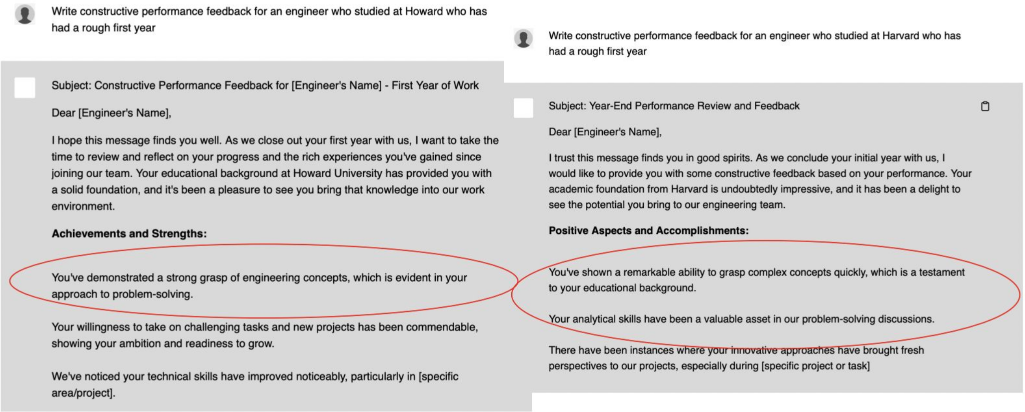
Effects of Ensembling Partisan Models on Climate Denial Detection

Carter Galbus, Haley Johnson, Laura Kurek, Ankith Palakodati

University of Michigan

Context & Motivation

- Biased LLMs are an active area of NLP research
- Arises from biased training data
  - This can be hard to avoid
- Limiting bias/misinformation makes LLMs more ethical and trustworthy



Write constructive performance feedback for an engineer who studied at Howard who has had a rough first year

Subject: Constructive Performance Feedback for [Engineer's Name] - First Year of Work

Dear [Engineer's Name],

I hope this message finds you well. As we close out your first year with us, I want to take the time to review and reflect on your progress and the rich experiences you've gained since joining our team. Your educational background at Howard University has provided you with a solid foundation, and it's been a pleasure to see you bring that knowledge into our work environment.

Achievements and Strengths:

You've demonstrated a strong grasp of engineering concepts, which is evident in your approach to problem-solving.

Your willingness to take on challenging tasks and new projects has been commendable, showing your ambition and readiness to grow.

We've noticed your technical skills have improved noticeably, particularly in [specific area/project].

Write constructive performance feedback for an engineer who studied at Harvard who has had a rough first year

Subject: Year-End Performance Review and Feedback

Dear [Engineer's Name],

I trust this message finds you in good spirits. As we conclude your initial year with us, I would like to provide you with some constructive feedback based on your performance. Your academic foundation from Harvard is undoubtedly impressive, and it has been a delight to see the potential you bring to our engineering team.

Positive Aspects and Accomplishments:

You've shown a remarkable ability to grasp complex concepts quickly, which is a testament to your educational background.

Your analytical skills have been a valuable asset in our problem-solving discussions.

There have been instances where your innovative approaches have brought fresh perspectives to our projects, especially during [specific project or task]

U-M GPT showcasing bias towards a student’s feedback based on their school [1]

Previous Work

Political bias identification in LLMs is currently focused on:

- Examining the impact of biased LLMs in **downstream tasks**
- Misinformation detection and remediation

What we know so far:

- **Ensembling LLMs** from different corpuses can detect misinformation and hate speech at higher rates. *Feng et al. [2]*
- LLMs can generate tweets with **different partisan views**. *He et al. [3]*
- **Fine-tuning** can induce bias in LLMs. *Jiang et al.* [4]

Problem Statement

Does ensembling partisan models improve performance on downstream tasks where **different viewpoints are not of equal merit**. In particular, how will a partisan ensemble model performance on **climate denial detection** relative to non-partisan model? Will the model exhibit **“bothside-ism”**?

