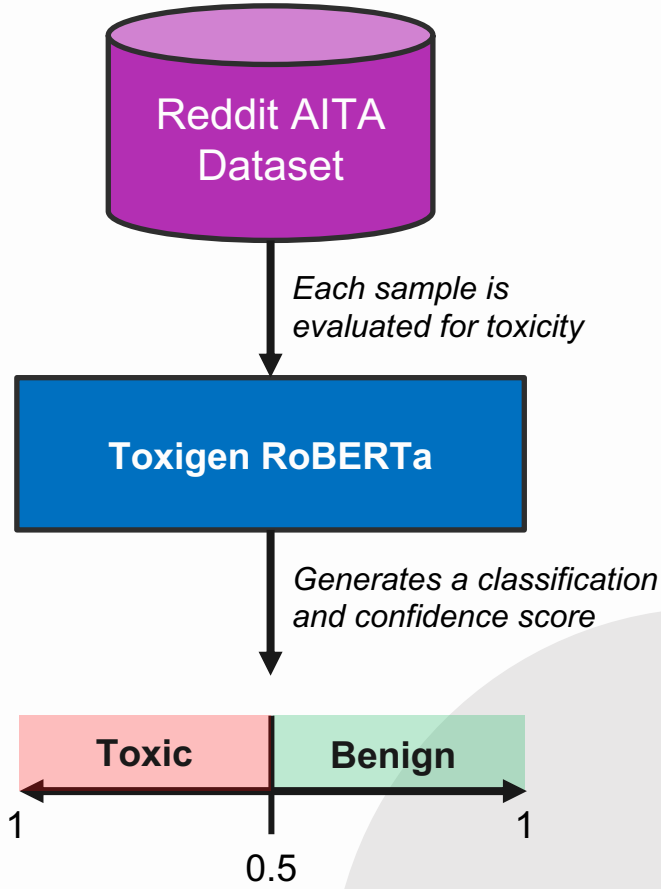
Warning: this paper discusses and contains content that can be offensive or upsetting.

Thomas Hartvigsen[♠] **Saadia Gabriel**[♡] **Hamid Palangi**[♣] **Maarten Sap**[▲△]
Dipankar Ray[◇] **Ece Kamar**[♣]

[♠]Massachusetts Institute of Technology [♡]University of Washington
[♣]Microsoft Research [▲]Allen Institute for AI [△]Carnegie Mellon University [◇]Microsoft

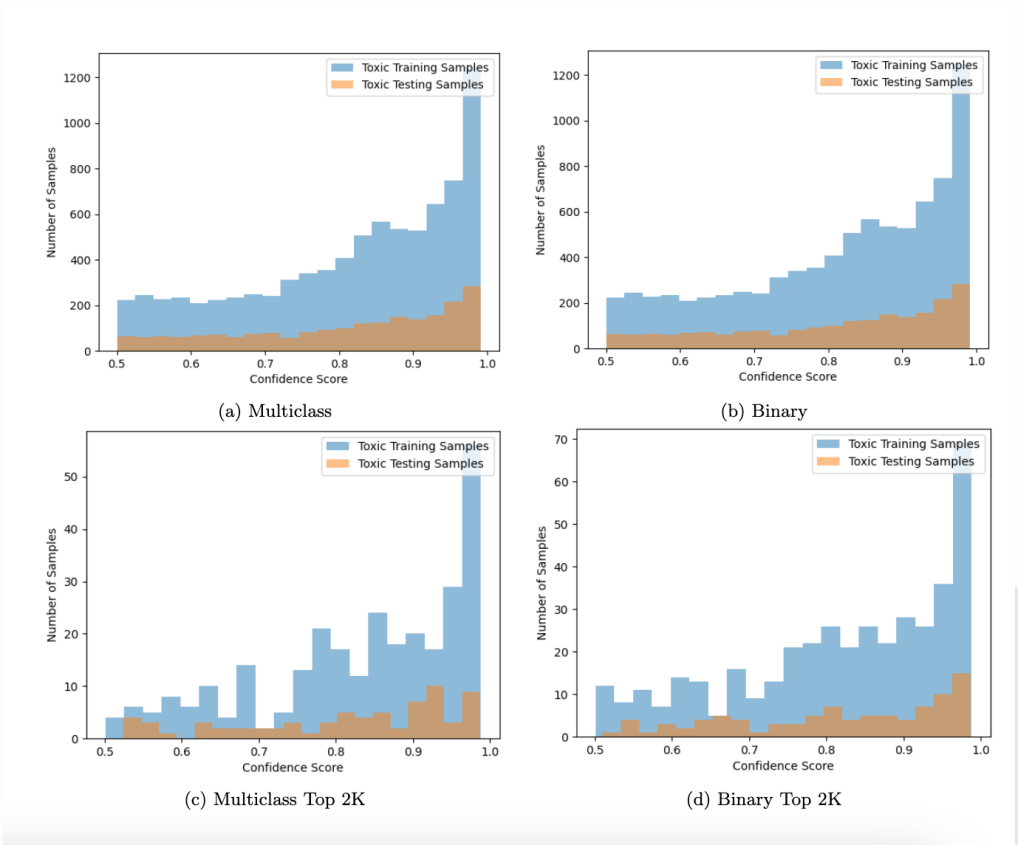
tomh@mit.edu, skgabrie@cs.washington.edu, hpalangi@microsoft.com, maartensap@cmu.edu
{diray,eckamar}@microsoft.com

Published at ACL (Association for Computation Linguistics) 2022



Approximately **a fifth of all samples** in each dataset were classified as **toxic**

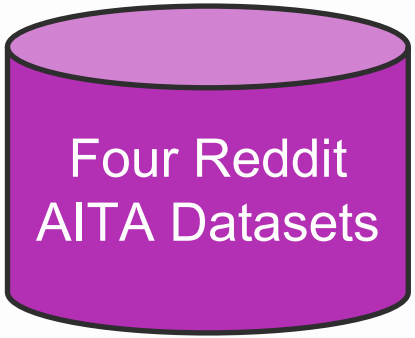
Dataset	Train Partition	Test Partition
Multiclass	0.219	0.224
Multiclass Top 2K	0.182	0.178
Binary	0.225	0.231
Binary Top 2K	0.253	0.232



(a) Multiclass (b) Binary

(c) Multiclass Top 2K (d) Binary Top 2K

Insights from Dataset Analysis



Are the behavior classifications ambiguous? **Yes**

Do the samples contain toxic language? **Yes**

These qualities imply that the samples are fair representatives of real-life interpersonal conflicts.

How can we mitigate the issues that appear when training on data with these qualities?

03

Finetuning Procedure

What considerations were taken to facilitate finetuning?

How was instruction finetuning completed for Flan-T5 (encoder-decoder) versus Llama-2-Chat (decoder-only)?

How methods were used to complete finetuning?

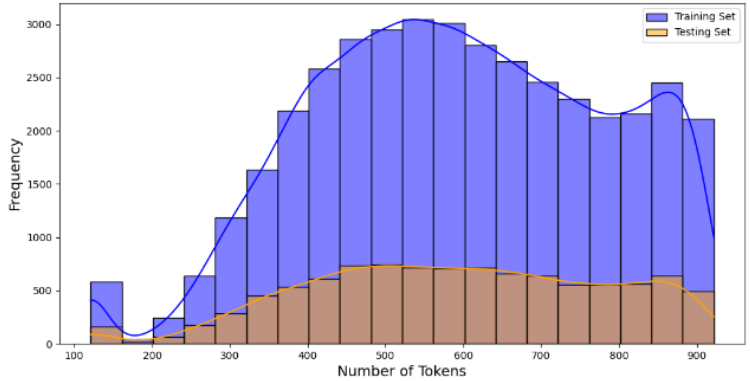
Context Window Consideration

The amount of information that is considered when generating the next token

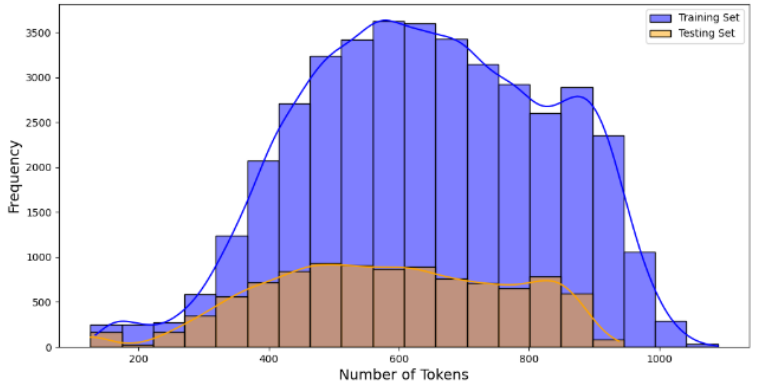
Encoder: 1024 tokens
Decoder: 256 tokens

4096 tokens

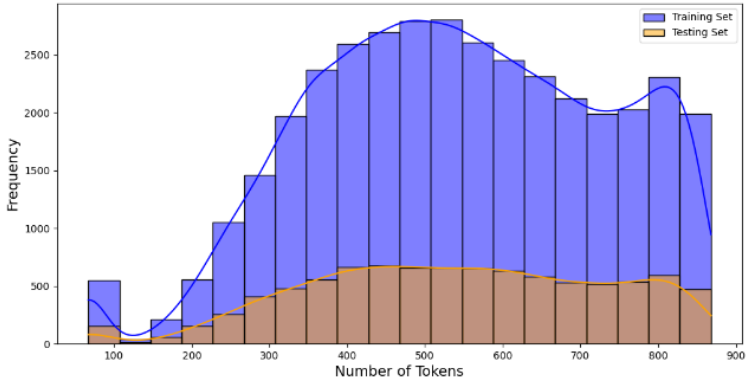
Do all the input sequences (instruction + submission texts) fit within these context windows?



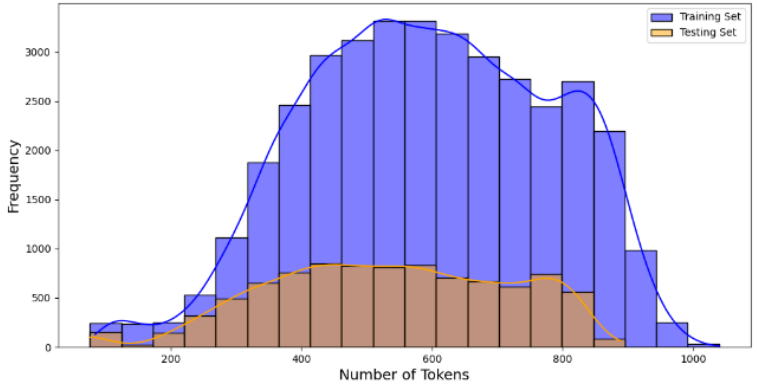
(a) Flan-T5 Token Counts (Multiclass)



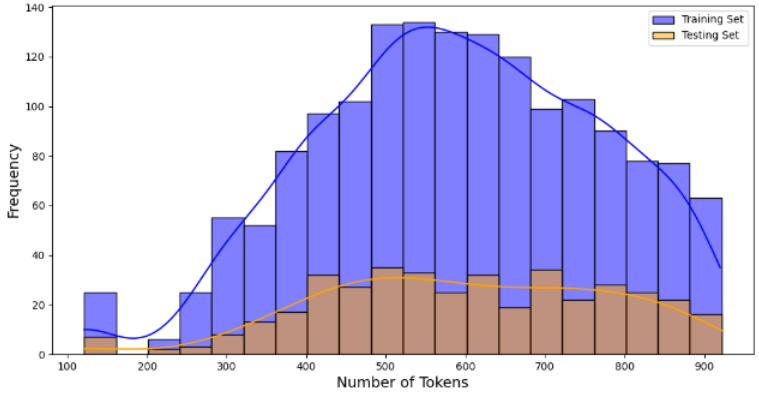
(b) Llama-2-Chat Token Counts (Multiclass)



