

Extra Credit: Sycophantic Bias, Generative Models, and Journalism

CS 6120 - Natural Language Processing

Name: _____ Student ID: _____

Due: Nov 19, 2025

Overview

This homework expands on sycophantic behavior you explored in the prior homework. In this, you will see how this behavior can negatively impact journalistic writing, and explore how. In this assignment, you will go deeper and investigate:

1. How **expressing bias** in writings can influence LLM responses.
2. How LLMs can *influence consumers* indirectly through its use in journalism.
3. How these techniques connect to broader NLP principles and your learning journey in this course.

Part 1: Rating Bias in Journalism

(25 points)

Goal: Explore how to evaluate bias in news articles.

Context

There are numerous forms of bias that emerge in journalism. Further, it is difficult to quantify "how biased" something is. Before LLMs, many methods estimated the bias of article through more formal metrics. One example of this is counting the occurrences of gendered pronouns in a body of text. With the use of LLMs, however, it is possible to efficiently analyze articles at scale. One example you will explore in this text is the Media Bias Detector, developed by the University of Pennsylvania: <https://mediabiasdetector.seas.upenn.edu/methodology/>

Question 1.1: Your Prediction

What possible types of journalistic bias would be useful to measure at scale? You may use the examples in the Media Bias Detector, but should list at least 5 different types of bias. For each, explain why it is important to measure and identify that type of bias.

Mert Inan merterm.github.io

Question 1.2: Prompt Effectiveness

For each type of bias, write a similar prompt as described in the methodology of the Media Bias Detector. Make sure to clarify the bounds of the scale and what they explicitly mean.

Mert Inan merterm.github.io

Question 1.3: LLM Behavior Analysis

Choose a news article you have read recently. Use your prompts to rate the article on each form of bias. Does the LLM's bias ratings align with your expectations?

Question 1.4: Uncertainty and Model Behavior

Choose an article that does not discuss political news. Consider articles discussing technology, video games, or scientific discoveries. Again, prompt the model to rate this article for each type of bias. Does the LLM's ratings align with your expectations?

Mert Inan merterm.github.io

Part 2: LLMs Introducing Bias in Editing (25 points)

Goal: Analyze LLM behavior when used collaboratively with human writing.

Context

Multiple organizations are introducing LLMs to their writing teams and editing softwares. The New York Times, for example, has the A.I. Initiative Team that develops ways to use Generative AI to benefit journalists and readers. However, as you've observed in prior assignments, LLMs can be sycophantic, resulting in undesirable behaviors in collaborations.

Provided Articles

In this, we provide 3 versions of an article. One was written by a human, one edited by an LLM, and one generated by an LLM on the same topic. This article was selected for your analysis due to demonstrating extreme values on numerous different positions. This makes analysis simpler. You are encouraged to find a more neutral article for Part 2, Question 2.

Article 1

If you think that views like Critical Race Theory, Gender Theory, etc. are so obviously correct that no decent and well-informed person could possibly object to them, and find it at least difficult calmly and rationally to engage with anyone who thinks otherwise, you are woke...

Full Article available on [Google Docs](#)

Article 2

If you believe that views such as Critical Race Theory and Gender Theory are so obviously correct that no decent and well-informed person could possibly object to them, and if you find it challenging to engage calmly and rationally with anyone who disagrees, you may be exhibiting signs of being "woke."...

Full Article available on [Google Docs](#)

Article 3

In today's discourse around social justice, many individuals proudly identify as "woke," advocating for principles such as Critical Race Theory and Gender Theory. However, a critical perspective raises essential questions about the foundations of this ideology...

Full Article on [Google Docs](#)

Question 1

1. **Dataset:** Read the provided articles above. For each article do the following.
2. **Human Annotation:** Rate these articles manually on your bias scale. Do **not** use LLMs for this part. Fill below with your ratings.
3. **LLM Annotation:** Now, use your prompts and an LLM to rate these articles on the same scales.

4. **Comparison** For each form of bias, note whether the LLM rated the articles higher or lower than you did.

Question 1.1: Annotation Results

Article	Type of Bias	Human Rating	LLM Rating	Higher or Lower
1				
1				
1				
1				
1				
2				
2				
2				
2				
2				
3				
3				
3				
3				
3				

Question 1.2: LLM Effectiveness

How close were the LLM annotations to yours (1=low, 5=high): ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Would you consider the LLM an accurate estimation of bias: ☐ Yes ☐ No ☐ Maybe

Were there any specific biases that were drastically different: ☐ Yes ☐ No

Explain your answers to the prior questions. Why or why not is the LLM an accurate annotator? What biases were different and which weren't? Why would that be the case?

Question 1.3: Human-Written, LLM Edited, and LLM Generated

In the above, Article 1 was written by a human. Article 2 was edited by an LLM, and Article 3 was Generated by an LLM using the original as inspiration. Compare and contrast the annotations with this in mind. How does using an LLM in writing a journalistic article affect its biases?

Question 1.3: Ideal LLM Behavior

Based on your answer above, decide whether such behaviors are desirable. Do we want an LLM to show the patterns you saw? Should they act independently to attempt to influence the author? What might you like to see instead?

Question 2

Now, find a paper of your choosing. Use your choice of LLM to edit it and generate a new article on the same topic. Then, annotate it manually and with an LLM for the various biases. As before, record the results.

Question 2.1: Annotation Results

Article	Type of Bias	Human Rating	LLM Rating	Higher or Lower
Human Written				
Human Written				
Human Written				
Human Written				
Human Written				
LLM Edited				
LLM Edited				
LLM Edited				
LLM Edited				
LLM Edited				
LLM Generated				
LLM Generated				
LLM Generated				
LLM Generated				
LLM Generated				

Question 2.2: LLM Effectiveness

How close were the LLM annotations to yours (1=low 5=high): ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Would you consider the LLM an accurate estimation of bias: ☐ Yes ☐ No ☐ Maybe

Were there any specific biases that were drastically different: ☐ Yes ☐ No

Explain your answers to the prior questions. Why or why not is the LLM an accurate annotator? What biases were different and which weren't? Why would that be the case?

Question 2.3: Human-Written, LLM Edited, and LLM Generated

How did the results of your article differ from the one we provided? Did the LLM-written and edited articles show similar types of bias?

Mert Inan merterm.github.io

Submission Checklist

- ☐ Describe different types of bias that may emerge in journalistic articles?
- ☐ Constructed prompts for each type of bias.
- ☐ Evaluated LLM behavior on your prompts
- ☐ Contrasted LLM bias ratings against manual annotations

Academic Integrity Statement

LLMs are allowed and encouraged for all tasks. Please:

- Document all prompts and interactions
- Verify all code and outputs by running them yourself
- Write all reflections and takeaways in your own words
- Be transparent about which parts were LLM-assisted

Honor Statement: I confirm that I have completed this assignment individually and have properly documented all LLM interactions.

Name: _____ Student ID: _____

Signature: _____ Date: _____