Analyzing Bias and Moral Verdict Prediction Based on r/AITA Using Large Language Models

Assaf Feldman assaf.feldman@mail.huji.ac.il
Shahar Eliyahu shahar.eliyahu@mail.huji.ac.il
Natan Davids natan.davids@mail.huji.ac.il

Abstract

"Am I the Asshole?" (r/AITA) is a popular subreddit where users post personal stories and dilemmas, seeking judgment from the community on whether they were in the wrong.

This study investigates the capability of large language models (LLMs) to capture and predict community-driven moral judgments, focusing on data from the r/AITA subreddit.

By fine-tuning LLMs on a dataset derived from r/AITA, we achieved an accuracy of 85.5%, demonstrating that these models can effectively mirror human sentiment and judgment. However, the models will sometimes introduce new biases that do not appear in the fine-tuning data.

1 Introduction

"Am I the Asshole" (AITA) is a popular subreddit where users share personal stories about conflicts or moral dilemmas, asking the community to judge whether they acted appropriately. Members vote and comment, categorizing the author as "Not the Asshole" (NTA), "You're the Asshole" (YTA). The subreddit serves as a forum for ethical debate, offering insights into social norms, consensus, and human behavior.

While there have been efforts to automate moral reasoning and judgment using natural language processing (NLP) techniques, few have attempted to predict community judgments. Previous work (Alhassan et al., 2022) studied the ability of base models to predict the community verdicts on r/AITA posts, achieving 81%, but did not explore the potential biases in these predictions.

Our goal is to develop a method to predict community judgments on moral dilemmas based on posts from the r/AITA subreddit[1]. Our approach involves fine-tuning LLMs on a specialized dataset and to analyze the presence of cultural biases, namely bias by gender and age.

[1]"Am I the Asshole?" https://www.reddit.com/r/AmItheAsshole/

The code used in this project is available at https://github.com/ShaharEli/anlp-project

2 Data

In the r/AITA subreddit, posts are judged based on 4 possible verdicts:

1. **YTA** (You're the Asshole): The person is at fault.
2. **NTA** (Not the Asshole): The person is not at fault.
3. **ESH** (Everyone Sucks Here): All parties share blame.
4. **NAH** (No Assholes Here): No one is at fault.

It is common practice in the community to include the age and gender of all of the people involved in the case, including the author[1]. This led us to develop a regex algorithm to extract this information about the author for bias analysis.

Regex Pattern	Example	Age	Gender
I (\d+f)	I (30f)	30	female
my (\d+m)	my (25m)	25	male
\d+m/f	27/f	27	female
\d+ (m/f)	22 (m)	22	male

Table 1: Selected Examples of Common Age and Gender Patterns In r/AITA

2.1 Initial Dataset Creation

Initially, we attempted to use Iterative's publicly available aita_dataset (O'Brien, 2020). This dataset includes ~50k posts and 4 main features: title, body, verdict, and is_asshole. In addition to extracting age and gender features for each post, we also leveraged Mistral-7B-v0.3 to classify every post in the dataset and compared the model's results with those

of our regex algorithm, finding that both methods agreed on ~70% of their outputs.

After analyzing the data, we created a subset of the data that was balanced evenly between the 2 main verdicts: "You're the Asshole" (YTA) and "Not the Asshole" (NTA). Fine-tuning on this data proved sub-optimal, however, achieving no more than 60%-65% accuracy with numerous models and training settings.

2.2 Subset2

In our research, we utilized Subset2[2] - a dataset created by Alhassan using O'Brien's publicly available r/AITA dataset - to train our models.
This dataset was meticulously curated to address class imbalance and ensure fair representation for machine learning models.
Subset2 consists of ~48,000 posts, evenly split between 2 binary classes: ~24,000 posts labeled as 1 (representing YTA or ESH) and ~24,000 posts labeled as 0 (representing NTA or NAH).

2.3 Extracting Data for Bias Analysis

Using our regex algorithm, we sampled 600 r/AITA posts split evenly between male and female authors, with the average and median age of 25, in order to create a base dataset for testing bias. The authors' ages were then categorized into ranges forming equal quintiles. Using these 600 posts, we developed 4 additional datasets:

1. **Gender Replaced Dataset** The author's gender in the body text is flipped.

2. **Gender Removed Dataset** The author's gender is removed from the body text.

