Language-Specific Political Bias in Large Language Models

Dakota Barnes and Nikolas Belle

1 Introduction

LLMs have become an integral part of many peoples' lives transforming how we communicate, access information, and make decisions. [1] With models becoming increasingly accessible through various platforms, like Anthropic's Claude, OpenAI's ChatGPT, and Cohere's multilingual Aya, which supports over 100 languages [2], users around the globe are becoming reliant on these AI tools across diverse linguistic contexts. As dependence on LLMs keeps growing, it is essential that we consider the implications of their inherent biases. Unrecognized biases have the ability to subtly guide our perceptions, opinions, and decisions, potentially exacerbating societal division if these biases differ significantly across different models and linguistic contexts. Political bias holds particular significance as it has the power to alter narratives around critical societal issues, like gender-inequality, body politics, and human rights. Despite extensive research into model and language-specific biases separately, existing studies have yet to deeply explore how these biases interact across both model architectures and languages simultaneously. Our paper addresses this critical gap by systematically evaluating political biases across multilingual LLMs, revealing that language and model architecture must be jointly considered to predict, understand, and mitigate these biases accurately. We present a framework enabling structured and generative bias evaluations across languages and models and demonstrate through our findings on political evaluations how LLM bias can change unpredictably when using this combined perspective.

2 Background and Motivation

As our reliance on LLMs continues to grow across a multitude of domains, the impact of dangers presented by inherent model bias scales as well. These biases can subtly influence individual and collective perceptions, decisions, and actions, potentially intensifying societal division if they differ significantly across various models or linguistic contexts.

2.1 Related Works

Previous research has provided foundational insights into biases exhibited by LLMs. Blodgett et al. (2020) conducted a critical survey to showcase the impact unchecked biases in language technology can have on society as they propitiate systemic inequalities [3].

Given that LLMs are being deployed across diverse linguistic contexts, it is imperative that we look at these biases across multilingual settings as well. XU et al. (2024) examined the influence of training corpora and alignment strategies on inherent bias through a comprehensive survey on Multilingual LLMs (MLLMs) [4]. They highlight that due to the high percentage of the Western language data in the training corpus of MLLMs, they typically reflect Western ideologies. Focusing on more specific types of biases, Aksoy et al. (2024) explored cultural biases in MLLMs, finding that moral foundations are represented differently across languages, meaning models may reflect cultural-specific moral biases [5]. Additionally, Zhu et al. (2024) studied the performance of LLMs developed in different countries, finding that language and cultural contexts have a strong influence on model behavior, leading to varying biases [6].

Political bias in LLMs is particularly dangerous as it can push forward conflicting ideologies and influence public stances on topics that are actively dividing nations. Yang et al. (2025) surveyed political biases across LLMs when being prompted with topics classified by different levels of polarization. They found that models display significantly more bias in their responses to topics that are highly polarized. Feng et al. (2023) administered the Political Compass Test to various LLMs and found that different models exhibit varying political leanings [7]. Rozado (2024) expanded on this work, testing LLMs with multiple political orientation tests, discovering that models tend to generate responses that align with left-of-center viewpoints [8]. Lunardi et al. (2024) challenged the efficacy of direct questioning techniques, like the Political Compass Test, highlighting that the lack of consistent responses make for a poor evaluation of an LLM's true political stance [9]. Röttger et al. (2024) similarly challenged the validity of multiple choice tests and suggested ways to overcome this inconsistency, such as requiring the LLMs to provide reasoning for its answer [10]. Although direct questioning is advantageous in some situations, generative analysis, such as directing the LLM to generate news articles, has shown promising results [11] [12].

While previous studies have explored political biases across LLMs offered from different providers, comprehensive analyses across multiple languages and models re-

main limited. Rettenberger et al. (2024) used the German Wahl-O-Mat to assess political biases in LLMs through German and English prompting and found that larger models typically aligned with more left-leaning political parties, while smaller models typically remained neutral, particularly in English [13]. Additionally, Zhou et al. (2024) analyzed inconsistencies and biases in GPT models when prompted in English and Chinese, shining light on the impact language has on political bias in LLMs [14].

2.2 Main Contributions

Despite these valuable insights, there is a lack of broad surveys analyzing political biases across multiple models and languages, specifically focusing on how language-specific biases vary between models. We aim to fill this gap by providing:

- A open-source framework for analyzing bias in multilingual LLMs across both structured and generative contexts.
- A cross-model and cross-language survey of political bias in multilingual LLMs using this framework.
- Empirical findings on stance flipping and varying political bias across models and languages.

3 Experimental Setup

Our framework is designed to systematically evaluate and quantify biases in MLLMs using both structured (multiple-choice) and generative (open-ended) contexts. More specifically, we aim to give researchers the ability to perform these structured and generative evaluations on any model with any language (assuming the language is supported by the model). Our experimental setup uses this framework to gather data and perform analysis on political bias across models and languages, which we will lay out below.

