Political bias in text generation of DeepSeek-R1-Distill-Llama-8B

Gianluca Cacciola

Computer Science Department, Università degli studi La Statale di Milano, Via Celoria 18, Milano, 20133, Italy, MI.

Abstract

The presence of political bias in large language models (LLMs) raises concerns about their objectivity and potential influence on public discourse. This study examines whether such bias exists in text generation in the model DeepSeek-R1-Distill-Llama-8B by analyzing responses from a set of political and non-political prompts. By assessing sentiment, stance, and ideological leanings in the generated texts, the research explores correlations that may indicate bias. The findings contribute to the broader discussion on fairness in AI, emphasizing the need for transparency and mitigation strategies in LLM development.

Keywords: Political bias, Large Language Models (LLMs), Sentiment analysis, Stance classification

1 Introduction

Large Language Models (LLMs) have become central to modern AI applications, including automated text generation. These models, trained on vast amounts of textual data, can produce fluent and contextually relevant responses. However, concerns have been raised about their potential political bias, which may influence their outputs in ways that favor specific ideological positions. This study investigates whether DeepSeek-R1-Distill-Llama-8B[1] exhibits measurable political bias in its generated text.

To assess this, a dataset of 100 prompts—both political and non-political—was submitted to the model, and the responses were analyzed using multiple Natural

Language Processing (NLP) techniques. The sentiment, stance, and ideological orientation of the generated text were quantified using specialized models. The core research question is: Does DeepSeek-R1-Distill-Llama-8B exhibit political bias in its text generation, and if so, can it be systematically measured through correlations between response attributes and political alignment scores?[2]

To answer this, correlations between different bias indicators—such as sentiment polarity, political stance classification, and entity recognition—were examined. By analyzing these relationships, the study aims to provide empirical evidence on whether the model generates systematically biased content.

2 Measuring Political Bias

To assess the presence of political bias in the text generated by DeepSeek-R1-Distill-Llama-8B, a structured pipeline was developed. The analysis consisted of multiple stages, including prompt dataset construction, response collection, and quantitative evaluation using state-of-the-art NLP models. The goal was to determine whether the model systematically favors or opposes certain political ideologies.

2.1 Dataset Construction and Response Collection

To systematically examine political bias in DeepSeek-R1-Distill-Llama-8B, a set of prompts was carefully designed to elicit responses on politically and economically relevant topics. Most of the questions address issues such as communism, capitalism, government policies, regulation, and political ideologies, allowing for an in-depth analysis of the model's position on contentious topics.

Prompts like ”What are the drawbacks of communism?” and ”Should more countries adopt capitalist policies?” aim to probe ideological viewpoints. In contrast, a subset of neutral prompts, such as ”What are the benefits of spending time in nature?” and ”How can you effectively manage stress?”, are included to investigate a possible tendency to answer in a political way to neutral questions. This dataset structure ensures a focused yet balanced evaluation of the model's behavior, helping to determine whether its outputs reflect systematic political bias.

The prompts were submitted to DeepSeek-R1-Distill-Llama-8B, and the generated responses were collected for analysis.

2.2 Sentiment Analysis

For this task,, the RoBERTa-based model(cardiffnlp / twitter-roberta-base-sentiment-latest) [3] was chosen since it is fine-tuned for sentiment classification and provides a robust method to determine whether responses express positive, negative, or neutral sentiment. This score is computed since political discourse often carries emotionally charged language, and this analysis helps quantify potential biases in how the model frames political topics.

Processing Steps: Each response was classified into three sentiment categories:

- **Positive** (score > 0.5)
- **Negative** (score < -0.5)
- **Neutral** (otherwise)

Use of Scores: The sentiment distribution across political and neutral prompts was analyzed to determine if the model consistently portrays certain ideologies in a more positive or negative light.

2.3 Named Entity Recognition (NER) for Entity Analysis

BERT-based Named Entity Recognition (NER) model (dslim/bert-base-NER)[4][5] is trained to identify political figures, countries, and organizations in text. Understanding which entities are frequently mentioned in responses can provide insights into potential biases in the model's focus and framing.

Processing Steps:

- Named entities were extracted from each response and categorized into **Person, Organization, Location, and Political Entity.**
- Frequency distributions of named entities were examined to identify trends, such as whether certain political figures or countries were disproportionately referenced.

Use of Scores: If a specific ideology, country, or political figure is consistently highlighted (or ignored), this could indicate implicit biases in the model's knowledge representation.

2.4 Zero-Shot Stance Classification

During the analysis, the challenge of determining whether a given text pertains to a specific political ideology emerged. Without the computational power and an appropriate dataset to train a traditional classifier, a pre-trained NLI models has been used as a ready-made zero-shot sequence classifier.

