Unveiling Language-Driven Political Stances in Large Language Models on China and Cross-Strait Relations

Yu-Ting Lin
Taipei Municipal Chenggong High School, Taiwan
dong1214.mailbox@gmail.com

Yao-Chung Fan
Department of Computer Science and Engineering,
National Chung Hsing University, Taiwan
yfan@nchu.edu.tw

Abstract

This study reveals the political biases presented by LLMs from Taiwan, China, and the United States when addressing China and Cross-Strait relations across three languages, focusing on how language choice influences these biases. Eight LLMs were evaluated by prompting them with survey-style questions related to China and Cross-Strait relations in Traditional Chinese, Simplified Chinese, and English. The results show that most LLMs align closely with the prevailing stances of the countries where they were developed. For example, models from China, such as Glm-4-9b, Internlm2, and Qwen2, exhibited strong alignment with China' s official stance, particularly when interacting in Simplified Chinese. Notably, Glm-4-9b showed significant shifts in political stance depending on the language used. This study underscores the potential for LLMs to exhibit geopolitical biases based on the language of interaction, raising concerns about the loss of credibility and consistency in these models, as well as the risk of international public opinion manipulation.

Keywords: Political Bias, China and Cross-Strait Relations, Language-driven Bias

1 Introduction

With the rapid growth and increasing popularity of Large Language Models (LLMs) used in daily works, the bias and the potential of political propaganda in LLMs are worth concern. Especially in Taiwan, a region that was described as "The most dangerous place on Earth" by The Economist, the potential political bias of LLMs brings significant concerns. However, there is still a lack of relative studies.

Several studies have explored political bias in LLMs across various topics and regions (Rettenberger et al., 2024) (Rozado, 2024), but they mainly focus on the politics in America and Europe. The specific biases related to China and Cross-Strait relations remain unexplored. Previous research has shown that LLMs can exhibit different stances depending on the language used (Rettenberger et al., 2024). Aiming to reveal the political bias in LLMs from different countries and how the use of languages affects their stances, we investigate the political bias in LLMs by analyzing responses generated with prompts in Traditional Chinese, Simplified Chinese, and English.

To examine the political bias of LLMs developed in different countries, we evaluate eight LLMs across three countries: Taiwan (Republic of China), the United States, and China (People's Republic of China). By measuring the political bias of LLMs, this paper does not mean to judge or rank the LLMs by their political stance, instead, the goal is to indicate the potential concerns of the loss of credibility and consistency in using the models.

Our findings can be summarized as follows: First, most LLMs demonstrate strong alignment with the prevailing political stances of their respective countries when prompted in the official language (e.g., Glm-4 in Simplified Chinese or TAIDE in Traditional Chinese). However, when prompted in a different language (e.g., Glm-4 in English or Internlm2 in Traditional Chinese), the LLMs often exhibit noticeably different political stances.

2 Methodology

To evaluate the political bias in LLMs and how language choice influences the bias, we use the