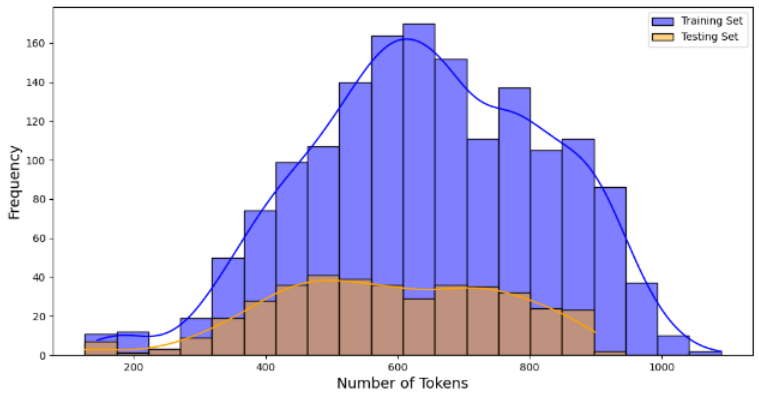
(c) Flan-T5 Token Counts (Binary)



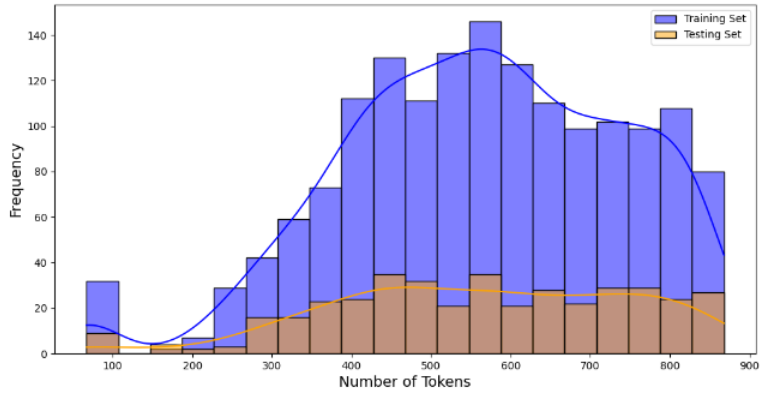
(d) Llama-2-Chat Token Counts (Binary)



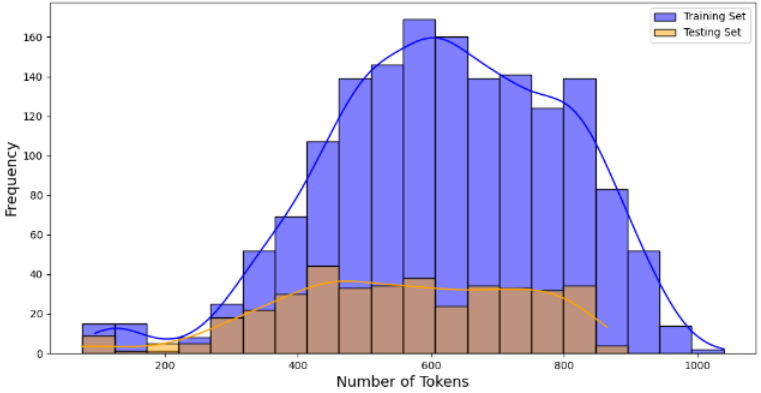
(e) Flan-T5 Token Counts (Multiclass Top 2K)



(f) Llama-2-Chat Token Counts (Multiclass Top 2K)



(g) Flan-T5 Token Counts (Binary Top 2K)



(h) Llama-2-Chat Token Counts (Binary Top 2K)

Instruction Finetuning

- To adapt a model to behave in a certain manner, an **instruction** can be added to each sample in the finetuning data

Flan-T5 Input: **Instruction** + **Sample Input**

Llama-2-Chat: <s>[INST] <<SYS>> {**Instruction**} <</SYS>> {**Sample Input**} [/INST]

Multiclass Instruction	Binary Instruction
Classify the interpersonal conflict and explain your choice. Your choices are limited to you're the a**hole (yta) when the writer is causing the conflict, not the a**hole (nta) when someone other than the writer is causing the conflict, no a**holes here (nah) when no one is causing the conflict, everyone sucks here (esh) when everyone is causing the conflict, more information needed (info) when a classification cannot be made with the context provided.	Classify the interpersonal conflict as either 'you're the a**hole (yta)' when the writer is causing the conflict or 'not the a**hole (nta)' when someone other than the writer is causing the conflict and explain your choice.

Example Flan-T5 and Llama-2-Chat Prompts

Flan-T5 Prompt	Llama-2-Chat Prompt
Classify the interpersonal conflict as either 'you're the a**hole (yta)' when the writer is causing the conflict or 'not the a**hole (nta)' when someone other than the writer is causing the conflict and explain your choice. Conflict context: we have a 4 month old daughter, and I'm a stay at home dad. My wife and I used to take the same train into the city for work, so we woke up at the same time. But since I don't work anymore, there's no reason to wake up that early. She wakes up around 6am every work day, and I now wake up around 7:45 or whenever our daughter wakes me up in the morning. She's been getting upset with me saying that it's not fair that I get to sleep in while she still has to wake up. I do most of the night care so it seems fair to me. She wants me to wake up with her every morning and I said no. She's getting more mad now. AITA?	<s>[INST] <SYS> Classify the interpersonal conflict as either 'you're the a**hole (yta)' when the writer is causing the conflict or 'not the a**hole (nta)' when someone other than the writer is causing the conflict and explain your choice. </SYS> We have a 4 month old daughter, and I'm a stay at home dad. My wife and I used to take the same train into the city for work, so we woke up at the same time. But since I don't work anymore, there's no reason to wake up that early. She wakes up around 6am every work day, and I now wake up around 7:45 or whenever our daughter wakes me up in the morning. She's been getting upset with me saying that it's not fair that I get to sleep in while she still has to wake up. I do most of the night care so it seems fair to me. She wants me to wake up with her every morning and I said no. She's getting more mad now. AITA? </s> [/INST]

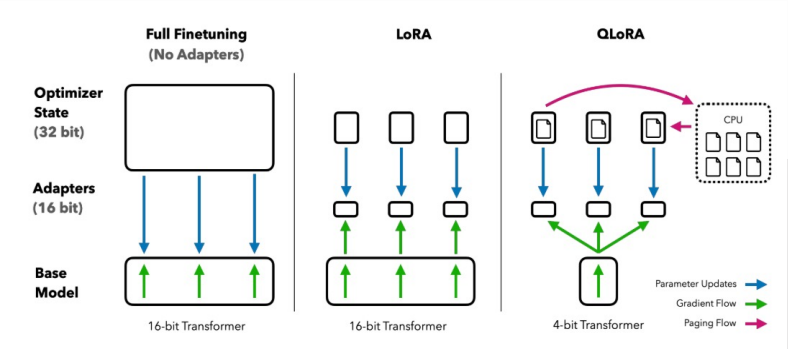
How was the finetuning completed?

QLoRA

QLoRA (Quantized Low Rank Adaption)

- Allows for the ***efficient finetuning*** of quantized 4-bit models without performance degradation
- During the training process, larger weight matrices are quantized into low-rank matrices. These are learned and subsequently applied to the original weight matrices via multiplication and addition. **The primary model weights are not trained.**
- Methods exist to merge these low rank matrices with the larger weight matrices
- Using QLoRA, one can fine-tune a 65B parameter model on a single 48GB GPU, a task which previously required 780 GB of memory.

Using QLoRA allowed us to train all the Flan-T5-XXL and Llama-2-13B-Chat models on a single Nvidia L40 GPU which has 48GB of memory.



Flan-T5-XXL: 11.3B parameters
Llama-2-13B-Chat: 13B parameters

04

Finetuning Results

Which model was more effective at learning the AITA interpersonal conflict resolution task?

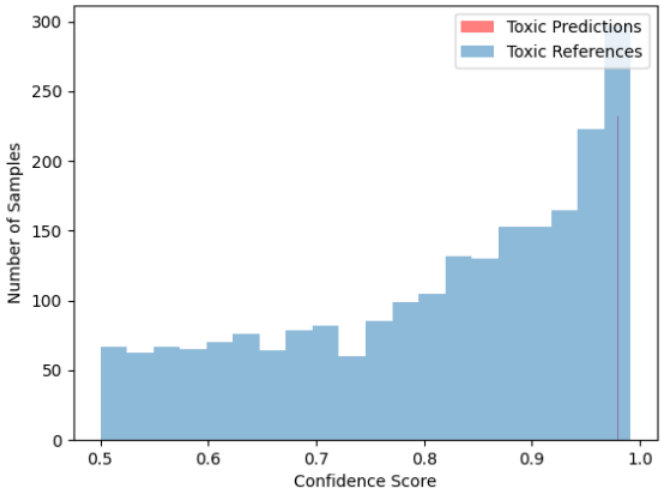
How resistive was each model to generating toxic language after finetuning?

Can the models effectively learn the task when finetuned on the smaller, two-thousand sample datasets?

Finetuning on Reddit AITA Multiclass

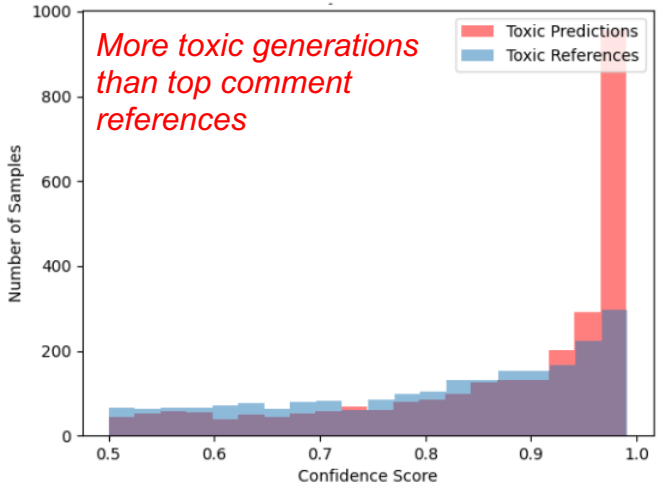
Reddit AITA Multiclass: Flan-T5 XXL

Change in Toxic Generations



(a) Zero-shot Toxic Generations

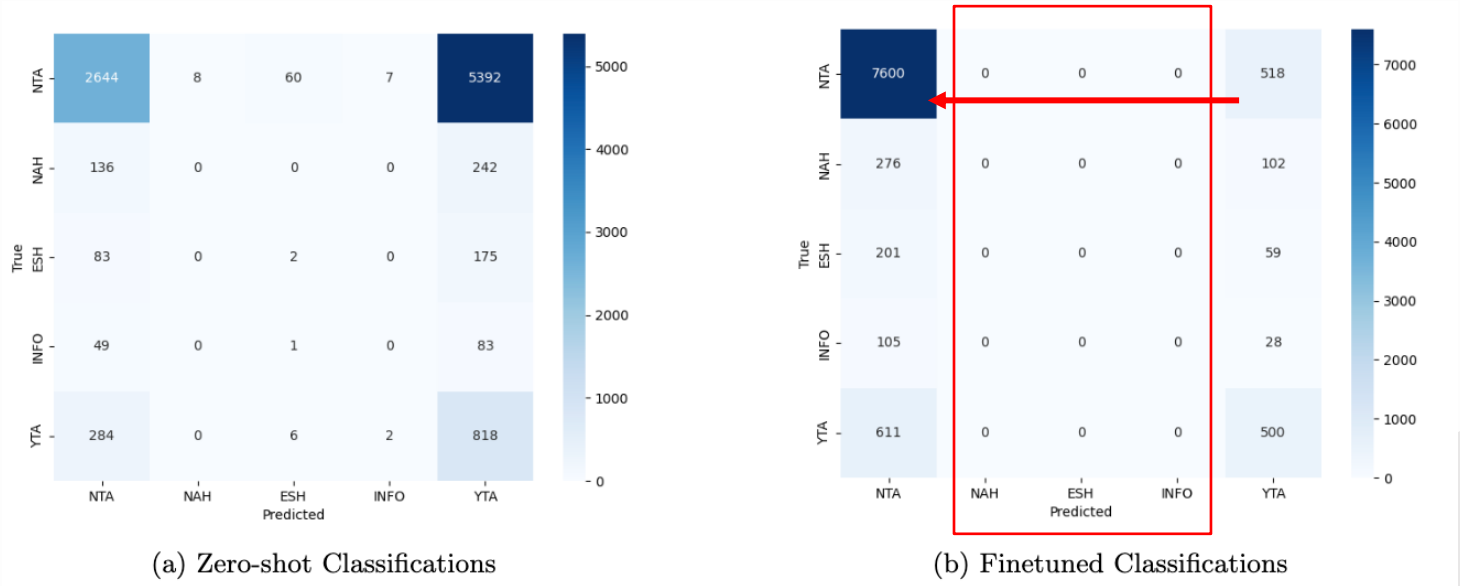
20.5%



(b) Finetuned Toxic Generations

Change in AITA Classifications

Reduced to only selecting NTA/YTA (binary)



