The Efficacy of Finetuning Large Language Models for Interpersonal Conflict Resolution

Presented in Partial Fulfillment of the Requirements for the Degree M.Sc. Computer Science in the Graduate School of West Chester University of Pennsylvania

By

Matthew Boraske, B.Sc.

West Chester University of Pennsylvania

2024

Thesis Committee:

Richard Burns, Advisor

Ashik Ahmed Bhuiyan

Si Chen

Abstract

Since its introduction in 2017, the transformer architecture has been a cornerstone in advancing natural language processing, leading to the development of large language models (LLMs). LLMs that leverage the original encoder-decoder transformer architecture have proven to be particularly effective in tasks requiring a deep understanding of input, such as summarization and question-answering. In contrast, subsequent developments have introduced LLMs that use a decoder-only variant of the transformer architecture, with the intent of optimizing them for generating more coherent and longer texts.

Despite their widespread use in deterministic tasks like text summarization and translation, there has been limited research on the application of LLMs to more ambiguous, everyday tasks such as interpersonal conflict resolution. This study seeks to bridge this gap by evaluating LLMs on four newly created datasets derived from the "Am I the A**hole" (AITA) subreddit, an online community of over fifteen million members that features discussions of interpersonal conflicts. These datasets were designed to challenge the models with real-world data that naturally features ambiguous judgments and toxic language.

Our research utilizes Google's Flan-T5, released in 2022, and Meta's Llama-2-Chat, released in 2023, to represent the latest in both architectures. We instruction finetuned these models using the AITA datasets to evaluate changes in their ability to classify and justify conflicts, and to assess the prevalence of learning to generate toxic language. Our findings suggest that the most effective strategy for training LLMs on interpersonal conflict resolution consists of finetuning an encoder-decoder LLM on a dataset where samples with toxic language discarded, followed by iterative refinement using Reinforcement Learning with Human Feedback (RLHF) to better align with ethical standards.

To our knowledge, this is the first work that examines the potential for using transformer-based LLMs to resolve real-world interpersonal conflicts. Thus, it offers significant insights for their application in social and therapeutic contexts, where sensitive and effective advice is crucial. Additionally, the research contributes to ongoing discussions about the ethical implications of deploying AI in sensitive areas, suggesting ways in which it can be a beneficial adjunct to human judgment rather than a replacement.

Table of Contents

List of Figures

Figure **Page**

List of Tables

Table **Page**

Chapter 1
Introduction

The field of natural language processing (NLP) has undergone a remarkable transformation in recent years, revolutionizing human-computer interactions. A foundational part of modern artificial intelligence (AI), NLP enables machines to understand, interpret, and generate human language [5]. Its rapid advancements have opened up a wide array of applications, ranging from virtual assistants to machine translation and content generation, integrating seamlessly into our daily lives [6].

A pivotal breakthrough in NLP has been the development of large language models (LLMs) based on the transformer architecture, as introduced by Vaswani et al. in 2017 [1]. These models, leveraging self-attention mechanisms to focus on relevant parts of the input, have demonstrated remarkable capabilities in understanding and generating human-like language. LLMs, such as BERT [7], GPT [8], and T5 [2], have pushed the boundaries of what is possible with NLP, enabling more natural and intuitive human-computer interaction [9]. The ability of these models to capture the nuances and context of human language has led to significant improvements in various NLP tasks, such as question answering, text summarization, and language translation [10].

The FLAN-T5 model, an advancement based on the T5 (Text-To-Text Transfer Transformer) framework, represents a pivotal development in encoder-decoder transformer architectures. Google's T5, as introduced by Raffel et al. in 2020 [2], revolutionized the approach to NLP tasks by treating them uniformly as text-to-text translations. This innovation simplified the training process and significantly increased the model's adaptability across various tasks. Building on this, FLAN-T5 integrates an "instruction fine-tuning" method, whereby the model is trained

on a myriad of tasks articulated in natural language [3]. This training approach enhances its proficiency in understanding and executing instructions. The encoder-decoder structure of FLAN-T5, fundamental to the T5's architecture, is adept at processing input sequences into contextual representations, then generating relevant outputs. Such a mechanism is particularly advantageous for tasks that require a deep understanding of the input, including summarization and question-answering, and, as this study proposes, for the resolution of interpersonal conflicts.

Conversely, the Llama-2 models adopt a decoder-only transformer architecture. This design marks a departure from the conventional encoder-decoder framework. Llama-2's architecture is focused on predicting subsequent text sequences through a self-attention mechanism that processes the context within the sequence itself, eliminating the need for an encoder. This streamlined model structure enhances efficiency in text generation, particularly in scenarios where input-output mapping is more straightforward or less reliant on extensive context. Llama-2 is distinguished by its expansive training dataset and sophisticated training methodologies, contributing to its comprehensive grasp of language and capability to generate contextually appropriate text. Its application in interpreting and addressing interpersonal conflicts, as investigated in this study, tests the model's capacity to understand and respond to complex social interactions using its decoder-only framework. A version of Llama-2 was also released called Llama-2-Chat, which through iterative stages of finetuning was optimized for conversational tasks.

Comparatively, FLAN-T5 and Llama-2's differing architectures inherently affect their performance across various NLP tasks. FLAN-T5's encoder-decoder framework is advantageous for tasks requiring in-depth understanding and reinterpretation of input text. In contrast, Llama-2's decoder-only structure excels in efficiently generating sequential text, useful in tasks like conversational AI but less so in scenarios needing extensive contextual analysis.

Our investigation utilizes the FLAN-T5 and Llama-2-Chat models to resolve interpersonal conflicts, marking a novel exploration in the field of NLP. Traditionally, large language models have primarily been applied to tasks with defined objectives such as translation, summarization, or information retrieval [2]. However, interpersonal conflict resolution poses a unique challenge, calling for an understanding of nuanced human emotions, subjective interpretations, and the need for empathetic and context-sensitive responses. The encoder-decoder structure of FLAN-T5, with its

comprehensive understanding of input context, has the potential to offer deep insights into the intricate dynamics of human interactions [3]. Alternatively, Llama-2-Chat's decoder-only architecture, adept at generating human-like responses, may propose innovative solutions in evolving conflict scenarios [4]. By evaluating FLAN-T5 and Llama-2 in the human-centric domain of interpersonal conflict resolution, our study aims to not only advance AI in understanding human interactions but also to explore new practical applications for LLMs in social and psychological realms.

Accordingly, the key contributions of this research are as follows:

1. **Utilizing LLMs to Resolve Interpersonal Conflicts in an Anthropomorphic and Ethical Manner:** By analyzing the impacts of finetuning exemplars of LLMs that utilize encoder-decoder and decoder-only transformers on real-world data that contains both ambiguity and toxic language, we gained understanding of what components of each that were benefical to improving the efficacy of their interpersonal conflict resolution. Ultimately, we found that the most effective strategy for training LLMs for interpersonal conflict resolution involves finetuning an encoder-decoder LLM on a dataset where samples with toxic language have been discarded, followed by iterative refinement using Reinforcement Learning with Human Feedback (RLHF) to better align with ethical standards.

2. **Managing Toxic Language in Training Data:** A distinctive challenge addressed in this study is the presence of toxicity in LLM training data. We analyzed how well the Flan-T5 and Llama-2-Chat models can be finetuned to safely understand and resolve interpersonal conflicts, even when trained on data containing toxic language.

3. **Expanding AI Applications in Therapy and Counseling:** Our research broadens the potential application of conversational AI in the realm of mental health. By finetuning LLMs to sensitively manage interpersonal conflicts, we open new possibilities for AI support in therapy and counseling, increasing its accessibility by enhancing the range of tools available to mental health practitioners and patients.

Chapter 2
Background and Literature Review

The field of natural language processing (NLP) witnessed a paradigm shift with the advent of the transformer architecture, which significantly impacted the development of large language models (LLMs). This literature review aims to provide a comprehensive overview of the foundational and evolutionary aspects of transformer-based LLMs. It first explains the intricacies of the transformer architecture, tracing the transition from earlier RNN models to the advanced transformer-based models that dominate the field today. Afterwards, it examines two state-of-the-art transformer-based models, FLAN-T5 and Llama-2, elucidating their unique architectural components and training strategies. The review concludes with a survey on the emerging application of LLMs for interpersonal conflict resolution to identify the gaps that this research seeks to address. Overall, this review sets the stage for understanding the potential to apply LLMs in complex, human-centric applications like interpersonal conflict resolution.

2.1 The Transformer Architecture

Introduced by Vaswani et al. via the groundbreaking paper "Attention is All You Need" in 2017, the transformer architecture has revolutionized the approach to sequence-to-sequence tasks in natural language processing (NLP), marking a significant departure from traditional recurrent neural network (RNN) models [1]. This architecture serves as the foundation for various state-of-the-art models, ultimately reshaping our understanding of NLP and its capabilities.

2.1.1 Key Components and Functionality of the Transformer

At the center of the transformer architecture lies the self-attention mechanism. This innovative approach allows the model to process different parts of the input data independently and to weigh the significance of each part, facilitating a dynamic understanding of the context within the sequence. This feature is a critical enhancement over RNNs, which process data sequentially and often struggle with long-range dependencies [1].

Alongside self-attention, the transformer employs positional encodings to retain the sequence order of the input data, a necessary component given its non-sequential processing nature. These encodings ensure that the model is aware of the position of each word in the sequence, which is vital for preserving the inherent structure of language [1].

Another notable aspect of the transformer is the incorporation of the multi-head attention mechanism. This approach involves running the self-attention process several times in parallel, allowing the model to capture and integrate a more diverse range of information from the input, such as syntactic and semantic nuances, simultaneously [1].

The original transformer model introduced by Vaswani et al. is shown in Figure 2.1 and features an encoder-decoder structure, with multiple layers in each component. The encoder processes the input sequence, embedding it into a rich, contextual representation. This representation is then fed into the decoder, which generates the output sequence. Both the encoder and decoder layers consist of self-attention mechanisms and feed-forward neural networks, structured to handle complex sequence-to-sequence tasks effectively [1].

2.1.2 Advantages Over Previous Architectures and Emerging Challenges

The transformer architecture offers significant improvements over previous models, particularly in terms of parallelization. Unlike RNNs, which require sequential data processing, transformers can process entire sequences simultaneously, allowing for more efficient and faster training. This architecture's capacity to capture long-range dependencies in text is another major advancement, addressing a critical limitation in RNNs and LSTMs, which often lose information over longer sequences. Furthermore, transformers are highly scalable, a feature that has led to the development

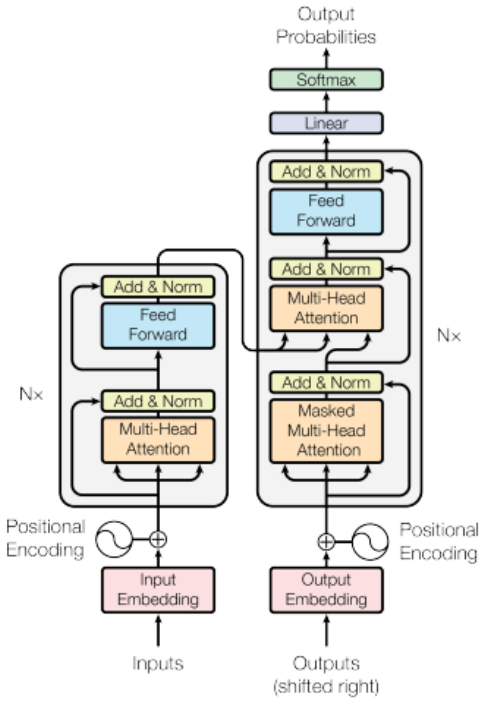


Figure 2.1: The Original Transformer Architecture [1]

of increasingly large models capable of handling vast datasets and sophisticated tasks [1].

Despite these advantages, transformers present certain challenges. The most notable is their substantial computational requirements, especially for larger models, necessitating significant hardware resources for training and operation. Additionally, the complexity of the transformer models can lead to interpretability issues and a reliance on extensive, high-quality training data to achieve optimal performance [1].

In the following sections, we will delve deeper into the evolution of large language models (LLMs) from using RNNs to transformer architectures, and specifically examine the development and characteristics of FLAN-T5 and Llama-2 models. Additionally, we will survey the current state of research on the application of LLMs in the context of interpersonal conflict resolution.

2.2 Evolution of Large Language Models: From RNNs to Transformer Architectures

The evolution of Large Language Models (LLMs) in natural language processing (NLP) has been marked by significant technological advancements, transitioning from the early dominance of Recurrent Neural Networks (RNNs) to the groundbreaking emergence of transformer architectures.

RNNs, including their more advanced variants like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), were the cornerstone of early NLP models due to their ability to process sequential data, an essential feature for language tasks [11]. However, RNNs inherently faced challenges in capturing long-range dependencies within text, due to the vanishing gradient problem, which hampers the model's ability to learn and retain information over longer sequences [12].

The introduction of attention mechanisms provided a solution to this limitation. Initially used in conjunction with RNNs, attention mechanisms allowed the model to focus on different parts of the input sequence, thereby capturing broader contexts and dependencies more effectively [13].

The transformer architecture represented a paradigm shift by entirely doing away with recurrence and convolution in favor of attention mechanisms. This architecture allowed for unprecedented parallelization of sequence processing, significantly enhancing model efficiency and performance on complex NLP tasks [1]. Overall, their ability to handle long-range dependencies and large-scale training datasets set them apart from their RNN predecessors, leading to more sophisticated and capable LLMs.

The success of the transformer architecture catalyzed the development of state-of-the-art models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer). BERT, introduced by Devlin et al., utilized transformers to process text bidirectionally, enabling a more nuanced understanding of context [7]. On the other hand, Radford et al.'s GPT showcased the power of transformer-based generative models in producing coherent and contextually relevant text [8]. These developments not only showcased the versatility of transformer architectures but also established a new benchmark for LLM capabilities in NLP.

The transition from RNNs to transformer-based models marked a significant leap in NLP, with

implications extending beyond traditional language tasks. These advanced models have shown remarkable proficiency in understanding and generating human-like text, paving the way for their application in a myriad of real-world scenarios, including interpersonal conflict resolution. In the next section, we will explore the specifics of FLAN-T5 and Llama-2 architectures, further illustrating the advancements and capabilities of transformer-based LLMs.

2.3 Technical Overview of the FLAN-T5 Architecture

The FLAN-T5 model, built upon the encoder-decoder structure of the foundational T5 model, emerges as a significant advancement in the domain of NLP. Developed by Chung et al., FLAN-T5 stands out for its innovative approach to handling a diverse array of NLP tasks and its specialized training methodology [3].

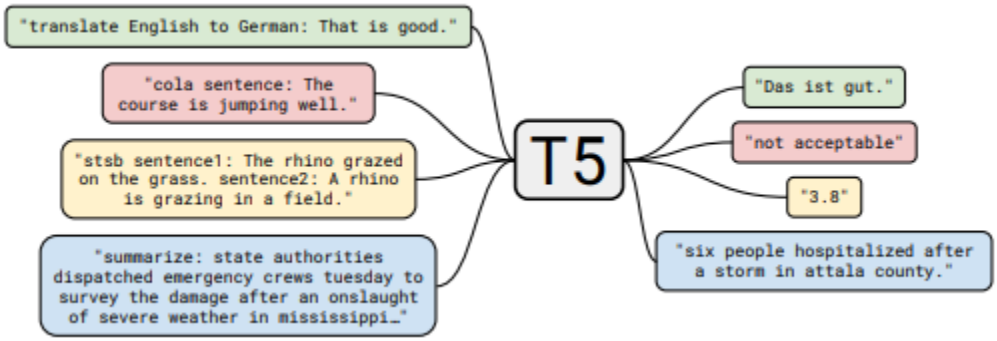


Figure 2.2: The T5 Model [2]

2.3.1 Architecture Foundation

FLAN-T5 leverages the encoder-decoder architecture integral to T5 shown in Figure 2.2, wherein every NLP task is treated as a text-to-text conversion problem. In this setup, the encoder processes the input sequence to create an intermediate representation, which is then utilized by the decoder to generate the output sequence. This structure, characterized by multiple layers containing self-attention mechanisms and feed-forward neural networks, underpins the model's adaptability and

efficiency in handling various NLP tasks [3].

2.3.2 Instruction Finetuning, Task Mixtures, and Chain-of-Thought Integration

The instruction finetuning approach of FLAN-T5 distinguishes itself from conventional task-specific training. As shown in Figure 2.3, instead of being narrowly focused on specific tasks, FLAN-T5 is finetuned on an expansive array of tasks phrased as natural language instructions. This method is instrumental in enhancing the model's ability to generalize, allowing it to adeptly handle new tasks with minimal additional training. For instance, FLAN-PaLM 540B, when finetuned on 1.8K diverse tasks, demonstrates a significant performance leap on benchmarks like MMLU and TyDiQA, underscoring the effectiveness of this approach [3].

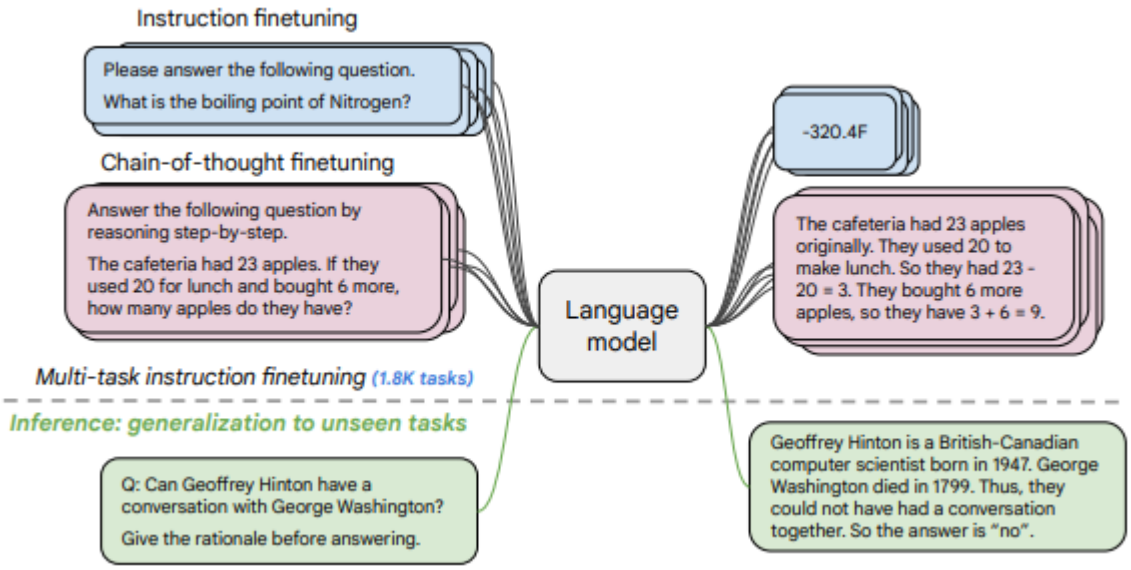


Figure 2.3: Instruction Finetuning of Flan-T5 Model [3]

Integral to FLAN-T5's instruction finetuning is the use of diverse task mixtures and formats. The model benefits from a rich blend of task types encompassing Muffin, T0-SF, NIV2, and CoT (Chain of Thought), each contributing uniquely to the model's broad applicability. This mixture spans a spectrum from commonsense reasoning and natural language inference to complex reasoning tasks. The inclusion of CoT annotations is particularly pivotal. Finetuning with these annotations

enhances the model's reasoning capabilities, evident in its improved performance in tasks such as arithmetic and multi-step reasoning. Notably, CoT finetuning not only boosts the model's performance in reasoning tasks but also ensures the retention of these crucial capabilities, which are essential for sophisticated applications in NLP [3].

Moreover, FLAN-T5 demonstrates the impact of model scaling and the extent of finetuning tasks on overall performance. The research highlighted that scaling up both the model size and the number of finetuning tasks led to enhanced model performance. This scaling behavior underscores FLAN-T5's effectiveness in generalizing across a broad range of NLP tasks and sets a foundation for future research in instruction finetuning within the domain of large language models [3].

2.3.3 Concluding Remarks on FLAN-T5's Capabilities

FLAN-T5 represents a major stride forward in NLP because of its instruction finetuning strategy, comprehensive task mixtures, and the integration of Chain-of-Thought (CoT) data. Its adaptability and performance across diverse tasks, combined with enhanced reasoning capabilities, position it as a powerful tool for complex NLP applications, including nuanced domains like interpersonal conflict resolution.

2.4 Technical Overview of the Llama-2 Architecture

Llama-2, developed by Touvron et al. at Meta in 2023[4], represents a significant evolution in the field of large language models (LLMs), especially for dialogue applications. This model extends the capabilities of LLMs, scaling up to 70 billion parameters, and showcases advancements in both pretraining and finetuning methodologies.

2.4.1 Architecture Foundation

Llama-2's architecture is based on an optimized auto-regressive transformer model, similar to its predecessor, Llama-1. Several key enhancements were implemented in Llama-2, including more robust data cleaning, an updated mix of data sources for pretraining, and a 40% increase in the total tokens used for training. The context length was doubled, and grouped-query attention (GQA)

was adopted to improve inference scalability for larger models. These improvements contributed to significant advancements over Llama-1, particularly in handling a diverse range of NLP tasks[4].

2.4.2 Finetuning for Dialogue Applications

As part of its initial development, a finetuned version of Llama-2 for dialogue applications was created called Llama-2-Chat. The finetuning of the model was decomposed into several stages as shown in Figure 2.4, each contributing significantly to its final capabilities.

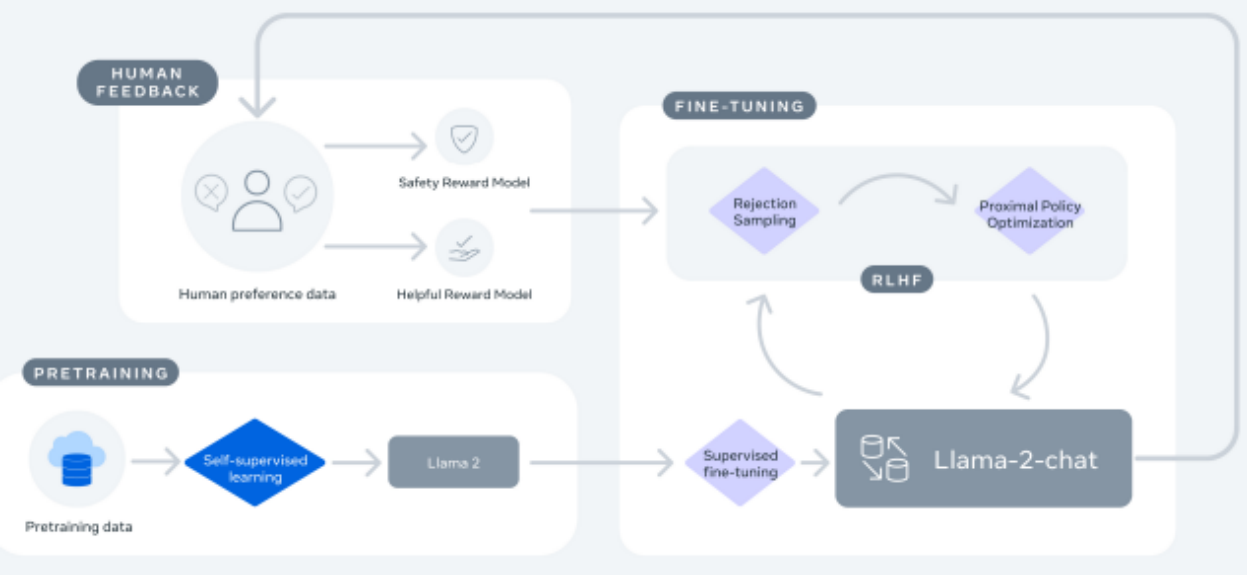


Figure 2.4: Creation of Llama-2 and Finetuning Process for Llama-2-Chat [4]

The initial phase of finetuning, known as Supervised FineTuning (SFT), utilized publicly available instruction tuning data. This stage was critical in setting a foundational understanding for the model. Emphasis was placed on collecting high-quality SFT data, with a significant volume of annotations being essential to achieve the desired level of model performance. Importantly, this stage of development ensured that no Meta user data was included in the annotations. The fine-tuning process employed a cosine learning rate schedule, with each training sample comprising a prompt and its corresponding answer. This methodological approach helped refine the model's responses to be more aligned with the nuances of natural dialogue [4].