For this reason, the BART-based model (facebook/bart-large-mnli)[6] was chosen since it can determine the stance of a response toward a given statement without requiring task-specific training.

Processing Steps:

- Each pair of prompt and response was evaluated against those predefined statements:
  - *"This text is about Capitalism"*
  - *"This text is about Communism"*
- The model assigned a probability score to **Agree, Disagree, or Neutral** categories for each statement.

Use of Scores: Stance classification scores were aggregated and compared across different political prompts to identify patterns in the model's ideological leanings.

2.5 Left-Right Political Bias Classification

This BERT-based model(bucketresearch/politicalBiasBERT)[7][8] is specifically pre-trained to classify text along the left-right political spectrum, making it a another useful tool to identify ideological bias in text generation.

Processing Steps:

- Each response was classified as **left wing, center, or right wing** based on its alignment with known political discourse.
- The proportion of responses classified into each category was analyzed across all prompts.

Use of Scores: If responses to prompts about socialism tend to be classified as "left" while responses about capitalism are more neutral then "right", this would suggest a systematic ideological bias in text generation.

2.6 Correlation Analysis of Bias Indicators

To quantify the relationship between sentiment, stance classification, and ideological classification, correlation analyses were conducted:

- **Sentiment Scores vs. Political Ideology:** Do positive sentiments correlate with certain political stances?
- **NER Entity Frequency vs. Bias Classification:** Are certain political figures associated with ideological biases and sentiment scores in responses?
- **Zero-Shot Stance Classification vs. Political Bias Scores:** Does the stance probability align with the overall left-right classification of responses?

The score of several labels combinations has been used to build a correlation matrix with the purpose of identify a systematic bias towards certain political frames.

3 Results

3.1 Label Distribution and Influence of Prompt Selection on Bias Analysis

The study first examines the distribution of generated responses across various categories. The histogram visualization of stance and sentiment labels indicates a non-uniform distribution, with a skew favoring certain ideological perspectives. Notably, responses aligned with centrist or mildly progressive ideologies appear more frequently than explicitly conservative or radical viewpoints.

The observed bias can be attributed to the composition of the prompt dataset, which likely reflects any imbalance in the generated text.

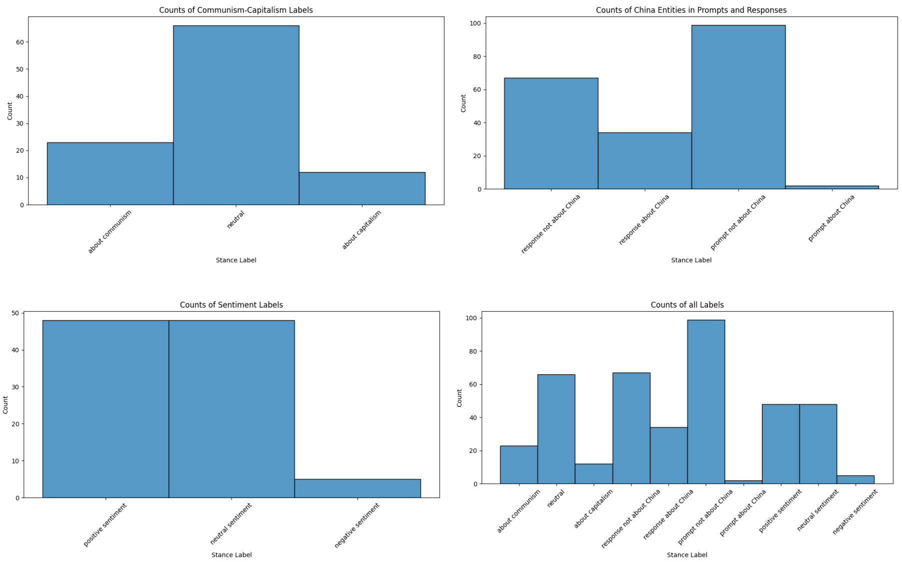

Fig. 1 counts of the occurences of labels

Comparing responses from political versus neutral prompts reveals that ideological biases are more pronounced when discussing governmental and economic systems. For instance, prompts such as ”What are the drawbacks of communism?” and ”Should more countries adopt capitalist policies?” tend to elicit responses with discernible stance preferences, while neutral questions such as ”What are the benefits of spending time in nature?” yield less polarizing content.

3.2 Analysis of Political and Sentiment Bias in Responses Related to China

An interest finding consists in the fact that a considerable portion of the responses mention the government of China or his ruling party. To better understand the meaning and significance of this phenomenon a further analysis was required.