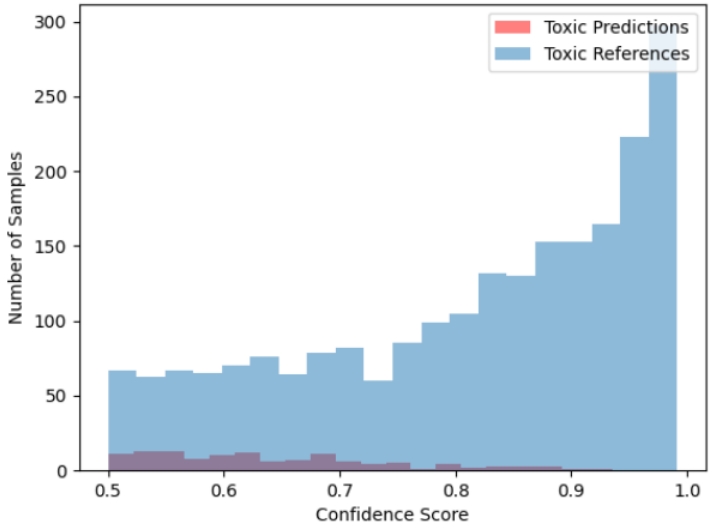
(a) Zero-shot Classifications

(b) Finetuned Classifications

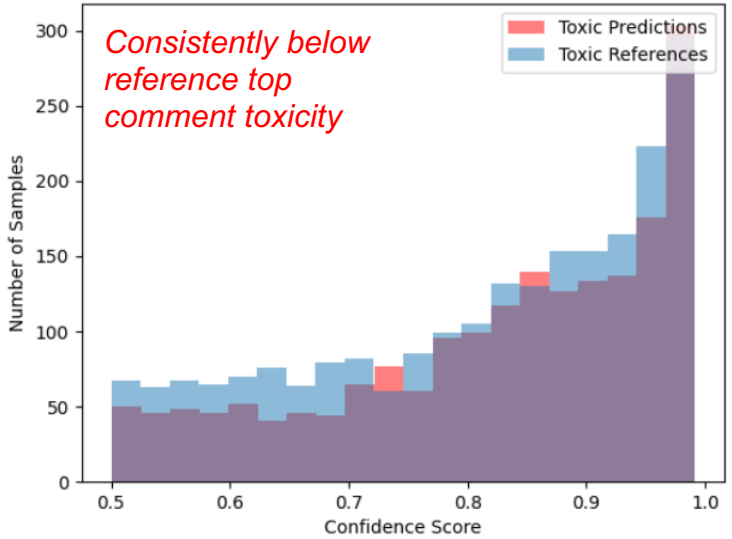
Reddit AITA Multiclass: Llama-2-13B-Chat

Change in Toxic Generations

17.8%



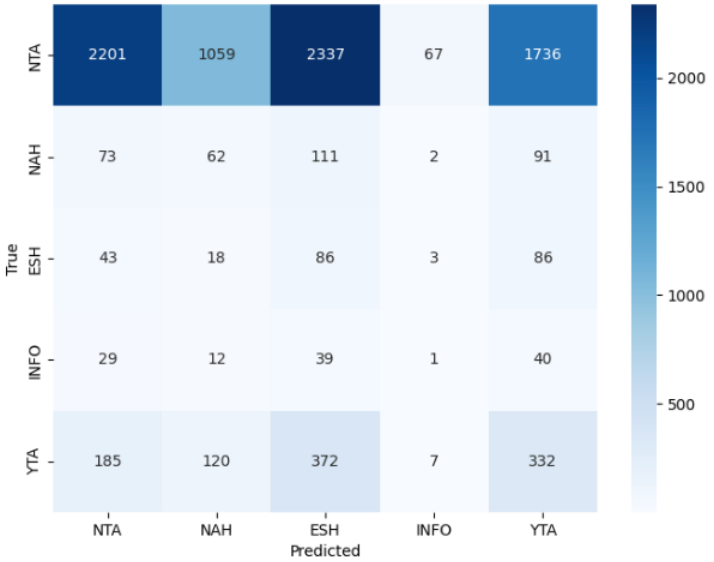
(a) Zero-shot Toxic Generations



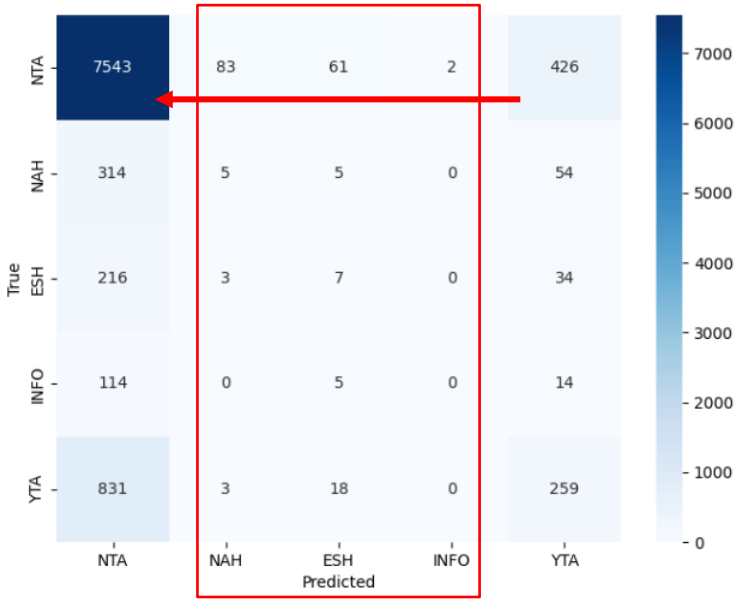
(b) Finetuned Toxic Generations

Change in AITA Classifications

Major, but not absolute, shift to NTA or YTA



(a) Zero-shot Classifications



(b) Finetuned Classifications

Reddit AITA Multiclass Performance Comparison

Reddit AITA Multiclass Reference Top Comment Toxicity Rate: 0.224

Finetuned Flan-T5 XXL is above this.

Finetuned Llama-2-13B-chat is below this.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$ [6]

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

Llama-2-13B-Chat

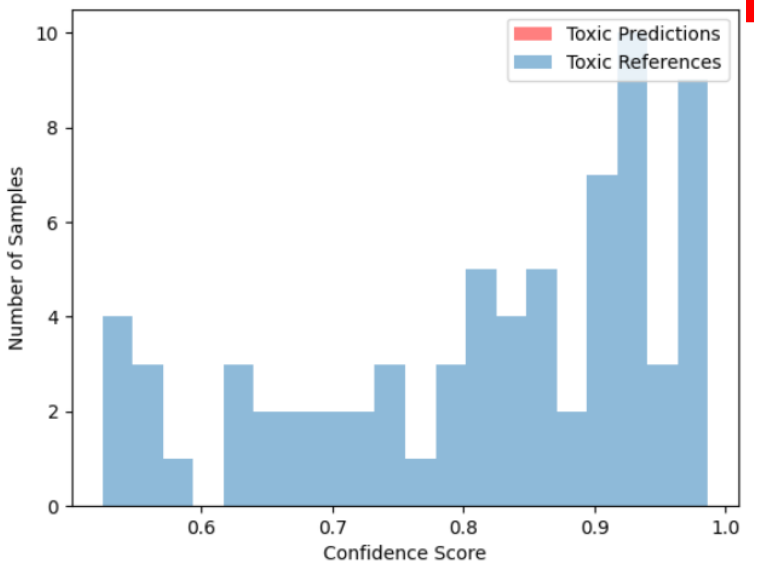
Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.136	0.573	0.012	0.73	0.29	0.39	0.055
Finetuned	0.122	0.514	0.190	0.72	0.78	0.75	0.165

Finetuning on Reddit AITA Multiclass Top 2K

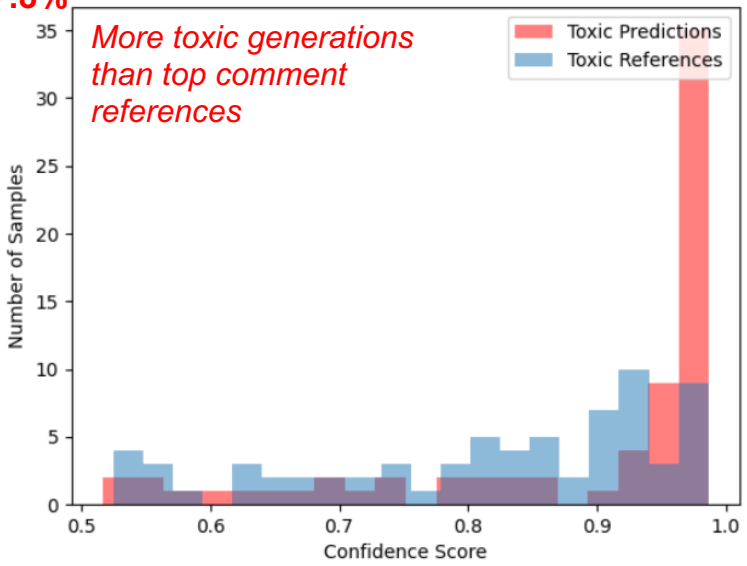
Reddit AITA Multiclass Top 2K: Flan-T5 XXL