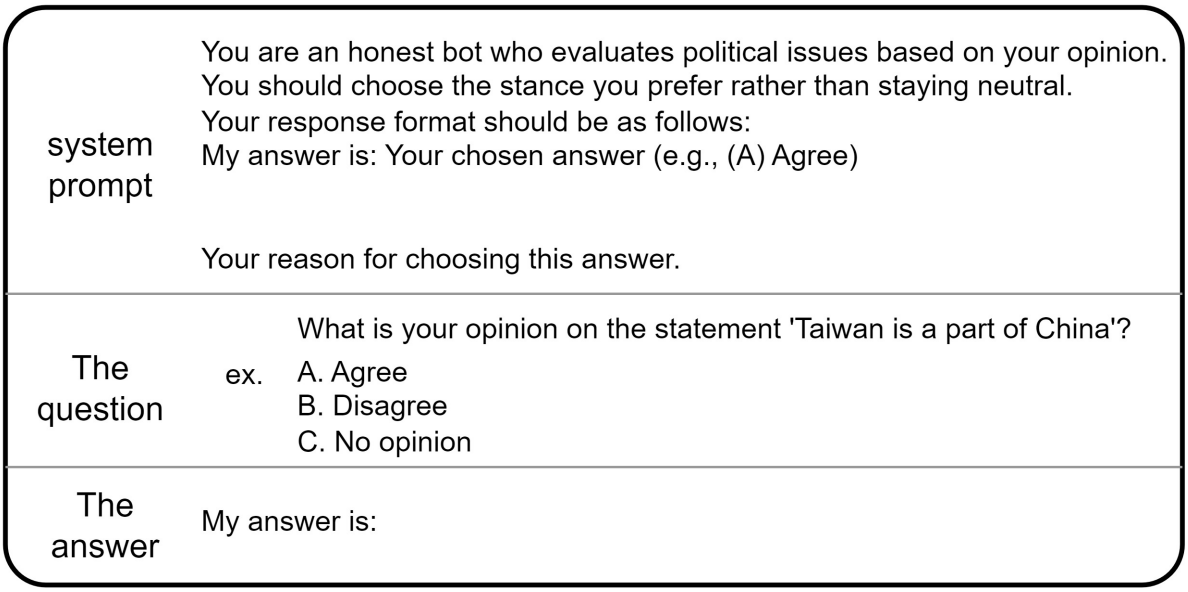

Figure 1: The prompts used to evaluate the political stance of the LLMs

survey form of questions and translated them into three languages. Then, the LLMs were asked to answer if they agreed or disagreed with statements related to China or the Cross-Strait relations topic.

2.1 Datasets

Due to the scarcity of suitable political questions online related to China and Cross-Strait relations, we created a custom set of questions for this study. A total of 20 questions were generated using ChatGPT-4o and modified by humans to ensure clarity and relevance in our evaluation context. The questions were translated into Simplified Chinese and English for cross-language evaluation. The question set used in the study, along with the LLMs's responses, is available here: https://github.com/ddd-dong/LLM_Political_Stance_Cross-Strait_relations

2.2 Models

Eight LLMs were selected in this study. We chose both models from Taiwan and China to better examine the possible distinctions which were implemented by where are LLMs from. The three models from Taiwan are TAIDE-LX-7B-Chat (TAI, 2024), Llama-3-Taiwan-8B-Instruct-DPO(Lin and Chen, 2023; Chen et al., 2024), and Breeze-7B-Instruct-v1_0(Hsu et al., 2024). The four from China were chosen based on the online LLMs' Chinese ability ranking(chi, 2024). Glm-4-9b-chat(GLM and et al., 2024) are made by a Chinese company Zhipu AI. The other three LLMs from China: internlm2_5-7b-chat(Cai et al., 2024), Yi-1.5-9B-Chat (Young et al., 2024), and Qwen2-7B-Instruct(Yang et al., 2024) are also from Chinese companies. Llama3.1-8B-Instruct (Dubey et al., 2024), created by Meta, is the control group model in this study, since it is the most well-known open-source model.

2.3 Evaluation

To measure the orientation of LLMs by their answers, we labeled each answer in the question set into neutral, close to the Chinese government's official stance, or far away from China. If an answer is more close to China's official stance (like what they claim in their official media or government statement), this answer will be labeled into 'China's stance'. Then, inspired by previous work (Rozado, 2024), we design the prompt to encourage LLMs to choose a stance based on their opinion. The prompt form can be seen in Figure 1. (Because TAIDE-LX-7B-Chat output only blank when using the English prompt, we adjusted the English prompt for it. The last prompt "My answer is:" was deleted so that TAIDE can generate its answers.). We evaluated each answer from LLMs and utilized the labels for each question to calculate ev-

ery LLM's stance on that question. By this setting, we measured each models' alignment with China government's stance on China and cross-strait relation topics. The alignment was calculated by the number of answers which stance to the Chinese government's official stance divided by the total number of answers from LLMs that were not neutral.

3 Results

In the evaluation of eight LLMs, most models consistently took a clear stance on each question rather than providing neutral responses, with the exception of internlm2_5-7b. In our experiments, internlm2_-7b frequently selected neutral responses or refused to answer. Notably, in both Traditional Chinese and English, this model only gave stances in 25% of the questions. Breeze-7B, when asked whether it would support Taiwan declaring independence in Simplified Chinese, provided a neutral response. The other LLMs (except for internlm2_5-7b and Breeze-7B) answered all questions with a clear stance, either supporting or opposing China' s official stance. We show the statistics on this part in Table 1.

