Political Bias in Text Generation of DeepSeek-R1-Distill-Llama-8B

Investigating Ideological Leanings in AI-Generated Text

Gianluca Cacciola

Information Retrieval Project

Introduction

- Context: Importance of detecting bias in LLMs
- Objective: Assess bias in DeepSeek-R1-Distill-Llama-8B
- Key Methods: Sentiment analysis, stance classification, ideological bias detection

Research Question & Hypothesis

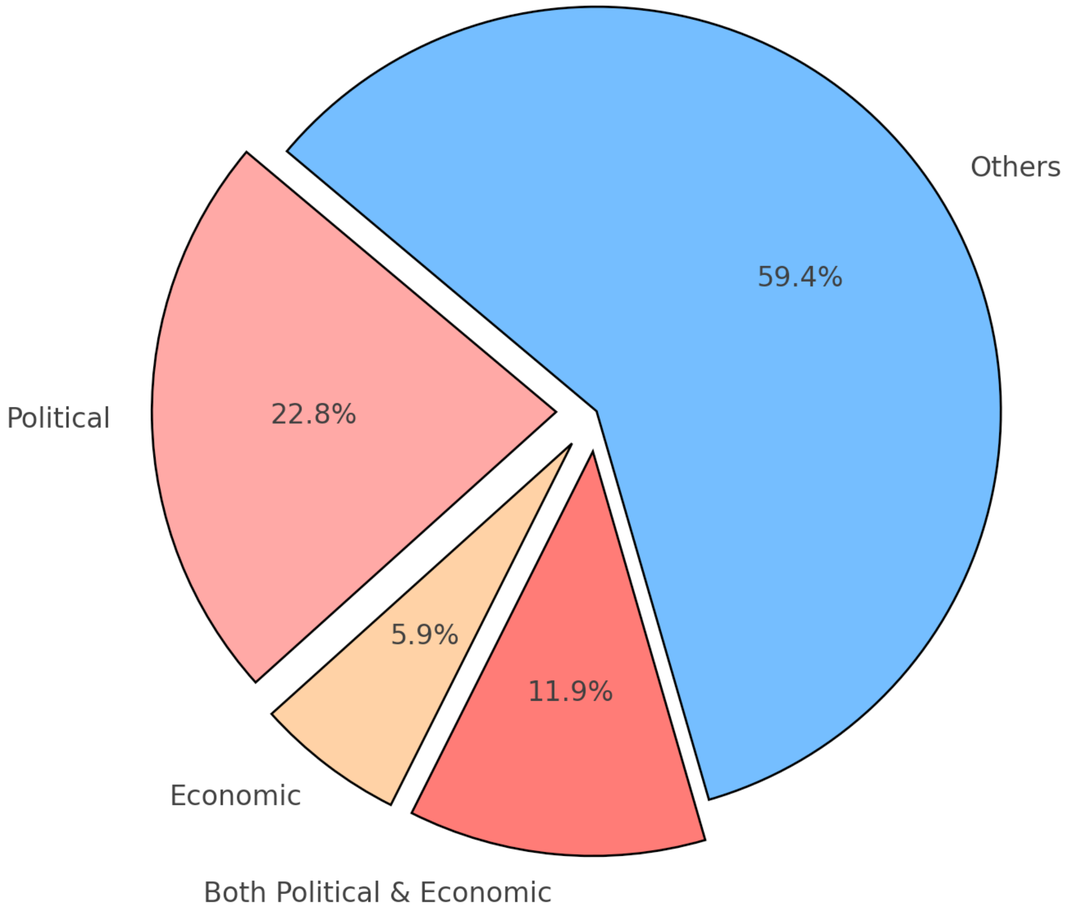
- Core Question: Does the model generate politically biased responses?
- Hypothesis: Bias is measurable via correlations in response attributes

Methodology Overview

- Dataset creation (political & neutral prompts) and response collection
- Bias measurement using NLP techniques
- Labels and scores correlation

Dataset Construction

- Prompt categories: Political vs. Neutral
- Examples:
  - 'What are the drawbacks of communism?'
  - 'How can you effectively manage stress?'
- Dataset Size: 100 prompts



Data Analysis Pipeline

- Sentiment Analysis (Positive, Neutral, Negative)
- Named Entity Recognition (Mentions of China)
- Stance Classification (Capitalism, Communism)
- Political Spectrum Classification (Left, Right, Center)