Change in Toxic Generations

17.8%



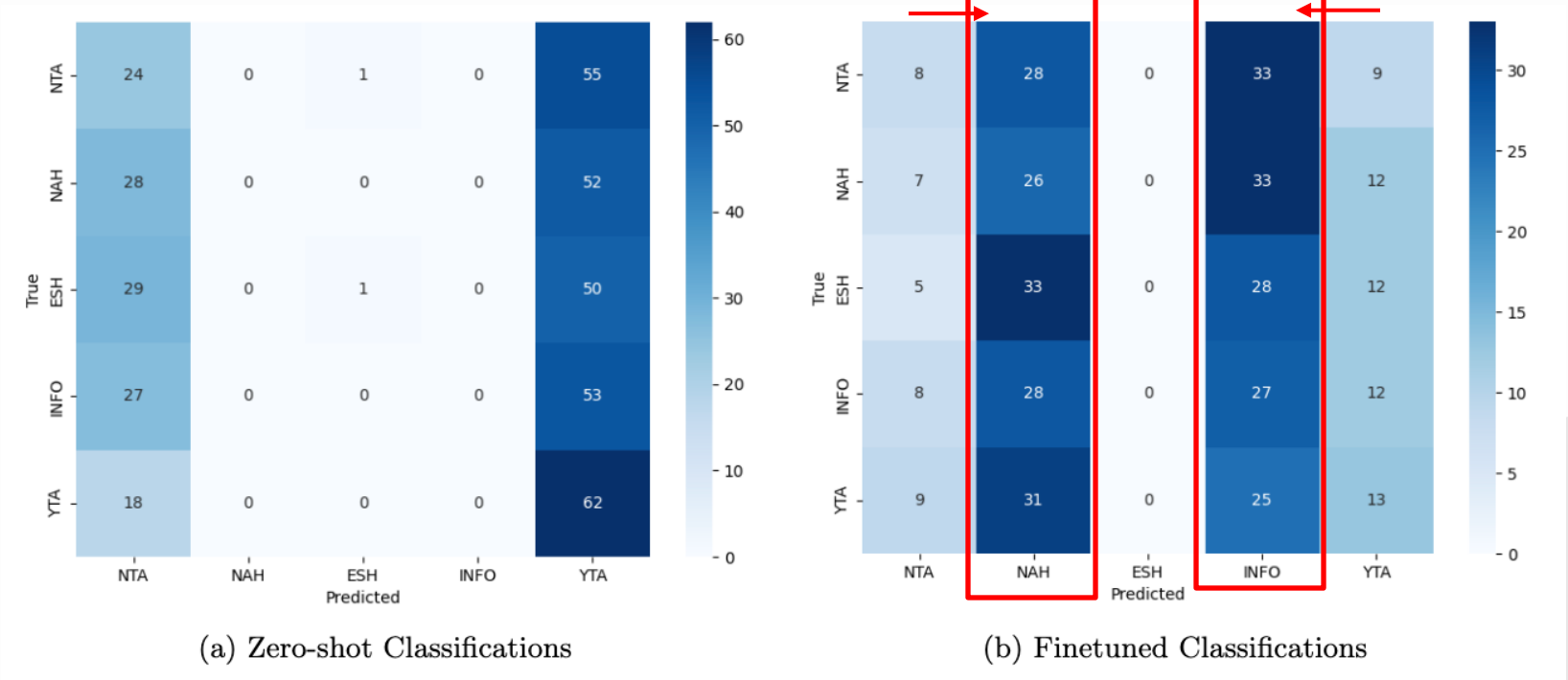
(a) Zero-shot Toxic Generations



(b) Finetuned Toxic Generations

Model learned to select NAH and INFO heavily, but not accurately

Change in AITA Classifications



(a) Zero-shot Classifications

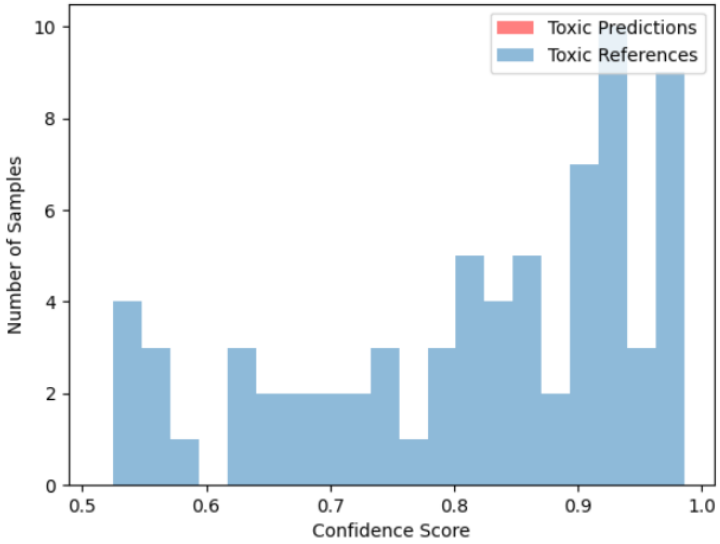
(b) Finetuned Classifications

Reddit AITA Multiclass Top 2K: Llama-2-13B-Chat

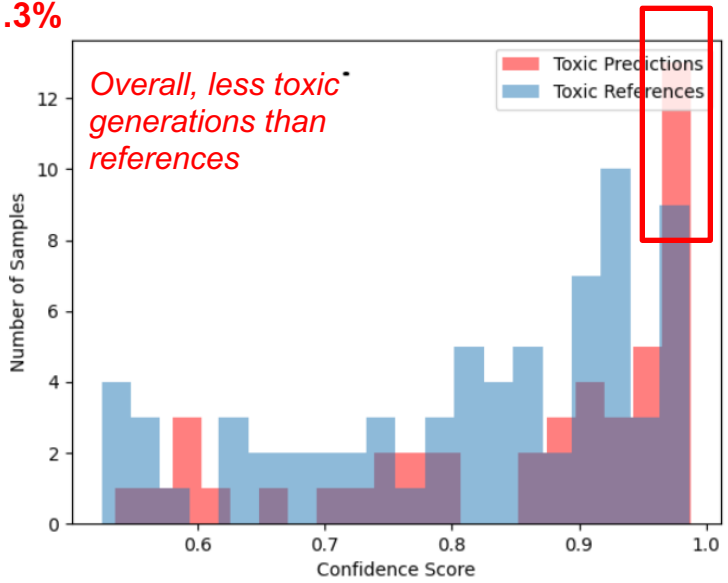
Change in Toxic Generations

However, does exceed in number of highly confident toxic generations

11.3%



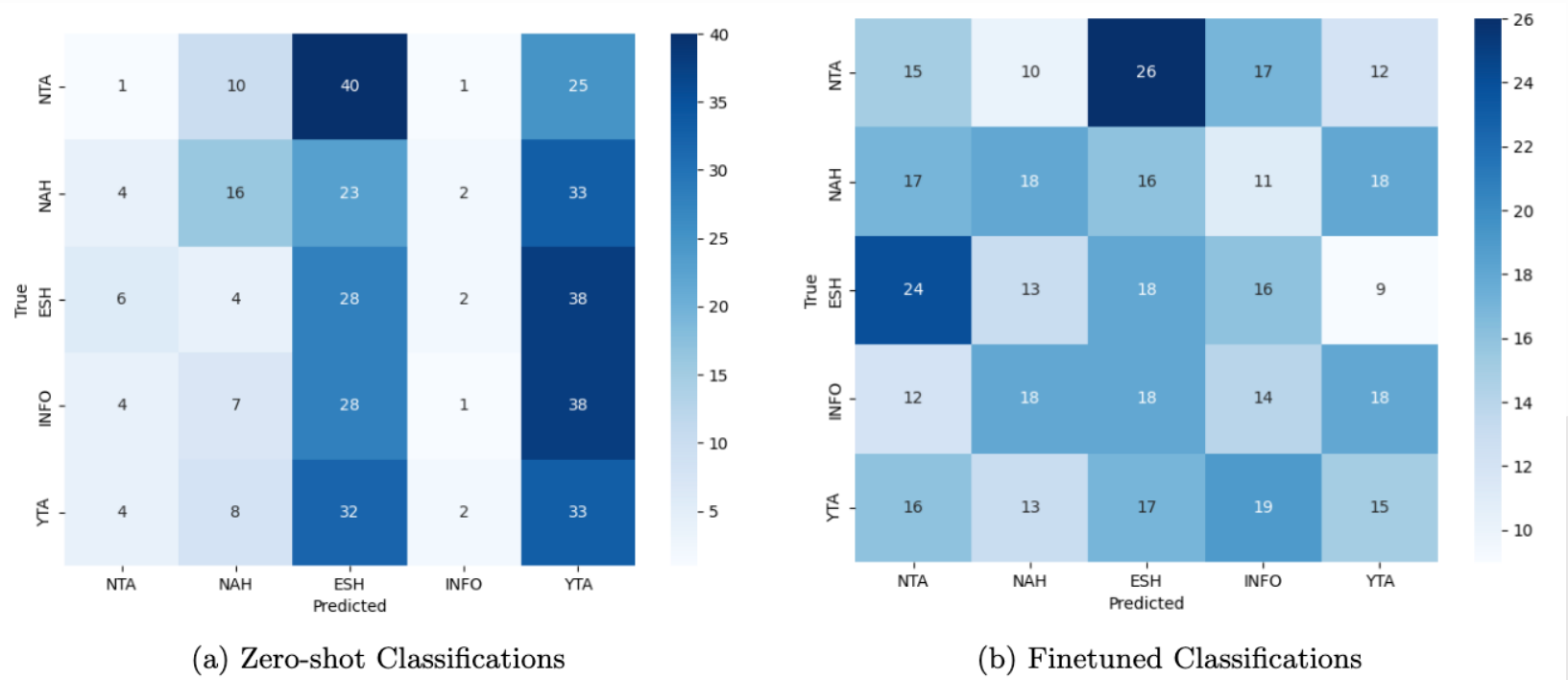
(a) Zero-shot Toxic Generations



(b) Finetuned Toxic Generations

Change in AITA Classifications

Classifications are scattered with no clear pattern



(a) Zero-shot Classifications

(b) Finetuned Classifications

Reddit AITA Multiclass Top 2K Performance Comparison

Reference Top Comment Toxicity Rate: 0.178

No clear signs of improvement at expense of learning toxic language

Flan-T5-XXL

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.016	0.308	0.000	0.18	0.22	0.12	0.030
Finetuned	0.101	0.448	0.178	0.16	0.19	0.16	-0.010

Llama-2-13B-Chat

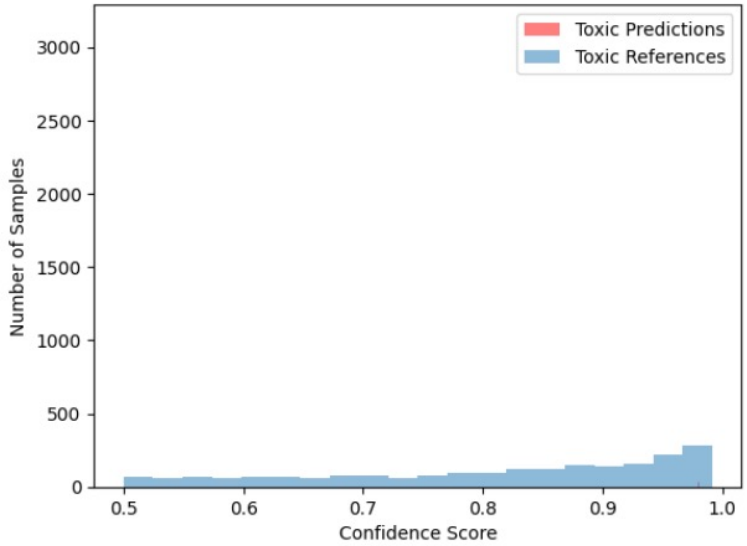
Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.108	0596	0.000	0.18	0.20	0.16	0.002
Finetuned	0.103	0.515	0.113	0.20	0.20	0.20	0.000