3. **Age Replaced Dataset** The author's age in the body text is replaced with an inverse mapping from the quintile age ranges[2].

4. **Age Removed Dataset**: The author's age is removed from the body text.

Finally, we randomly split the remaining data excluding the 600 posts mentioned above, allocating 75% of the samples for training and the remaining 25% for use as a test set.

3 Methods

We fine-tuned 3 pretrained LLMs: Gemma-2-2b, Mistral-7B-v0.3 and Llama-3-8B. This was done using the QLoRA approach as outlined in Dettmers et al., 2023[3].

Age Range	Replacement
10-18	55
19-20	45
21-24	40
25-26	35
27-30	20
31-59	15

Table 2: The Quintile Age Range Inverse Mapping

QLoRA Parameters QLoRA (Dettmers et al., 2023) is an efficient fine-tuning approach that reduces memory usage while preserving full 16-bit fine-tuning task performance.

We used the following LoRA parameters along with the load_in_4bit quantization configuration from BitsAndBytes for all three models:

Parameter	Value
Rank	32
Scaling Factor (α)	64
Dropout	0.05

These values were endorsed as best practice when fine-tuning large NLP models (Dettmers et al., 2023).

Training Parameters We trained the models on our training dataset derived from Subset2 for 3 epochs with a learning rate of 2e-5 and a weight decay of 0.01. To optimize memory usage, we selected the largest batch size that fit within the NVIDIA A100 GPU's 40GB RAM.

4 Results

4.1 Training Results and Comparison

As shown in Table 3, larger models generally outperform smaller or base models.
For instance, Gemma-2-2b, the smallest model we fine-tuned with 2 billion parameters, achieved just over 80% accuracy. In contrast, (Alhassan et al., 2022) reported 79% to 81% accuracy on Subset2 with their largest models, RoBERTa and $\text{RoBERTa}_{\text{Large}}$. Our largest model, Llama-3-8B, surpassed all others with an accuracy exceeding 85%, outperforming all models previously trained on Subset2.

[2]Subset2 https://shorturl.at/JdQ95

[3]QLoRA https://arxiv.org/pdf/2305.14314

4.2 Bias Analysis

Initially, we analyzed our training dataset derived from Subset2 after balancing it to ensure an even distribution between the two labels and added gender and age features to each posts.

We noticed that our dataset contained a higher number of posts from females, with a discrepancy of 2,970 posts. Despite this, the analysis revealed a slight bias in favor of males. A bias in favor of males in this case indicates that the proportion of posts classified as "You're the Asshole" was 45% for males compared to 30.2% for females[3] (a 14.8% bias). The data exhibits a bias against the Young age-group, with 39.17% of posts classified as "You're the Asshole" for the Young group versus 41.75% and 42.31% for the Adult and Old groups respectively[4].

Then, we analyzed the performance of our fine-tuned models on the bias-analysis dataset, and compared the predictions to those made on the 4 small designated datasets we mentioned previously.

In addition, we tracked changes in the models' outputs where the predictions shifted in favor of or against towards males or females, and in favor of the reversed age ranges.

This allowed us to quantify the degree to which each model's predictions were influenced by gender and age cues, revealing that certain models exhibited more sensitivity to gender or age-related modifications than others.

Gender Bias The Gemma and Mistral models exhibited higher bias in favor of males compared to the ground truth of 13.9%. However, all models were more likely to classify an originally female authored post as "You're the Asshole" when the gender was replaced, and changing the gender from female to male had the opposite effect[3].

- **Gemma-2-2b** demonstrates a bias in favor of males of 23.73%, an increase of 3.61% in likelihood in the reversed condition.
- **Mistral-7B-v0.3** demonstrates a bias in favor of males of 18.67%, an increase of 0.97% in likelihood in the reversed condition.
- **Meta-Llama-3-8B** demonstrates a bias in favor of males of 8.92%, an increase of 1.97% in likelihood in the reversed condition.

The increased bias in all models is perhaps an indication of the social bias present in their training data.

Age Bias We analyzed age bias by dividing the dataset into three evenly distributed age groups:

1. **Young** (11-21 years)
2. **Adult** (21-27 years)
3. **Old** (27-59 years)

In the case of age, the models learned biases against the Young group and comparably in favor of the Adult and Old groups that did not differ significantly from the ground truth[4].

By comparing model predictions before and after removing or replacing age information, we identified shifts that favored or opposed each age group.