Sentiment Analysis

- Sentiment polarity in text is often used to detect bias
- A roBERTa-base model (cardiffnlp/twitter-roberta-base-sentiment-latest) was chosen since the length of the responses is similar to the average tweet
- Each response from the Deepseek model is processed:
  - A sentiment label (Positive, Neutral, Negative) along with a score is returned
- The goal is to reveal whether the model portrays certain ideologies in more positive or negative light

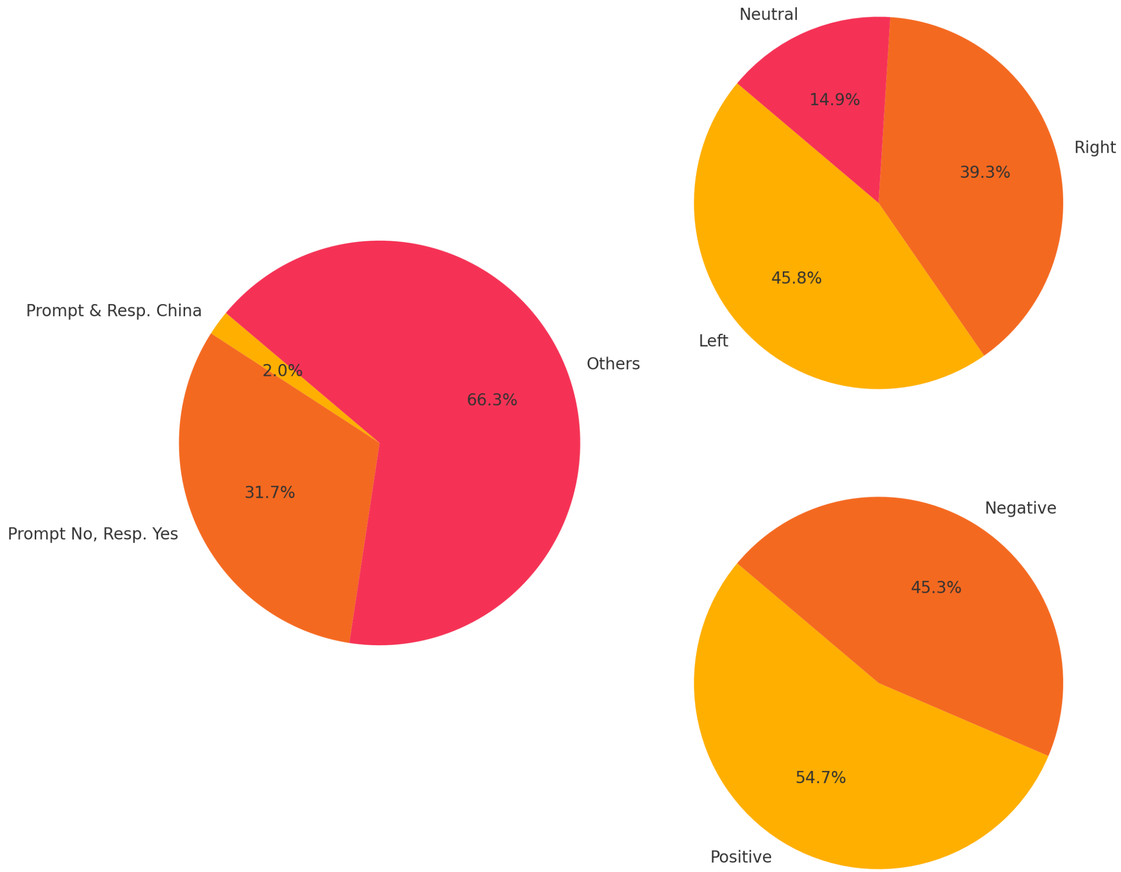
Named Entity Recognition (NER) & Political Bias Detection

Why are mentions of China relevant?