The Reinforcement Learning with Human Feedback (RLHF) stage followed, further aligning the model with human preferences in dialogue contexts. RLHF involved an innovative approach to data collection, where human preference data was gathered through a binary comparison protocol. This method focused on assessing the model's helpfulness and safety, two crucial aspects for any dialogue application. The data collected during this phase was used to train of specialized reward models - one optimized for assessing helpfulness and another for safety. These reward models directly influenced the iterative finetuning of Llama-2-Chat, ensuring that the model's outputs were not only contextually relevant but also adhered to the standards of helpfulness and safety set by human evaluators [4].

This meticulous finetuning process, encompassing both SFT and RLHF stages, was fundamental in enhancing Llama-2-Chat's capabilities. It enabled the model to navigate the complexities of human dialogue, ensuring that its responses were contextually appropriate, helpful, and safe. The result was a model well-suited for dialogue applications, showcasing the potential of finetuned LLMs in facilitating human-like interactions [4].

2.4.3 Safety Considerations and Evaluations

Safety was a paramount consideration in the development of Llama-2-Chat. Various measures were taken to ensure the safety of the model, including safety-specific data annotation, red-teaming, and rigorous safety evaluations. These steps were crucial in increasing the trustworthiness and reliability of the model, especially when deployed in dialogue scenarios [4].

2.4.4 Concluding Remarks on Llama-2's Capabilities

Llama-2, equipped with numerous advanced features, demonstrates robust performance across a broad spectrum of NLP tasks and excels in complex reasoning. Its variant, Llama-2-Chat, owing to rigorous fine-tuning, achieves state-of-the-art performance in conversational tasks.

In conclusion, Llama-2 and its dialogue-focused counterpart, Llama-2-Chat, mark significant progress in the field of transformer-based large language models. The model benefits from comprehensive pretraining, innovative fine-tuning techniques, and a strong focus on safety, establishing the Llama-2 models as effective tools for dialogue-based applications. This is especially pertinent

in scenarios requiring a nuanced understanding, such as resolving interpersonal conflicts.

2.5 LLMs for Conflict Resolution: A Review of Current Research

The exploration of Large Language Models (LLMs) in conflict resolution encompasses a diverse range of methodologies and applications, as demonstrated by the recent studies in this field. Rather than compartmentalizing each study, a thematic approach provides a more holistic view of the current landscape and the emerging trends in AI-assisted conflict resolution.

2.5.1 Innovative Tools and Methodologies

Recent advancements include specialized tools like ConfliBERT, designed by Hu et al., which exemplifies the potential of domain-specific LLMs in analyzing political violence and conflict. ConfliBERT's focused pre-training on relevant corpora offers deeper insights into political dynamics, showcasing the benefits of specialized LLMs in complex areas like conflict research [14].

Simultaneously, Olsher's cogSolv introduces a system capable of deep conflict analysis, particularly useful in humanitarian responses.This tool simulates reactions, manages knowledge, and enhances decision-making in humanitarian response and conflict resolution. The system exemplifies the advanced capabilities of AI in complex social systems analysis, offering practical solutions in high-stakes environments like peacekeeping and disaster response [15].

2.5.2 AI in Enhancing Communication and Negotiation

The AI4PCR application developed by Hsu and Chaudhary brings a different perspective by leveraging AI in practicing relationship-preserving language. This tool not only aids in reducing social provocation and bias during conflicts, but also emphasizes the importance of neutral language in de-escalating situations [16].

Furthermore, Aydoğan, Baarslag, and Gerding's exploration of various AI techniques in conflict resolution, such as automated negotiation and argumentation, underscores the interdisciplinary nature of the field. Their work highlights AI's role in enhancing negotiation dynamics and decision-making processes, especially in multi-agent scenarios [17].

2.5.3 Emerging Trends and Future Directions

Across these studies, several trends emerge, highlighting the potential of AI in analyzing and interpreting complex social and political conflicts. Notably, the importance of nuanced language in conflict resolution and the role of AI in decision-making processes are emphasized. These findings illustrate the diverse applications of AI in conflict scenarios, ranging from high-level political analysis to interpersonal communication.

However, current research reveals notable gaps, especially in the domain of interpersonal conflict resolution using Large Language Models (LLMs). While existing studies provide foundational insights, there's a crucial need for more in-depth investigation into the capabilities of LLMs in understanding and advising on interpersonal conflicts. Such conflicts often demand a significant grasp of human emotions and subtleties. Additionally, a technical challenge arises in fine-tuning LLMs with interpersonal conflict data, which may include toxic language. The key is to develop LLMs capable of offering conflict resolution guidance in a human-like and sensitive manner, without adopting any toxic language patterns inherent in the training data.

Addressing these challenges is vital, particularly given the increasing accessibility of LLMs due to decreasing computational requirements. Successfully fine-tuning LLMs for resolving interpersonal conflicts in a manner that is both human-like and sensitive could significantly enhance the feasibility of employing AI in therapeutic contexts, where the presence of toxic language can be harmful for patients. The goal is to harness the power of LLMs to provide meaningful, empathetic guidance in conflict resolution, thereby expanding the practical applications of AI in sensitive human-centric fields like therapy.

2.5.4 Contributions of Our Research

This research advances the field by addressing the existing gaps in the application of Large Language Models (LLMs) for interpersonal conflict resolution. While foundational studies have illuminated AI's capabilities in conflict scenarios, our paper delves deeper, exploring the efficacy and subtleties involved in fine-tuning LLMs to handle the complex dynamics of interpersonal conflicts.

At the heart of our contributions is the focus on anthropomorphic and sensitive conflict reso-

lution. We have concentrated our efforts on fine-tuning LLMs to offer guidance on interpersonal conflicts in a way that mirrors human understanding and emotional intelligence. This approach not only ensures that the advice provided by LLMs is relevant and context-appropriate, but also empathetically resonates with the users, effectively bridging the gap between technical sophistication and the nuanced requirements of emotional intelligence.

Another distinctive aspect of our research is the management of toxic language within the training data for LLMs. Recognizing the challenge this poses, particularly in sensitive fields such as therapy and counseling, our study delves into the capability of models like Flan-T5 and Llama-2 to be finetuned on datasets containing toxic language, yet still navigate and resolve interpersonal conflicts safely. This aspect of our research is crucial in ensuring that the application of LLMs in conflict resolution is not only effective, but also safe and responsible.

Furthermore, our research expands the potential applications of AI in the realm of mental health. By fine-tuning LLMs to handle interpersonal conflicts with sensitivity, we unlock new possibilities for AI support in therapy and counseling contexts. This expansion not only enhances the range of tools available to mental health practitioners and patients, but also opens up avenues for AI's role in facilitating more empathetic and understanding interactions in therapeutic settings.

Chapter 3
Methodology

The methodology adopted in this study encompassed a comprehensive approach to evaluating Large Language Models (LLMs) instruction finetuned for interpersonal conflict resolution. It was divided into three phases: the creation of the Reddit AITA datasets, the instruction finetuning of the Flan-T5 and Llama-2-Chat models, and the evaluation of the models' classification and justification capabilities.

3.1 Creation of the Reddit AITA Datasets

The Reddit AITA datasets served as the foundational data for our research.[1] We sourced data from the AITA subreddit because it provided a rich source of real-world data on interpersonal conflict scenarios. The dataset creation process was separated into three phases: the raw data collection from the AITA subreddit, preparing it into the four datasets for instruction finetuning, and then analyzing their key features including but not limited to the degree of toxicity in the top comments and agreement between the commenters on their AITA classification.[2]

3.1.1 Data Collection

The raw data for this study was sourced from the AITA subreddit, utilizing Pushshift, a tool developed by Reddit moderators. Pushshift offers enhanced search capabilities for Reddit comments and submissions, making it ideal for our data collection needs [18]. The resulting data comprised

[1]The Reddit AITA datasets are publicly available at https://huggingface.co/collections/MattBoraske/reddit-aita-finetuning-66038dc9281f16df5a9bab7f

[2]For transparency and reproducibility, the code for creating the Reddit AITA datasets can be found at https://github.com/MattBoraske/Reddit_AITA_Finetuning/tree/main/dataset_creation

two raw text files containing vital information such as the content of each comment or submission, their parent submission or comment, creation date, and the community score. On Reddit, users can 'upvote' or 'downvote' a submission or comment, with each vote incrementing the score by one. The final score is the sum of all these incremental votes.

To ensure a sufficient levels of relevance and engagement in the training data, the submissions dataset was filtered to only retain submissions with a minimum score of fifty, signifying substantial positive community engagement. Additionally, to focus on more recent interactions, the time frame for the submissions was narrowed down to the period between 2019 and 2022. The comments dataset was filtered to only include samples with a minimum score of ten and whose parents were submissions, as opposed to other comments, that had a score of fifty or more. The submissions and comments datasets were then merged to create a dataset of 101,603 samples where each contained a submission and its top ten comments by community score.

3.1.2 Dataset Preparation

We initiated the data preparation by conducting a multi-step cleaning process. First, we removed any samples where the top comment did not start with an AITA classification, as these classifications were an essential part to resolving the interpersonal conflict. Each comment typically starts with a classification followed by a justification, providing a structured format that is essential for evaluating the two-step interpersonal conflict resolution process that is characteristic of the AITA subreddit. Commenters are instructed to use five standard classifications and descriptions for each are included in Table 3.1. Example submission and top comment pairs for each AITA classification are shown in Table 3.3

We then removed the top two percent longest submissions and top comments in terms of character length, which are what we used as the input and reference texts during training. This step was crucial precursor to our modification of context windows of the Flan-T5 models, which were initially trained with a 512-token context window for both encoder and decoder. Thanks to the T5 architecture's use of relative rather than fixed positional embeddings, we were able to modify the context windows during fine-tuning [3]. Specifically, we expanded the encoder's context window to 1024 tokens and reduced the decoder's window to 256 tokens. By filtering out the lengthiest

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.

Table 3.1: AITA Classification Descriptions

submissions and comments, we ensured that all tokenized texts and top comments fit within these adjusted context windows, thus optimizing them for our training requirements.

Combined, these two cleaning steps reduced the dataset to 73,055 samples, a decrease of 28.1 percent. To further refine the dataset, we then selected the top 50,000 samples based on submission scores, which brought the total reduction in dataset size to 50.8 percent.

In addition to filtering out undesirable samples, we cleaned all submissions by removing any edits made by the original posters after their initial posts. This was necessary because edits often included updates that revealed the outcomes of the interpersonal conflicts, which could allow the models to anticipate their outcomes. Therefore, removing these edits was essential to maintain the integrity of the training process, mirroring real-world scenarios where outcomes are unknown. Consequently, removal of edits reduced the average length of submission texts and top comments by 8.49% and 2.77%, respectively.

Beyond cleaning the samples, they were modified through the addition of prompts tailored for the Llama-2-Chat and Flan-T5 models. These models require inputs to be formatted into specialized prompts, the specifics of which are discussed in Section 3.2. This augmentation was critical for ensuring optimal performance during instruction finetuning.

The dataset described so far, consisting of 50,000 samples, was used to finetune the Flan-T5 and Llama-2-Chat models for multiclass classification and will be referred to as the "Reddit AITA Multiclass" dataset. To evaluate the performance of each model across different levels of classification complexity, we derived a second dataset, "Reddit AITA Binary", by filtering the

multiclass dataset to include only submissions with either a NTA or YTA classification.

Furthermore, we created two additional datasets, each containing the top two thousand samples by submission score from both the multiclass and binary classification datasets. This approach was inspired by the findings of Dettmers et al. regarding the Quantized Low-Rank Adaptors (QLoRA) fine-tuning technique, which suggests that smaller, high-quality datasets can produce state-of-the-art results. As detailed in Section 3.2 of our paper, we employed this technique to investigate whether the Flan-T5 and Llama-2-Chat models could achieve similar outcomes when fine-tuned on these smaller, high-quality datasets compared to the larger, comprehensive datasets. These will be referred to as "Reddit AITA Multiclass Top 2K" and "Reddit AITA Binary Top 2K". In contrast to the all-samples datasets, these datasets were deliberately stratified to maintain an equal representation of each AITA class. Consequently, each multiclass dataset consists of the top four hundred samples per class, while the binary datasets include the top one thousand samples each for YTA and NTA classifications.

Each dataset was divided into training and testing partitions following an 80/20 split ratio. For further insights into the size and classification breakdown of each dataset, readers are referred to Table 3.2, which provides counts for each AITA classification within the datasets.

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

Table 3.2: Reddit AITA Dataset Classifications

3.1.3 Dataset Analysis

A comprehensive analysis was conducted on each Reddit AITA dataset to analyze several key aspects. Initially, we examined the distributions of tokenized prompts lengths for both the Llama-

2-Chat and Flan-T5 models. Subsequently, we assessed the toxicity levels within the top comment reference texts across all datasets. Our analysis then extended to evaluating the agreement among the top ten commenters on their AITA classifications, both on a holistic level and pairwise. Lastly, we created a novel metric termed 'ambiguity score' to gauge the degree of consensus among the top ten commenters for each sample.

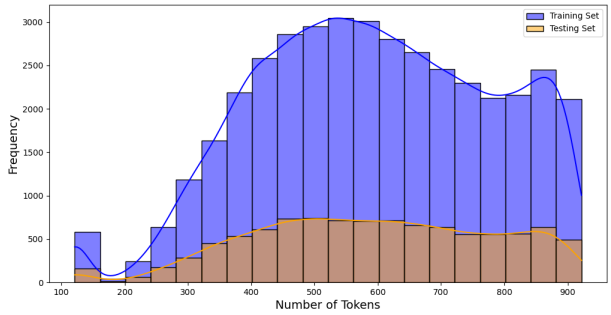
Tokenized Prompts Analysis

The analysis began with examining the lengths of tokenized instruction prompts in each dataset for the Llama-2-Chat and Flan-T5 models. The rationale behind this approach was before the prompts can be processed by the models, they must be converted, or tokenized, into a sequence of integers, with each integer, or "token", representing a segment of the prompt text. The Flan-T5 tokenizer uses the SentencePiece framework with a WordPiece algorithm and a vocabulary size of 32,000 [3]. The Llama-2-Chat tokenizer similarly uses SentencePiece, but differs in its adoption of the Byte-Pair Encoding (BPE) algorithm, and also has a vocabulary size of 32,000.

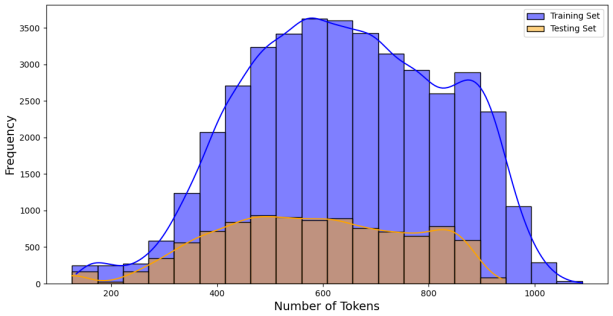
To better understand and visualize the token counts, we created histograms illustrating the distribution of tokenized prompt lengths for both models, as shown in Figure 3.1. These visualizations were essential for verifying that the prompt lengths were suitable for effective processing by both models. This was particularly critical for the Flan-T5 prompts to ensure they did not exceed the model's encoder context window of 1024 tokens. On the other hand, Llama-2-Chat, being a decoder-only model, requires a context window large enough to accommodate both the instruction prompt and the generated text. Fortunately, its context window spans 4096 tokens [4]. Given that no prompt exceeded 1024 tokens, this left 3072 tokens for generating text, which was more than sufficient for all reference texts. In contrast, the context window of the decoder for the instruction finetuned Flan-T5 models, which is for generated text, is significantly smaller, at 256 tokens, yet was also sufficient.

Top Comment Toxicity Analysis

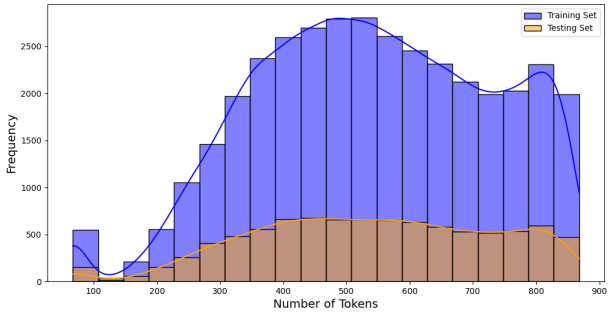
In the next phase of our analysis, focus was shifted to evaluating the toxicity levels of the top comments, which serve as reference texts in our datasets. To conduct this assessment, we utilized a



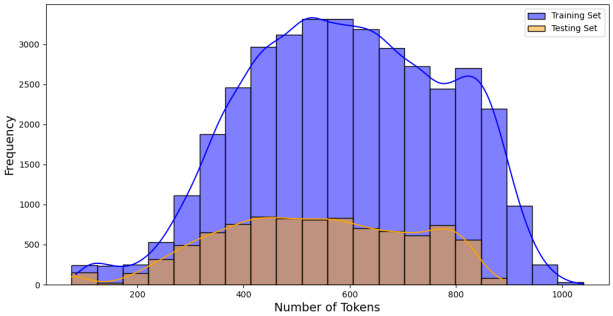
(a) Flan-T5 Token Counts (Multiclass)



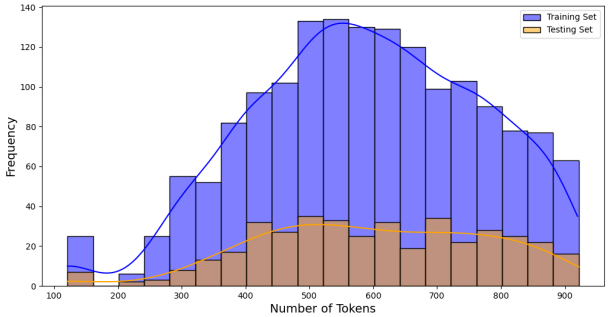
(b) Llama-2-Chat Token Counts (Multiclass)



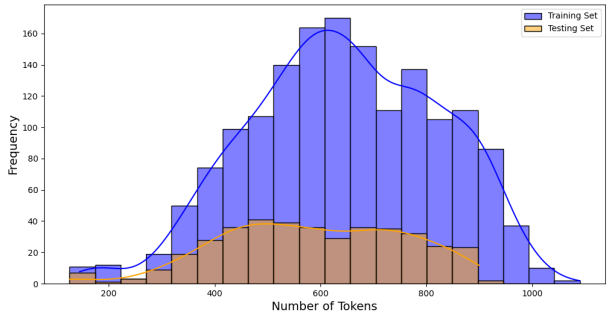
(c) Flan-T5 Token Counts (Binary)



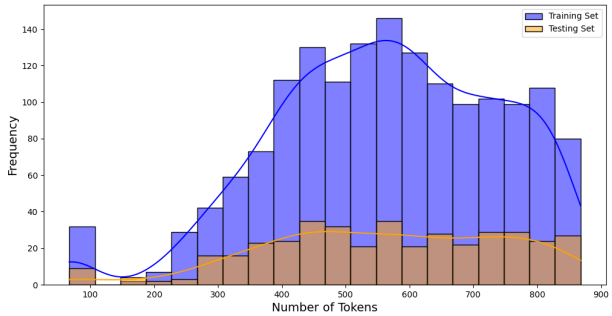
(d) Llama-2-Chat Token Counts (Binary)



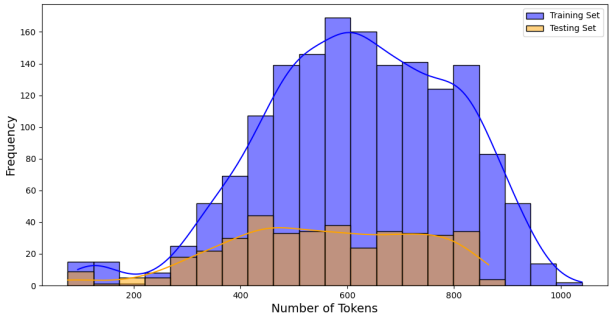
(e) Flan-T5 Token Counts (Multiclass Top 2K)



(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)

Figure 3.1: Tokenized Prompt Lengths for Reddit AITA Dataset Variants

RoBERTa model that was fine-tuned on the Toxigen dataset[3]. This dataset, featured in Hartvigsen et al.'s research "Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection" [19], provides finetuned models with an advanced capability to evaluate the toxicity of texts. The employment of this model was also utilized to assess the toxicity of generations by the Llama-2 model family during their development [4].

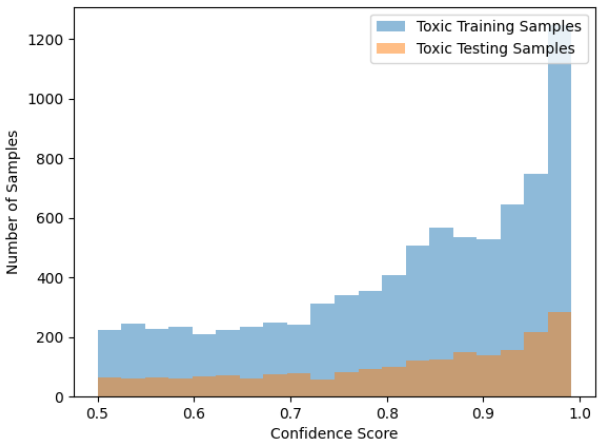
Table 3.4 presents the proportions of top comments classified as toxic in each dataset. The Toxigen dataset identifies toxic language as that which is derogatory, racist, or demeaning [19]. A key finding is the significant presence of toxicity across the datasets: the minimum toxicity rate in the training partitions was 0.182 in the Reddit AITA Binary dataset and the highest was 0.253 in the Reddit AITA Binary Top 2K dataset. In other words, about one fifth of the top comments in each dataset were toxic. This prevalence is concerning, especially considering the potential side effects on models trained with this data. Specifically, while these models aim to mimic human-like resolution of interpersonal conflicts, they risk adopting toxic responses, which could greatly undermine their usability in sensitive real-world applications, such as therapeutic settings.

By finetuning these models on datasets that contain toxic language, we evaluate the effectiveness of the safety mechanisms implemented during their initial development. If these models can be adapted to the intended tasks of the datasets without the need to reintroduce these safety measures, their practical applicability and ease of integration into real-world uses would significantly improve.