Finetuning on Reddit AITA Binary

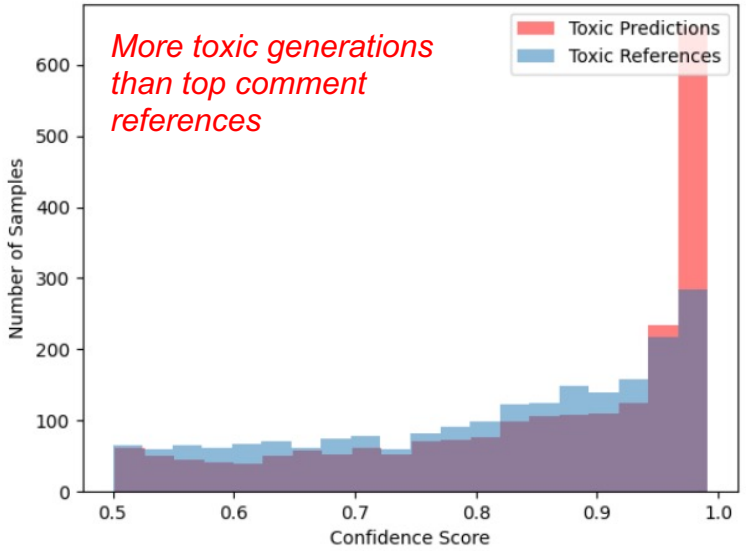
Reddit AITA Binary: Flan-T5 XXL

Change in Toxic Generations

23.5%



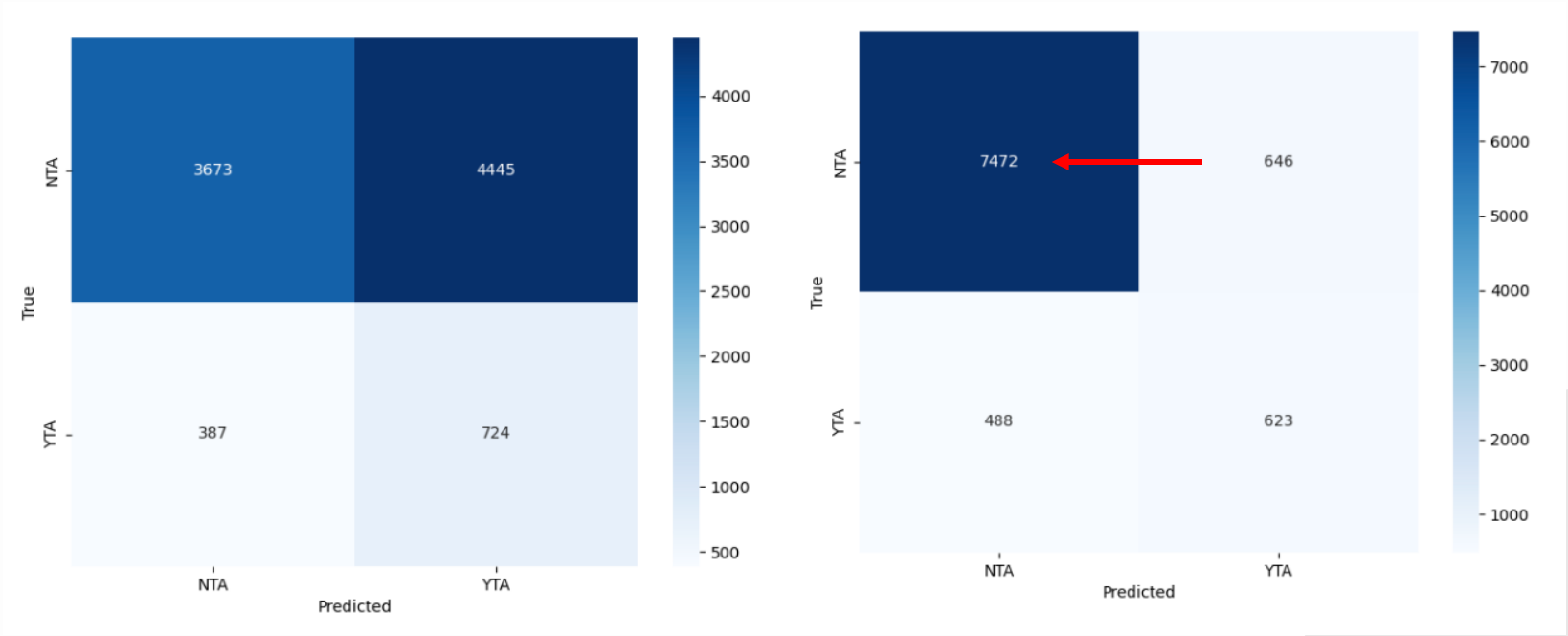
(a) Zero-shot Toxic Generations



(b) Finetuned Toxic Generations

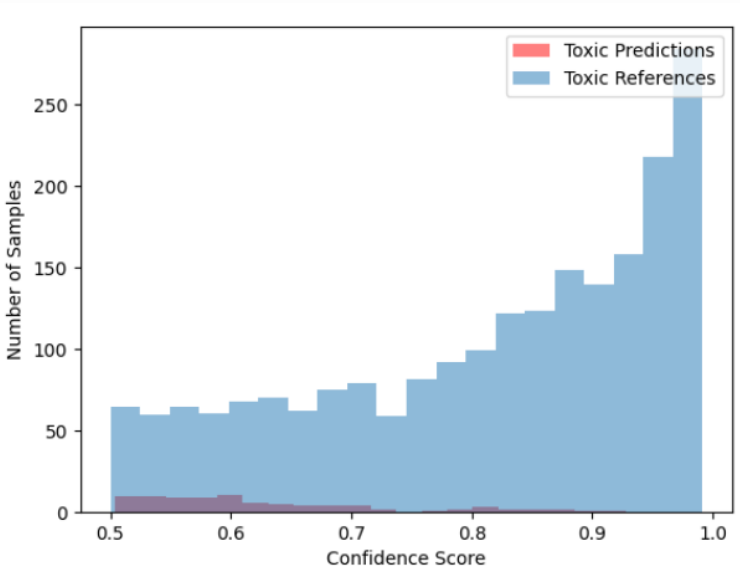
Change in AITA Classifications

Majority of improvement came from learning to correctly identify NTA conflicts instead of misclassifying them as YTA



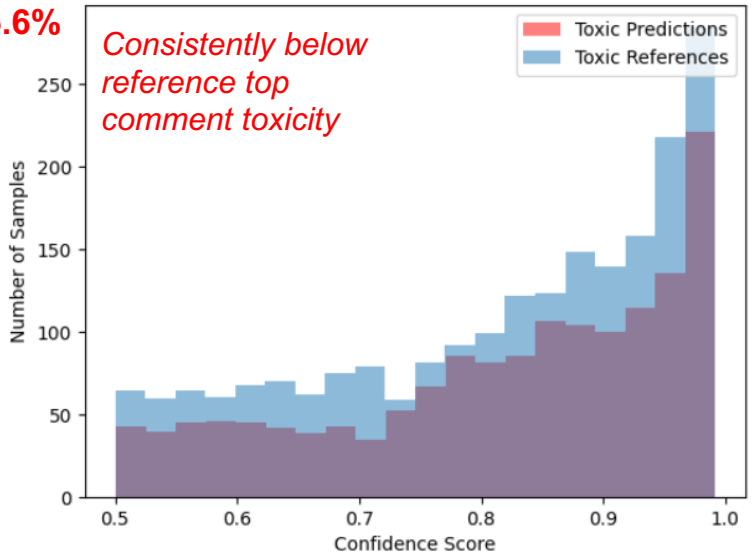
Reddit AITA Binary: Llama-2-13B-Chat

Change in Toxic Generations



(a) Zero-shot Toxic Generations

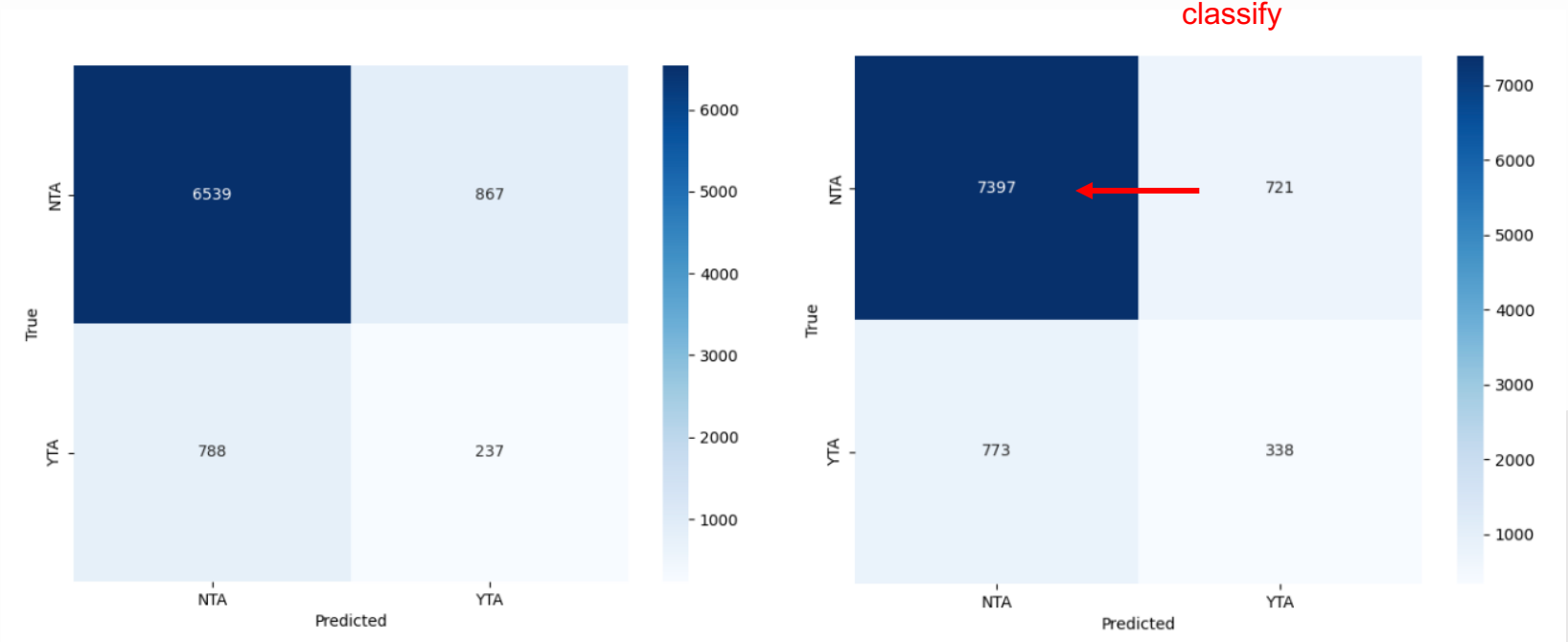
15.6%



(b) Finetuned Toxic Generations

Change in AITA Classifications

1. Correctly identifying NTA conflicts instead of misclassifying them as YTA
2. Giving NTA classifications to conflicts it previously failed to classify



Reddit AITA Binary Performance Comparison

- Best performance across any model in classification was Flan-T5-XXL on this dataset

- Flan-T5-XXL achieved higher ROUGE Lsum, but base Llama-2-13B-Chat still maintained highest COMET score