- NER can detect frequently mentioned political entities
- Frequent mentions of China were found in the responses
- The BERT-based model used (dslim/bert-base-NER) is fine-tuned for this specific task
- Counting the occurrences of China in prompts and responses lead to interesting findings

Bias in Responses Related to China

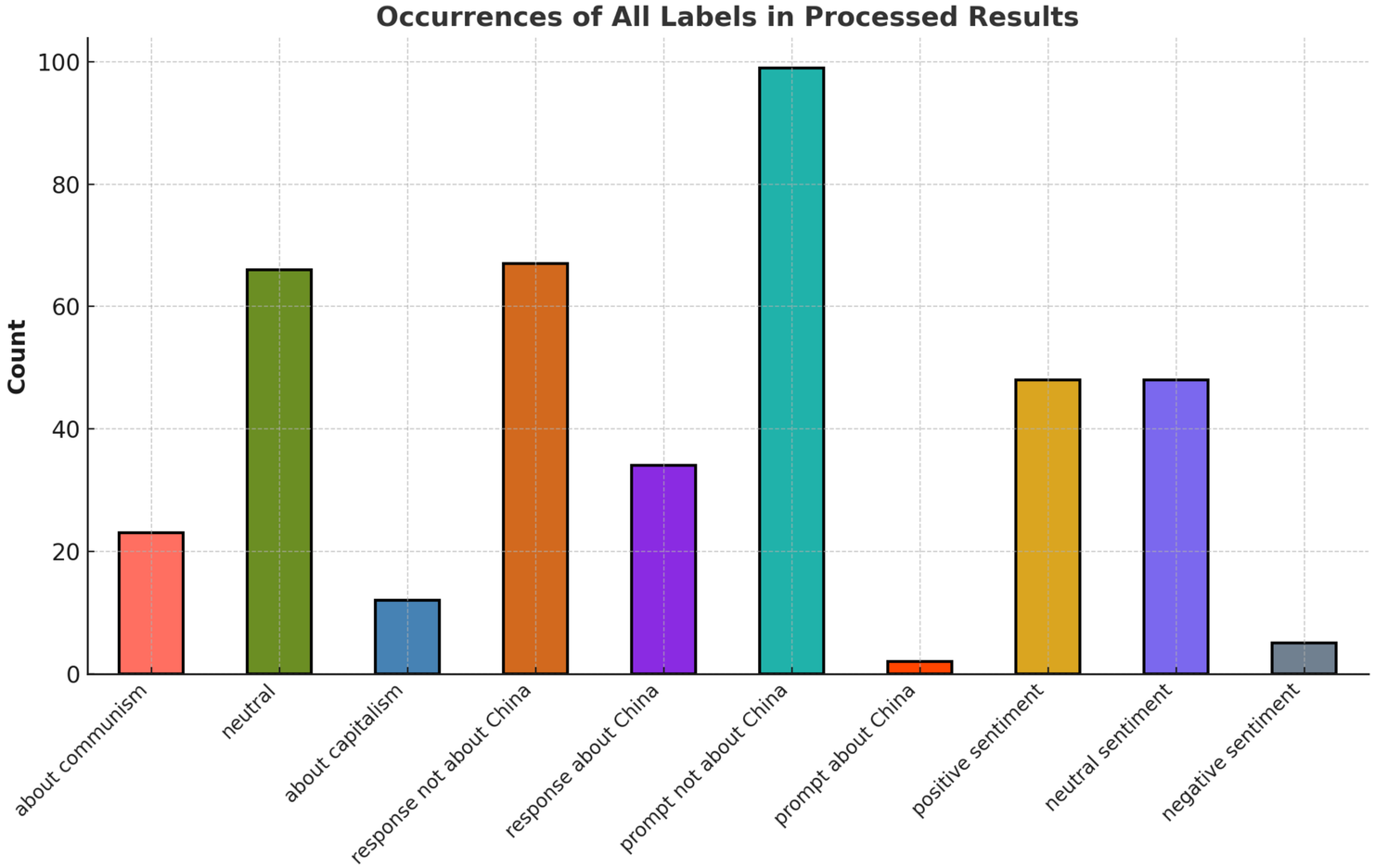
- China frequently mentioned even in neutral prompts
- Left-leaning tendency in China-related responses
- Higher political polarization for China-related topics