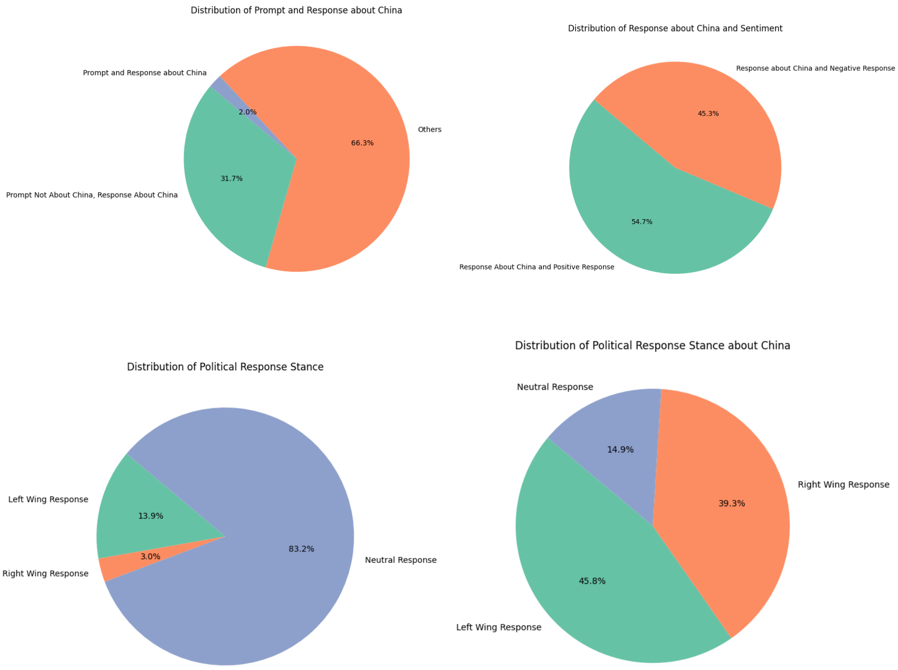

Fig. 2 distribution skews of labels occurency in China related responses

The first pie chart illustrates the frequency with which prompts and responses involve discussions about China. Notably, only 2.0% of the interactions explicitly mention China in both prompt and response, while a significant 31.7% of the responses bring up China despite the prompt not referring to it.

The last pie chart categorizes responses regarding China into left-wing, right-wing, and neutral positions. The data suggests a slight lean toward left-wing responses (45.8%), followed closely by right-wing responses (39.3%), while neutral responses make up 14.9%. The relatively high proportion of left-wing and right-wing stances indicates that responses about China tend to be politically charged rather than purely neutral.

A broader view of political stance across all responses (not just those about China) is displayed in the third pie chart. Interestingly, 83.2% of responses are neutral, while left-wing responses account for 13.9% and right-wing responses for only 3.0%. This stark contrast with the China-specific data suggests that discussions about China are significantly more polarized than other topics.

3.3 Correlation Analysis of Political and Sentiment Biases

To further assess the presence of political bias in text generation, we analyzed the correlations between various prompt and response labels, including political stance and sentiment. The correlation matrix (Fig. 3) highlights significant patterns in how the model generates responses across different political topics.