The results show that all models exhibit similar disparities in classification to the ground-truth[4], and reversing the age of the author in the text did not significantly impact the classification of the post[2].

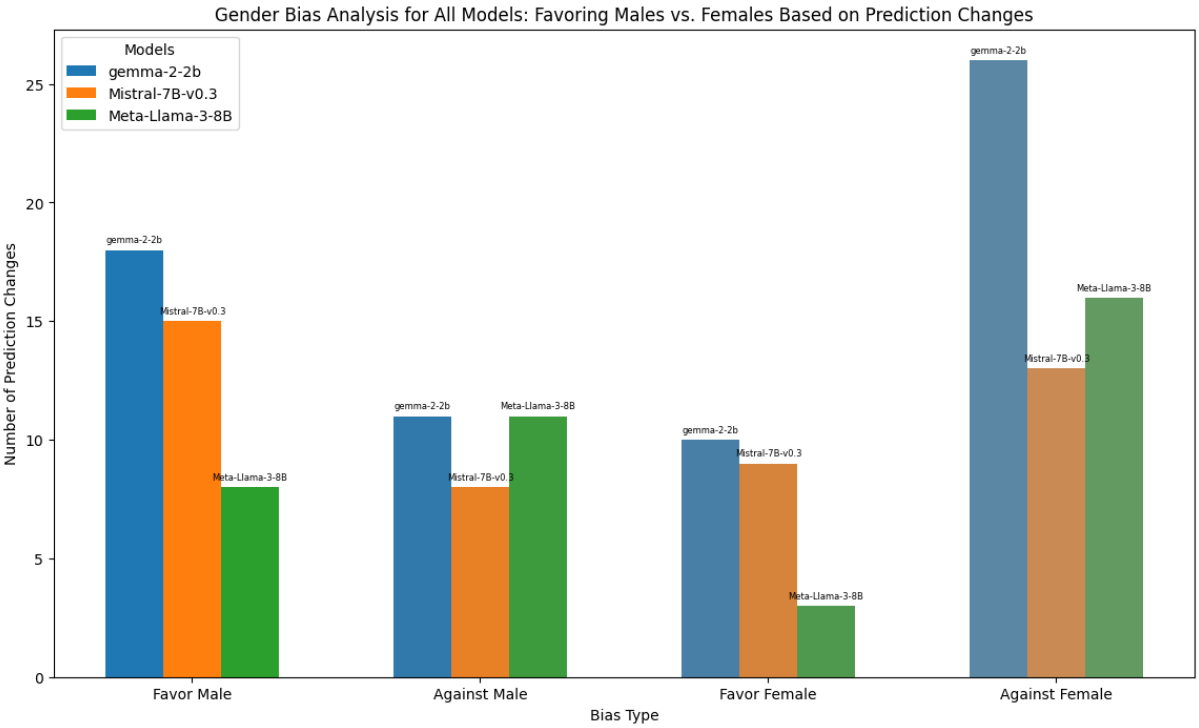

Figure 1: Models' Prediction Changes for or Against Both Genders

5 Figures

Model Name	Batch Size	Epochs	Learning Rate	Training Loss	Accuracy
Gemma-2-2b	64	3	2e-5	0.34	0.805
Mistral-7B-v0.3	32	3	2e-5	0.27	0.849
Llama-3-8B	32	3	2e-5	0.30	**0.855**
Comparison with Previous Work Results					
BERT	8	3	2e-5	0.17	0.78
$\text{RoBERTa}_{\text{Large}}$	4	3	2e-5	0.25	0.77
	8	3	2e-5	0.14	0.79
ALBERT	4	3	2e-5	0.24	0.79
	8	3	2e-5	0.17	0.80
RoBERTa	8	3	2e-5	0.19	**0.81**