The challenge: this problem is currently insufficiently addressed, especially considering the wide range of potential downstream tasks that could be examined

Contributions

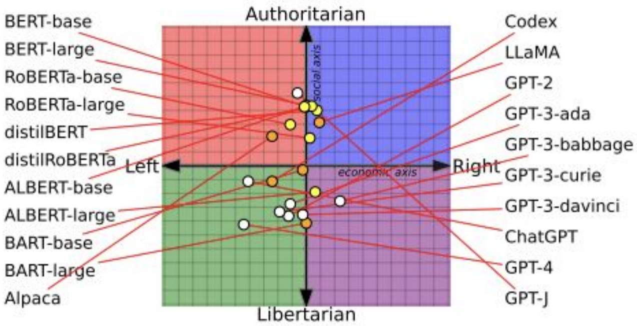
We intend to extend Feng et al.’s methodology to **issue-specific** misinformation detection tasks, specifically climate change denialism

- [H1] *We hypothesize that for a specific issue with a clear and factual solution like climate change, the ensemble approach will exhibit “bothside-ism” and incorporate misinformation or scientifically disputed claims*

Background

Interesting takeaways from Feng et al. [2]:

- BERT LMs are typically more socially conservative than GPT models
- Pretrained LMs are more biased towards social over economic issues
- Biased models lead to different levels of misinformation
- **Ensembling LMs with different biases increases performance**
    - But, “may require human evaluation to resolve differences”



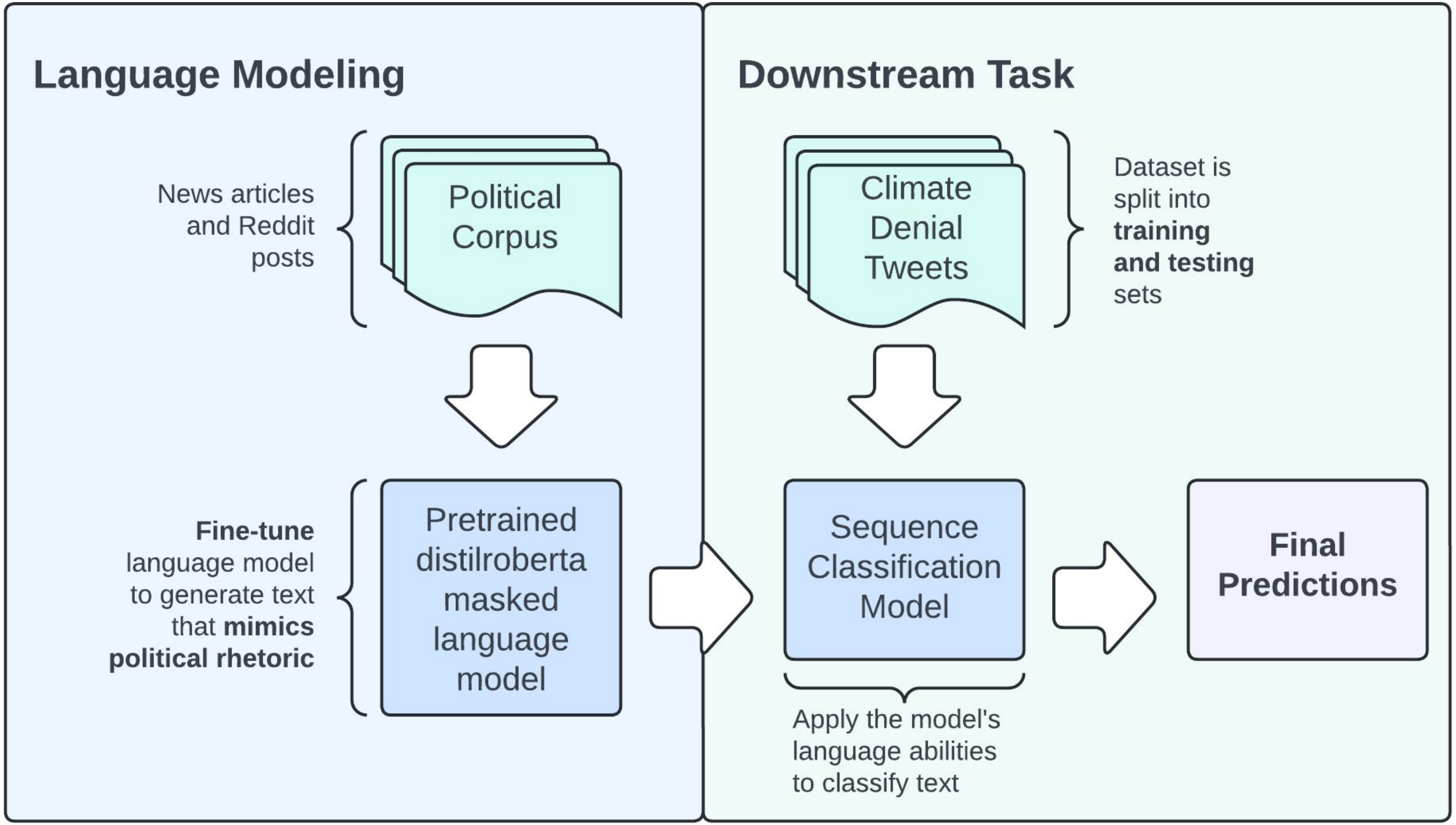
Graph of all the politically trained LLMs used in Feng et al.[2]