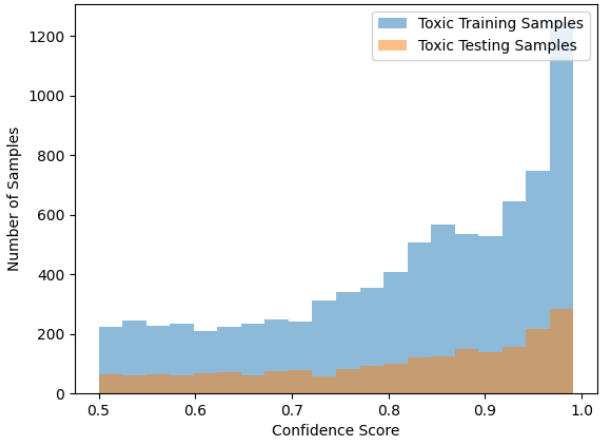
Holistic Commenter Agreement Analysis

We applied Krippendorff's alpha to assess the overall consensus among the top ten commenters in each dataset. This statistical measure is widely utilized across various fields such as communication, sociology, and content analysis to gauge the consistency and reliability of coding or rating processes. Krippendorff's alpha compares the observed agreement among raters with the agreement that would be expected by chance alone, offering a quantitative measure of agreement beyond chance. It evaluates all potential pairwise comparisons between raters or coding schemes, with weighting based on the analyzed categories and the number of raters involved. The resulting alpha coefficient

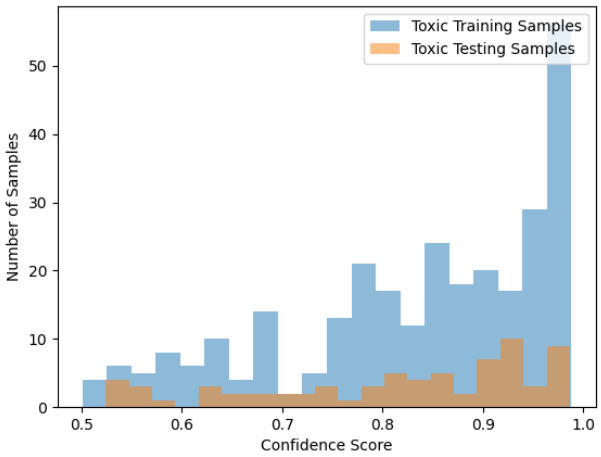
[3]The RoBERTa model finetuned on the Toxigen dataset is publicly available for use on https://huggingface.co/tomh/toxigen_roberta



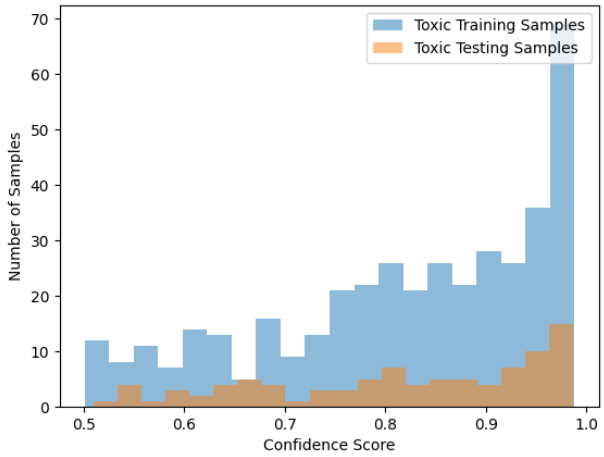
(a) Multiclass



(b) Binary



(c) Multiclass Top 2K



(d) Binary Top 2K

Figure 3.2: Toxic Samples in Reddit AITA Datasets

ranges from zero, indicating no agreement, to one, representing perfect agreement. Notably, values below 0.8 signify increasing levels of disagreement [20].

The Krippendorff's alpha coefficients for both the train and test partitions of each Reddit AITA dataset are contained in Table 3.5. Notably, three out of the four datasets exhibited values falling below the 0.8 threshold, indicating a statistically significant level of disagreement. The sole exception was the Reddit AITA Binary Top 2K dataset, albeit with values closely approaching the threshold at 0.844 and 0.832 for the train and test partitions, respectively. Additionally, an interesting observation emerged from the top two thousand sample datasets: while the formation of the Reddit AITA Multiclass Top 2K dataset from the overall multiclass dataset resulted in a decrease of Krippendorff's alpha by 11.62% and 11.80% for the train and test partitions, respectively, the creation of the Reddit AITA Binary Top 2K dataset from the overall binary dataset led to an increase of 12.23% and 9.62% for the corresponding partitions. This disparity suggests greater disagreement among commenters regarding the non-binary AITA classifications of NAH, ESH, and INFO compared to the binary ones of NTA and YTA, as their exclusion improves agreement. Furthermore, intentionally stratifying the data to ensure that 60% of all samples possess a non-binary classification, as done in the Reddit AITA Multiclass Top 2K dataset, appears to degrade agreement.

Pairwise Commenter Agreement Analysis

We utilized Cohen's kappa to evaluate the pairwise consensus among the top ten commenters in each dataset. Unlike Krippendorff's alpha, Cohen's kappa measures the agreement between two specific raters while considering the probability of chance agreement. Cohen's kappa values range from negative to positive one, with higher values indicating greater levels of agreement. A value of zero suggests agreement equivalent to that expected by random chance. Typically, values above 0.75 indicate high agreement, 0.4 to 0.75 denote moderate agreement, and those below 0.4 signify poor agreement (McHugh, 2012). Overall, calculating Cohen's kappa allowed us assess the relationships between commenters in their approach to interpersonal conflict resolution, thus complementing the broader perspective offered by Krippendorff's alpha.

The Cohen's kappa scores for each pair of top comments in each Reddit AITA dataset are

presented in Tables 3.6, 3.7, 3.8, and 3.9. Among the four Reddit AITA datasets, three of them: Multiclass, Multiclass Top 2K, and Binary, exhibited scores ranging from 0.4 to 0.75, indicating only moderate agreement. In contrast, the Binary Top 2K dataset displayed several scores surpassing the high agreement threshold of 0.75, although the maximum was 0.85, suggesting complete agreement was never achieved. Overall, these findings regarding pairwise agreement were consistent with the holistic agreement conclusions drawn from their Krippendorff's alpha scores.

Notably, the scores for the Reddit AITA Multiclass and Binary datasets for all commenter pairs remained identical, suggesting that removing the non-binary classifications did not affect the pairwise agreement. However, significant changes were observed in the scores for the two-thousand-sample datasets. Specifically, the Cohen's kappa scores for the Reddit AITA Multiclass Top 2K dataset were, on average, 30.23% lower than those of the Reddit AITA Multiclass dataset. This reduction in agreement is likely attributed to the intentional stratification of the data, with the two-thousand-sample multiclass dataset featuring a higher proportion of non-binary classifications (60% vs. 26% in the original multiclass dataset). This decline in agreement corresponds with the change observed in Krippendorff's alpha between these two datasets. Conversely, the Cohen's kappa scores for the Reddit AITA Binary Top 2K dataset were, on average, 17.78% higher than those of the Reddit AITA Binary dataset. This suggests that as the level of engagement, as indicated by the submission score for a binary interpersonal conflict, increased, the consensus between commenters improved.

This observation was consistent with expectations, as it's reasonable to anticipate that for increasingly popular subreddit submissions, users may be influenced to align with earlier commenters. This phenomenon, known as "conformity", has been extensively studied. Solomon Asch's classic experiments in the 1950s demonstrated how individuals often yield to group pressure, even when it contradicts their own judgment [21]. Since Asch's pioneering work, numerous studies have delved into the factors influencing conformity, such as group size and the presence of a dissenter. Social psychologists like Stanley Milgram and Muzafer Sherif have also made significant contributions to our understanding of conformity. Milgram's obedience studies examined the extent to which individuals obey authority figures, even when it involves potentially harming others [22]. Similarly, Sherif's Robbers Cave experiment provided insights into how group dynamics and competition can

shape attitudes and behaviors [23]. These studies collectively illustrate that humans frequently adjust their opinions or behaviors to match those of the majority, driven by desires to fit in, fear of rejection, or beliefs in the correctness of the group.

Individual Sample Commenter Agreement Analysis

Calculating Krippendorff's alpha and Cohen's kappa scores for each dataset offered both pairwise and holistic perspectives on commenter agreement. However, these measures couldn't assess the level of agreement between each commenter on an individual sample basis. To address this gap, we developed a novel metric termed the "ambiguity score". This metric evaluates the level of agreement among individual commenters for each sample and as a result provides a unique perspective on the degree of ambiguity between the each sample's AITA classifications.

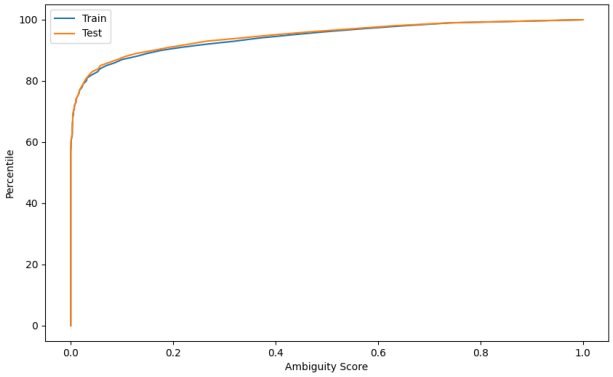
Equation 3.1 represents the formula for computing the ambiguity score. We initially assigned a numerical value ranging from one to five to the classifications provided by the top ten comments: one for 'YTA', two for 'ESH', three for 'INFO', four for 'NAH', and five for 'NTA'. This scheme was deliberately chosen to reflect the relationships between these classifications, particularly emphasizing the direct opposition between YTA and NTA. Subsequently, the ambiguity score was calculated by applying a parabolic function to the mean (μ) and standard deviation (σ) of these numerical representations. This function was tailored to amplify the measure of disagreement, particularly in cases where classifications leaned heavily towards the YTA or NTA extremes.

$$\text{Ambiguity Score} = \sigma \times (2 - |3 - \mu|)^2 \tag{3.1}$$

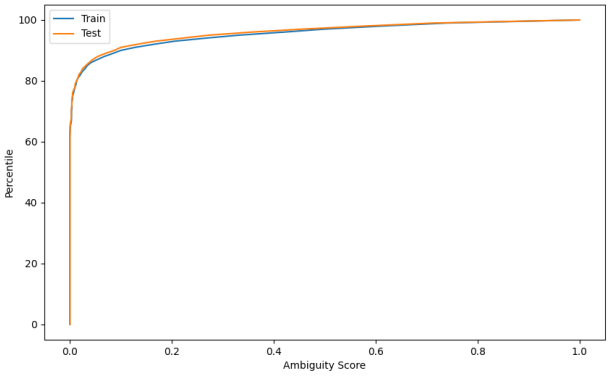
Figure 3.3 illustrates the cumulative frequency of ambiguity scores for each Reddit AITA dataset. To complement this, the proportions of unambiguous samples were calculated and are presented in Table 3.10. The Multiclass, Binary, and Multiclass Top 2K datasets exhibited similar proportions of unambiguous samples, ranging from 0.576 to 0.623, indicating that approximately 40% of samples contained ambiguity in their AITA classification. In contrast, only 0.242 of the samples in the Multiclass Top 2K dataset had zero ambiguity, indicating that approximately 75% of samples contained ambiguity in their AITA classification. This observation aligns with the findings

from analyzing commenter agreement holistically via Krippendorff's alpha and pairwise via Cohen's kappa, both of which indicated increased levels of disagreement in the Reddit AITA Multiclass Top 2K dataset. Thus, this discrepancy is also likely a result of the intentional data stratification to include equivalent proportions of each of the five AITA classifications.

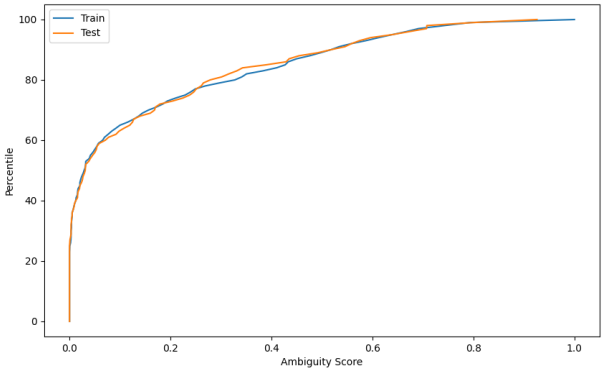
Nevertheless, computing the ambiguity scores confirmed that each dataset contained at least a moderate amount of ambiguity on an individual sample basis. This supports the notion that interpersonal conflict resolution in a real-world context is non-deterministic. Therefore, for optimal performance, models trained on this data must learn that the correct response isn't always uniform and can depend on complex social factors.



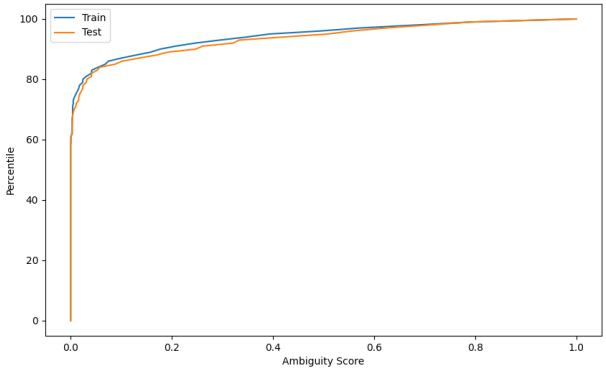
(a) Multiclass



(b) Binary



(c) Multiclass Top 2K



(d) Binary Top 2K

Figure 3.3: Cumulative Frequency of Ambiguity Scores in Reddit AITA Datasets

3.2 Instruction Finetuning Procedure

The instruction finetuning process was applied to four base models: Flan-T5 XL, Flan-T5 XXL, Llama-2-Chat 7B, and Llama-2-Chat 13B. Each model was finetuned for one epoch on all four of the Reddit AITA datasets, culminating in sixteen unique finetuned models. To facilitate the finetuning, each submission text was augmented with either the binary or multiclass AITA instruction in Table 3.11, and then formatted into prompts that are compatible with the respective architectures of Flan-T5 and Llama-2-Chat models.

To instruction finetune the models, we used QLoRA (Quantized Low Rank Adaption), which is a method that allows for the efficient finetuning of quantized 4-bit models without performance degradation. Quantization is the process of reducing the precision or bit-width of the weights within a model. QLoRA was introduced by Dettmers et al. and focuses on backpropagating gradients through a frozen, 4-bit quantized pretrained language model into Low Rank Adapters (LoRA), thereby enabling fine-tuning of massive models on relatively modest hardware [24]. Central to QLoRA is the concept of Low Rank Adaptors (LoRA). LoRA was invented by Hu et al. and involves decomposing the larger weight matrices found in the transformer architectures of LLMs into smaller, more manageable low-rank matrices [25]. During the training process, these low-rank matrices are learned and subsequently applied to the original weight matrices via multiplication and addition. This approach enables the model to adapt to specific tasks without altering the pre-trained weights, thereby maintaining the extensive knowledge embedded in them while still allowing for task-specific adaptability through a smaller subset of parameters. In terms of memory efficiency, QLoRA introduces several notable techniques, including using the 4-bit NormalFloat (NF4) data type for optimally handling normally distributed weights and paged optimizers to improve the management of memory spikes.

QLoRA's testing across various model types, including LLaMA and T5, has shown that it can achieve state-of-the-art results on a small, high-quality dataset even with models smaller than those used in previous benchmarks. A particularly impressive feat of QLoRA is its ability to fine-tune a 65B parameter model on a single 48GB GPU, a task previously requiring over 780 GB of GPU memory [24]. The advent of QLoRA thus marked a significant advancement in making the fine-

tuning of large language models more practical and accessible. By reducing the computational and memory overhead, QLoRA opens new possibilities for instruction fine-tuning at scales previously considered impractical.

Our use of QLoRA inspired the creation of the Reddit AITA Top 2K datasets. Dettmer et al. demonstrated that QLoRA significantly enhances performance when fine-tuning on a small, high-quality dataset consisting of a few thousand samples. We designed the Top 2K datasets to exemplify such datasets by only including submissions with the highest community scores. Additionally, we explored stratifying the AITA classifications within these datasets to determine whether maintaining equal distributions of interpersonal conflict types, rather than natural distributions, enhances dataset quality. We hypothesized that equal distributions could prevent the model from overlearning dominant classifications like NTA and underrepresenting others like INFO, which could otherwise be dismissed as noise due to their rarity in the training data.

By using QLoRA for the instruction finetuning process, all finetunings were completed with a single Nvidia L40 GPU, which has 48 GB of VRAM. QLoRA's memory optimization capabilities allowed us to accelerate training by using a batch size of sixteen for all models while staying within the memory limits of the L40 GPU. [4]

3.2.1 Flan-T5 Instruction Finetuning

We configured the instruction finetuned Flan-T5 models with an encoder context window of 1024 tokens and a decoder context window of 256 tokens. This adjustment was feasible despite the original Flan-T5 model being trained with a context window of 512 tokens for both encoder and decoder. Such flexibility in context window size is a feature of the T5 architecture, which employs relative, as opposed to fixed, positional embeddings[3].

The QLoRA configuration for the Flan-T5 models involved several key settings: an attention dimension of 64, an alpha scaling parameter set at 16, and a dropout probability for the LoRA layers of 0.05. For quantization, we employed a 4-bit NormalFloat (NF4) data type, and during training, a 16-bit bfloat (brain float) was used as the computation data type. We maintained a

[4]For transparency and reproducibility, the code for finetuning the Llama-2-Chat and Flan-T5 models on the Reddit AITA datasets can be found at https://github.com/MattBoraske/Reddit_AITA_Finetuning/tree/main/finetuning

batch size of 16 throughout the training.

In terms of optimization, gradient clipping was implemented with a maximum gradient norm of 0.3. We chose Adafactor as the optimizer, set the initial learning rate to 5E-4, and used a constant scheduler, aligning with the original training approach of the Flan-T5 models[3].

For the Flan-T5 models, we adopted a sequence-to-sequence translation approach by prefixing each instruction to the submission text. This was consistent with the original training paradigm of Flan-T5, wherein each training task was constructed as a text-to-text translation, utilizing an instruction prefix to guide the model's responses[3]. An example of a binary classification prompt for the Flan-T5 models can be found in Table 3.12

3.2.2 Llama-2-Chat Finetuning

In the fine-tuning process for the Llama-2-Chat models, we maintained consistency with the Flan-T5 models in aspects such as gradient clipping, batch size, and the application of QLoRA. However, there were several crucial adaptations made to suit the architecture and training requirements of Llama-2 models.

A key distinction lies in the architecture of Llama-2, which is decoder-only. Consequently, we redefined the fine-tuning task to focus on causal language modeling. This task involves predicting the next token in a given sequence, wherein the model's attention is restricted to tokens on its left, effectively rendering it unable to foresee future tokens in the sequence. To facilitate casual language modeling, the Llama-2 family of models uses a specific prompt template that includes both a system and user prompt portion which are separated by a set of delimiter tokens [4]. System prompts act as a framework or template for the model's responses. They are used to standardize outputs, incorporate specific response styles or structures, and align the model's outputs with desired outcomes or objectives. The user prompt serves as the starting point or the query for the model. It sets the context and guides the model on what kind of information or response is expected. Therefore, we used the Redit AITA classification instruction as the system prompt and the sample submission text as the user prompt. An example of a binary classification prompt for the Llama-2-Chat models can be found in Table 3.12

To stay aligned with the original training methodology of the Llama-2 models, we used the

AdamW optimizer, initiated the finetuning with a learning rate of 2E-4, and employed a cosine annealing scheduler [4].

3.3 Evaluation Process

To assess each models' efficacy in resolving interpersonal conflicts within the Reddit AITA subreddit framework, we evaluated both their ability to correctly classify the writer's behavior into one of the five categories outlined in Table 3.1 and generate anthropomorphic yet safe justifications. Table 3.3 offers a typical example of how users undertake this task on the AITA subreddit. Given a submission, a user posts a comment where their AITA classification is presented first before being followed by their justification. This is enforced by the moderators who regulate the AITA subreddit and was key in our evaluation as it provided a clear method to parse the classification from the text generated by the models.

Our evaluation approach was thus twofold: firstly, an analysis of the classifications made by each model and, secondly, an analysis of the justifications, benchmarked against the reference texts sourced from top comments. The performance of each model was evaluated in a zero-shot scenario as well as after being instruction finetuned on the Reddit AITA datasets. This comparison was instrumental in understanding the impact of finetuning on each model's abilities to accurately classify interpersonal conflicts and generate contextually appropriate, anthropomorphic, and safe justifications. [5]

3.3.1 Evaluation of Classifications

In our evaluation of the finetuned models' proficiency in classifying interpersonal conflicts, a critical factor to account for was the imbalanced nature of the Reddit AITA datasets, as shown in Table 3.2. To address this, we calculated two complementary metrics: the F1 score and the Matthews correlation coefficient (MCC), each particularly suited for imbalanced data scenarios.

The F1 score is a harmonized measure of precision and recall. Precision represents the ratio of true positive identifications to all positive identifications made by the model, while recall quantifies

[5]For transparency and reproducibility, the code for evaluating the Llama-2-Chat and Flan-T5 models, both in a zero-shot context and after being finetuned, is available at https://github.com/MattBoraske/Reddit_AITA_Finetuning/tree/main/evaluation

the proportion of actual positives that were correctly identified. The F1 score, being the harmonic mean of these two, offers a balanced view of a model's accuracy, effectively accounting for both false positives and false negatives prevalent in imbalanced datasets.

Complementing the F1 score, the MCC provided a second, robust dimension to our evaluation of the classifications. It measures the correlation between the model's predictions and actual classifications, with values ranging from negative one for completely inverse predictions, through zero for random predictions, to positive one for perfect predictions. Its robustness lies in its ability to factor in true and false positives, as well as negatives, thus offering a comprehensive and balanced evaluation regardless of dataset size and class distribution [26].

By calculating both the F1 score and the MCC, we engineered redundancy into our evaluation process. This dual approach allowed us to thoroughly assess the models' classification abilities, mitigating the challenges posed by the imbalanced nature of the datasets.

Additionally, to better understand the disparities between model predictions and correct AITA classifications, we generated confusion matrices. These were powerful tools for visualizing the classification performance of the models as they highlight not only the accuracy, but also the types of errors made. Each matrix illustrates how often each predicted class corresponds to each actual class, thereby helping with identifying misclassification patterns and areas where the models require further improvement.

3.3.2 Evaluation of Justifications

We started our assessment of the quality of the justifications by calculating the ROUGE LSum for each model. Calculating ROUGE-Lsum is valuable as it provides a comprehensive measure of the overlap between the longest common subsequences in generated text and a set of reference texts, thereby assessing their similarity. It is also particularly useful because it captures not just the presence of key phrases, but also their arrangement and cohesion in the summary [27].

To collect a complementary statistic to ROUGE-LSum, we calculated a COMET (Crosslingual Optimized Metric for Evaluation of Translation) score for each model. COMET is a framework for training multilingual machine translation (MT) evaluation models. It leverages recent advancements in cross-lingual pretrained language models to create MT evaluation models that are highly

multilingual and adaptable, using information from both the source input and a target-language reference translation. COMET models are trained to estimate different types of human judgments on MT quality, such as Direct Assessments (DA), Human-mediated Translation Edit Rate (HTER), and Multidimensional Quality Metrics (MQM). COMET's significance lies in its ability to correlate strongly with human judgments, addressing the challenges of segment-level correlation and differentiating high-performing MT systems. It surpasses traditional metrics like BLEU, which struggle with semantic similarity beyond the lexical level. For the purposes of this study, the COMET model we used was WMT22-Comet-Da, the recommended and most recently released version by the authors of the COMET framework at the time of this writing [28].