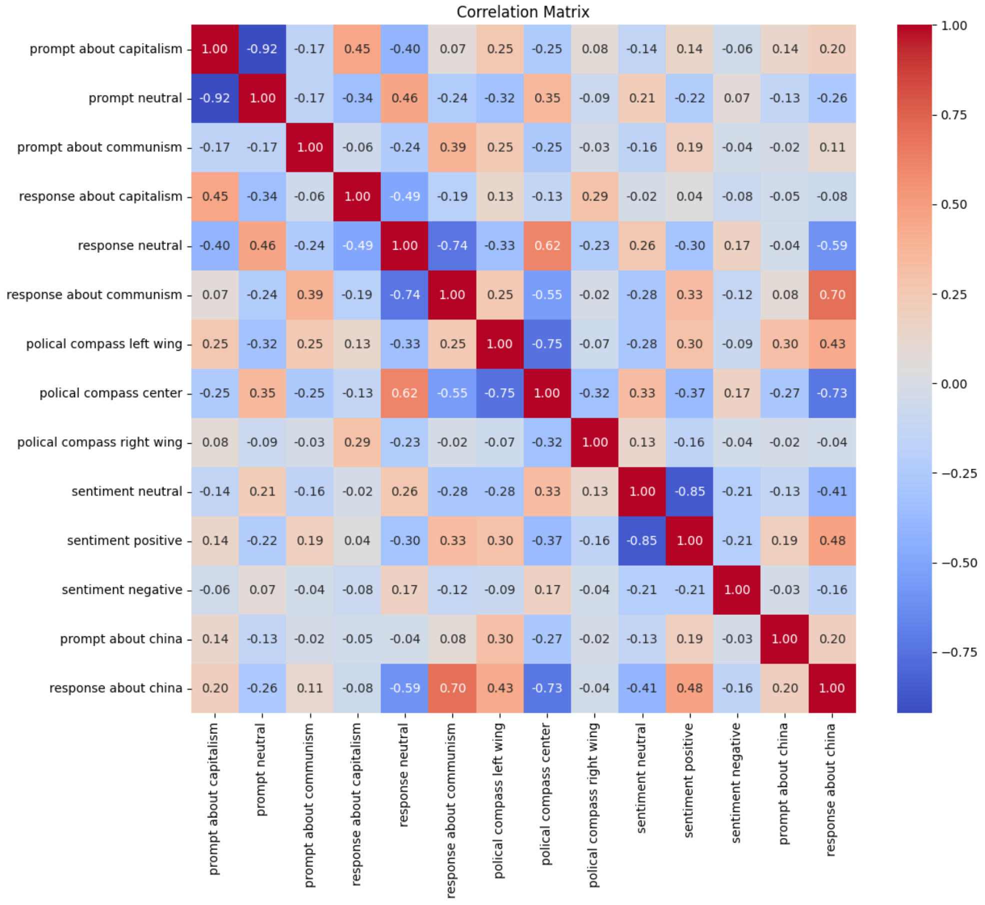

Fig. 3 correlatin matrix of all possible labels

3.3.1 Topic Polarization and Non-Neutrality

One of the most notable findings is the strong negative correlation between prompts about capitalism and neutral prompts (-0.92). This indicates that discussions on capitalism tend to be opinionated, rather than neutral. Similarly, responses discussing communism exhibit a strong negative correlation with neutrality (-0.74), suggesting that communism-related topics elicit stronger political stances.

Interestingly, responses about China exhibit a moderately strong negative correlation with neutrality (-0.59), indicating that responses tend to take a stance rather than remain neutral.

3.3.2 Political Leanings in Responses

The data indicates subtle tendencies toward political leanings in responses. Left-wing responses show a slight positive correlation with discussions of communism (0.25), while right-wing responses exhibit a weak correlation with prompts about capitalism (0.29). Although these correlations are not strong, the model's responses might reflect ideological biases when discussing economic systems. Additionally, neutral sentiment is negatively correlated with both left-wing and right-wing perspectives, while centrist perspectives correlate positively with neutrality (0.62). This suggests that politically charged responses—whether left-leaning or right-leaning—are less likely to be neutral.

Responses about China strongly correlate with left-wing political stance (0.70) and negatively correlate with centrism (-0.73), indicating that discussions on China tend to align with left-leaning narratives rather than centrist or balanced viewpoints.

These findings suggest that while the model does not overwhelmingly favor one ideology, it tends to frame China-related discussions from a left-leaning perspective.

3.3.3 Sentiment Bias and Political Stance

Sentiment analysis offers further insights into how political bias may manifest in AI-generated responses. The data suggests that left-leaning responses tend to be slightly more positive in tone, as indicated by a modest correlation of 0.30. This could point to a subtle tendency for left-wing narratives to be framed in a more favorable light compared to other perspectives.

Similarly, discussions about China show a moderate correlation with positive sentiment (0.48), suggesting that responses related to China are more likely to adopt a positive framing.

4 Conclusion

The study highlights the presence of political bias in the response of DeepSeek-R1-Distill-Llama-8 model, particularly in discussions on economic ideologies and geopolitics.

Bias becomes more pronounced in China-related discussions, showing heightened polarization, with the topic frequently emerging even when not explicitly prompted. Sentiment analysis further reveals that left-leaning responses tend to be more positive, while discussions on China are moderately correlated with positive sentiment.

Bias in stance distribution, topic selection, and sentiment framing raises concerns about AI's impact on political discourse. Future research should focus on improving bias detection though a better prompt selection and mitigating these biases through dataset balancing, reinforcement learning, and improved transparency.

References

[1] DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W.L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X.Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y.X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., Zhang, Z.: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning (2025). https://arxiv.org/abs/2501.12948

[2] Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., Slonim, N.: Stance classification of context-dependent claims. In: Lapata, M., Blunsom, P., Koller, A. (eds.) Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 251–261. Association for Computational Linguistics, Valencia, Spain (2017). https://aclanthology.org/E17-1024/

[3] Rosenthal, S., Farra, N., Nakov, P.: Semeval-2017 task 4: Sentiment analysis in twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 502–518 (2017)

[4] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018) arXiv:1810.04805

[5] Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pp. 142–147

(2003). https://www.aclweb.org/anthology/W03-0419

[6] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. CoRR **abs/1910.13461** (2019) 1910.13461

[7] Baly, R., Da San Martino, G., Glass, J., Nakov, P.: We can detect your bias: Predicting the political ideology of news articles. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). EMNLP '20 (2020)

[8] Bucket Research