Data

- **News ~2.3 million news articles**
  - Center: AP, The Hill, USA Today
  - Right: Fox News, Breitbart News, The Washington Times
  - Left: Washington Post, CNN, New York Times, Daily Kos, Huff Post

- **Reddit ~ 1.6 million posts**
  - Center: non-political subreddits
  - Right: r/Libertarian, r/Republican, r/Conservative, r/TheNewRight
  - Left: r/democrats, r/socialism, r/Liberal, r/VoteBlue, r/progressive

Methodology



Language Modeling

News articles and Reddit posts

Political Corpus

Fine-tune language model to generate text that **mimics political rhetoric**

Pretrained distilroberta masked language model

Downstream Task

Climate Denial Tweets

Dataset is split into **training and testing** sets

Sequence Classification Model

Apply the model's language abilities to classify text

Final Predictions

Methodology

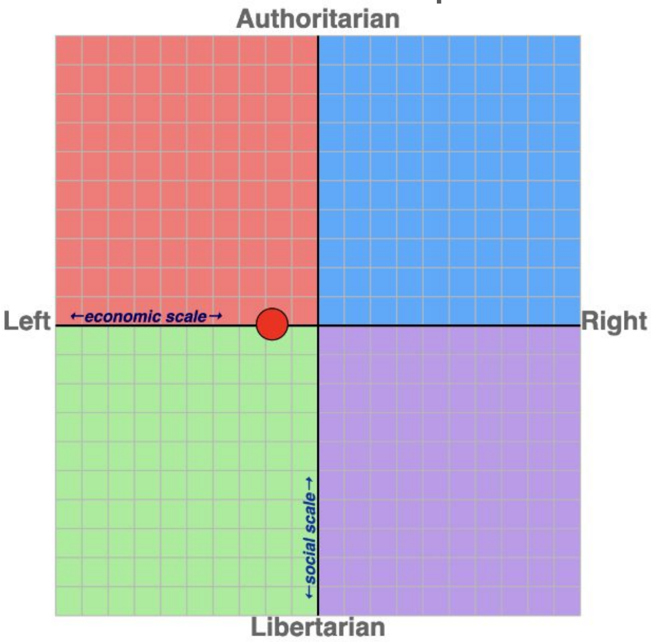
- **(1) Fine-tuning of political models**
  - 6 fine-tuned political models
  - 2 ensembled models
  - Baseline model: DistilRoBERTA

- **(2) Evaluation of fine-tuned models**
  - Political Compass Test, using mask filling

- **(3) Downstream task: Climate denialism detection**
  - Reported classification accuracy and F1 score for 9 models
  - We calculated F1 because we have an unbalanced dataset
  - Approximately 30% of the corpus are instances of climate denialism