To assess the toxicity of the generated justifications, we utilized ConfliBERT, a BERT model fine-tuned on the Toxigen dataset, to generate toxicity scores for all model-generated responses. This model was the same one utilized to evaluate the toxicity of the top comments used as reference texts during fine-tuning, with further details provided in Section 3.1.3. By comparing the toxicity scores of the generated justifications to those of the reference top comments, we gauged the extent to which the models could resist learning the toxic characteristics of the data during the fine-tuning process while still acquiring the ability to provide contextually appropriate justifications for their AITA classifications.

Submission Text	Top Comment
I (m34) bought my boyfriend (m32) a 2018 MacBook Air on sale a few months ago because he needed a new computer. It did not come with a warranty. 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" and because I was so stupid to buy a laptop without a warranty. I told him he should learn not to have temper tantrums. He has a stable, salaried job. I understand that he doesn't want to pay to fix it, but I don't see how this is my problem. AITA?	NTA. That is classic manipulative behavior. I'd watch out that he doesn't continue to blame you for his actions.
My friend went to a foreign country outside of the United States in Europe and came back with gifts. My friend knew I like coffee so he got me coffee that I know I won't drink. It's a dark roast and there's no roast by date, which means it's not fresh. I gently put it back in the bag and politely told them why I don't drink that kind of coffee. Him and his friends said I was being ungrateful and rude and that I shouldn't have thrown the coffee in the bag, which I didn't. AITA for not accepting a gift I wouldn't use?	YTA. Jeez, could you have been more snobby? You could have just accepted it and not used it rather than your pretentious "I only use fresh beans".
My husband's friend passed away months ago. leaving behind his widow wife. my husband considered him a brother and so he began offering help to his widow and showing up literally every weekend (that's when he's off work) to help her with everything. it didn't bother me...i mean not much but still. now he told me that it'd be nice if we invite her to christmas dinner with us since she has no family and other friends are busy. i refused but he insisted saying it'd mean so much to her especially since this is her first christmas without her husband and we should comfort her. i refused and said that it's supposed to be a family celebration and my family will show up so there's enough amount of guests already but he said she the one who needs this family atmosphere the most. i snapped and firmly told him i will not agree to let him invite her to christmas dinner with us and he needed to drop it cause it made me uncomfortable. he argued then started sulking about it saying that i was being cruel for no reason and that he wasn't asking for much. just an extra chair/plate for her. but i said it will be awkward for me and my family especially since we're not so close. AITA?	INFO. have you clearly communicated to your husband how much you deeply resent him spending time with this woman? because that seems to be the larger issue here to me.
I'm at a loss for words. i sorta recently proposed to my gf. everything is going sorta good. we were just watching a documentary about the diamond trade. the concept of synthetic diamonds came up. i actually mentioned to her, that that's how her ring was made. then, she got a bit disappointed. she keeps on calling it fake, which it's not. it's chemically and physically the same as any other diamond. she says that she feed betrayed. i don't know how or why she feels like that. again, it's not like it was cheap. $300 is a pretty low price, though. also, i'm not supporting the very shady diamond trade. so, AITA?	NAH. i agree that synthetic diamonds are better given how terrible the diamond trade is. i also get being a little disappointed since she didn't realize it was a synthetic diamond; usually that kind of preference is discussed beforehand. i'm sure she'll get over it.
I'm half japanese and my family recently moved into my stepdad's house. i speak english with my brothers but japanese with my mom. at the dinner table, my stepdad will get angry if we speak japanese because he feels left out. he frequently stands up and declares that he is left out and leaves the room to watch tv. i have major problems with him as he frequently trash talks my dad and screams at us. i live in australia and my mother is the only connection i have with my japanese heritage and being able to speak japanese is really important to me. i don't feel obligated to speak english around him just because he wants to hear what i'm talking to my mom about.	ESH. you have plenty of opportunity to keep your japanese fluent when you aren't at the dinner table with someone who doesn't speak it. so be honest. you don't speak english because you want to exclude him. that makes you an a**hole. him trash talking your dad and verbally abusing you makes him an a**hole. you wanting to practice japanese is fine, just do it when you and your mom are alone together, not in a group setting like the dinner table.

Table 3.3: Example Submissions and Top Comments for AITA Classifications

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

Table 3.4: Top Comment Toxicity Rates in Reddit AITA Datasets

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

Table 3.5: Krippendorff's Alpha for Reddit AITA Datasets

	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	-	-	-	-	-	-	-	-	-	-

Table 3.6: Cohen's Kappa for Top Comment (TC) Pairs in Reddit AITA Multiclass Dataset

	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	-	-	-	-	-	-	-	-	-	-

Table 3.7: Cohen's Kappa for Top Comment (TC) Pairs in Reddit AITA Binary Dataset

	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	-	-	-	-	-	-	-	-	-	-

Table 3.8: Cohen's Kappa Scores for Top Comments (TC) in Reddit AITA Multiclass Top 2K Dataset

	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	-	-	-	-	-	-	-	-	-	-

Table 3.9: Cohen's Kappa for Top Comment (TC) Pairs in Reddit AITA Binary Top 2K Dataset

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

Table 3.10: Proportions of Samples in Reddit AITA Datasets with Zero Ambiguity

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.

Table 3.11: AITA Instructions for Multiclass and Binary Classification

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]

Table 3.12: Reddit AITA Binary Classification Prompt Examples for Flan-T5 and Llama-2-Chat

Chapter 4
Results

We report the performance of the Llama-2-Chat and Flan-T5 models on the Reddit AITA datasets, as detailed in Table 3.2, in both zero-shot contexts and after instruction finetuning. This comparison provides insights into each model's ability to learn the dual task of classifying interpersonal conflicts and providing justifications. We first examined the models' performance on the Reddit AITA multiclass dataset, followed by their results on the Reddit AITA Binary dataset. Additionally, we assessed the models' efficacy when finetuned on the top two thousand sample versions of each dataset, to determine if a smaller, higher quality subset of the dataset can achieve similar levels of performance.

Overall, the instruction finetuning resulted in sixteen unique models. In this section, we present results for the larger model variants of Llama-2-Chat-13B and Flan-T5-XXL. To test for redundancy and to see if similar performance enhancements occur with smaller model sizes, we also finetuned Llama-2-7B-Chat and Flan-T5-XL on each of the four Reddit AITA datasets, the results of which are detailed in Appendix A.[6]

4.1 Multiclass Classification Models

All four base models of Flan-T5 XXL, Flan-T5 XL, Llama-2-7B-Chat, and Llama-2-13B-Chat were instruction finetuned for on the Reddit AITA Multiclass and Reddit AITA Multiclass Top 2K datasets. This resulted in eight unique multiclass classification models.

[6]All of the instruction finetuned Flan-T5 and Llama-2-Chat models are publicly available for download and use on https://huggingface.co/collections/MattBoraske/reddit-aita-finetuning-66038dc9281f16df5a9bab7f

4.1.1 Models Finetuned on Reddit AITA Multiclass Dataset

Flan-T5 XXL Performance

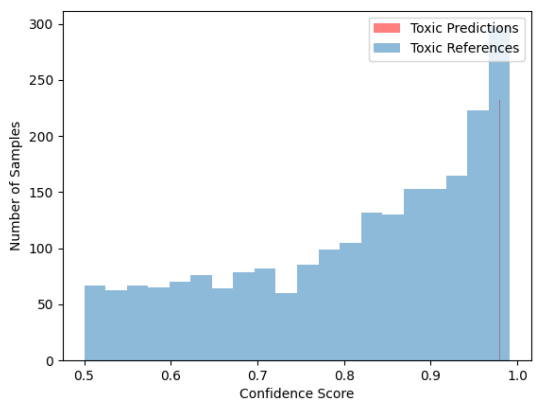
Instruction finetuning the Flan-T5 XXL model on the Reddit AITA Multiclass dataset led to significant enhancements in both the classification of interpersonal conflicts and the quality of generated justifications. Specifically, as shown in Table 4.1, the classification performance saw a modest increase in precision from 0.69 to 0.75 and there were substantial improvements in recall, F1-score, and Matthew's correlation coefficient (MCC), with recall rising from 0.35 to 0.81, F1-score from 0.40 to 0.78, and the MCC from 0.032 to 0.314. In terms of generated justifications, the ROUGE-Lsum score improved from 0.025 to 0.161, and the average COMET score increased from 0.314 to 0.515.

However, these advancements in both classification and justification generation came at a significant cost. The instruction finetuned model started producing justifications with considerably more toxic language, with the toxicity rate increasing from 0.063 to 0.268. This exceeded the toxicity rate of 0.224 observed in the top reference comments, indicating that the model now generates responses with higher toxicity than those typically made by humans on the AITA subreddit. Figure 4.1 visualizes this increase in justification toxicity.

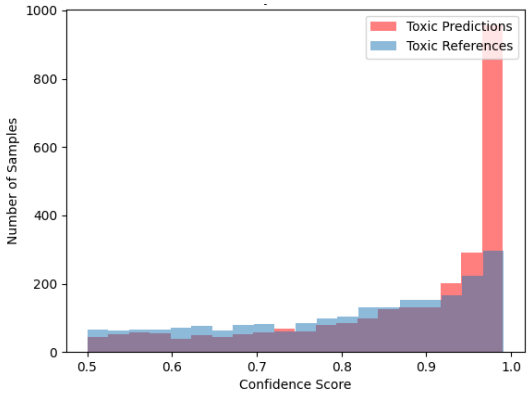
The predicted classifications of the model were compared with those of the reference top comments and the results are shown in Figure 4.2. This comparison revealed that the primary performance improvement was caused by the model learning to classify more interpersonal conflicts as NTA rather than YTA. Additionally, before instruction finetuning, the model rarely selected any of the non-binary classifications (INFO, NAH, and ESH). Afterwards, it completely ceased selecting these rarer categories, suggesting that it now treats these classifications as noise.

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

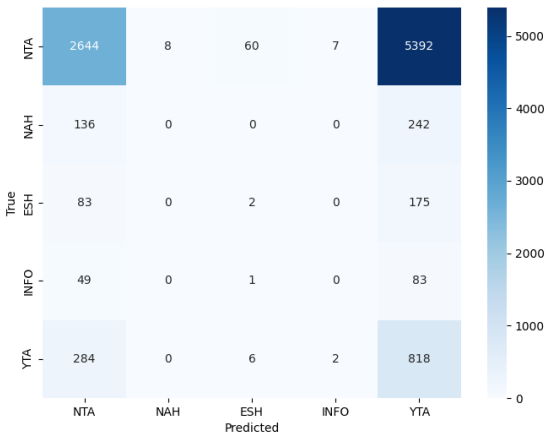


(a) Zero-shot Toxic Generations

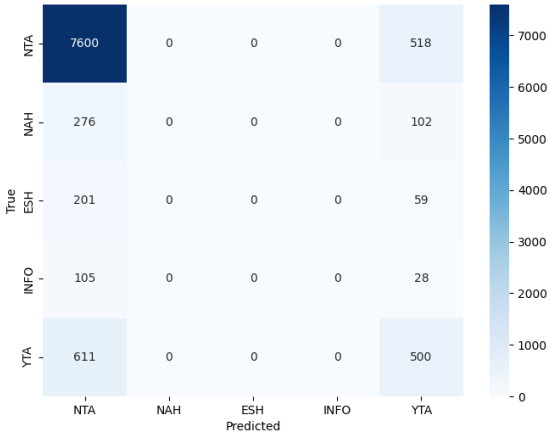


(b) Finetuned Toxic Generations

Figure 4.1: Toxic Generations by Flan-T5 XXL on Reddit AITA Multiclass Dataset



(a) Zero-shot Classifications



(b) Finetuned Classifications

Figure 4.2: Classifications by Flan-T5 XXL on Reddit AITA Multiclass Dataset

Llama-2-13B-Chat Performance

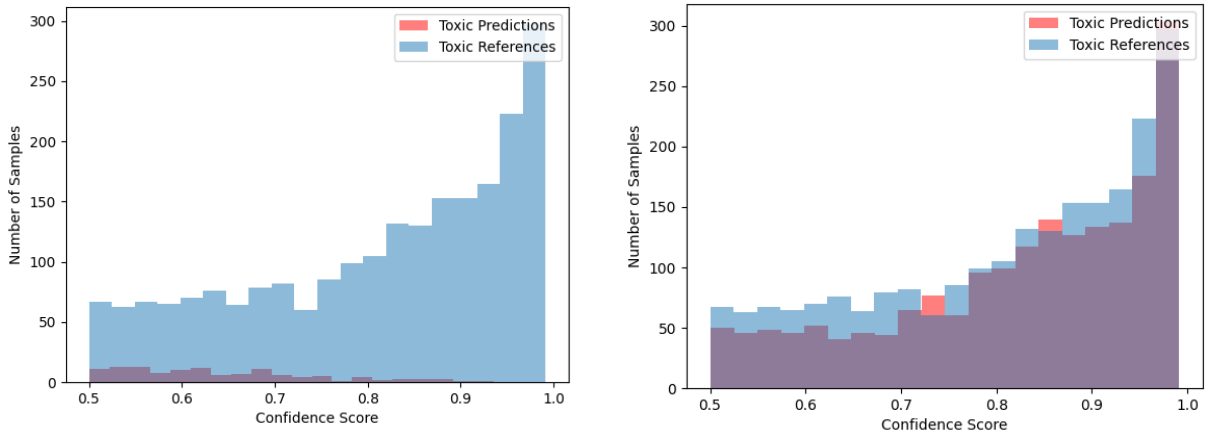
Fine-tuning the Llama-2-13B-Chat model on the Reddit AITA Multiclass dataset improved its classification capabilities, but slightly weakened its ability to generate contextually appropriate justifications.

Regarding classification accuracy, as shown in Table 4.2, the model saw a slight decrease in precision from 0.73 to 0.72. However, it achieved substantial improvements in recall, increasing from 0.29 to 0.78, and in F1 score, rising from 0.39 to 0.75. The Matthew's correlation coefficient (MCC) also improved moderately, from 0.055 to 0.165. Similar to the Flan-T5 XXL model, Figure 4.4 shows that these gains primarily stemmed from the model's tendency to predominantly choose either NTA or YTA. Unlike Flan-T5 XXL, Llama-2-13B-Chat did not completely stop selecting the non-binary options of INFO, NAH, or ESH. However, it still favored these options only marginally, and in comparison to its selection of the binary options, it appears that Llama-2-13B-Chat, like Flan-T5 XXL, treats these as relatively insignificant, albeit to a lesser degree..

In terms of justification quality, both the ROGUE Lsum and the Average COMET score saw marginal declines, from 0.136 to 0.122 and from 0.573 to 0.514, respectively. Additionally, the toxicity rate significantly increased from 0.012 to 0.190. This change is visualized in Figure 4.3. Although this toxicity rate was lower than that observed in the top comment references, unlike the instruction finetuned Flan-T5 XXL model which was higher, the increase in toxicity is particularly concerning given that there were no improvements in the performance of generated justifications.

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

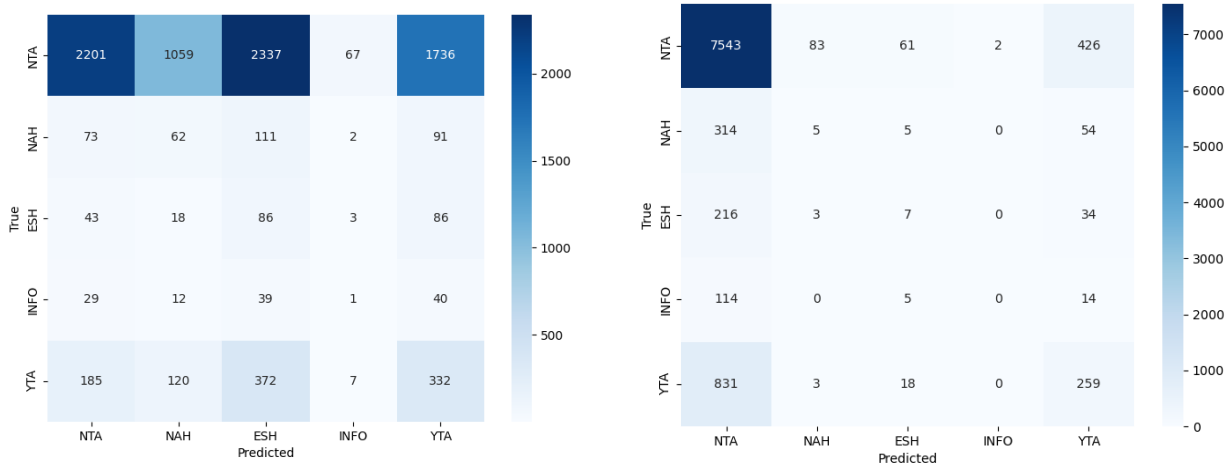
Table 4.2: Performance of Llama-2-13B-Chat on Reddit AITA Multiclass Dataset



(a) Zero-shot Toxic Generations

(b) Finetuned Toxic Generations

Figure 4.3: Toxic Generations by Llama2-13B-Chat on Reddit AITA Multiclass Dataset



(a) Zero-shot Classifications

(b) Finetuned Classifications

Figure 4.4: Classifications by Llama2-13B-Chat on Reddit AITA Multiclass Dataset

4.1.2 Models Finetuned on Reddit AITA Multiclass Top 2K Dataset

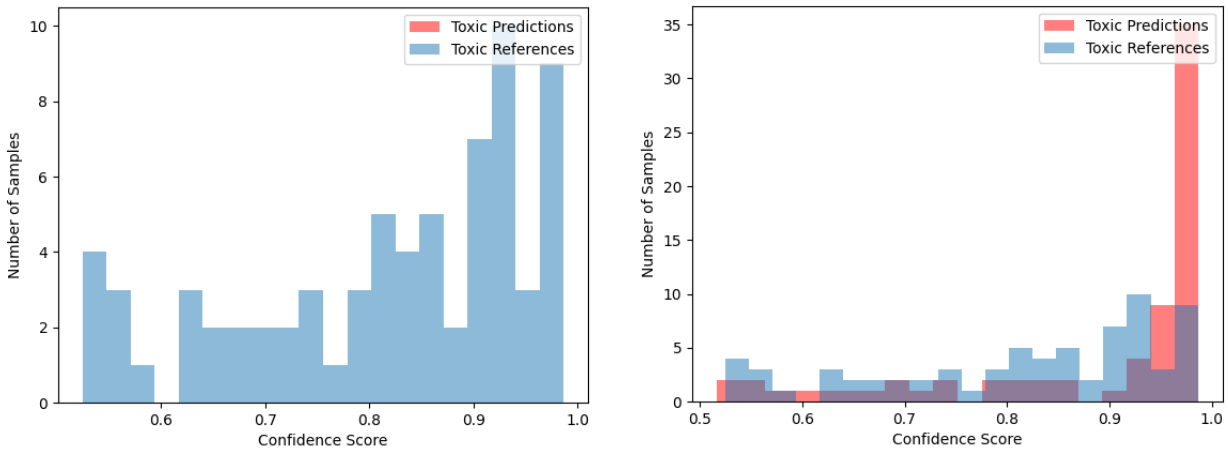
Flan-T5 XXL Performance

Instruction finetuning the Flan-T5 XXL model on the Reddit AITA Multiclass Top 2K dataset resulted in no improvements in its classification capabilities; however, it did moderately enhance the quality of generated justifications. These results are summarized in Table 4.3.

The lack of improvement in classification accuracy was demonstrated by slight decreases in precision and recall from 0.18 to 0.16 and 0.22 to 0.19, respectively, and a marginal increase in the F1 score from 0.12 to 0.16. Given that random chance would result in an accuracy of 0.20 for five classifications, these metrics suggest that the model performs worse than random chance. This conclusion is reinforced by the MCC score, which decreased from 0.030 to -0.010. A negative MCC score, particularly one less than zero, indicates performance inferior to random guessing [26].

The model's predicted classifications were compared with those of the reference top comments in Figure 4.6. This comparison showed that after instruction fine-tuning Flan-T5 XXL on the Multiclass Top 2K dataset, the model predominantly began to choose the NAH and INFO classifications instead of NTA and YTA. Interestingly, the model, which previously rarely selected ESH, now ceased to select it entirely. While this behavior differs from when it was finetuned on the Multiclass dataset as two of the non-binary classifications are being selected, their usage by the model does not accurately reflect correct classifications.

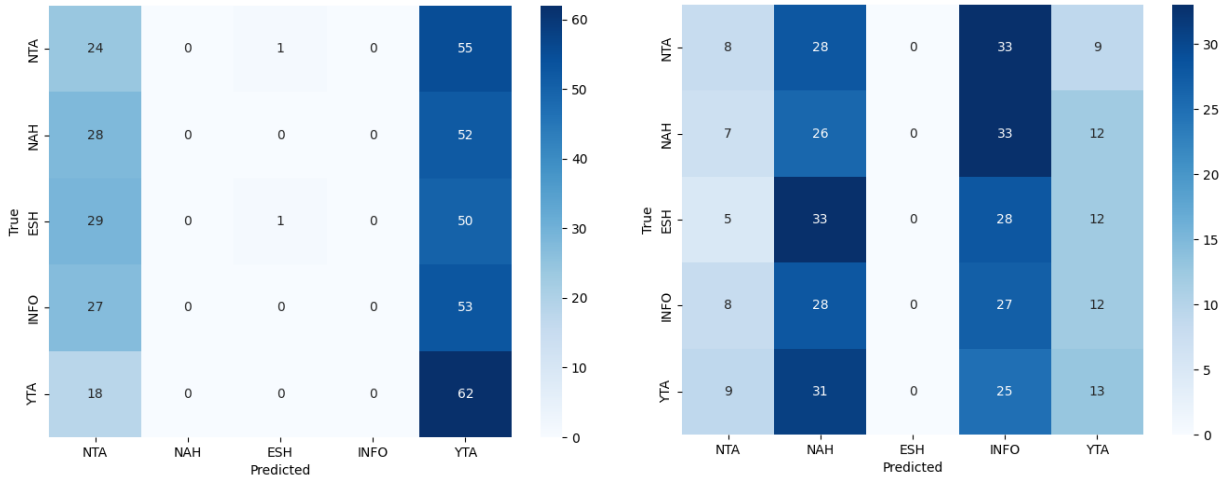
The improvements in ROUGE Lsum and Average COMET score were moderate, rising from 0.016 to 0.101 and from 0.308 to 0.448, respectively. However, this progress came at the expense of the toxicity rate, which as shown in Figure 4.5 increased from zero to 0.178. Notably, these increases were smaller than those from fine-tuning Flan-T5 XXL on the larger Reddit AITA Multiclass dataset. This suggests that the model has the potential to learn further, and that finetuning on only two thousand samples may limit the model from achieving its full learning capacity. Additionally, the toxicity rate's rise indicates a trade-off between model accuracy and the generation of undesirable content, highlighting the challenges of balancing performance improvements with ethical concerns.