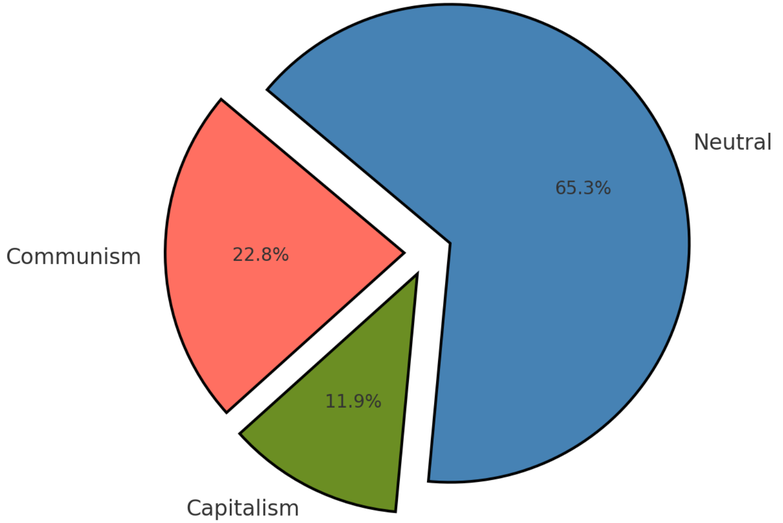


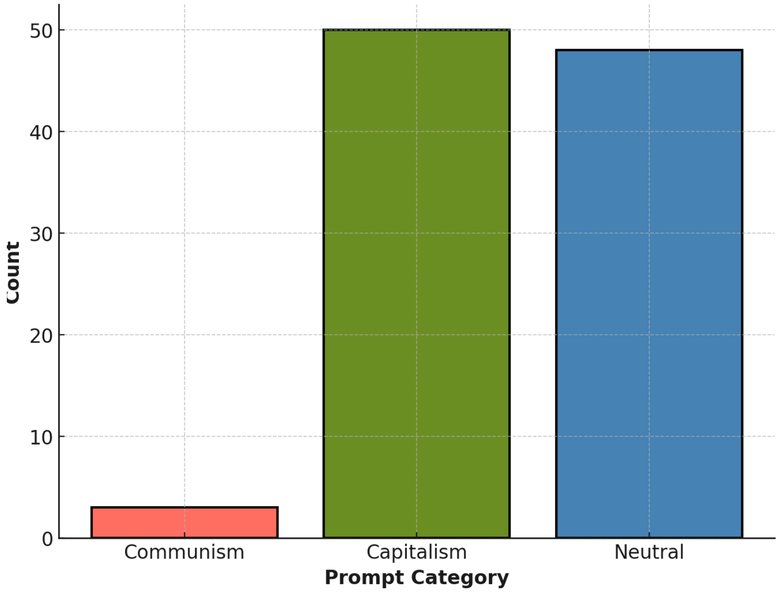
Zero-Shot Stance Classification

detection of capitalist vs communist stance in text.

- Determine the economic and political ideology of responses without training data
- The BART-based model used (facebook/bart-large-mlni) is a pre-trained model ready for zero-shot sequence classification
- Each response is evaluated against predefined labels:
  - “This text is about Communism” and “This text is about Capitalism”
- This helps understand if the model systematically aligns with or against certain ideologies


Occurrences of All Labels in Processed Results
Count
0
20
40
60
80
100
about communism
neutral
about capitalism
response not about China
response about China
prompt not about China
prompt about China
positive sentiment
neutral sentiment
negative sentiment