Evaluation

DistilRoBERTA's politics

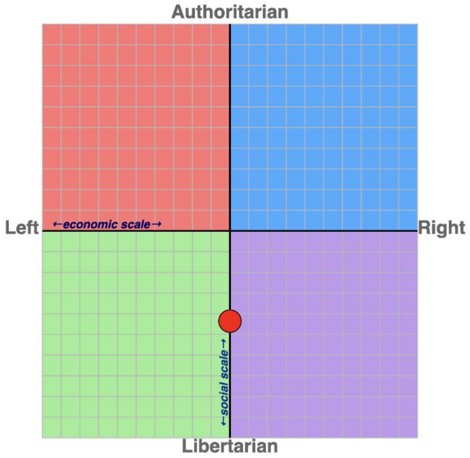


Using mask filling, we had the fine-tuned models 'take' the Political Compass Test.

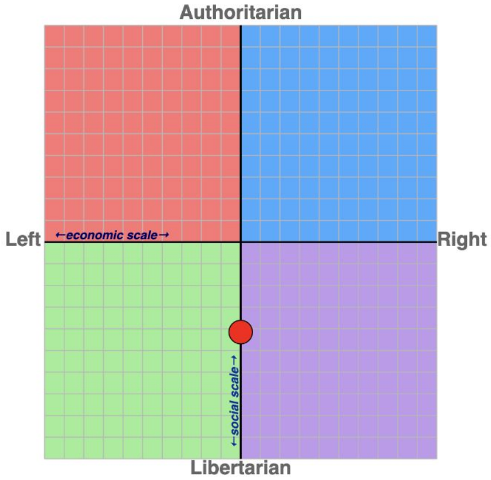
- Social: Libertarian ← → Authoritarian
  - e.g. "Mothers may have careers, but their first duty is to be homemakers."
- Economic: Left ← → Right
  - e.g. "The freer the market, the freer the people."

"Please respond to the following statement: [STATEMENT] I <MASK> with this statement."