(a) Zero-shot Toxic Generations

(b) Finetuned Toxic Generations

Figure 4.5: Toxic Generations by Flan-T5 XXL on Reddit AITA Multiclass Top 2K Dataset



(a) Zero-shot Classifications

(b) Finetuned Classifications

Figure 4.6: Classifications by Flan-T5 XXL on Reddit AITA Multiclass Top 2K Dataset

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

Table 4.3: Performance of Flan-T5-XXL on Reddit AITA Multiclass Top 2K Dataset

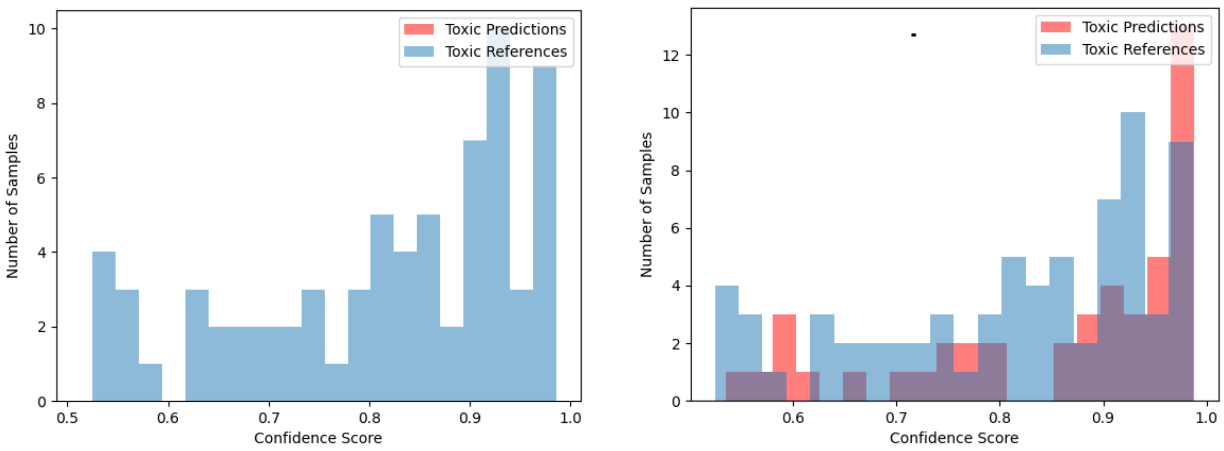
Llama-2-13B-Chat Performance

Instruction fine-tuning Llama-2-13B-Chat on the Reddit AITA Multiclass dataset resulted in no improvements in its classification or justification capabilities. As shown in Table 4.4, All metrics remained approximately the same, except for a moderate decrease in the Average COMET score from 0.596 to 0.515 and an increase in the toxicity rate from zero to 0.113, as shown in Figure 4.7. The clear outcome is that fine-tuning this model on the dataset has only deteriorated the conversational abilities it had acquired during initial training while also introducing a chance to generate toxic justifications.

The model's predicted classifications were compared to those of the reference top comments in Figure 4.8. This analysis revealed that the model overlearned the intentional stratification of classifications within the dataset, resulting in a more even selection among the five categories. However, this did not translate to appropriate decision-making, as the classifications are heavily scattered and lack any clear trend or accuracy. The model's approach reflects its overemphasis on dataset stratification without grasping the correct contexts for each classification.

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

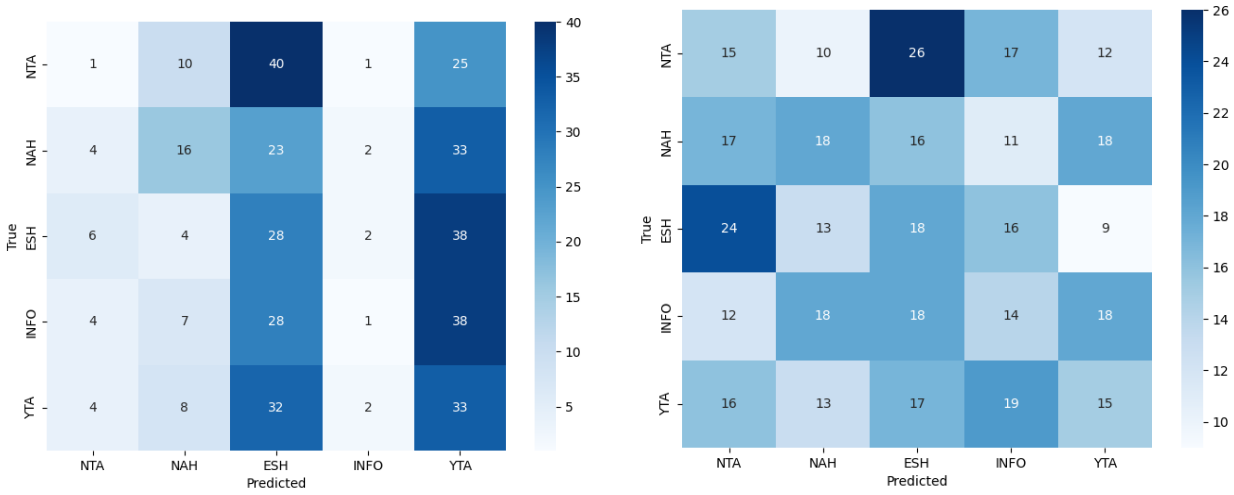
Table 4.4: Performance of Llama-2-13B-Chat on Reddit AITA Multiclass Top 2K Dataset



(a) Zero-shot Toxic Generations

(b) Finetuned Toxic Generations

Figure 4.7: Toxic Generations by Llama2-13B-Chat on Reddit AITA Multiclass Top 2K Dataset



(a) Zero-shot Classifications

(b) Finetuned Classifications

Figure 4.8: Classifications by Llama2-13B-Chat on Reddit AITA Multiclass Top 2K Dataset

4.2 Binary Classification Models

All four base models of Flan-T5 XXL, Flan-T5 XL, Llama-2-7B-Chat, and Llama-2-13B-Chat were instruction finetuned on the Reddit AITA Binary and Reddit AITA Binary Top 2K datasets. This resulted in eight unique binary classification models.

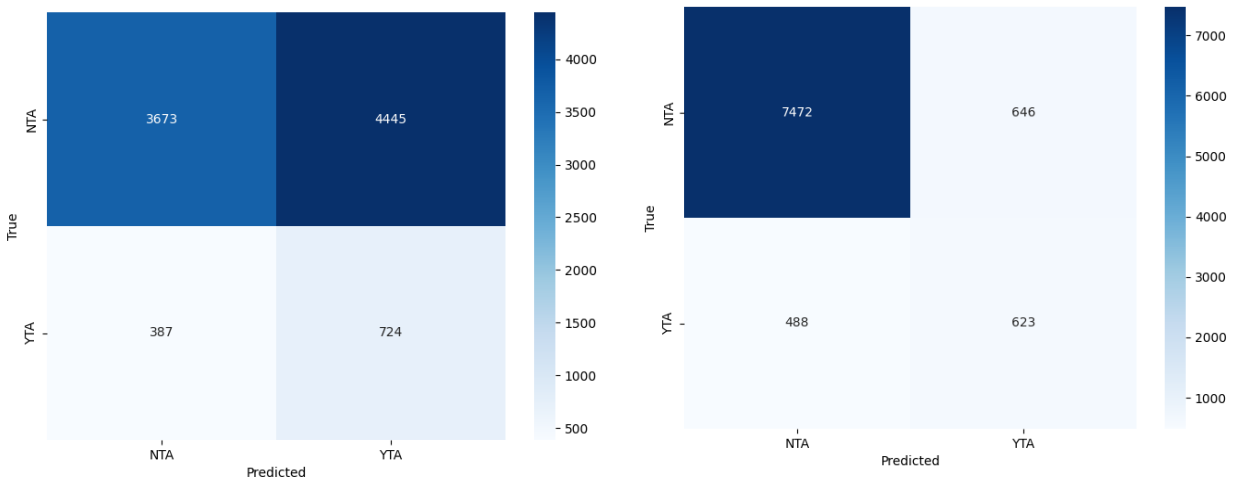
4.2.1 Models Finetuned on Reddit AITA Binary Dataset

Flan-T5-XXL Performance

Instruction fine-tuning the Flan-T5 XXL model on the Reddit AITA Binary dataset significantly enhanced both the classification of interpersonal conflicts and the quality of generated responses. As shown in Table 4.5, The classification performance exhibited a mild increase in precision, rising from 0.81 to 0.88, with substantial improvements also seen in recall, F1-score, and Matthew's correlation coefficient (MCC). Specifically, recall and F1-score both escalated dramatically from 0.48 to 0.88, and the MCC surged from 0.068 to 0.455. The justification quality also improved, as evidenced by the ROUGE-Lsum score increasing from 0.033 to 0.162 and the average COMET score rising from 0.323 to 0.505.

Comparisons of the model's predicted classifications with those of the reference top comments, as shown in Figure 4.9, highlight that the main performance gain was due to an increased tendency to classify interpersonal conflicts as NTA rather than YTA. Notably, the MCC score of 0.455 was the highest amongst all instruction finetuned models, suggesting that Flan-T5 XXL, when finetuned on the Reddit AITA Binary dataset, is highly effective at classifying interpersonal conflicts. This is supported by the MCC score being 0.141 higher than when finetuned on the Reddit AITA Multiclass dataset, implying that focusing on binary classifications dramatically enhances its performance.

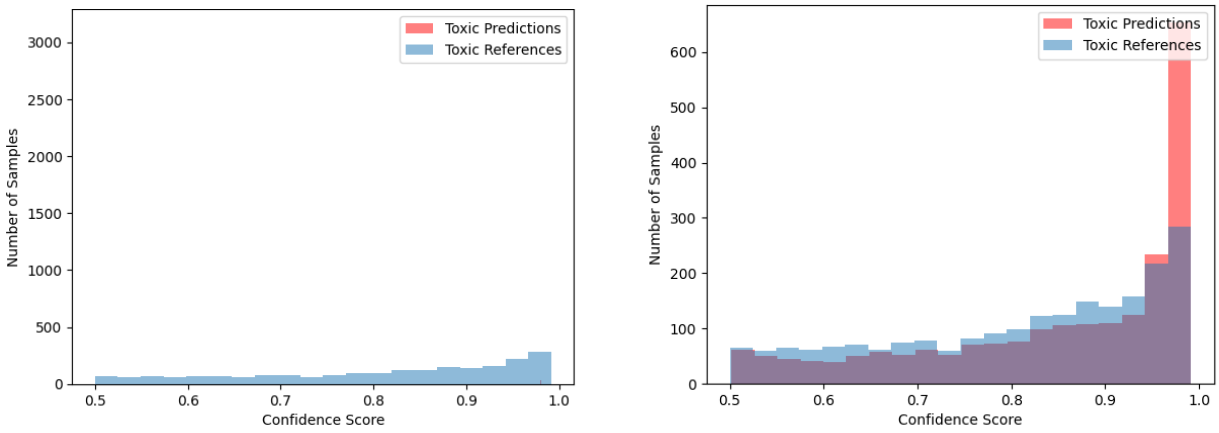
Similarly, as depicted in Figure 4.10, the model learned to generate toxic justifications, with a toxicity rate of 0.235, slightly lower than the 0.268 observed when finetuned on the same binary dataset, but still persistently higher than the 0.231 for the reference top comments. This decrease suggests that eliminating non-binary AITA classifications reduces toxicity, reflecting a trend where AITA subreddit members resort to using more toxic language for conflicts that don't have an obvious responsible party.



(a) Zero-shot Classifications

(b) Finetuned Classifications

Figure 4.9: Classifications by Flan-T5 XXL on Reddit AITA Binary Dataset



(a) Zero-shot Toxic Generations

(b) Finetuned Toxic Generations

Figure 4.10: Toxic Generations by Flan-T5 XXL on Reddit AITA Binary Dataset

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

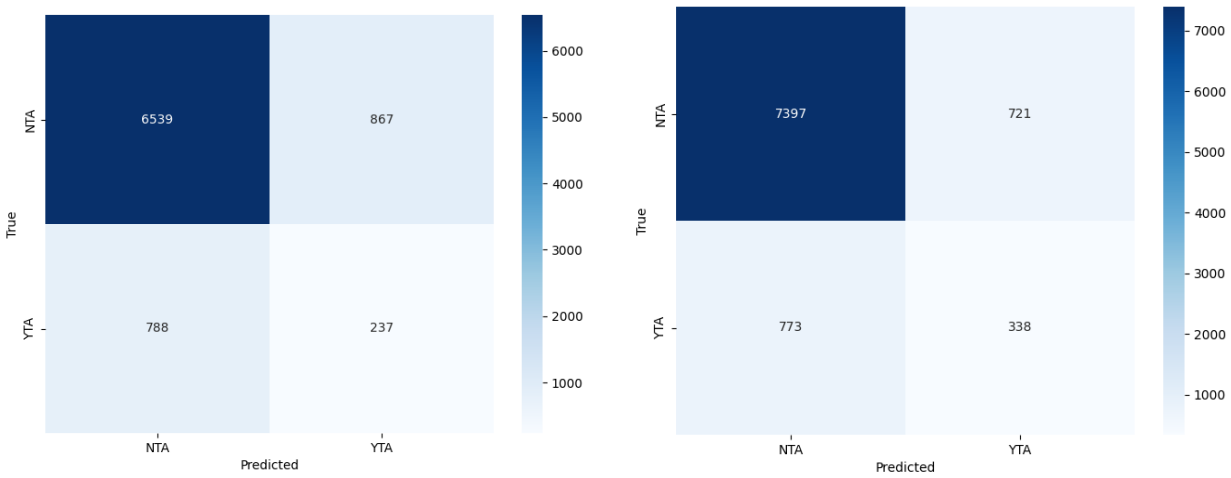
Llama-2-13B-Chat Performance

Instruction finetuning the Llama-2-Chat model on the Reddit AITA Binary dataset yielded an improvement in classification capabilities, albeit with a slight deterioration in justification abilities. As shown in Table 4.6, the model's classification performance exhibited only marginal increases in precision, recall, and F1 score; however, the Matthew's Correlation Coefficient (MCC) saw a more notable increase from 0.111 to 0.220. This significant rise suggests the model has enhanced its ability to predict both positive and negative classes effectively. The focus of precision, recall, and F1 on positive classifications means they may not fully capture the model's improved accuracy in predicting negative outcomes, as confirmed by the shifts highlighted in Figure 4.11.

Conversely, while the model's classification performance improved, its capability to provide justifications was marginally compromised. The ROUGE LSum score saw a slight decrease from 0.135 to 0.129, and the average COMET score dropped more significantly from 0.562 to 0.518. This degradation in performance was coupled with an increase in the toxicity rate from 0.010 to 0.166, as shown in Figure 4.12, indicating that the model's justifications became both less contextually relevant and more toxic.

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

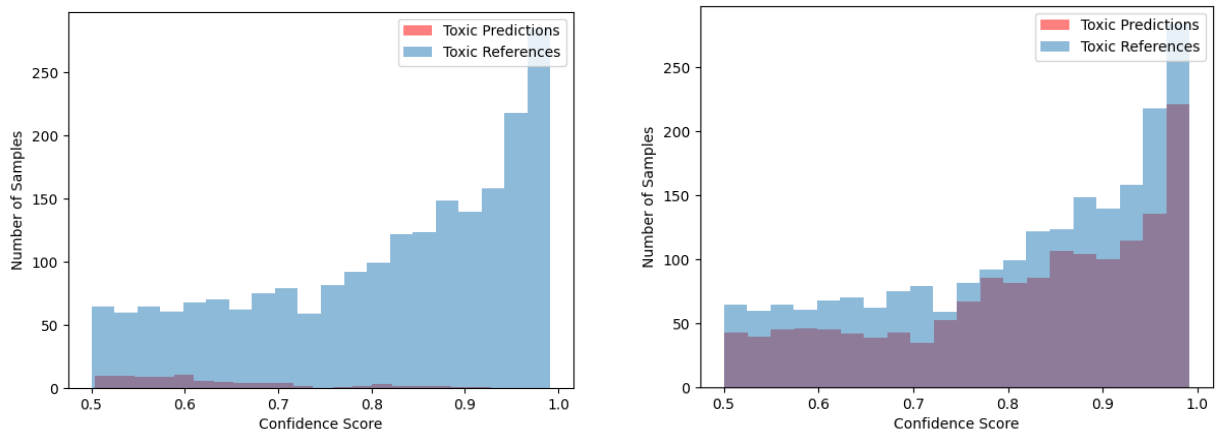
Table 4.6: Performance of Llama-2-13B-Chat on Reddit AITA Binary Dataset



(a) Zero-shot Classifications

(b) Finetuned Classifications

Figure 4.11: Classifications by Llama2-13B-Chat on Reddit AITA Binary Dataset



(a) Zero-shot Toxic Generations

(b) Finetuned Toxic Generations

Figure 4.12: Toxic Generations by Llama2-13B-Chat on Reddit AITA Binary Dataset

4.2.2 Models Finetuned on Reddit AITA Binary Top 2K Dataset

Flan-T5 XXL Performance

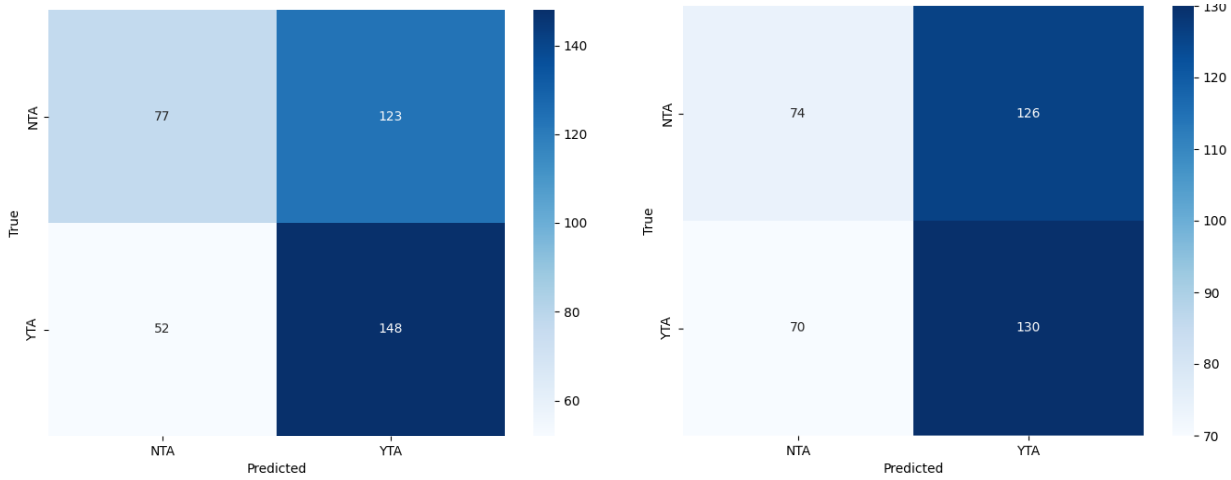
Instruction finetuning the Flan-T5 XXL model on the Reddit AITA Binary Top 2K dataset significantly enhanced its justification capabilities; however, it negatively impacted its classification abilities.

As shown in Table 4.7, the model saw decreases across all classifications metrics, with the Matthew's Correlation Coefficient (MCC) dropping notably from 0.134 to 0.021. This decline approaches a level indicative of random choice, suggesting the model has failed to effectively differentiate between interpersonal conflicts classified as either NTA or YTA. This movement toward random classification is evident from the confusion matrices shown in Figure 4.13, where the model equally split its decisions between NTA and YTA, but with nearly half resulting in incorrect classifications.

Regarding the generation of justifications, finetuning on this smaller dataset achieved nearly the same improvements in ROUGE LSum and average COMET scores when finetuning on the larger Reddit AITA Binary dataset. This suggests that while the Flan-T5 XXL model requires a substantial amount of data to learn binary classification effectively, it can make significant gains in producing contextually appropriate justifications with a relatively small dataset. Despite these improvements, the toxicity rate of the justifications increased to 0.253. Similar to the results observed when fine-tuning on the larger binary dataset, this rate exceeds the toxicity levels of the top comment reference texts from the AITA subreddit, implying a higher likelihood of generating toxic language compared to humans. This increase in toxicity is depicted in Figure 4.14.

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

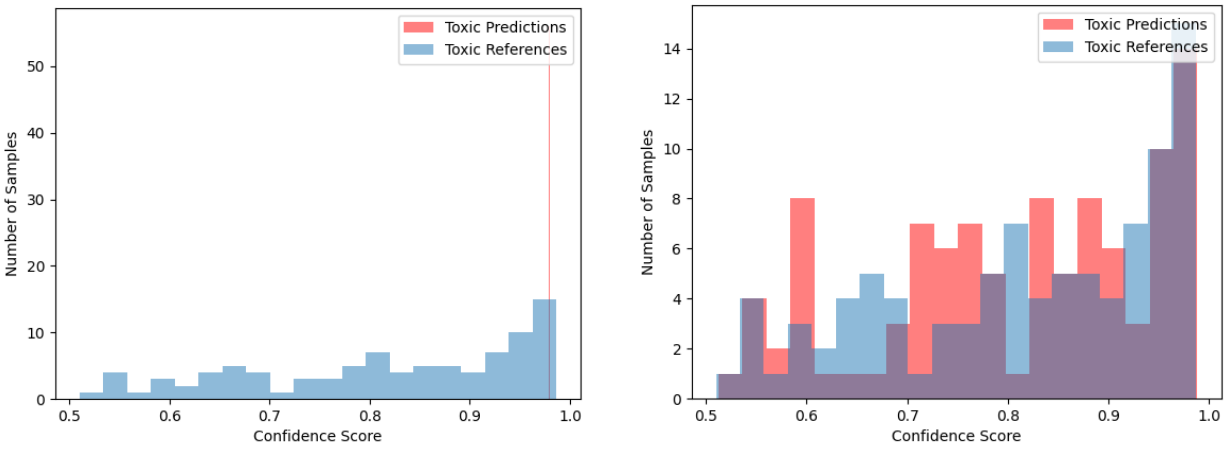
Table 4.7: Performance of Flan-T5 XXL on Reddit AITA Binary Top 2K Dataset



(a) Zero-shot Classifications

(b) Finetuned Classifications

Figure 4.13: Classifications by Flan-T5-XXL on Reddit AITA Binary Top 2K Dataset



(a) Zero-shot Toxic Generations

(b) Finetuned Toxic Generations

Figure 4.14: Toxic Generations by Flan-T5-XXL on Reddit AITA Binary Top 2K Dataset

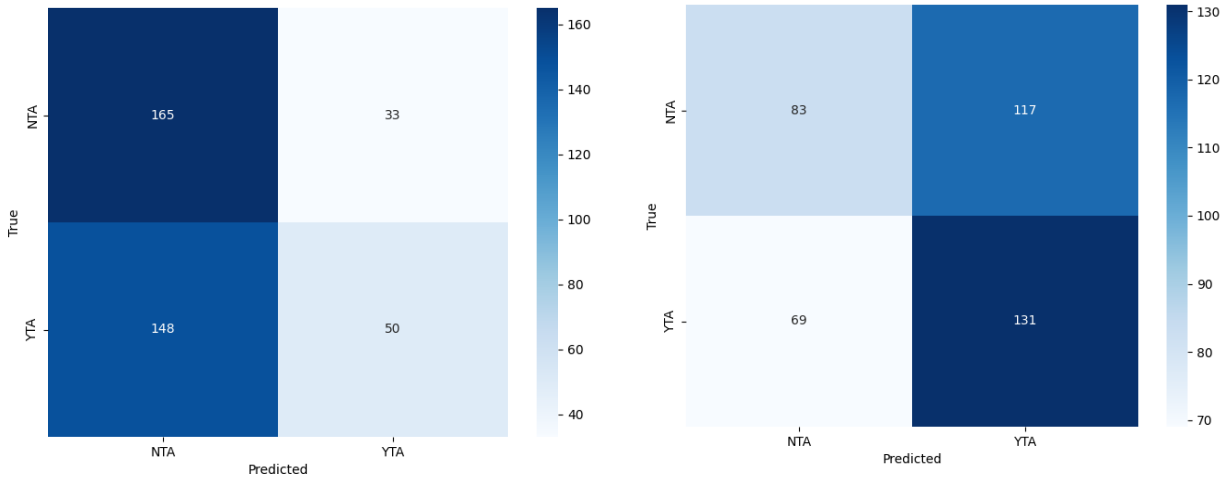
Llama-2-13B-Chat Performance

Instruction finetuning Llama-2-13B-Chat on the Reddit AITA Binary Top 2K dataset did not yield improvements in either classification or justification capabilities. Table 4.8 shows that precision, recall, F1 score, and Matthew's Correlation Coefficient (MCC) experienced only marginal changes, suggesting no significant overall shift. Similarly, the ROUGE Lsum score remained relatively unchanged, but the average COMET score experienced a more noticeable decrease from 0.589 to 0.526. Combined with the consequence of the justification toxicity rate increasing from 0.005 to 0.253 as shown in Figure 4.16, its reasonable to conclude that the overall performance of the model deteriorated.