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

Llama-2-13B-Chat

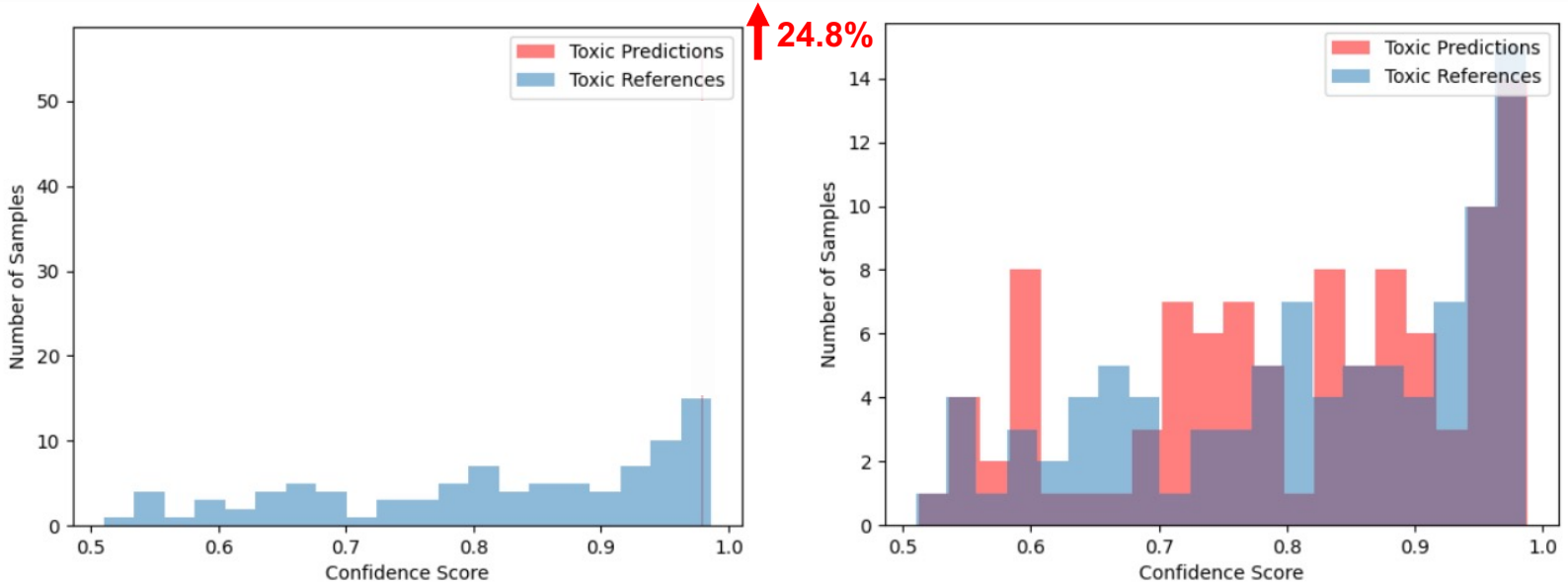
Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.135	0.562	0.010	0.81	0.80	0.81	0.111
Finetuned	0.129	0.518	0.166	0.83	0.84	0.84	0.220

Reference Top Comment Toxicity Rate: 0.231

Finetuning on Reddit AITA Binary Top 2K

Reddit AITA Binary Top 2K: Flan-T5 XXL

Change in Toxic Generations

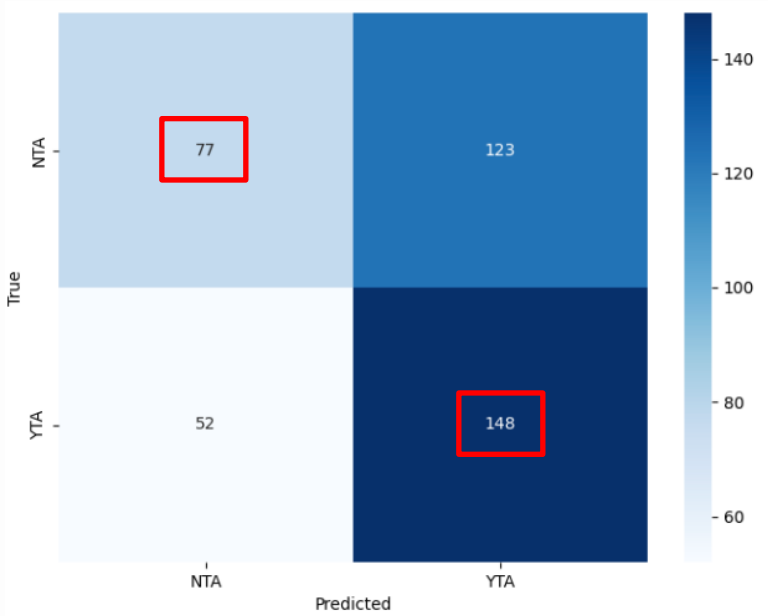


(a) Zero-shot Toxic Generations

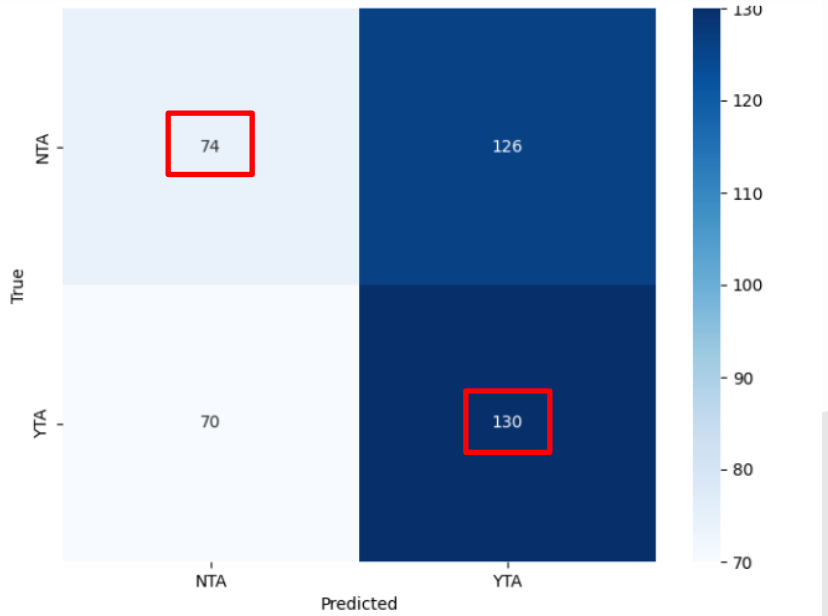
(b) Finetuned Toxic Generations

Change in AITA Classifications

Decreases in accuracy for both NTA and YTA classifications



(a) Zero-shot Classifications

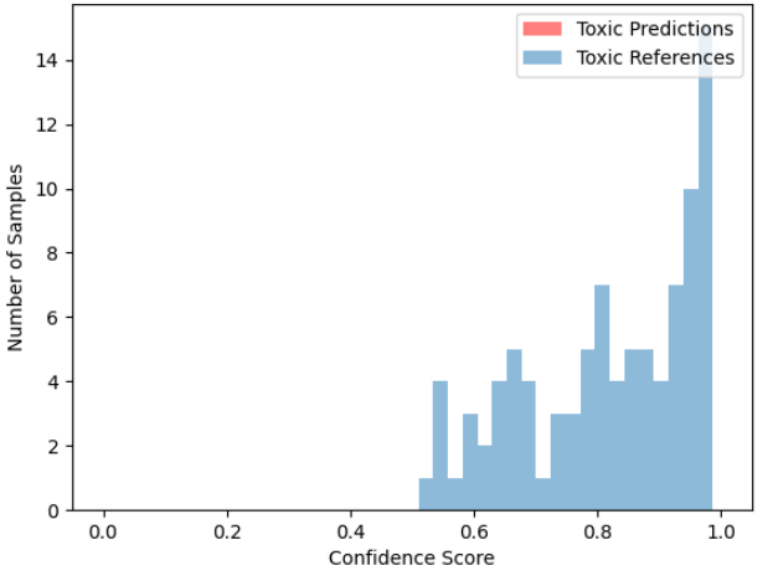


(b) Finetuned Classifications

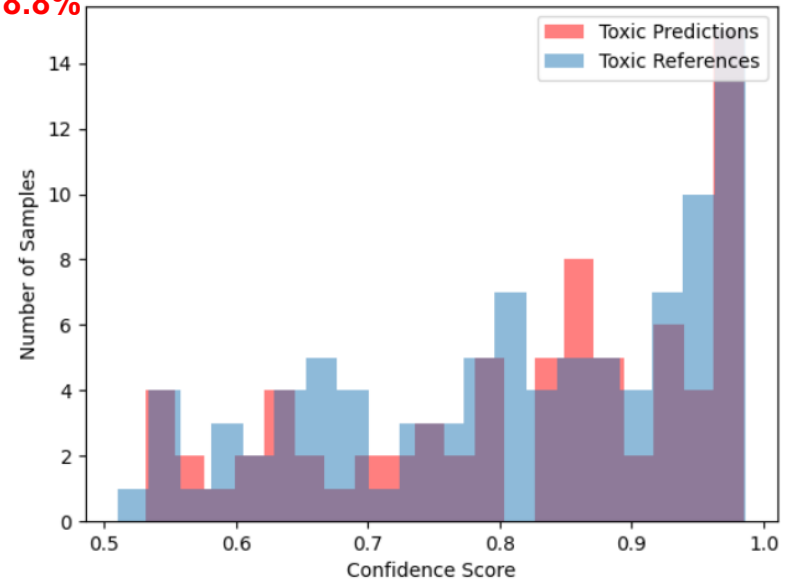
Reddit AITA Binary Top 2K: Llama-2-13B-Chat

Change in Toxic Generations

18.8%



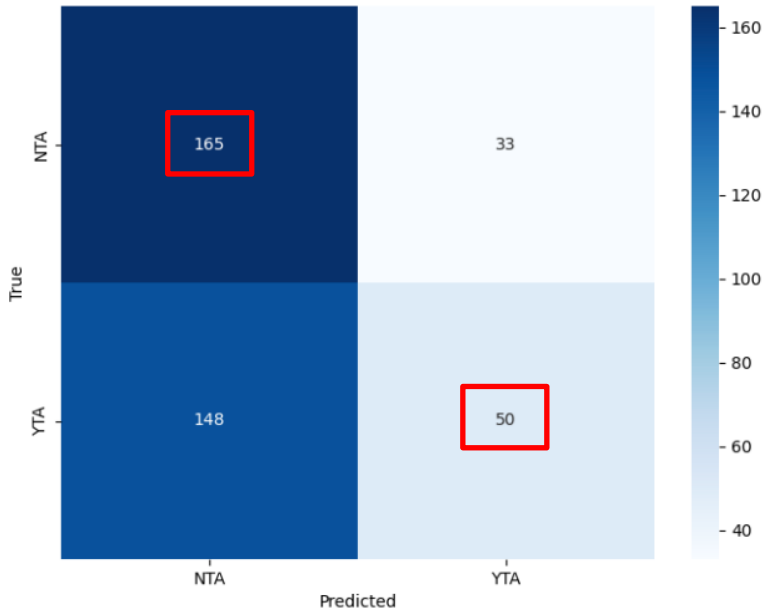
(a) Zero-shot Toxic Generations



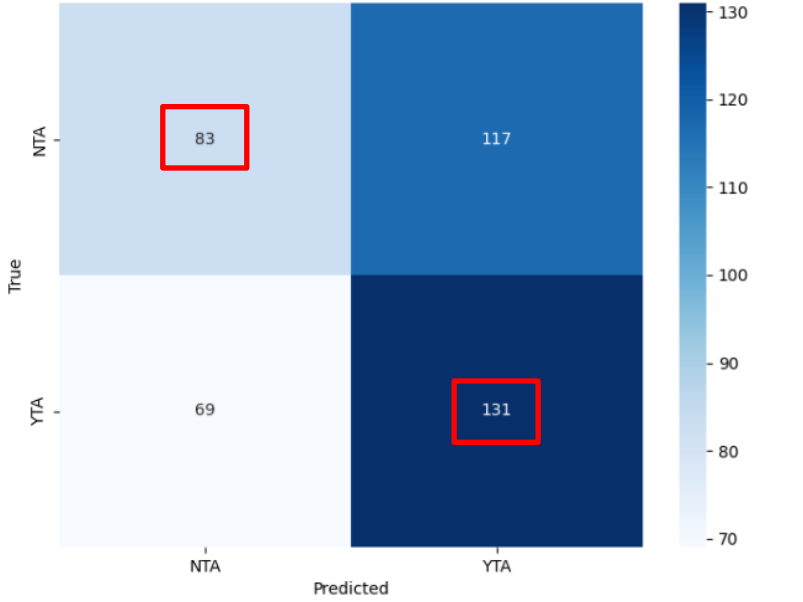
(b) Finetuned Toxic Generations

Change in AITA Classifications

Decrease in accuracy for NTA classifications but increase for YTA ones



(a) Zero-shot Classifications



(b) Finetuned Classifications

Reddit AITA Binary Top 2K Performance Comparison

Reference Top Comment Toxicity Rate: 0.232

No clear signs of improvement at expense of learning toxic language

Flan-T5-XXL

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.040	0.318	0.005	0.57	0.56	0.55	0.134
Finetuned	0.150	0.539	0.253	0.51	0.51	0.50	0.021

Llama-2-13B-Chat

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.103	0.589	0.000	0.56	0.54	0.50	0.105
Finetuned	0.119	0.526	0.188	0.54	0.54	0.53	0.072

Key Insights From Finetuned Models

- Both Flan-T5 and Llama-2-13B-Chat are **not good at identifying** the more nuanced AITA classifications of **NAH, INFO, and ESH.**
  - Either fail or upon finetuning learn to not select these classes.
  - Since most of interpersonal conflicts fall within NTA or YTA, **are these other classes simply noise**?
- Higher ROUGE Lsum scores suggest that **Flan-T5 is superior at classifications and learning the textual relationships** on a lexical level.
- Higher COMET scores and lower toxicity rates suggest that **Llama-2-13B-Chat is superior at generating safer responses** that are more agreeable with human judgements.
- Lack of performance improvements on Top 2K datasets implies they are **too small** for quality learning.