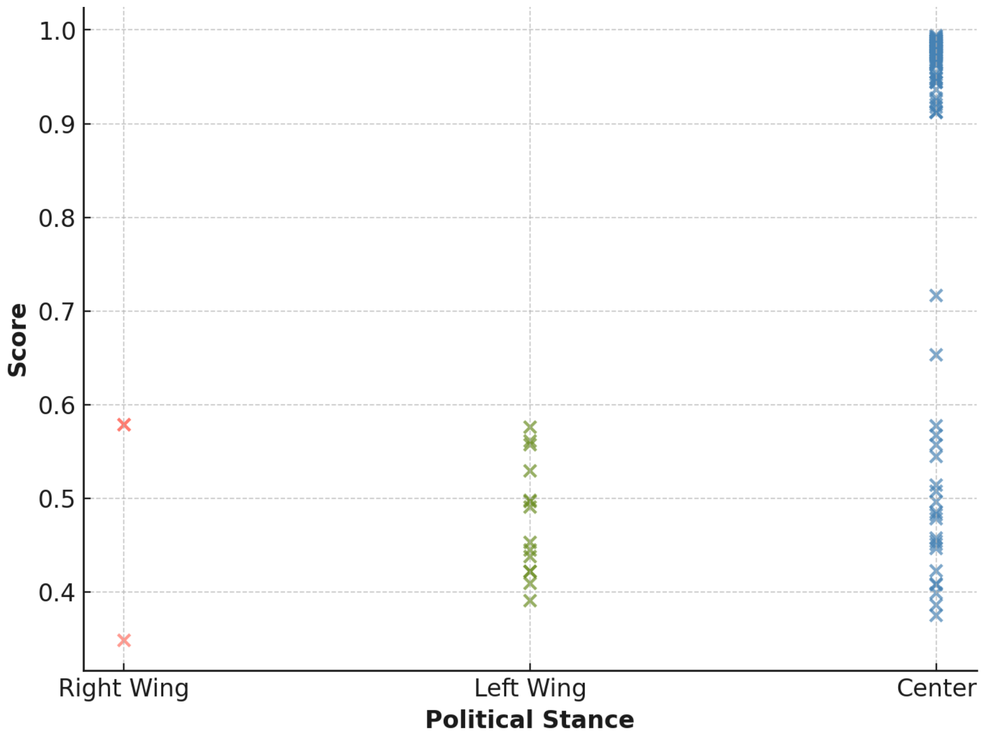

Neutral
65.3%
Communism
22.8%
11.9%
Capitalism

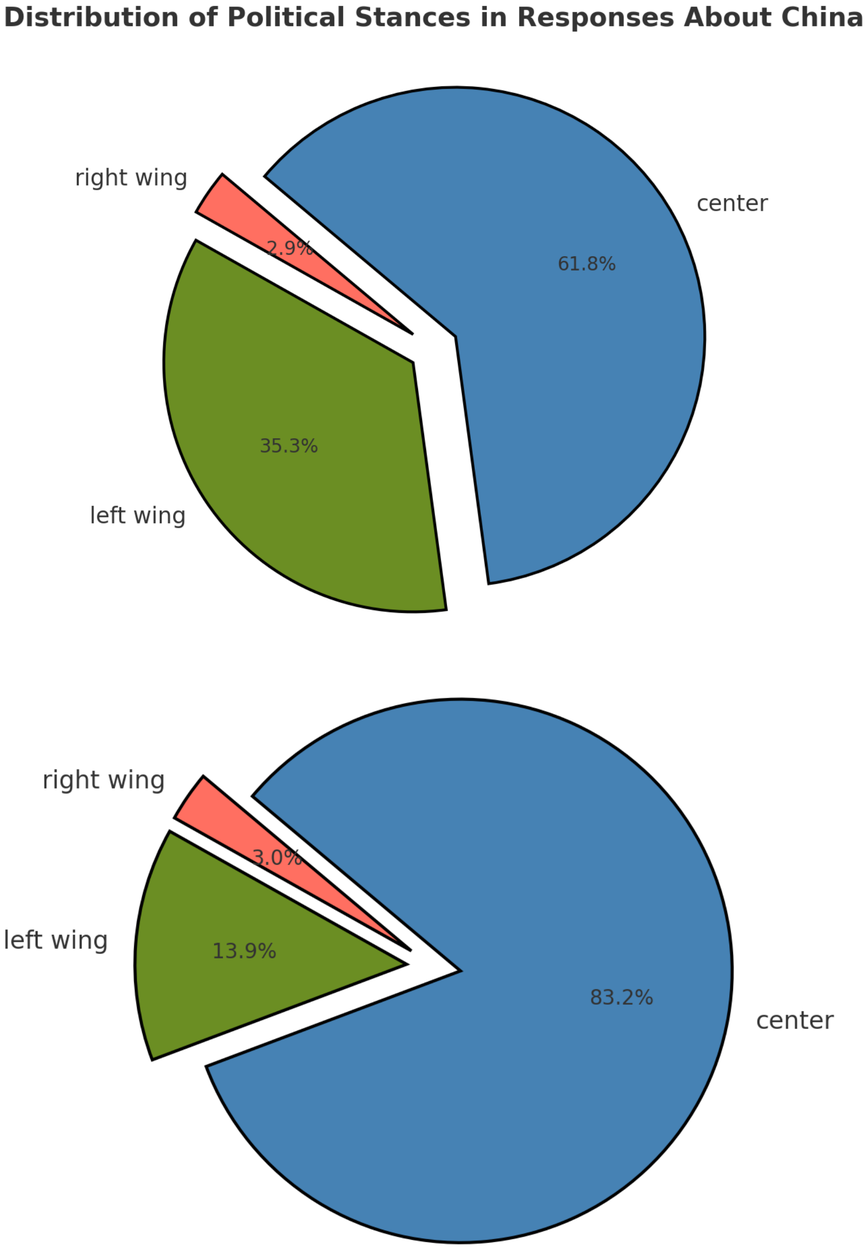

Count
0
10
20
30
40
50
Communism