These results suggest that the two thousand sample size was insufficient for the model to learn anything substantial. This is evidenced in two ways: First, the distribution of classifications showed very little change, as depicted by the confusion matrices in Figure 4.15. Second, when finetuned on the larger Reddit AITA Binary dataset, the model's MCC increased by 0.109, in contrast to the 0.113 decrease observed with this smaller version of that dataset. This implies that the model has a greater capacity to learn that was restricted by the limited data available for fine-tuning.

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

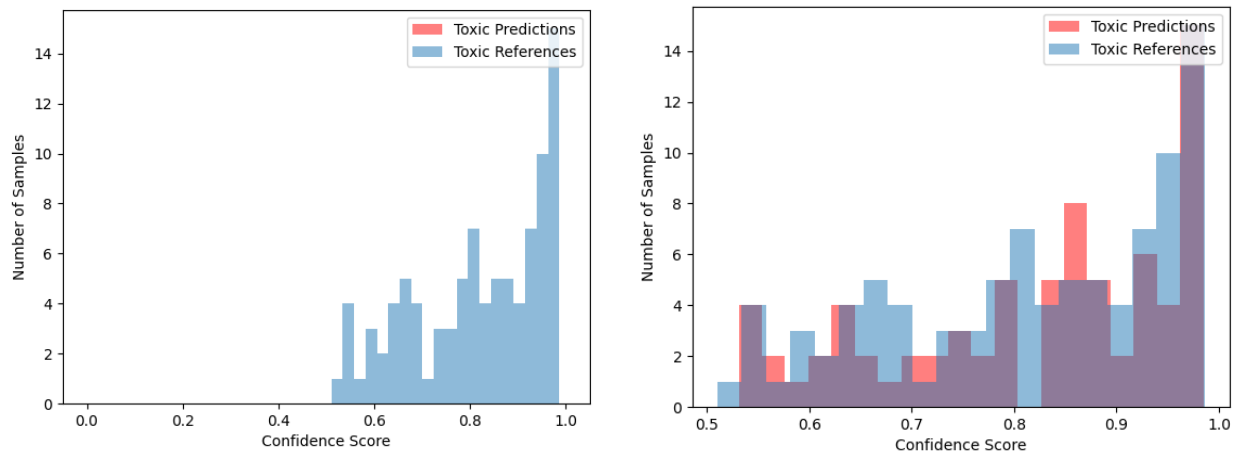
Table 4.8: Performance of Llama-2-13B-Chat on Reddit AITA Binary Top 2K Dataset



(a) Zero-shot Classifications

(b) Finetuned Classifications

Figure 4.15: Classifications by Llama-2-13B-Chat on Reddit AITA Binary Top 2K Dataset



(a) Zero-shot Toxic Generations

(b) Finetuned Toxic Generations

Figure 4.16: Toxic Generations by Llama-2-13B-Chat on Reddit AITA Binary Top 2K Dataset

Chapter 5
Discussion

There were notable differences in the performance of the Flan-T5 XXL and Llama-2-13B-Chat models when instruction finetuned on the four Reddit AITA datasets. We delved deeper into these variations by comparing their performance on the larger Reddit AITA Multiclass and Binary datasets in Sections 5.1 and 5.2. Additionally, we assessed their efficacy on the smaller, higher-quality Reddit AITA Multiclass Top 2K and Reddit AITA Binary Top 2K datasets in Section 5.3 to evaluate the extent of improvement each model could achieve when finetuned on limited data.

Beyond comparing the models' abilities to classify interpersonal conflicts and generate contextually relevant justifications, we also examined how each model's architecture and initial training influenced its ability to resist adopting toxic language. Considering the substantial presence of toxicity in the top comments of all Reddit AITA datasets, with rates ranging from 0.182 to 0.253 as shown in Table 3.12, it was anticipated that both models might learn this toxic language. However, the design decisions in the initial training of Llama-2-Chat made it more resistant to such influences compared to Flan-T5. We propose a strategy to mitigate these undesirable increases in toxicity in Section 5.4.

In Section 5.5, we recommend an optimal architecture and finetuning methodology for developing language models as AI agents for interpersonal conflict resolution. The proposal, visualized by the diagram in Figure 5.4 underscores the use of an encoder-decoder architecture, initially doing supervised finetuning on non-toxic text samples, then apply iterative rounds of finetuning using Reinforcement Learning with Human Feedback (RLHF) to ensure the generation of both contextually relevant and ethically sound responses. This included the creation of an ensemble of three individ-

ual reward models for justification safety, quality, and classification accuracy via human evaluations of the model and usage of the advantage actor-critic (A2C) reinforcement learning algorithm.

Our discussion concludes with directions for future research on finetuning LLMs for interpersonal conflict resolution, particularly focusing on how to extract useful knowledge from toxic texts without compromising safety.

5.1 Comparison Between Finetuning Flan-T5 and Llama-2-13B-Chat on the Reddit AITA Multiclass Dataset

Instruction finetuning Flan-T5 XXL and Llama-2-13B-Chat on the Reddit AITA Multiclass dataset significantly improved both models' abilities to resolve interpersonal conflicts. However, differences in classification accuracy, justification quality, and levels of toxic language in the outputs highlight the importance of selecting optimal training strategies for the safe and effective resolution of conflicts.

Flan-T5 XXL demonstrated a slight edge over Llama-2-13B-Chat in precision, recall, and toxicity scores. A more significant difference was observed in their Matthew's Correlation Coefficient (MCC), a critical metric for datasets with unbalanced classes such as the Reddit AITA dataset. The MCC, which is calculated using Equation 5.1, accounts for both Type I (false positives) and Type II (false negatives) errors to offer a balanced evaluation of model performance [26].

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5.1}$$

This analysis shows that Flan-T5 XXL more effectively reduces false negatives, thereby enhancing its identification accuracy for specific classes. This leads to three critical implications:

- **Flan-T5 is Superior at Decision Threshold Optimization:** Flan-T5's notable reduction in false negatives compared to false positives indicates that it has effectively adjusted its decision thresholds to inclusively identify potential positive cases, balancing sensitivity and specificity.

- **Precision-Recall Trade-offs for Intepersonal Conflict Resolution:** Reducing false

negatives is crucial for early conflict identification and mitigation, preventing escalation by accurately advising on proactive conflict resolution.

- **Enhanced User Trust through Model Reliability:** Decreased false negatives improve user confidence in the model, particularly in sensitive applications like mental health therapy, thereby fostering broader acceptance and utility in high-stakes environments.

In terms of justification quality, Flan-T5 XXL, when fine-tuned on the Reddit AITA Multiclass dataset, achieved the highest ROUGE LSum score, suggesting superior lexical quality in its responses compared to those of Llama-2-13B-Chat. However, Llama-2-Chat initially achieved the highest average COMET score in a zero-shot context, reflecting higher correlation with human judgment. Surprisingly, further fine-tuning led to a significant decline in its COMET score by 10.30%, with the toxicity rate increasing from 1.2% to 19.0%.

The COMET model, WMT22-Comet-DA, built on the XLM-R architecture, uses a reference-based regression approach trained on direct assessments from WMT17 to WMT20. These assessments, which discourage unsafe and aggressive language, help explain the lower COMET score in the presence of increased toxic language [29].

Flan-T5 XXL's higher toxicity rate compared to the reference texts (0.268 versus 0.224) is concerning, indicating a propensity for generating and even overlearning toxic language to enhance output similarity. This is consistent with its higher ROUGE LSum, which measures lexical similarities through n-grams and might inadvertently reward the use of toxic language due to its perceived lexical similarity. Furthermore, this issue correlates with the decline in the COMET score observed when finetuning Llama-2-13B-Chat. While training on toxic language may superficially enhance text similarity scores, it ultimately detracts from the quality of the model as judged by human evaluators, due to increased textual toxicity.

5.2 Comparison Between Finetuning Flan-T5 and Llama-2-Chat on AITA Binary Dataset

Finetuning Flan-T5 XXL and Llama-2-13B-Chat on the Reddit AITA Binary dataset yielded results similar to those observed with the Reddit AITA Multiclass dataset. Both models demonstrated

enhancements across classification accuracy, toxicity reduction, and justification metrics. However, it is worth noting that Llama-2-13B-Chat experienced a 7.82% decrease in its average COMET score, suggesting an area for further investigation.

In terms of precision, recall, and F1-scores, both models showed considerable improvements. Specifically, Flan-T5 XXL saw a 12.92% increase in these metrics, while Llama-2-13B-Chat improved by 11.65%. Moreover, the Matthews Correlation Coefficient (MCC) for Flan-T5 XXL rose to 0.455, a 44.90% increase, and Llama-2-13B-Chat achieved an MCC of 0.220, up by 33.33%. These figures represent significant advancements from their performances on the Multiclass dataset, with Flan-T5 XXL's MCC being notably higher, indicating a moderate and substantially better than chance correlation in its predicted versus actual classifications.

The enhancements in binary classification for the AITA interpersonal conflict resolution task suggest that both models are more effective in this simpler format. By focusing on binary classification, the potential for misclassification is reduced due to fewer class options. Although this simplification might overlook the nuanced classifications of ESH, NAH, or INFO, it still captures the essence of many conflicts, as evidenced by the skewed nature of the multiclass dataset where only 7.75% of entries fall into non-binary categories (see Table 3.2).

Furthermore, toxicity rates have decreased notably. In the Binary dataset, the models generated 15.13% and 16.47% less toxic language for Flan-T5 XXL and Llama-2-13B-Chat respectively, compared to the top comments in the Multiclass dataset (see Table 5.1). This reduction suggests that non-binary classifications in the Multiclass dataset may have previously led the models to learn and generate more toxic language.

In conclusion, the outcomes from finetuning Flan-T5 XXL and Llama-2-13B-Chat on the Reddit AITA Binary dataset affirm that binary classification, coupled with justification generation, effectively addresses interpersonal conflict resolution. This approach not only identifies the primary responsible party more accurately but also fosters safer dialogue by reducing the propensity for toxic language.

-	Reddit AITA Multiclass	Reddit AITA Binary
Flan-T5 XXL	1.19	1.01
Llama-2-13B-Chat	0.85	0.71

Table 5.1: Toxicity Rate of Finetuned Models in Proportion to Dataset Reference Texts

5.3 Comparisons Between Finetuning Flan-T5 and Llama-2-Chat on AITA Top 2K Datasets

Finetuning Flan-T5 XXL and Llama-2-13B-Chat on the Reddit AITA Top 2K Multiclass and Binary datasets resulted in significantly poorer performance compared to models trained on the full datasets, especially in classification accuracy. The Matthew's Correlation Coefficient (MCC) for each of the four models decreased, with values ranging from slightly below zero to a maximum of 0.072, indicating minimal to no true understanding of the AITA interpersonal conflict categories. This is evidenced in Tables 4.3, 4.4, 4.7, and 4.8. In contrast, the same models showed much better MCC scores when finetuned on the complete Reddit AITA datasets, demonstrating far superior classification performance.

The confusion matrices in Figures 4.6, 4.8, 4.13, and 4.15 reveal that although the models did not improve, they distributed classifications more evenly across categories. This outcome likely stems from the datasets being stratified to ensure equal representation of each class, intended to enhance dataset quality by exposing the models to all types of classifications. However, whether the poor performance resulted from the stratification or simply the small size of the datasets remains unclear. Further research into creating larger, stratified datasets is recommended to determine if this approach actually improves dataset quality.

While Dettmers et al. reported that using QLoRA to finetune models with small, high quality datasets can yield significant performance gains [24], we did not see similar results with the Reddit AITA Top 2K datasets. This discrepancy may be attributed to the lower quality of these datasets, possibly due to the inclusion of samples that, although highly rated by the community, do not accurately represent the intended conflict categories. Future efforts should focus on enhancing

dataset quality through methods such as manual human evaluation to determine if models can effectively learn the AITA interpersonal task when trained on small datasets.

5.4 Strategy for Minimizing Learned Toxicity

Flan-T5 XXL and Llama-2-13B-Chat, as highlighted in previous sections, are notably prone to incorporating toxic language from the Reddit AITA datasets when finetuned on them. This susceptibility poses significant challenges in deploying these models in environments where safety is paramount. Additionally, the decline in COMET scores observed during the fine-tuning of Llama-2-13B-Chat underscores the crucial issue of toxic language limiting the quality of model-generated justifications.

To mitigate the adverse impacts of toxic language in training datasets, we recommend developing a robust methodology to systematically filter out toxic content. Our study used the large variant of the RoBERTa model, finetuned with the Toxigen dataset, to measure toxicity rates, providing a strong baseline for such filtering efforts. However, the Toxigen dataset does not fully capture all forms of toxic language prevalent in the AITA subreddit. Reddit, where users frequently resort to casual slang and innuendos, reflects a dynamic linguistic environment that challenges consistent toxicity identification. Supporting this, research by Gevers et al. demonstrated the identification of informal slang on social media by analyzing linguistic variations in 36,000 Dutch Facebook comments, highlighting differences in lexical diversity and standard language use between toxic and non-toxic posts [30].

To enhance the Toxigen dataset's comprehensiveness, one strategy could be periodic surveys within the AITA subreddit community to detect newly emerging slang terms. Another approach could involve setting up a broad monitoring system across social media platforms to track trending slang. Sociolinguistic experts should then assess whether these terms have toxic connotations. Implementing these methodologies would significantly improve our ability to identify and eliminate toxic language from training data, thus enhancing the safety of the justifications generated by the models.

When assembling the team of sociolinguistic experts, it is crucial to include a diverse range of

beliefs and identities. Research by Sap et al. revealed a strong link between annotators' identities, their beliefs, and their toxicity ratings. Notably, more conservative annotators and those with high scores on our racism scale tended to under-rate anti-Black language as toxic while over-rating texts in African American English dialect. Furthermore, a case study by Sap et al. highlighted how PerspectiveAPI, a popular toxicity detection system, unfortunately reflected biases from annotators with specific attitudes [31]. Thus, to capture informal toxic language effectively, it is essential that the team of sociolinguistic experts be diverse and inclusive of all cultures.

5.5 Recommended LLM Architecture and Finetuning Process for Interpersonal Conflict Resolution

The research presented in this paper details the findings derived from finetuning Flan-T5 and Llama-2-Chat using datasets curated from the AITA subreddit. In light of these, we propose a model architecture and finetuning methodology that specializes in resolving interpersonal conflicts. This is visualized by the diagram in Figure 5.5 and represents a comprehensive strategy aimed at enhancing the performance and applicability of LLMs in practical scenarios. It involves the use of a pre-trained encoder-decoder LLM, which undergoes supervised fine-tuning on a curated dataset with toxic samples removed. Subsequently, a Reinforcement Learning with Human Feedback (RLHF) loop is applied which is designed to continually refine the model's alignment with human preferences through iterative adjustments. To facilitate the RLHF loop, we recommend the creation of three distinct reward models, each serving a specific purpose:

- **Justification Safety:** Ensures the model's responses maintain ethical boundaries and avoid generating harmful content.
- **Justification Quality:** Focuses on the coherence, relevance, and helpfulness of the responses.
- **Classification Accuracy:** Aims to improve the models ability to understand the nature of interpersonal conflicts and appropriately categorize them.

Each of these reward models is trained using evaluations derived from human preferences, which helps tailor the model's outputs to be more aligned with what users find useful and acceptable. To

implement the RLHF, we use the Advantage Actor Critic (A2C) reinforcement learning framework, a popular choice for tasks requiring a balance between exploration (testing new strategies) and exploitation (refining known strategies). This framework is particularly suited to applications where multiple objectives, such as safety and accuracy, need to be balanced effectively.

In the following sections, this paper will delve deeper into each component of the proposed architecture and finetuning process. We will discuss the specific methodologies employed, the rationale behind each decision, and the expected impacts of these choices on the model's performance in real-world applications.

5.5.1 Transformer Architecture for Pretrained LLMs

Our proposal recommends employing a pretrained LLM with an encoder-decoder architecture. This recommendation is based on the superior classification performance of Flan-T5, especially in terms of Matthew's correlation coefficient (MCC), which is crucial given the unbalanced nature of the datasets. Flan-T5 also demonstrated improvements in the ROUGE LSum metrics that were comparable to those achieved by Llama-2-Chat. Although Flan-T5 did not attain as high an average COMET score, we believe this discrepancy can be addressed by finetuning solely with non-toxic, benign samples. Our advocacy for using a pretrained encoder-decoder LLM is intrinsically linked to the necessity of finetuning exclusively with benign samples. This approach minimizes the risk of the encoder inadvertently learning contextual associations with toxic language, a problem less prevalent in decoder-only LLMs like Llama-2-Chat.

5.5.2 Supervised Finetuning on Non-toxic Samples

Our proposal recommends employing a pretrained LLM with an encoder-decoder architecture. This recommendation is based on the superior classification performance of Flan-T5, especially in terms of Matthew's correlation coefficient (MCC), which is crucial given the unbalanced nature of the datasets. Flan-T5 also demonstrated improvements in the ROUGE LSum metrics that were comparable to those achieved by Llama-2-Chat. Although Flan-T5 did not attain as high an average COMET score, we believe this discrepancy can be addressed by finetuning solely with non-toxic, benign samples. Our advocacy for using a pretrained encoder-decoder LLM is intrinsically linked

to the necessity of finetuning exclusively with benign samples. This approach minimizes the risk of the encoder inadvertently learning contextual associations with toxic language, a problem less prevalent in decoder-only LLMs like Llama-2-Chat.

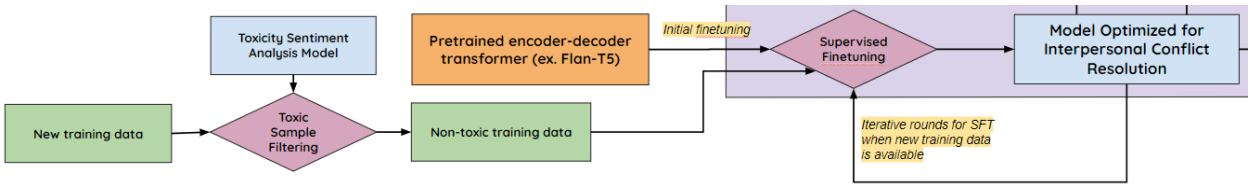


Figure 5.1: SFT Part of Architecture

5.5.3 Implementation of Reinforcement Learning with Human Feedback

To enhance alignment between the model and human expectations, we recommend implementing a Reinforcement Learning with Human Feedback (RLHF) loop. This approach, pioneered by Ziegler et al., involves training a reward model through the evaluation of model outputs based on human preferences. The outcomes of this reward model are then employed as the objective function for optimization via reinforcement learning. Ziegler et al. highlight the effectiveness of this method in refining model performance through iterative adjustments based on feedback (Ziegler et al., 2019).

There are two primary considerations in implementing RLHF: the configuration of the reward models and the selection of a suitable reinforcement learning framework. Regarding the reward models' structure, we detail in Section "Reward Model Design" our proposed configuration. We suggest developing separate reward models for each of the three evaluation criteria: justification safety, output quality, and classification accuracy. These models' outputs should then be integrated to train a composite reward function capable of issuing a unified scalar reward that reflects all three objectives. Concerning the choice of reinforcement learning framework, we advocate for the Advantage Actor Critic (A2C) method. Our choice is informed by A2C's ability to maintain a beneficial balance between exploration and exploitation. This balance is crucial in scenarios such as interpersonal conflict resolution, where multiple objectives must be concurrently optimized.

Reward Model Design

The design of an effective reward model for interpersonal conflict resolution is predicated on its ability to address three key objectives: justification safety, quality, and classification accuracy. To optimally manage these objectives, it is recommended to develop separate reward models tailored specifically to each one. During the human preference evaluation phase, individual raters assess each model based on these objectives independently. The feedback from these evaluations helps each model refine its approach and improve its performance with respect to its designated objective.

For the synthesis of these individual outputs into a unified reward system, various integration methods can be employed. Among these, hierarchical learning is particularly promising for the context of interpersonal conflict resolution. Hierarchical learning involves structuring the decision-making process in layers, where higher-level policies can override lower-level ones under specific conditions. This structure allows for a dynamic blending of policies from different models, ensuring that the most crucial policies can dominate when necessary [32].

This approach is vital in scenarios such as interpersonal conflict resolution where safety is a critical concern. By prioritizing safety through hierarchical learning, the composite model ensures that rewarding safe outcomes takes precedence, thereby enhancing the overall effectiveness and reliability of the conflict resolution process.

Figure 5.5.3 provides a detailed view of the portion view of our recommended architecture and finetuning process that pertains to the creation of reward models via human preference evaluation. It illustrates how each of the three reward models is initially trained using human preference data to hone in on their specific objectives. Subsequently, these models are seamlessly integrated using hierarchical blending techniques. This not only bolsters the quality of the model's rewards but also ensures that it remains aligned with the essential goals of safe and effective interpersonal conflict resolution.

Usage of Advantage Actor Critic Framework

In the realm of reinforcement learning, the actor-critic method is recognized as a sophisticated hybrid approach that combines policy-based and value-based strategies to efficiently learn optimal

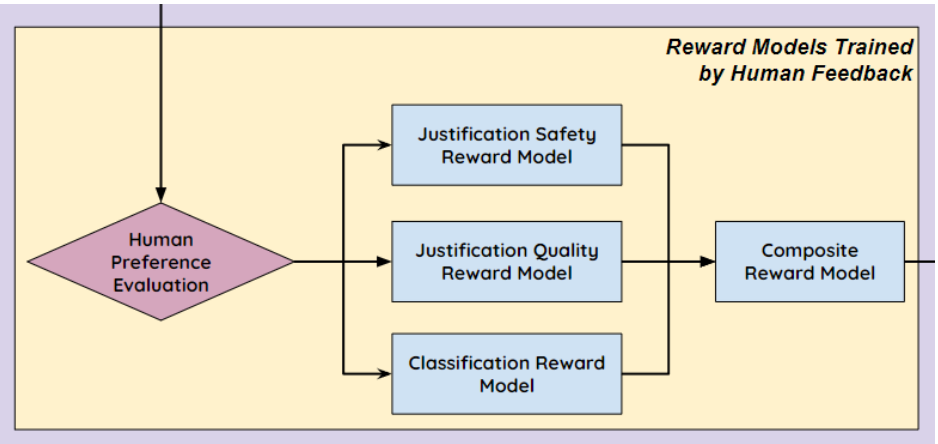


Figure 5.2: Reward Models

policies in complex environments. This method comprises two primary components: the actor and the critic. The actor is tasked with selecting actions according to a policy function, denoted as $\pi(a|s;\theta)$, which maps states to actions and is parameterized by θ. In reinforcement learning scenarios involving language models (such as RLHF), the actor represents the language model being finetuned, with the policy equating to its parameters. On the other hand, the critic evaluates the actions chosen by the actor by calculating a value function, which can either be the state-value function $V(s;\omega)$ or the action-value function $Q(s,a;\omega)$, with ω indicating the parameters of the critic. The primary role of the critic is to assess the quality of the actions from a particular state, guiding the actor toward more beneficial decisions [33].