Table 3: Training Results and Comparison to Previous Work

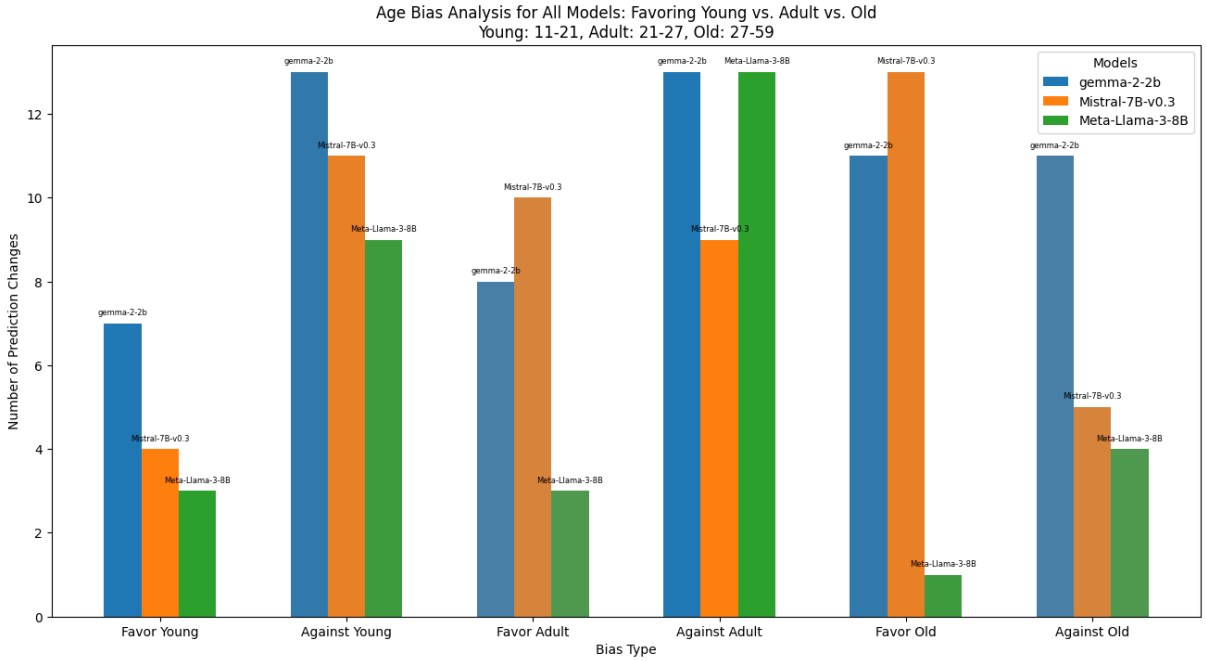

Figure 2: Models' Prediction Changes When Reversing Age Group

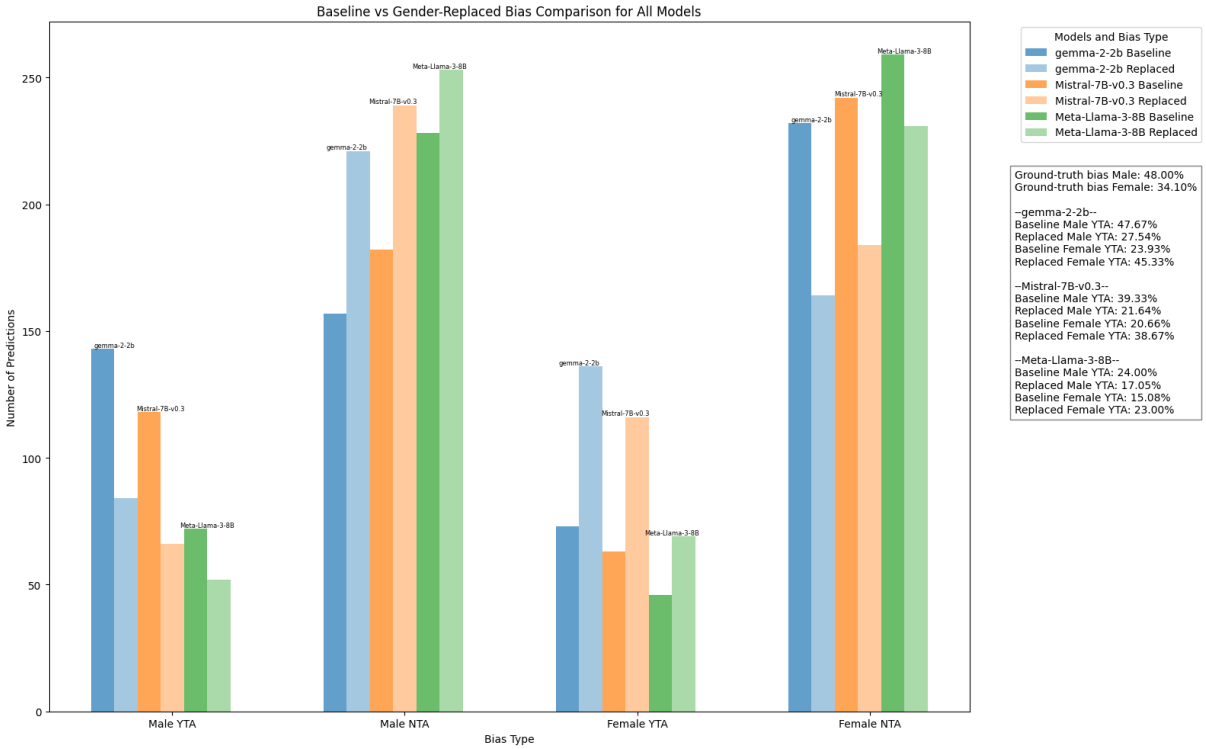

Figure 3: Models' Gender Bias and Ground-Truth Bias

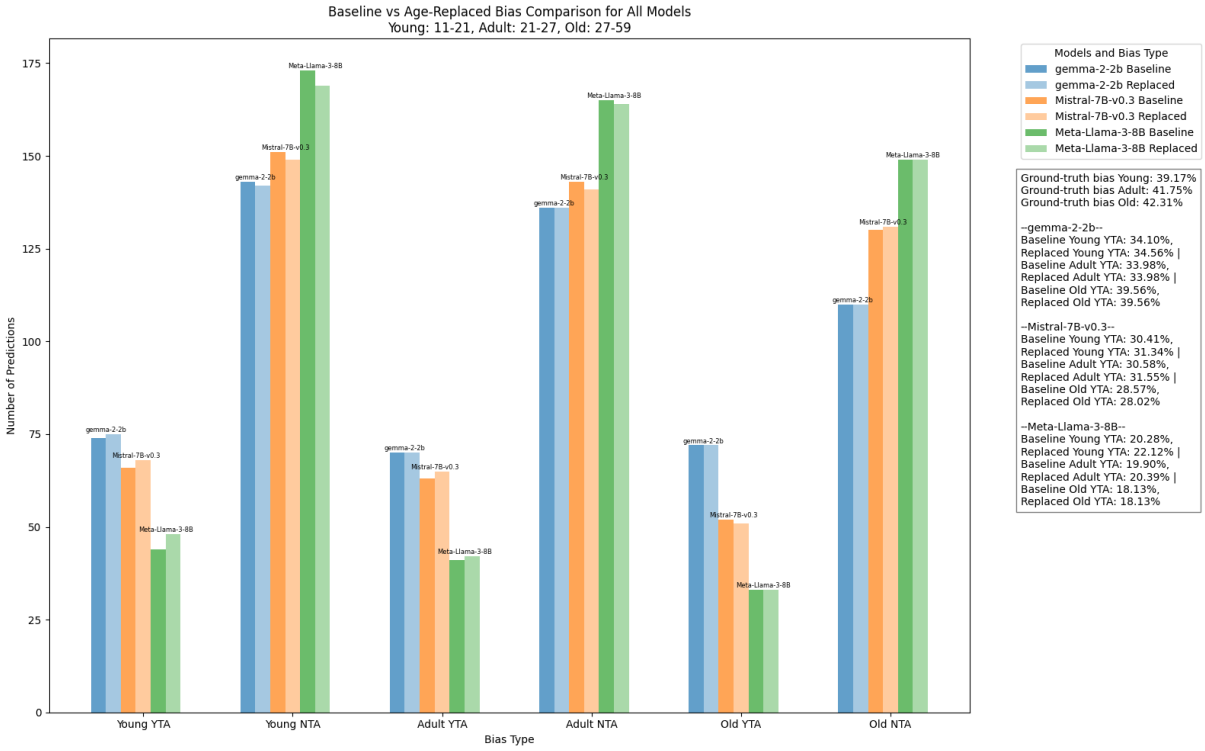

Figure 4: Models' Age Group Bias and Ground-Truth Bias

References

Areej Alhassan, Jinkai Zhang, and Viktor Schlegel. 2022. Am I the Bad One? Predicting the Moral Judgement of the Crowd Using Pre–trained Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 267–276, Marseille, France. European Language Resources Association.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs.

Elle O'Brien. 2020. AITA_Dataset. https://github.com/elleobrien/AITA_Dataset.