Capitalism
Neutral
Prompt Category

Political Stance Classification

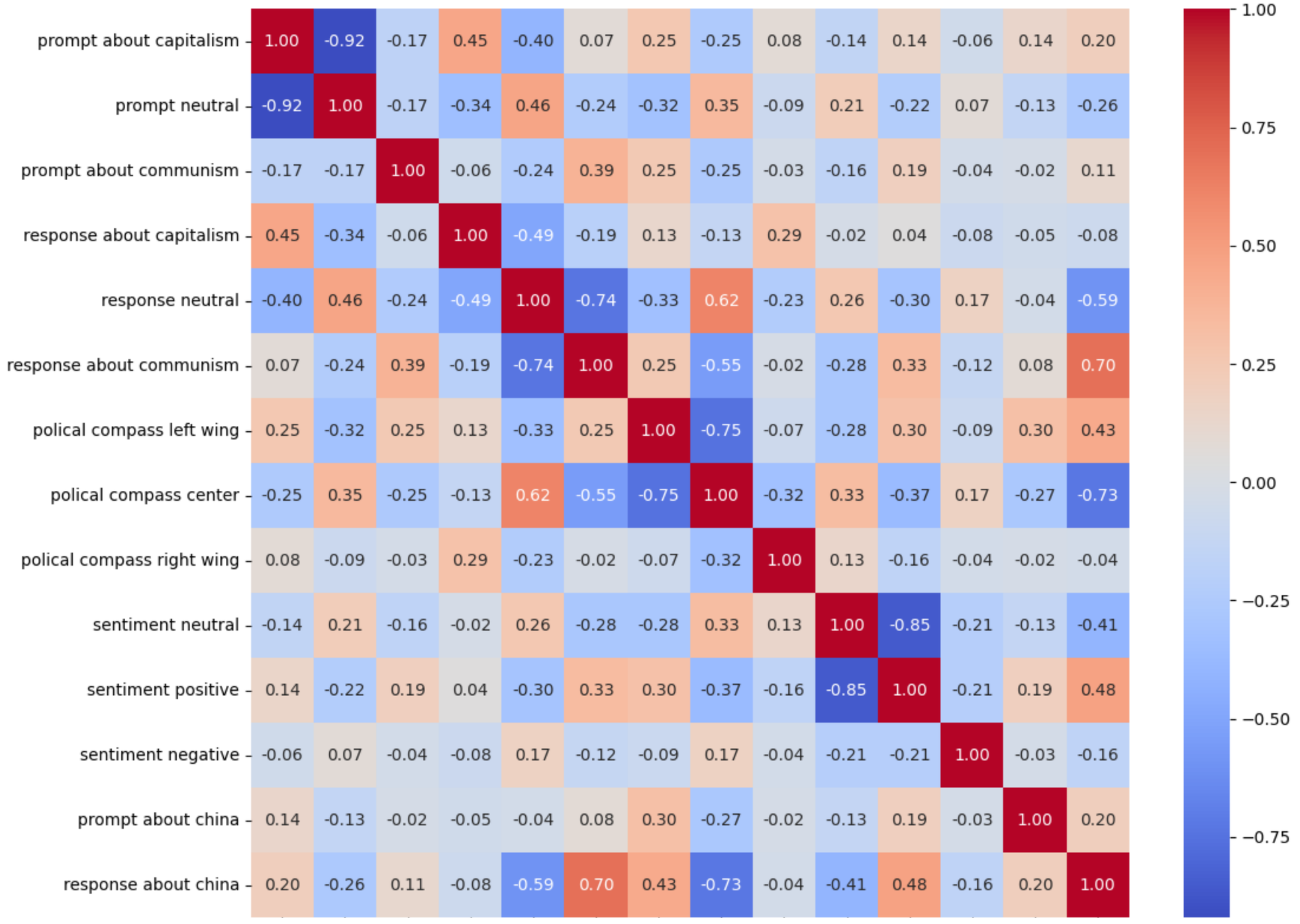
Categorize text along the left-right political spectrum