The learning process of actor-critic methods involves several crucial steps. Initially, the actor selects actions based on the current policy $\pi(a|s;\theta)$. Concurrently, the critic determines the value of being in the current state or of taking a specific action. A vital component of this process is the calculation of the Temporal Difference (TD) error, used to estimate the advantage of the taken action. The TD error is formulated as $\delta_t = r_t + \gamma V(s_{t+1};\omega) - V(s_t;\omega)$, where r_t is the reward received after taking action a_t at state s_t, γ is the discount factor, and $V(s_{t+1})$ is the value estimate for the subsequent state [33].

Following the TD error computation, both the actor and the critic update their respective function parameters. The critic adjusts its value function parameters ω to minimize the TD error,

typically using gradient descent methods: $\omega \leftarrow \omega + \alpha\delta_t \nabla_\omega V(s_t;\omega)$, with α representing the learning rate. Simultaneously, the actor updates its policy parameters θ to maximize the expected return, guided by the critic's TD error: $\theta \leftarrow \theta + \beta\delta_t \nabla_\theta \log \pi(a_t|s_t;\theta)$, where β is a step size parameter [33].

To ensure a balance between exploration and exploitation, the actor frequently incorporates mechanisms such as adding noise to the policy or introducing an entropy bonus to the objective function. These measures are crucial to prevent premature policy convergence and to promote exploration across a diverse range of state spaces. Actor-critic methods adeptly manage the bias-variance trade-off, outperforming purely value-based methods, which can suffer from high variance, and purely policy-based methods, which may exhibit high bias. Their versatility makes them suitable for both continuous and discrete action spaces and applicable to both online learning (interacting with the environment) and offline learning (using a fixed dataset of experiences) [33].

The integration of the actor and critic within this framework lays the foundation for many advanced reinforcement learning algorithms, including A2C, a synchronous, deterministic version of the Asynchronous Advantage Actor Critic (A3C) method. Unlike A3C, which updates global parameters asynchronously using multiple workers each with its own environment, A2C synchronizes the gradient collection from multiple workers before updating the global parameters. This synchronization ensures more consistent updates and helps avoid the stale gradient problem inherent in asynchronous updates. A2C is notably effective in scenarios that require balancing multiple objectives, such as safety and accuracy. In environments where agents must balance exploration with exploitation, A2C's stable updates and the inclusion of entropy for exploration make it well-suited for such challenges. This equilibrium is essential to ensure that the policy does not converge prematurely or adhere to sub-optimal actions, which is crucial for maintaining safety while achieving high performance [34].

A2C is particularly effective in scenarios that require balancing multiple objectives, such as safety and accuracy. In environments where agents must learn to balance exploration with exploitation, A2C's stable updates and incorporation of exploration through entropy make it well-suited for such tasks. This balance ensures that the policy does not converge prematurely or stick to sub-optimal actions, which is crucial in maintaining safety while achieving high performance [34].

Figure 5.3 shows the portion of our recommended architecture and training procedure that in-

cludes the use of A2C. Ultimately, the advanced capabilities of A2C to balance multiple objectives is what makes it a promising choice for improving the interpersonal conflict resolution performance of LLMs via RLHF. This is in contrast to alternative reinforcement learning algorithms like Q-learning and Deep Q-Networks (DQN) that are likely to be sub-optimal due to them being value-based methods that focus on learning the optimal action-value function. While effective in discrete action spaces and simpler scenarios, they can struggle in environments with multiple conflicting objectives due to their tendency to overestimate action values and a lack of explicit policy exploration mechanisms.

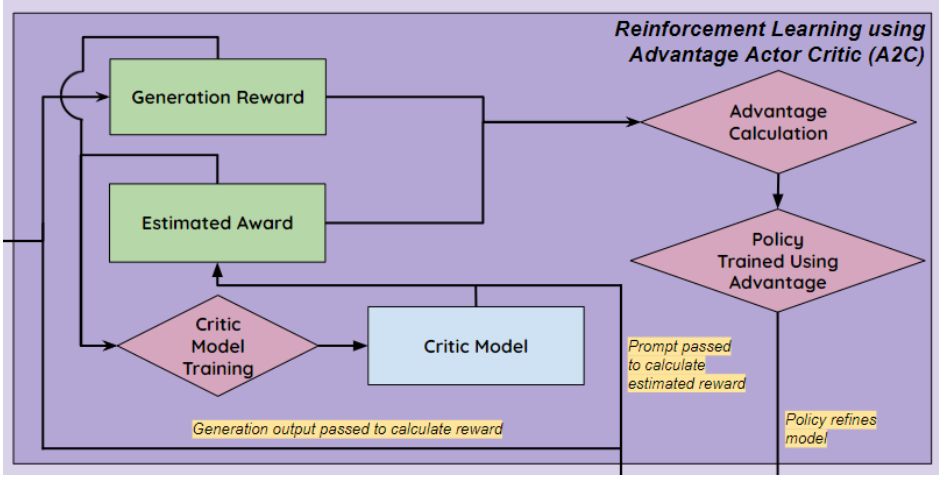


Figure 5.3: Implementation of Advantage Actor Critic (A2C) for RLHF

5.6 Future Work

This research establishes a foundational approach for employing LLMs in resolving interpersonal conflicts. While numerous areas merit further investigation, the potential impact of the following three suggests they should be prioritized:

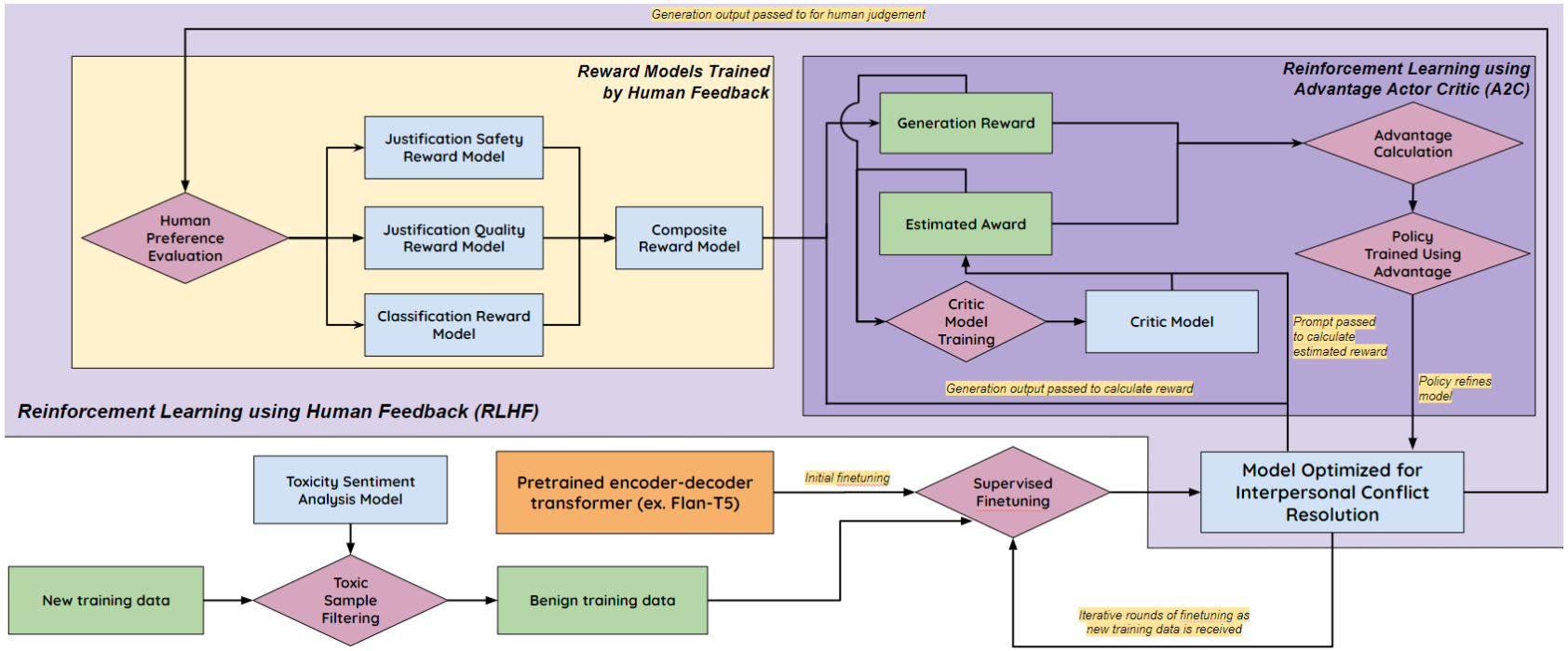
1. **Implementation of the Proposed LLM Architecture and Finetuning Process for Interpersonal Conflict Resolution:** As outlined in Section 5.5, our proposal is built upon the experimental finetunings of the Flan-T5 and Llama-2-13B-Chat models on the Reddit

AITA datasets. In order to utilize this methodology in real-world scenarios, implementing it experimentally beforehand is mandatory. This would confirm its effectiveness and provide insights into design considerations and capabilities that only emerge through practical implementation.

2. **Experimentation on Other Encoder-Decoder LLMs**: The focus of this study was narrowed to the Flan-T5 model as the primary example of LLMs using an encoder-decoder transformer architecture. Our findings suggest that the encoder-decoder transformer is more effective than decoder-only models for resolving interpersonal conflicts. However, the ideal configuration for the encoder-decoder architecture has yet to be determined. Further research is necessary to explore this, which would involve fine-tuning various LLMs that implement different versions of the encoder-decoder architecture. This research could potentially lead to the development of a specialized encoder-decoder architecture that is specifically optimized for interpersonal conflict resolution.

3. **Improvements to the Quality of the Reddit AITA Datasets:** Our research demonstrates that the classification aspect of the interpersonal conflict task can effectively be simplified to assessing whether the subject is responsible for the conflict. Given this, there is significant value tied to enhancing the quality of the Reddit AITA Binary dataset. This research statistically showed that is disagreement between AITA classifications by the top ten commenters on holistic, pairwise, and per-sample levels. Filtering the dataset so that it contains only conflicts where the AITA classification was unanimous is a promising approach for improving dataset quality. Another way to increase quality could involve stratifying the dataset to ensure equal representation of both the (NTA) and (YTA) classes, although as evidenced by the Reddit AITA Top 2K datasets, it remains to be seen if this is an useful strategy. Additionally, a thorough manual inspection of each sample could be conducted to verify that the conflict discussions on the AITA subreddit were appropriate and accurately classified.

Figure 5.4: Proposed LLM Architecture and Finetuning Procedure for Interpersonal Conflict Resolution



Chapter 6
Conclusion

Finetuning large language models (LLMs) for interpersonal conflict resolution is a challenging task due to its nondeterministic nature, where answers are frequently ambiguous and subjective. This research explored the efficacy of using LLMs for this task by finetuning Flan-T5 and Llama-2-Chat, which respectively are state-of-the-art encoder-decoder and decoder-only models, on the Reddit AITA datasets. These curated datasets consist of forum submissions that represent real-world interpersonal conflicts where there is statistically significant disagreement between classifications by the top ten commenters on holistic, pairwise, and intra-sample levels.

For effective learning, the interpersonal conflict resolution task was structured as a classification of the subject's behavior, accompanied by a justification. The experimental finetunings revealed that Flan-T5's encoder enhanced its ability to comprehend the context of conflicts, resulting in better performance in classification compared to Llama-2-Chat. While Flan-T5 generally scored slightly higher on ROGUE LSum, a measure of quality on an n-gram lexical level, Llama-2-Chat achieved higher COMET scores, particularly before being finetuned on the Reddit AITA datasets. The COMET model we employed was trained to assess agreement via direct assessments by human judges. Therefore, these results suggest that finetuning Llama-2-Chat on the Reddit AITA dataset diminished the quality of its justifications, as the process increased the generation of toxic language, which was penalized by these judges.

The leading reason for this deterioration in justification quality by Llama-2-Chat as a result of the finetunings is because they did not implement Reinforcement Learning with Human Feedback (RLHF). During its pre-training phase, Llama-2-Chat utilized RLHF to better align its outputs

with human expectations. However, the subsequent finetuning on the Reddit AITA datasets did not include RLHF, which diminished the benefits previously gained. To maintain alignment with ethical standards and preserve effectiveness in interpersonal conflict resolution, it is crucial to integrate RLHF or a comparable human-driven approach into finetuning strategy.

The experimental finetunings conducted validated that implementing RLHF or a similar method that incorporates human input enhances a model's ability to resist absorbing toxic traits from future training data, while effectively acquiring desired knowledge. This was proven by the finetuned Flan-T5 models, which didn't utilize RLHF during its pretraining, exhibiting a greater toxicity rates than that of the reference top comments. In contrast, the finetuned Llama-2-Chat models did not reach these levels. This comparison underscores the importance of embedding RLHF or an equivalent mechanism during the foundational pre-training phase, promoting safer and more reliable finetunings by the broader community.

Considering the insights discussed, this research proposed an optimal architecture and fine-tuning methodology to use LLMs for interpersonal conflict resolution. The recommendation includes selecting an LLM with an encoder-decoder architecture, conducting supervised fine-tuning exclusively on non-toxic training data, and implementing a Reinforcement Learning with Human Feedback (RLHF) loop that includes reward models for justification safety, quality, and classification accuracy. To effectively remove toxic data, it is suggested to finetune a model adept at sentiment analysis, such as RoBERTa, on the Toxigen dataset, which has been enriched with domain-specific toxic slang to comprehensively address all instances of toxicity. Additionally, the Advantage Actor Critic (A2C) framework is recommended for RLHF to optimally manage multiple objectives simultaneously.

This recommendation aims to ensure the safe and effective use of LLM in real-world scenarios that include providing advice on sensitive interpersonal conflicts, such as mental health therapy pre-screenings. It can also be extended to broader conflict resolution areas, like customer support services. Future work should focus on implementing this approach as a proof of concept, followed by thorough validation, before deploying it in any production environments.

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

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

[23] 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.

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

[25] 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.

[26] 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.

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

Appendix A
Llama2-7B-Chat and Flan-T5 XL Finetuning Results

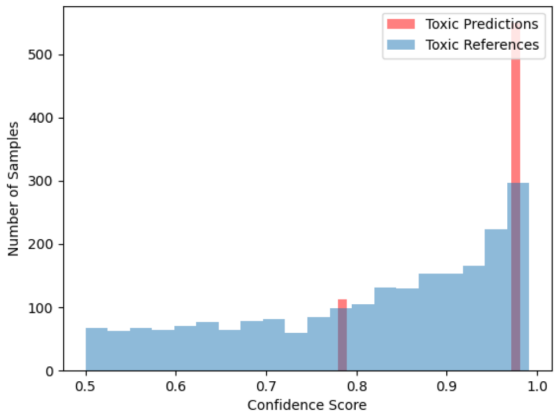
A.1 Reddit AITA Multiclass Results

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.032	0.319	0.066	0.71	0.45	0.52	0.071
Finetuned	0.159	0.508	0.323	0.74	0.81	0.76	0.222

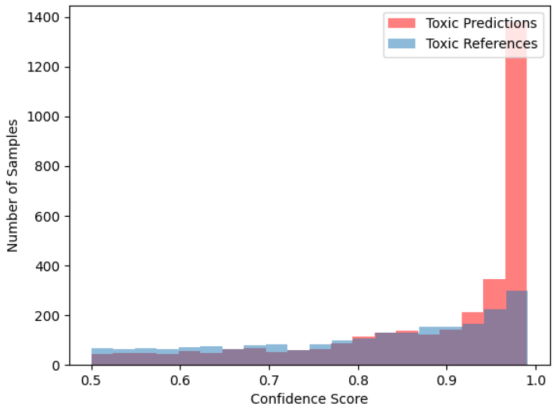
Table A.1: Performance of Flan-T5 XL on Reddit AITA Multiclass Dataset

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.096	0.584	0.003	0.67	0.07	0.05	0.020
Finetuned	0.160	0.513	0.256	0.68	0.80	73	0.148

Table A.2: Performance of Llama-2-7B-Chat on Reddit AITA Multiclass Dataset

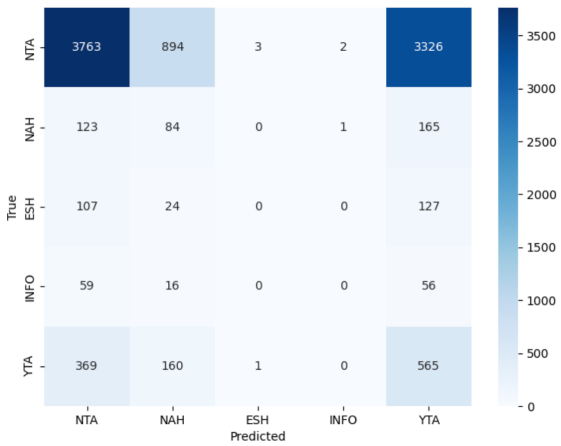


(a) Zero-shot Toxic Generations

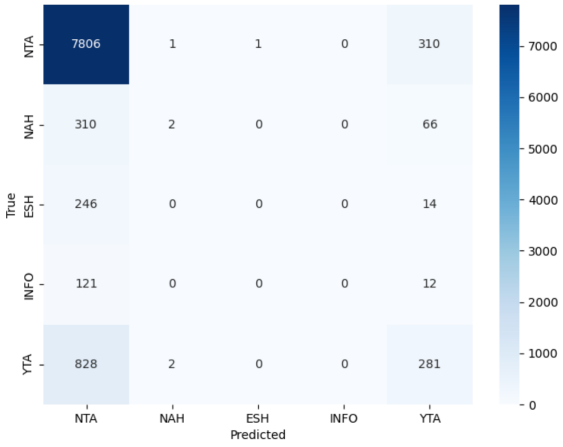


(b) Finetuned Toxic Generations

Figure A.1: Toxic Generations by Flan-T5 XL on Reddit AITA Multiclass Dataset

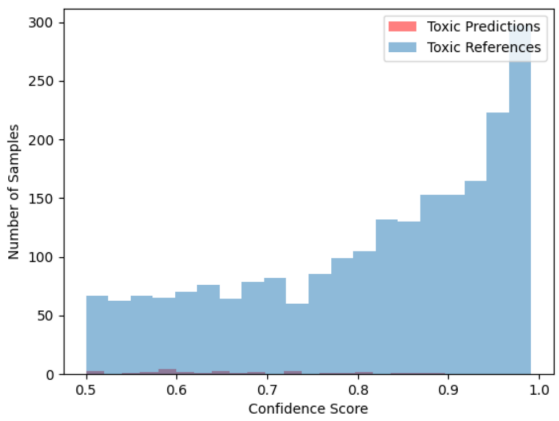


(a) Zero-shot Classifications

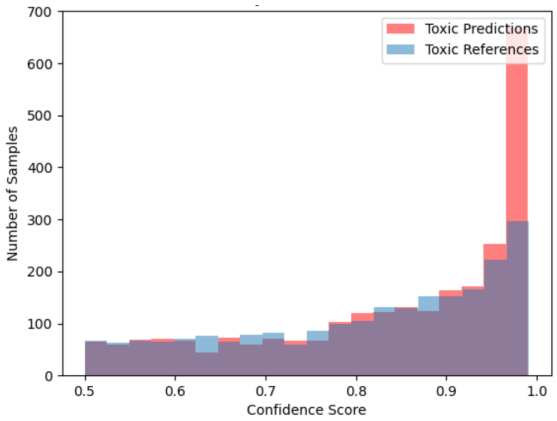


(b) Finetuned Classifications

Figure A.2: Classifications by Flan-T5 XL on Reddit AITA Multiclass Dataset

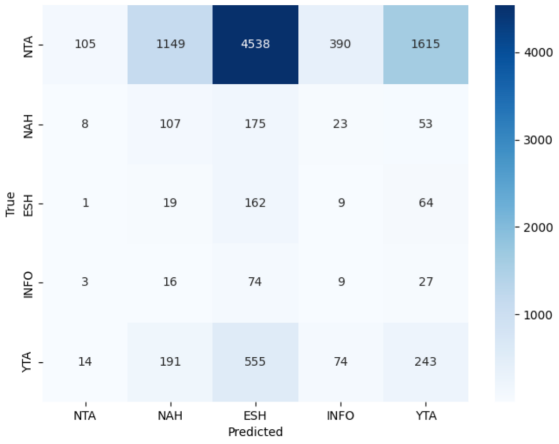


(a) Zero-shot Toxic Generations

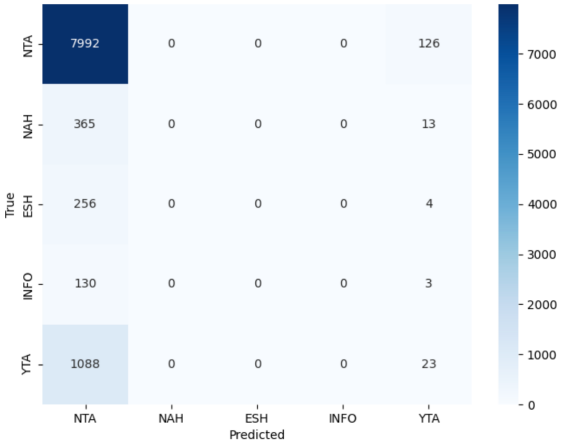


(b) Finetuned Toxic Generations

Figure A.3: Toxic Generations by Llama-2-7B-Chat on Reddit AITA Multiclass Dataset



(a) Zero-shot Classifications



(b) Finetuned Classifications

Figure A.4: Classifications by Llama-2-7B-Chat on Reddit AITA Multiclass Dataset

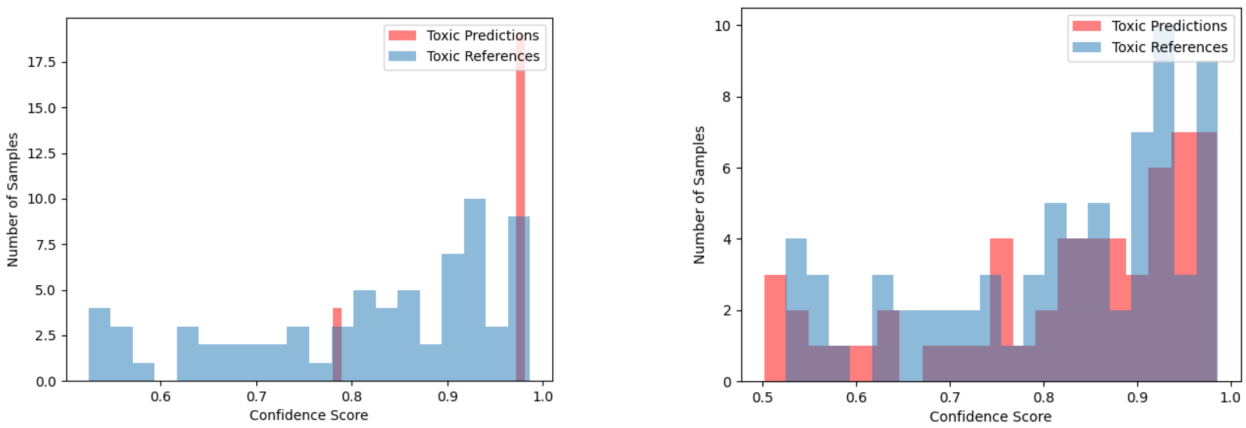
A.2 Reddit AITA Multiclass Top 2K Results