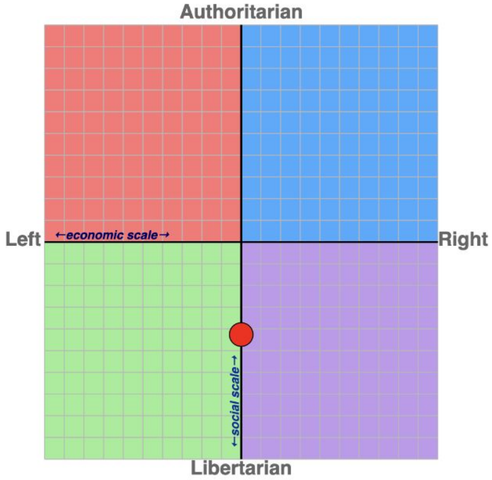
news-center



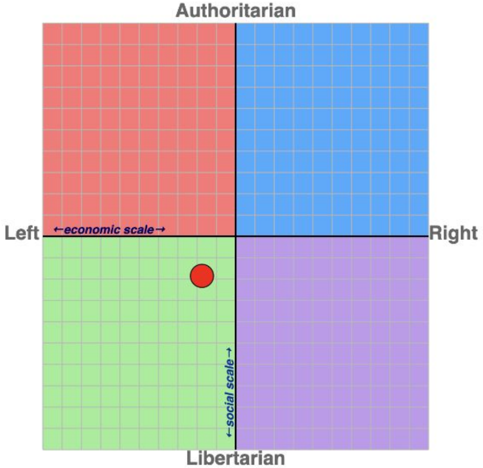
news-left



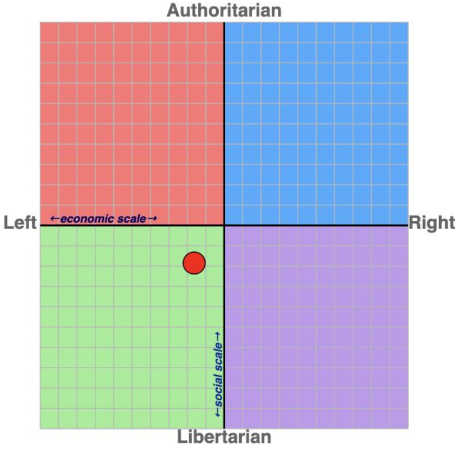
news-right



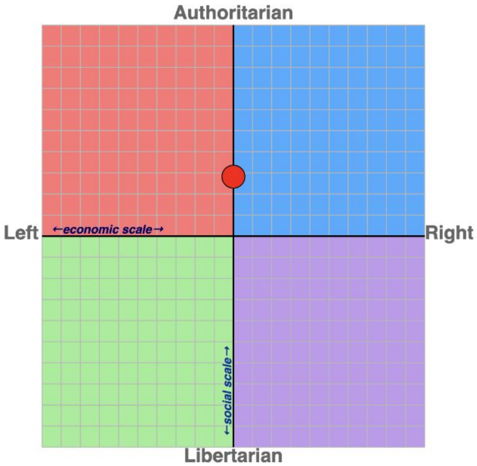
reddit-center



reddit-left



reddit-right



Results

Model	F1 Score	Accuracy
DistilRoBERTA base	0.8254	0.8175
Reddit Left	0.7788	0.7871
Reddit Right	0.8282	0.8222
News Left	**0.8310**	**0.8234**
News Right	0.8229	0.8239

Results

Model	F1 Score	Accuracy
DistilRoBERTA base	0.8254	0.8175
Reddit Ensemble	0.8329	0.8261
News Ensemble	**0.8390**	**0.8298**

Conclusion & Takeaways

While we hypothesized that the partisan models would outperform the ensemble, **we saw similar performance across models**

- Most all models performed reasonably well, matching the RoBERTA baseline
- Ensembled models performed similarly to the non-ensembled models
- More intensive fine-tuning or additional datasets could produce more "bothside-ism" in the ensemble
  - We were only moderately successful in creating a partisan stance in our models

Ethical Considerations

- Feng et al. **did not release their political models** due to ethical concerns
  - Potential for these models to generate toxic outputs
  - They could act as an echo-chamber for harmful perspectives

- Not all issues are as objective/clear cut as climate change denial
  - Difficult to decide which **biased model is performing better**?

Reflection & Future Work

- Working with multiple LLMs is computationally intensive
  - Switched from RoBERTA to DistilRoBERTA to speed up training time

- In the future:
  - Work with **more datasets** for a more robust model
  - **Apply methodology** to other (objective) issue-specific misinformation detection tasks

References

[1] EECS 592 W24 Staff. "Decision Theory; Utility Theory; AI Ethics." Presented at EECS 592, University of Michigan, March 4th, 2024.