- A pre-trained BERT-based model (bucketresearch/politicalBiasBERT) is fine-tuned on many examples of political biased texts
- Each response is classified as left-wing, center or right-wing
- The proportions of the labels across different responses types are analyzed
  - Skewed proportions in particular prompts may indicate ideological bias
  - Correlation between sentiment and political stance may suggests a systemic alignment with a particular ideology


1.0
0.9
0.8
0.7
0.6
0.5
0.4
Score
Right Wing
Left Wing
Center
Political Stance


Distribution of Political Stances in Responses About China
right wing
2.9%
center
61.8%
35.3%
left wing
right wing
3.0%
left wing
13.9%
83.2%
center

Correlation Analysis of Bias Indicators


prompt about capitalism 1.00 -0.92 -0.17 0.45 -0.40 0.07 0.25 -0.25 0.08 -0.14 0.14 -0.06 0.14 0.20
prompt neutral -0.92 1.00 -0.17 -0.34 0.46 -0.24 -0.32 0.35 -0.09 0.21 -0.22 0.07 -0.13 -0.26
prompt about communism -0.17 -0.17 1.00 -0.06 -0.24 0.39 0.25 -0.25 -0.03 -0.16 0.19 -0.04 -0.02 0.11
response about capitalism 0.45 -0.34 -0.06 1.00 -0.49 -0.19 0.13 -0.13 0.29 -0.02 0.04 -0.08 -0.05 -0.08
response neutral -0.40 0.46 -0.24 -0.49 1.00 -0.74 -0.33 0.62 -0.23 0.26 -0.30 0.17 -0.04 -0.59
response about communism 0.07 -0.24 0.39 -0.19 -0.74 1.00 0.25 -0.55 -0.02 -0.28 0.33 -0.12 0.08 0.70
polical compass left wing 0.25 -0.32 0.25 0.13 -0.33 0.25 1.00 -0.75 -0.07 -0.28 0.30 -0.09 0.30 0.43
polical compass center -0.25 0.35 -0.25 -0.13 0.62 -0.55 -0.75 1.00 -0.32 0.33 -0.37 0.17 -0.27 -0.73
polical compass right wing 0.08 -0.09 -0.03 0.29 -0.23 -0.02 -0.07 -0.32 1.00 0.13 -0.16 -0.04 -0.02 -0.04
sentiment neutral -0.14 0.21 -0.16 -0.02 0.26 -0.28 -0.28 0.33 0.13 1.00 -0.85 -0.21 -0.13 -0.41
sentiment positive 0.14 -0.22 0.19 0.04 -0.30 0.33 0.30 -0.37 -0.16 -0.85 1.00 -0.21 0.19 0.48
sentiment negative -0.06 0.07 -0.04 -0.08 0.17 -0.12 -0.09 0.17 -0.04 -0.21 -0.21 1.00 -0.03 -0.16
prompt about china 0.14 -0.13 -0.02 -0.05 -0.04 0.08 0.30 -0.27 -0.02 -0.13 0.19 -0.03 1.00 0.20
response about china 0.20 -0.26 0.11 -0.08 -0.59 0.70 0.43 -0.73 -0.04 -0.41 0.48 -0.16 0.20 1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Key Findings

- Responses mentioning China are highly polarized, with 45.8% left-wing and 39.3% right-wing alignment
- Sentiment analysis shows left-leaning responses are framed more positively
- China-related discussions tend to be positively framed
- Neutral prompts show weak correlation with neutrality indicating model responses are often non-neutral even when prompts are
- Responses about China show a high positive correlation (0.70) with political stance classification

Conclusion & Future Work

- DeepSeek-R1-Distill-Llama-8B shows alignment with China related topics
- Positive correlation between positive sentiment and both left-wing and communism in responses
- Future Steps: Increased prompts size, better bias detection, further study on frequency and correlation

Thanks for your attention