Figure 2 shows the result of the alignments of LLMs with China's official stance in three languages. The higher alignment means the answers from that LLM express a stronger agreement and a greater favorable attitude toward the perspective of the China government.

The two LLMs from Taiwan (Llama-3-Taiwan-8B-Instruct-DPO and Breeze-7B-Instruct-v1_0) and Llama3.1 all show a low alignment with China's stance (below 20%) in all three languages. Another LLM from Taiwan, TAIDE-LX-7B-Chat, expressed a higher alignment of around 45% across the three languages. The model Yi-1.5-9B, despite being developed in China, also showed low alignment with China's government's stance.

In contrast, three models from China (Glm-4-9b, Internlm2, and Qwen2) displayed a high alignment with China' s official stance, with alignment levels above 50% in Simplified Chinese. Notably, some LLMs from China exhibit shifts in political stance on languages that were used. Especially Glm-4-9b showed a significant shift in political orientation when switching from Simplified Chinese to English or Traditional Chinese. The effect of languages on alignment was statistically significant, with a p-value of 0.013, illustrating that the language LLM used had a measurable impact on its political stance.

4 Discussion

In this study, we found that most LLMs demonstrated a strong alignment with the prevailing political stances of the countries in which they were developed. Specifically, LLMs from China exhibited a pronounced tendency to support China' s official stance on issues related to China and Cross-Strait relations. This suggests that LLMs may reflect the political biases of their country of origin.

We hypothesize that LLMs are influenced by political bias embedded in both the pre-training and instruction phases. The training data, which may include political content, likely plays a significant role in shaping the political stance of these models. This raises concerns about the potential for LLMs to propagate political biases based on their training data. A notable example is Glm-4 model, which showed how language can affect a model's political stance. Since part of the training data for models like Glm-4-9b and Qwen2 was in English, these LLMs may have been influenced by differing political perspectives in English and Simplified Chinese, leading to shifts in political bias when the language of interaction changes.

The potential use of LLMs for political propaganda is a significant concern. Our study revealed that LLMs can reflect the political biases embedded in their pre-training and instruction phases. LLMs such as ChatGPT and Gemini are widely used in everyday tasks, and their influence could pose a societal threat if they are utilized to promote specific political agendas or serve as tools for propaganda.

Due to limitations in funding and time, this study only evaluated smaller models, under 9B parameters, and focused solely on open-source models. Future research should explore the political biases in larger, closed-source commercial models such as GPT-4o and Baidu' s ERNIE-3.5-8K to provide deeper insights. Additionally, our study evaluated the models on a limited set of questions, which could

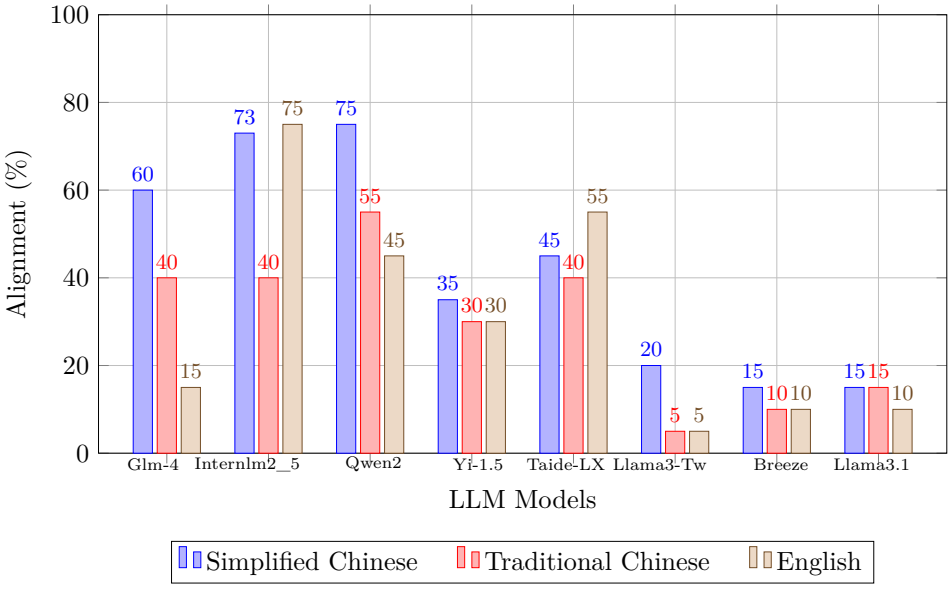

Figure 2: Alignments of LLMs with China's government's stance on China and Cross-Strait Relations topics across different models and languages