Dataset	Total Samples	YTA	NTA	ESH	NAH	INFO
Multiclass	50000	5576	40549	1331	1887	657
Multiclass-Top-2K	2000	400	400	400	400	400
Binary	46125	5576	40549	0	0	0
Binary-Top-2K	2000	1000	1000	0	0	0

Only 7.75% of samples are ESH, NAH, or INFO

How can we use this insights to to create ideal AI agents for interpersonal conflict resolution?

Encoder-decoder transformer architecture

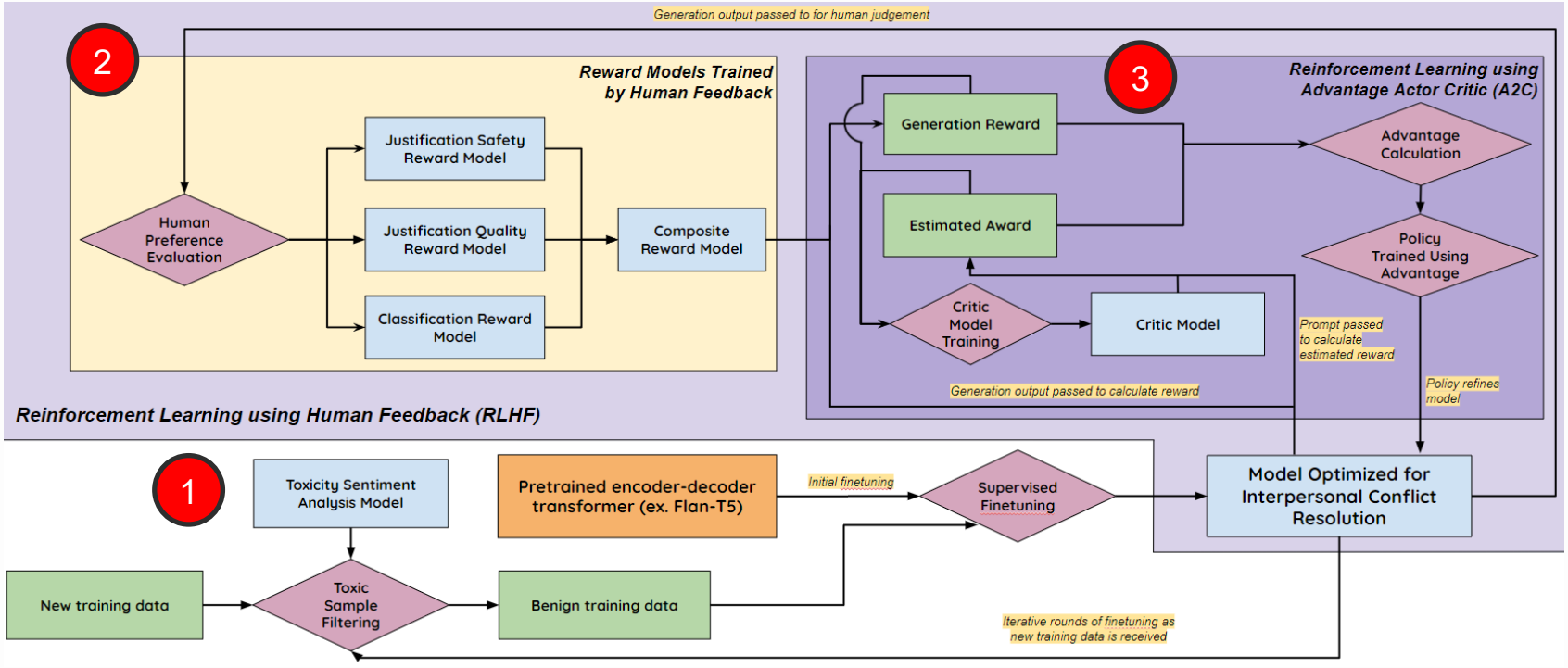
Llama-2-Chat training methods

05

Proposed LLM Architecture and Finetuning Methodology for Interpersonal Conflict Resolution

What type of pretrained LLM should we use and what finetuning techniques should we use to optimize performance?

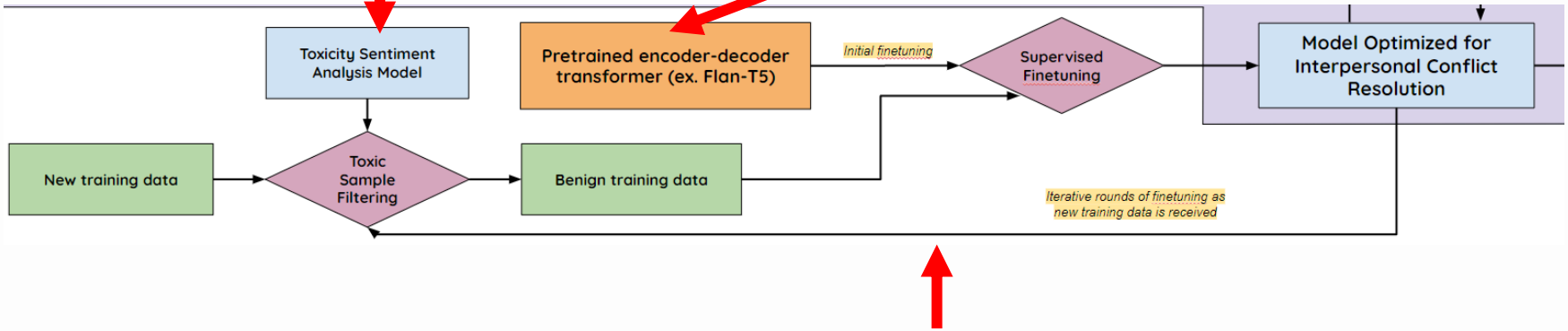
Prioritizes **three** objectives: 1. Justification Safety 2. Justification Quality 3. Classification Accuracy



Supervised Finetuning (SFT) on Benign Samples

Removal of toxic samples using a model finetuned using Toxigen dataset and domain-specific toxic slang

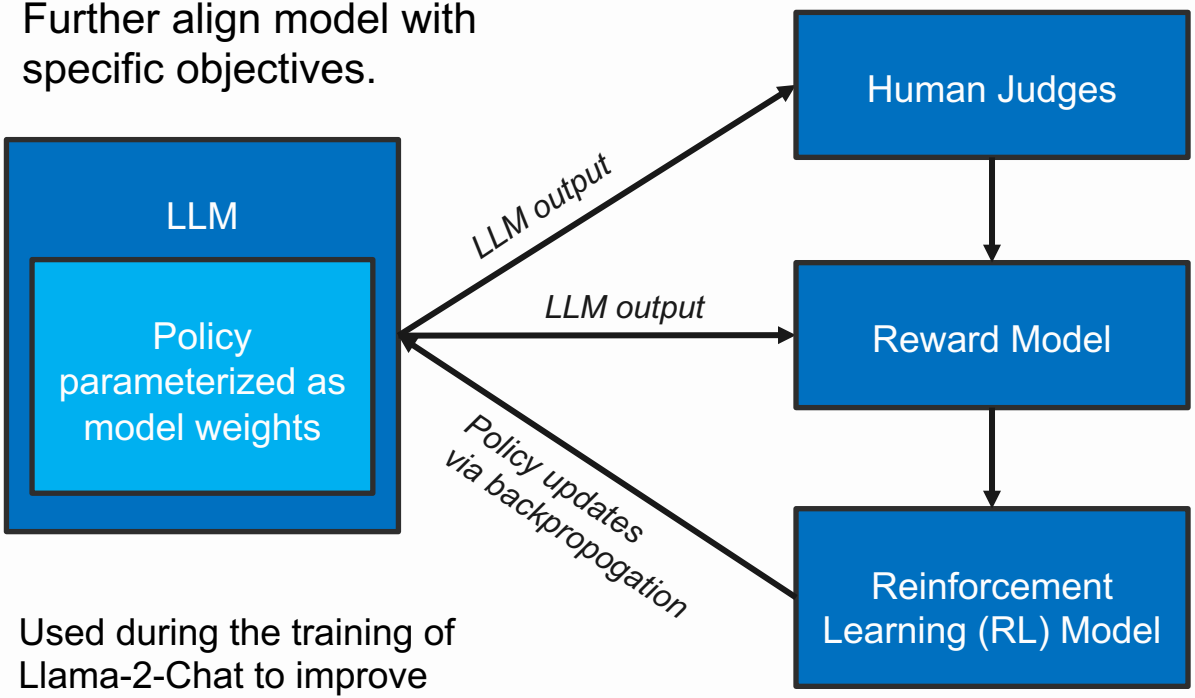
Encoder-decoder architecture for pretrained LLM



New training is also cleaned of toxic samples before new rounds of SFT

RLHF (Reinforcement Learning with Human Feedback)

- Further align model with specific objectives.



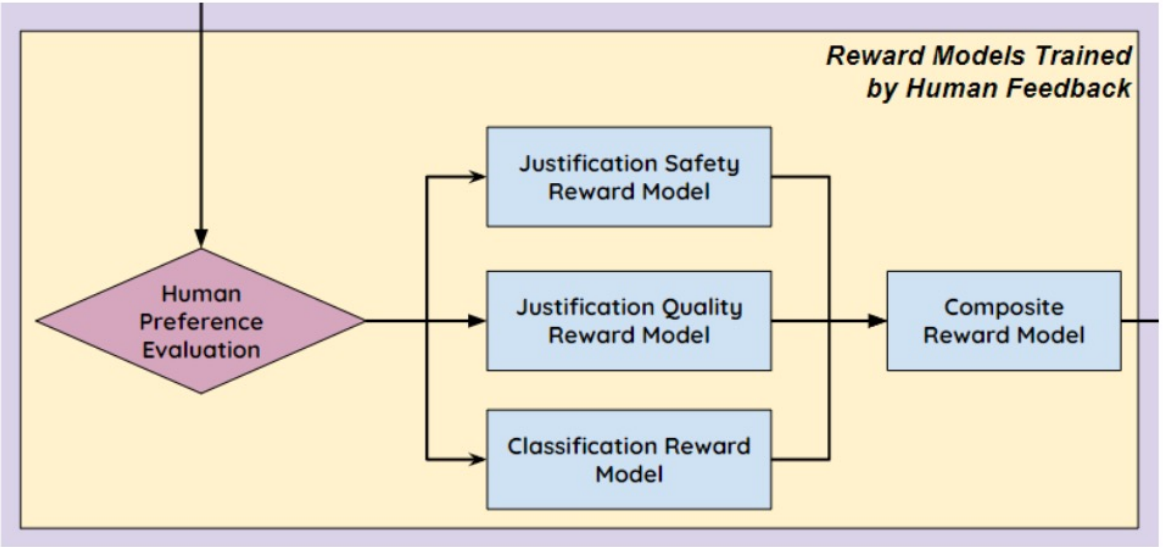
1. Evaluate LLM on defined metrics and these are used to train reward model

2. Reward model generates a reward for the LLM output

3. RL Model uses reward to calculate updates to LLM policy

- Used during the training of Llama-2-Chat to improve safety.