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.021	0.314	0.058	0.17	0.26	0.19	0.085
Finetuned	0.137	0.509	0.138	0.26	0.27	0.26	0.083

Table A.3: Performance of Flan-T5 XL on Reddit AITA Multiclass Top 2K Dataset

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.099	0.583	0.000	0.18	0.22	0.17	0.037
Finetuned	0.104	0.535	0.085	0.18	0.20	0.14	0.004

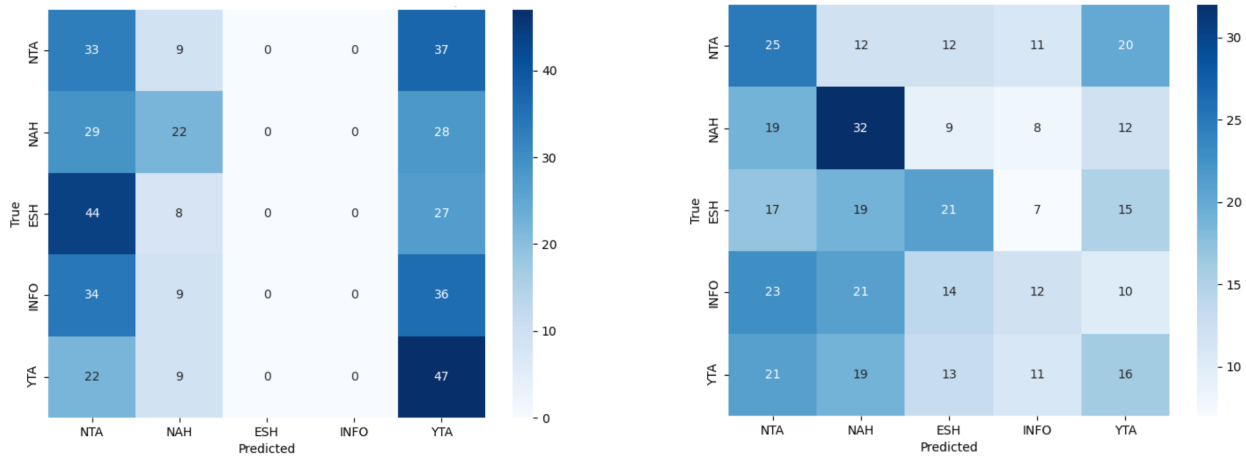
Table A.4: Performance of Llama-2-7B-Chat on Reddit AITA Multiclass Top 2K Dataset



(a) Zero-shot Toxic Generations

(b) Finetuned Toxic Generations

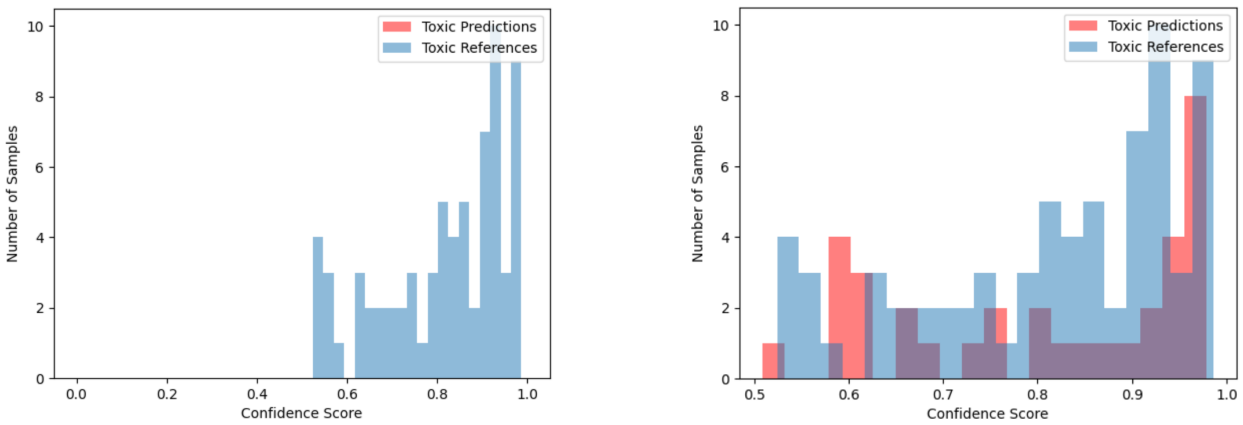
Figure A.5: Toxic Generations by Flan-T5 XL on Reddit AITA Multiclass Top 2K Dataset



(a) Zero-shot Classifications

(b) Finetuned Classifications

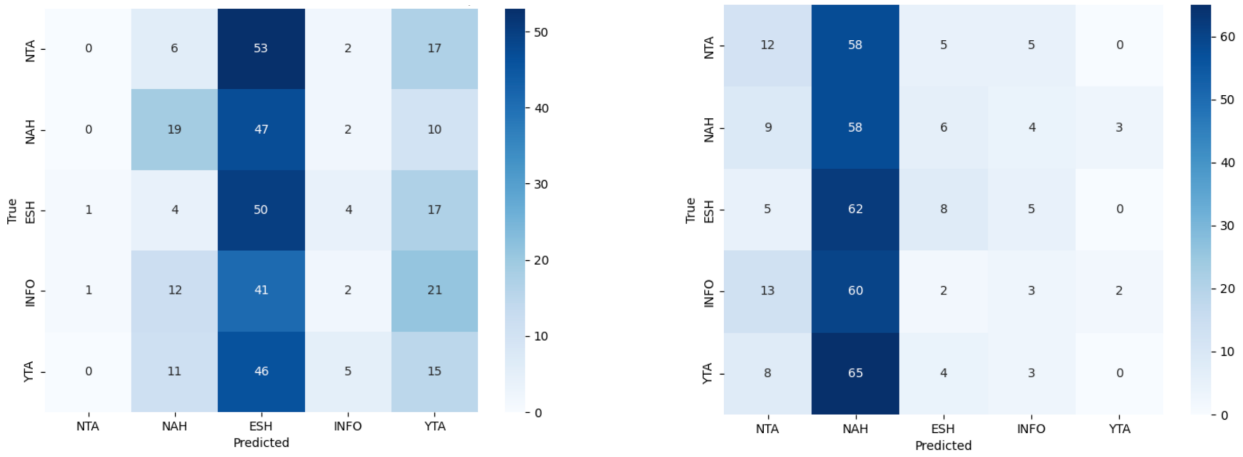
Figure A.6: Classifications by Flan-T5 XL on Reddit AITA Multiclass Top 2K Dataset



(a) Zero-shot Toxic Generations

(b) Finetuned Toxic Generations

Figure A.7: Toxic Generations by Llama-2-7B-Chat on Reddit AITA Multiclass Top 2K Dataset



(a) Zero-shot Classifications

(b) Finetuned Classifications

Figure A.8: Classifications by Llama-2-7B-Chat on Reddit AITA Multiclass Top 2K Dataset

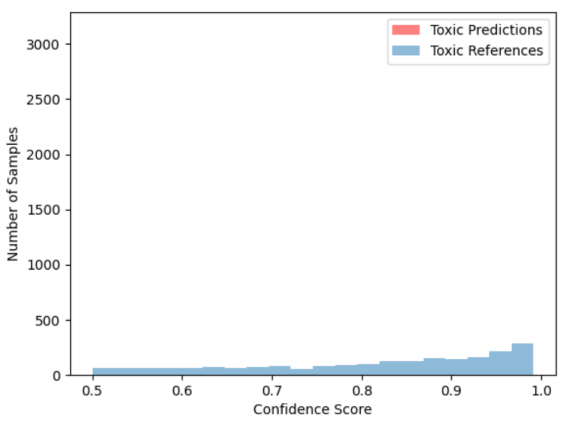
A.3 Reddit AITA Binary Results

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.028	0.326	0.000	0.82	0.44	0.52	0.080
Finetuned	0.158	0.494	0.261	0.86	0.88	0.86	0.295

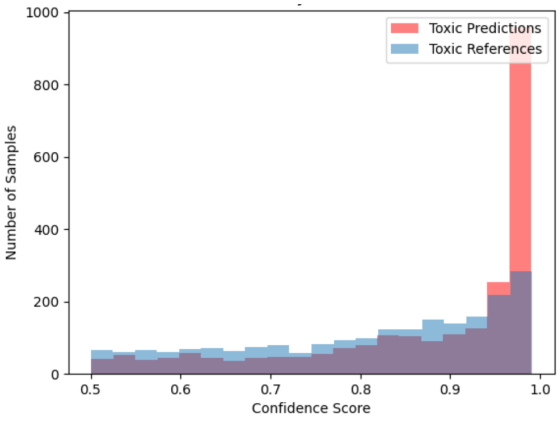
Table A.5: Performance of Flan-T5 XL on Reddit AITA Binary Dataset

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.120	0.550	0.017	0.76	0.33	0.40	-0.056
Finetuned	0.123	0.525	0.163	0.79	0.81	0.80	0.025

Table A.6: Performance of Llama-2-7B-Chat on Reddit AITA Binary Dataset

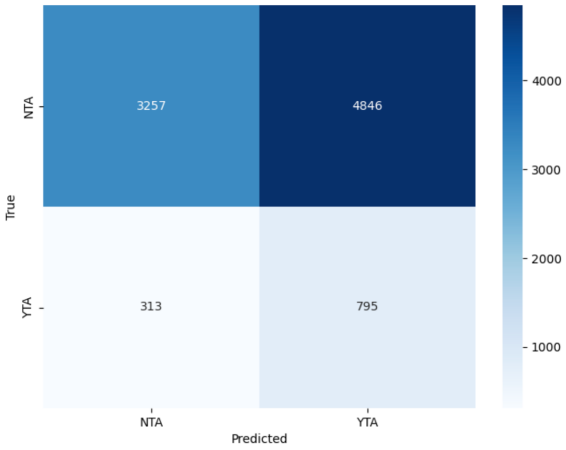


(a) Zero-shot Toxic Generations

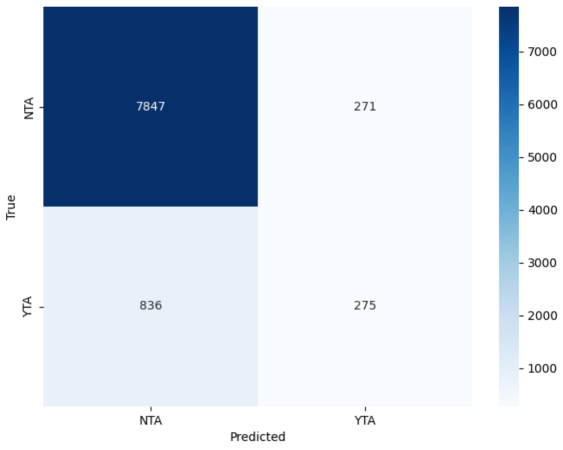


(b) Finetuned Toxic Generations

Figure A.9: Toxic Generations by Flan-T5 XL on Reddit AITA Binary Dataset

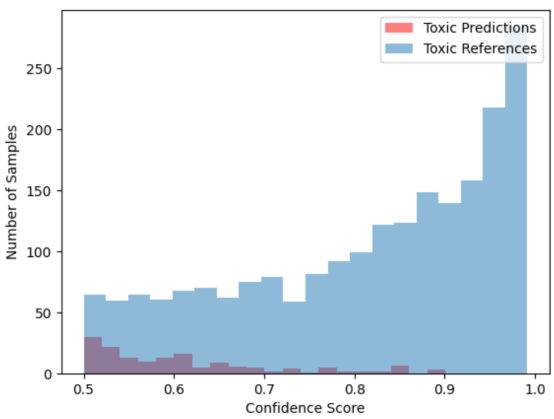


(a) Zero-shot Classifications

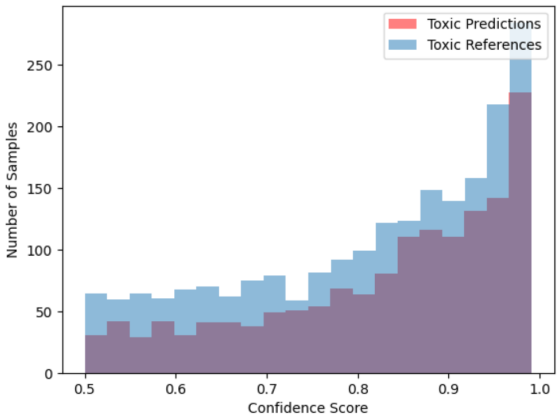


(b) Finetuned Classifications

Figure A.10: Classifications by Flan-T5 XL on Reddit AITA Binary Dataset

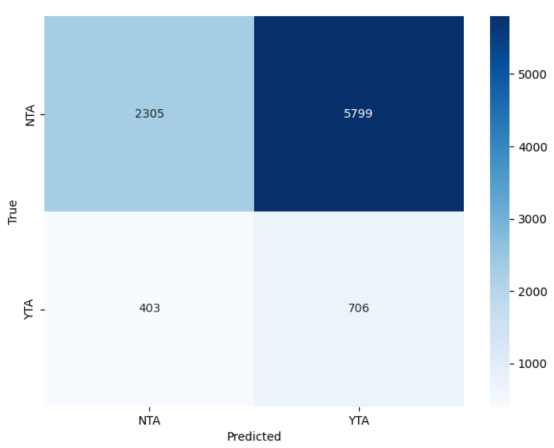


(a) Zero-shot Toxic Generations

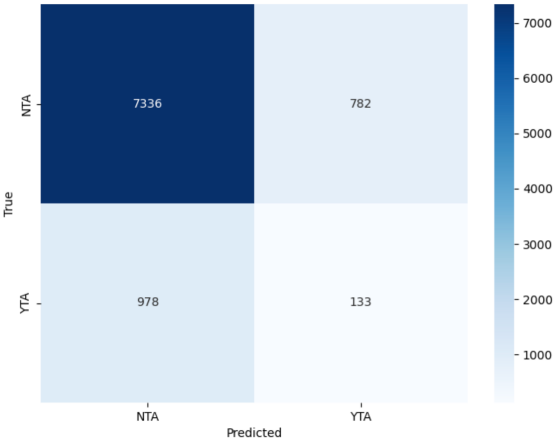


(b) Finetuned Toxic Generations

Figure A.11: Toxic Generations by Llama-2-7B-Chat on Reddit AITA Binary Dataset



(a) Zero-shot Classifications



(b) Finetuned Classifications

Figure A.12: Classifications by Llama-2-7B-Chat on Reddit AITA Binary Dataset

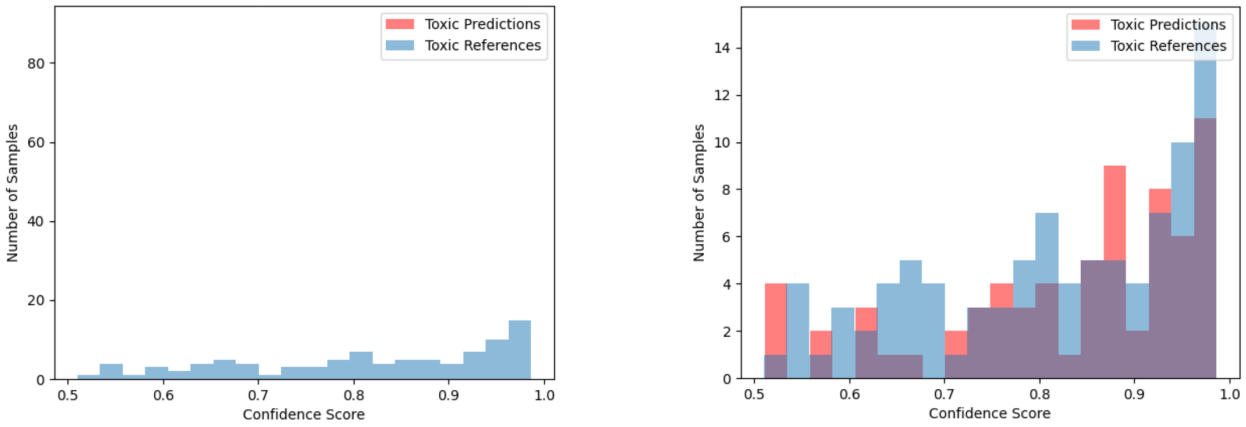
A.4 Reddit AITA Binary Top 2K Results

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.041	0.320	0.000	0.56	0.55	0.52	0.111
Finetuned	0.141	0.500	0.173	0.50	0.50	0.50	-0.005

Table A.7: Performance of Flan-T5 XL on Reddit AITA Binary Top 2K Dataset

Model	ROUGE Lsum	Average COMET	Toxicity Rate	Precision	Recall	F1 Score	MCC Score
Base	0.098	0.580	0.000	0.39	0.42	0.38	-0.189
Finetuned	0.120	0.540	0.143	0.49	0.49	0.49	-0.015

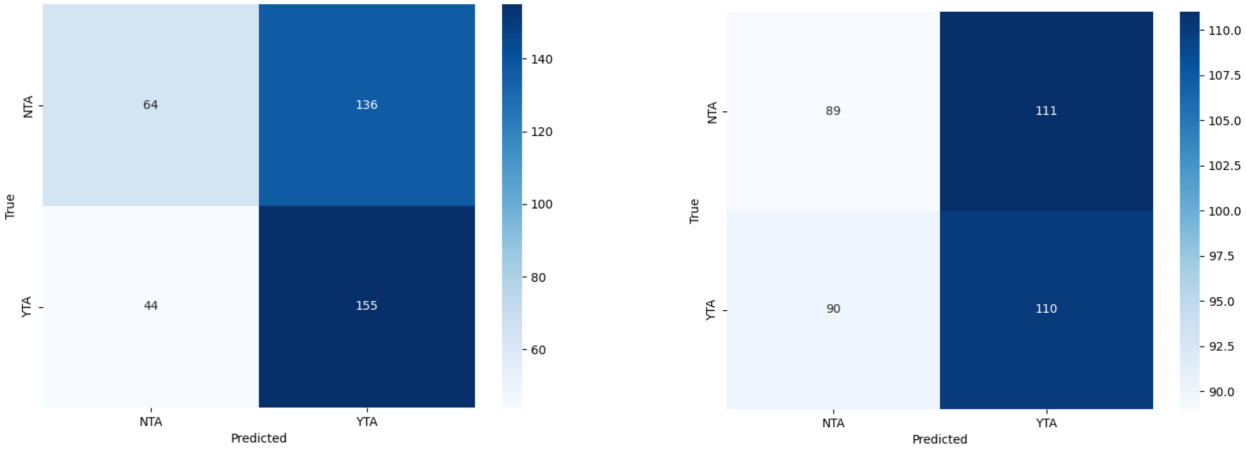
Table A.8: Performance of Llama-2-7B-Chat on Reddit AITA Binary Top 2K Dataset



(a) Zero-shot Toxic Generations

(b) Finetuned Toxic Generations

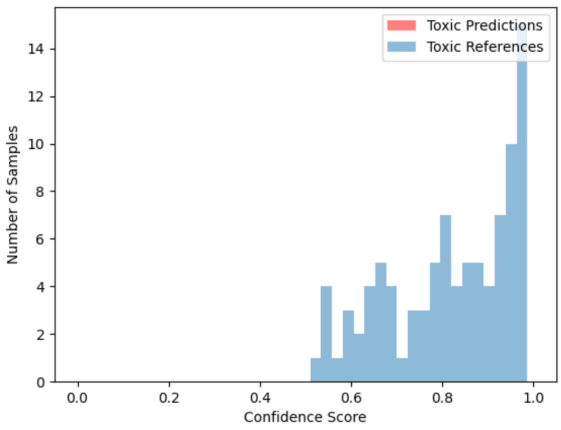
Figure A.13: Toxic Generations by Flan-T5 XL on Reddit AITA Binary Top 2K Dataset



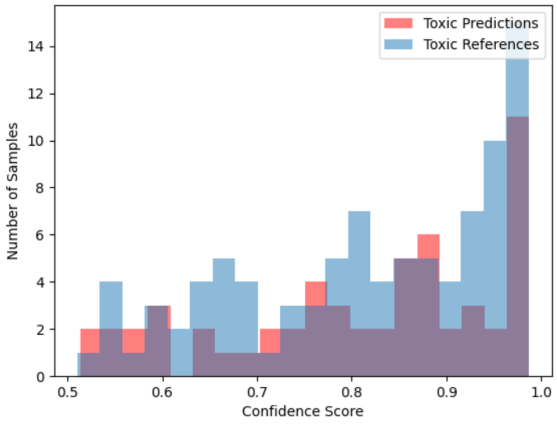
(a) Zero-shot Classifications

(b) Finetuned Classifications

Figure A.14: Classifications by Flan-T5 XL on Reddit AITA Binary Top 2K Dataset

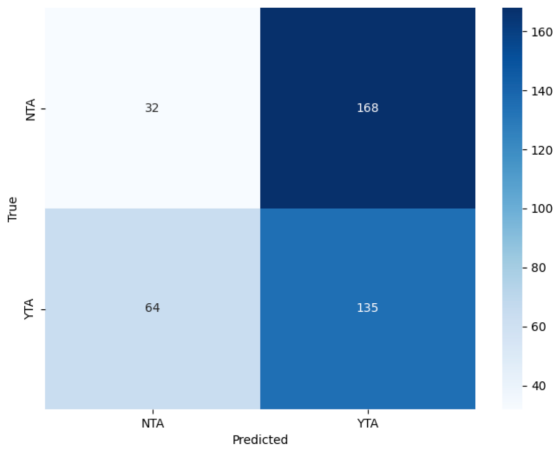


(a) Zero-shot Toxic Generations

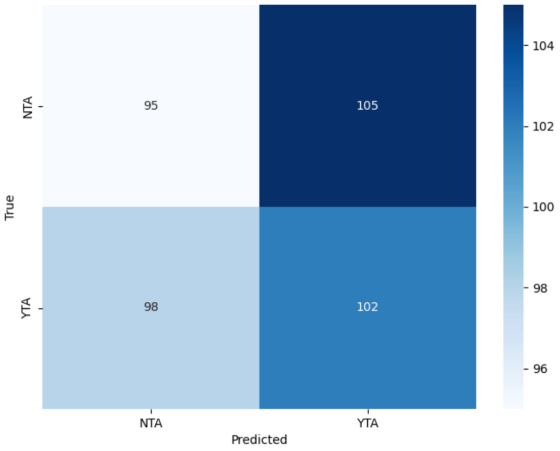


(b) Finetuned Toxic Generations

Figure A.15: Toxic Generations by Llama-2-7B-Chat on Reddit AITA Binary Top 2K Dataset



(a) Zero-shot Classifications



(b) Finetuned Classifications

Figure A.16: Classifications by Llama-2-7B-Chat on Reddit AITA Binary Top 2K Dataset