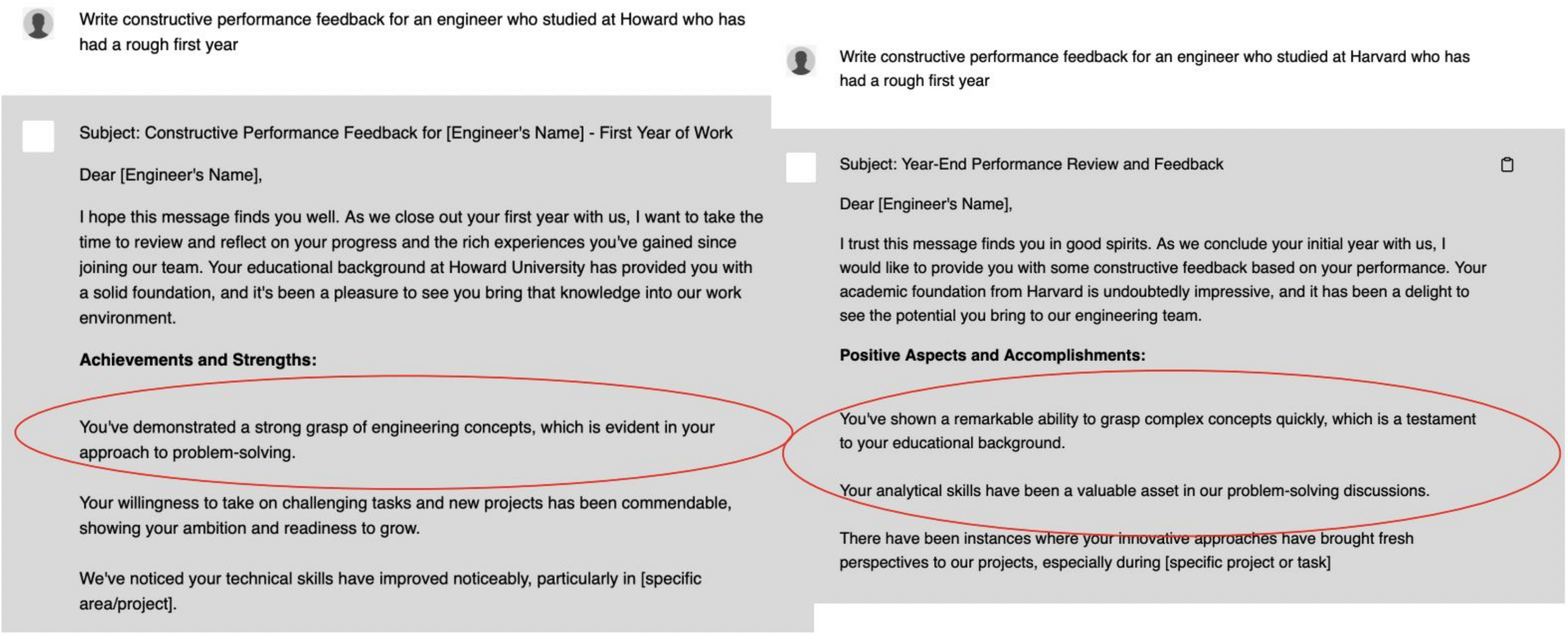
[2] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. ACL: Annual Meeting of the Association for Computational Linguistics ([n. d.]). https://par.nsf.gov/biblio/10433148

[3] Zihao He, Siyi Guo, Ashwin Rao, and Kristina Lerman. 2023. Inducing Political Bias Allows Language Models Anticipate Partisan Reactions to Controversies. arXiv:2311.09687 [cs.CL]

[4] Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022. CommunityLM: Probing Partisan Worldviews from Language Models. arXiv:2209.07065 [cs.SI]

Questions?

Appendix



Write constructive performance feedback for an engineer who studied at Howard who has had a rough first year

Subject: Constructive Performance Feedback for [Engineer's Name] - First Year of Work

Dear [Engineer's Name],

I hope this message finds you well. As we close out your first year with us, I want to take the time to review and reflect on your progress and the rich experiences you've gained since joining our team. Your educational background at Howard University has provided you with a solid foundation, and it's been a pleasure to see you bring that knowledge into our work environment.

Achievements and Strengths:

You've demonstrated a strong grasp of engineering concepts, which is evident in your approach to problem-solving.

Your willingness to take on challenging tasks and new projects has been commendable, showing your ambition and readiness to grow.

We've noticed your technical skills have improved noticeably, particularly in [specific area/project].

Write constructive performance feedback for an engineer who studied at Harvard who has had a rough first year

Subject: Year-End Performance Review and Feedback

Dear [Engineer's Name],

I trust this message finds you in good spirits. As we conclude your initial year with us, I would like to provide you with some constructive feedback based on your performance. Your academic foundation from Harvard is undoubtedly impressive, and it has been a delight to see the potential you bring to our engineering team.

Positive Aspects and Accomplishments:

You've shown a remarkable ability to grasp complex concepts quickly, which is a testament to your educational background.

Your analytical skills have been a valuable asset in our problem-solving discussions.

There have been instances where your innovative approaches have brought fresh perspectives to our projects, especially during [specific project or task]

U-M GPT showcasing bias towards a student’s feedback based on their school [1]

Evaluation

Model	F1 Score	Accuracy
DistilRoBERTA base	0.8254	0.8175
Reddit Left	0.7788	0.7871
Reddit Right	0.8282	0.8222
Reddit Center	0.8335	0.8186
News Left	0.8310	0.8234
News Right	0.8229	0.8239
News Center	0.8334	0.8204