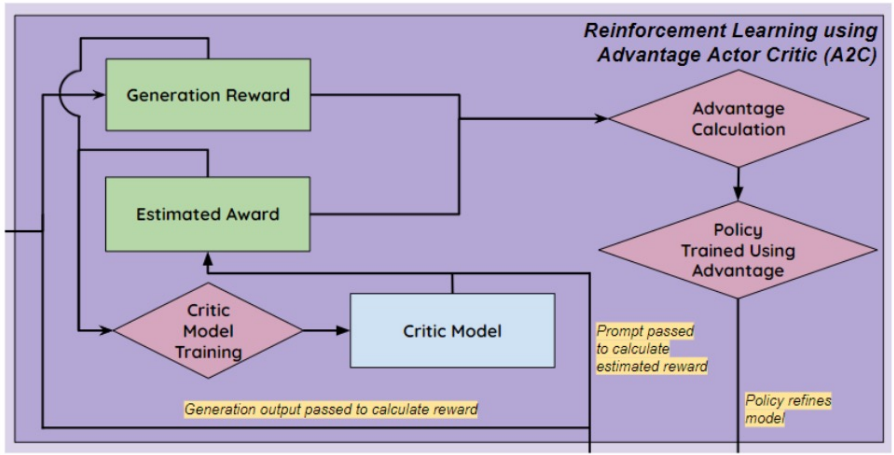
Reward Model for RLHF



- Separate reward model for each objective
- Synthesized together using **hierarchical learning**

higher-level (safety) policies can override lower-level (justification quality and classification accuracy) when necessary

RLHF using Advantage Actor Critic (A2C)



- Advantage Actor Critic (A2C) is an of **actor-critic** RL method, meaning that it combines the benefits of both value-based methods and policy-based methods.
- A2C uses the **advantage function A(s,a)** to evaluate actions taken by the LLM.
- The advantage function compares the value of taking a specific action in a particular state **according to the reward model** to the value of being in that state **according to the critic**.
- Ideal for when **multiples objectives** are being optimized during RLHF.

'actor' = policy function that selects actions → parameterized by LLM

'critic' = another model (ex. Deep Neural Network) evaluates the utility of the action

$$A(s,a) = Q(s,a) - V(s)$$

q value for action a in state s

average value of that state

$$A(s,a) = Q(s,a) - V(s)$$

$$r + \gamma V(s')$$

$$A(s,a) = r + \gamma V(s') - V(s)$$

TD Error

06

Conclusion

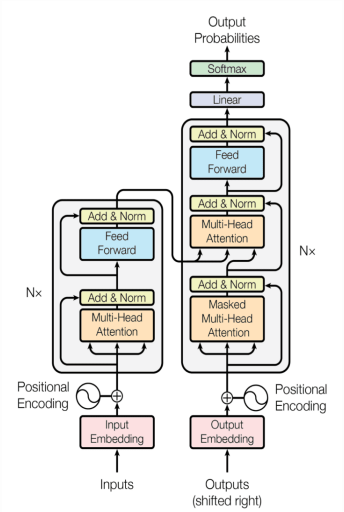
What did this research contribute?

What future work can be built upon it?

Two Types of LLMs...

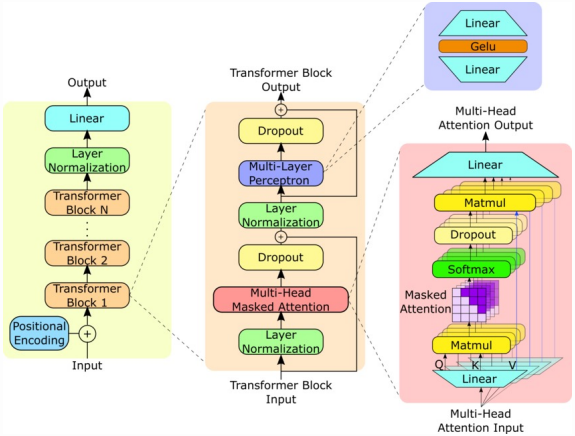
which one is optimal for interpersonal conflict resolution?

Encoder-Decoder



Flan-T5 is superior at classifications and equivalent in justification quality

Decoder-only



Llama-2-13B-Chat is superior at generating safer, more ethical responses

Binary Classification is Sufficient for Interpersonal Conflict Resolution

- Both LLMs when finetuned on Reddit AITA Multiclass considered **ESH, NAH, and INFO** as **noise**

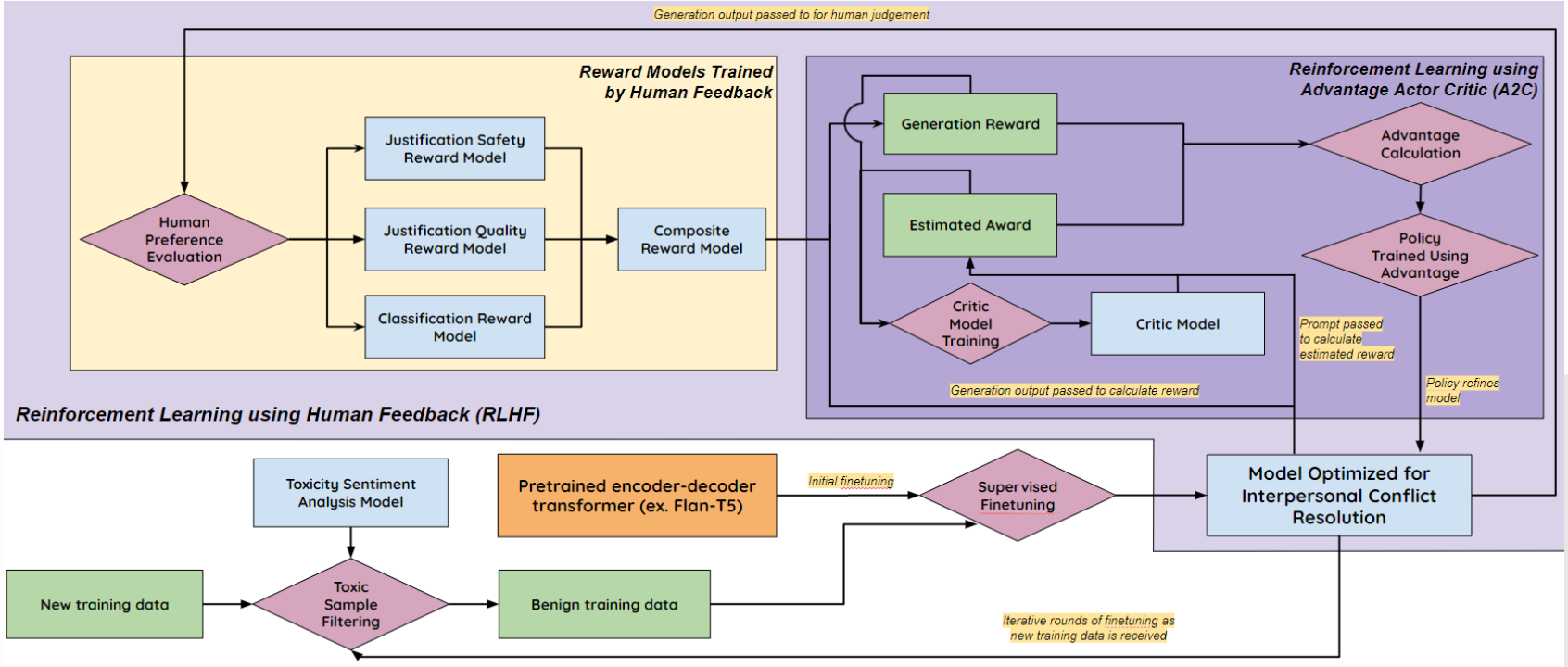
Flan-T5-XXL finetuned on Reddit AITA Binary

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

Dataset	Total Samples	YTA	NTA	ESH	NAH	INFO
Multiclass	50000	5576	40549	1331	1887	657
Multiclass-Top-2K	2000	400	400	400	400	400
Binary	46125	5576	40549	0	0	0
Binary-Top-2K	2000	1000	1000	0	0	0

Only 7.75% of samples are ESH, NAH, or INFO

Optimal Architecture and Finetuning Process for Interpersonal Conflict Resolution



Future Work

- Implementation of the Proposed LLM Architecture and Finetuning Process for Interpersonal Conflict Resolution
- Experimentation on Other Encoder-Decoder LLM
- Improvements to the Quality of the Reddit AITA Datasets

<u>Potential Methods to Improve Dataset Quality</u>

- Filtering the dataset to remove samples with ambiguous classifications
- Stratification to ensure equal representation of NTA and YTA classes
- Manual inspection to verify that the classifications were appropriate.

Other popular encoder-decoder LLMs

Thank you!

Questions?

Contact me at…

mb1020923@wcupa.edu

or

mattboraske@gmail.com

Figure and Equation References

[1] https://aiml.com/compare-the-different-sequence-models-rnn-lstm-gru-and-transformers/

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[3] Yang, S. D., Ali, Z. A., & Wong, B. M. Fluid-gpt (fast learning to understand and investigate dynamics with a generative pre-trained transformer): Efficient predictions of particle trajectories and erosion. Industrial & Engineering Chemistry Research, 62(37), 2023.

[4] Klaus Krippendorff. Computing krippendorff's alpha-reliability, 2011.

[5] Mary L McHugh. Interrater reliability: the kappa statistic. Biochemia medica, 22(3):276–282, 2012.

[6] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC genomics, 21:1–13, 2020.

Thesis Bibliography

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[3] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[4] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[5] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[6] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744, 2023.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of

deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

[9] Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. Better language models and their implications. *OpenAI blog*, 1(2), 2019.

[10] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[12] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. Pmlr, 2013.

[13] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[14] Yibo Hu, MohammadSaleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick Brandt, and Vito D'Orazio. Conflibert: A pre-trained language model for political conflict and violence. In *Proceedings of the Association for Computational Linguistics Conference*. Association for Computational Linguistics, 2022.

[15] Daniel J Olsher. New artificial intelligence tools for deep conflict resolution and humanitarian response. *Procedia Engineering*, 107:282–292, 2015.

[16] Anne Hsu and Divya Chaudhary. Ai4pcr: Artificial intelligence for practicing conflict resolution. *Computers in Human Behavior: Artificial Humans*, 1(1):100002, 2023.

[17] Reyhan Aydoğan, Tim Baarslag, and Enrico Gerding. Artificial intelligence techniques for conflict resolution. *Group Decision and Negotiation*, 30(4):879–883, 2021.

[18] J. Baumgartner. pushshift/api. https://github.com/pushshift/api, 2019. Accessed: insert-date-here.

[19] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.

[20] Klaus Krippendorff. Computing krippendorff's alpha-reliability, 2011.

[21] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[22] Solomon E Asch. Effects of group pressure upon the modification and distortion of judgments. In *Organizational influence processes*, pages 295–303. Routledge, 2016.

[23] Stanley Milgram. Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4):371, 1963.

[24] University of Oklahoma. Institute of Group Relations and Muzafer Sherif. *Intergroup conflict and cooperation: The Robbers Cave experiment*, volume 10. University Book Exchange Norman, OK, 1961.

[25] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[27] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13, 2020.

[28] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[29] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*, 2020.

[30] Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, 2022.

[31] Ine Gevers, Ilia Markov, and Walter Daelemans. Linguistic analysis of toxic language on social media. In *Computational Linguistics in the Netherlands*, volume 12, pages 33–48, 2022.

[32] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*, 2021.

[33] Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35, 2021.

[34] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

[35] Yuhuai Wu, Elman Mansimov, Roger B Grosse, Shun Liao, and Jimmy Ba. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. *Advances in neural information processing systems*, 30, 2017.