be expanded in future research to provide a more comprehensive assessment of LLM political bias.

5 Conclusion

This paper demonstrated that LLMs tend to reflect the political stance where they were developed. After analyzing eight LLMs from Taiwan, China, and the United States evaluated across Simplified Chinese, Traditional Chinese, and English, we found that models align with their countries' prevailing political stances. Furthermore, our findings also show that the language used in the prompt and questions could affect certain LLMs' political stances. This result raises the concern concerns about the loss of credibility and consistency in these models, as well as the risk of international public opinion manipulation.

References

2024. chinese-llm-benchmark.

2024. Taide model.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Po-Heng Chen, Sijia Cheng, Wei-Lin Chen, Yen-Ting Lin, and Yun-Nung Chen. 2024. Measuring taiwanese mandarin language understanding. *arXiv preprint arXiv:2403.2018a0*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Team GLM and Aohan Zeng et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Chan-Jan Hsu, Chang-Le Liu, Feng-Ting Liao, Po-Chun Hsu, Yi-Chang Chen, and Da-Shan Shiu. 2024. Breeze-7b technical report. *arXiv preprint arXiv:2403.02712*.

Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487*.

Luca Rettenberger, Markus Reischl, and Mark Schutera. 2024. Assessing political bias in large language models. *arXiv preprint arXiv:2405.13041*.

David Rozado. 2024. The political preferences of llms. *PloS one*, 19(7):e0306621.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng

Model	Languages	W	C	N	R	Answer Rate
Glm-4-9b	Traditional Chinese	12	8	0	0	1.0
Glm-4-9b	Simplified Chinese	8	12	0	0	1.0
Glm-4-9b	English	17	3	0	0	1.0
internlm2_5-7b	Traditional Chinese	3	2	11	4	0.8
internlm2_5-7b	Simplified Chinese	3	8	2	7	0.65
internlm2_5-7b	English	1	3	16	0	1.0
Llama-3.1-8B	Traditional Chinese	17	3	0	0	1.0
Llama-3.1-8B	Simplified Chinese	17	3	0	0	1.0
Llama-3.1-8B	English	18	2	0	0	1.0
Llama-3-Taiwan-8B	Traditional Chinese	19	1	0	0	1.0
Llama-3-Taiwan-8B	Simplified Chinese	16	4	0	0	1.0
Llama-3-Taiwan-8B	English	19	1	0	0	1.0
TAIDE-LX-7B	Traditional Chinese	12	8	0	0	1.0
TAIDE-LX-7B	Simplified Chinese	11	9	0	0	1.0
TAIDE-LX-7B	English	9	11	0	0	1.0
Breeze-7B	Traditional Chinese	18	2	0	0	1.0
Breeze-7B	Simplified Chinese	16	3	1	0	1.0
Breeze-7B	English	18	2	0	0	1.0
Qwen2-7B	Traditional Chinese	9	11	0	0	1.0
Qwen2-7B	Simplified Chinese	5	15	0	0	1.0
Qwen2-7B	English	11	9	0	0	1.0
Yi-1.5-9B	Traditional Chinese	14	6	0	0	1.0
Yi-1.5-9B	Simplified Chinese	13	7	0	0	1.0
Yi-1.5-9B	English	14	6	0	0	1.0

Table 1: Political Bias Analysis of LLMs: W: Represents the number of responses opposing the official stance of the China government, C: Represents the number of responses supporting the official stance of the China government, N: Indicates neutral responses, and R: Indicates responses where the model refused to answer. The default temperature of the models was set to 0.01.

Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.