3.1 Models

The models that we choose for our experiments are Mistral-Large-Latest (v24.11), Mistral-Small-Latest (v25.01), Gpt-3.5-turbo (0125), DeepSeek-chat (v3_24/12/26). We choose these models because of their diverse countries of origin, namely France, the United States, and China. The models are accessed using each company's API.

3.2 Languages

The languages that we used in our evaluations include English, Chinese, French, German, Italian, Japanese, Korean, and Spanish. For the structured tests we utilized the cheaper Mistral Small to translate the large number of questions and answers. Because the generative evaluation required less translations, we utilized the Google Translate API to translate the LLM prompts.

3.3 Structured Evaluation

The structured approach to model evaluation involves administering a multiple choice test to an Oracle LLM in hopes of illuminating some broader ideology that couldn't be extracted from the LLM by explicitly prompting for it. The multiple choice tests were run on a MacBook Pro Apple M1 Max 32 GB memory, which took an average of 30 minutes per test.

3.3.1 The Political Compass Test

We chose to use the Political Compass Test for our structured political bias evaluation due to its popularity, length, and simplicity to visualize. The test consists of 62 statements and the test-taker must select from the options "Strongly Agree", "Agree", "Disagree", and "Strongly Disagree" for each statement. After taking the test, an economic and social score is calculated. The economic score ranges from left to right, while the social score ranges from authoritarian (top) to libertarian (bottom) on the political compass graph.

3.3.2 Prompting

Despite the advantage explainability when evaluating an LLM's political standing through a series of multiple choice questions, Röttger et al. pointed out some of the difficulties with eliciting responses that can be trusted as the model's true stance [10]. The most prominent issue noted was the lack of consistency in a model's response when prompted with only its answer choice. To overcome this issue we prompt the LLM to always include reasoning that explains its choice. While this significantly improved consistency, there were some cases where the model provided reasoning that contradicted its answer choice. This issue required a slightly more complex solution and some judgement calls, as we can't confirm that either the reasoning or the answer is the model's true stance. We decided on the integration of an Evaluator LLM for the most natural approach to extracting the model's true answer, which will be explained in greater detail in the following section.

3.3.3 Question Answering Framework

The pipeline for retrieving an answer from the Oracle LLM (Figure 1), our test-taker, begins by fetching a question from the question bank. This question bank stores the questions and answer options for a single test and is filled in by the user, based on our provided structure. The fetched question, along with a general prompt for how the oracle should respond, is given to the Translator, which can be connected to the Google Translate API or any LLM. We decided to use the Mistral Small model for our structured evaluation experiments due to its minimal cost. This translated prompt, which includes the question and answer choices, is then given to the Oracle LLM. To ensure that there is no unwanted English-specific bias

impacting the model's response, we made sure that no previous context is passed in with the query and that the entire prompt was in the desired language. Next, we fetch the answer and reasoning from the JSON returned by the Oracle LLM and call the Translator to translate them back into English. The original question, answer, and reasoning (all in English) are then passed to the Evaluator LLM, which is prompted to determine if the reasoning is consistent with the answer with respect to the question. It responds with a Boolean value representing if they are consistent, along with an explanation for its answer in JSON format. If they are consistent, the answer is added to a results JSON file. If not, the explanation from the Evaluator LLM is translated back into the respective language, appended to the Orcle LLM prompt as feedback, and the Oracle prompting loop is repeated for a number of times defined by the user. We decided to set this maximum repeats value to 5, to conserve budget and also because we recognized the model staying firm on its believe when it doesn't fix the inconsistency after the first set of feedback. In this rare case, we chose to use the answer as the LLM's true stance, as the feedback as the reasoning did not influence the model to change it. While this framework could be optimized by keeping the Oracle LLMs output in its respective language for the Evaluator LLM, we chose to prompt the Evaluator LLM in English so that we can understand how it makes decisions. Finally, we run each model-language combination for five trials and take the most consistent answer for each question and append it to an average results JSON file. This provided further assurance that we were retrieving the model's true position on each question.

3.4 Generative Evaluation

The generative evaluation involved prompting models to create politically charged news headlines in multiple languages. These headlines were embedded using a Sentence Transformer to quantify stance and reveal each model's implicit bias. The topics we tested, along with their respective stances are in Table 2. The full generative test was run on a MacBook Pro I9 Intel Core 16GB memory, which took a total estimate of 8 hours.

3.4.1 Prompting

The prompt for querying the model followed the format, *"Write {number} {stance}news headlines about the topic of '{topic}'."* The {number} specifies the number of headlines generated per query, the {topic} is replaced from the list of political topics, and the {stance} is replaced with an item from the proponent (pro) column, opponent (opp) column, or nothing to represent the test stance. For our results, we prompted each model, on each language, on each topic 5 times with the pro/opp prompt, and 3 times for the test prompt. The models occasionally failed to generate articles that could be seperated and parsed, most noticeably in Japanese and Chinese with the model gpt-3.5-turbo.

3.4.2 Sentence Transformer

The generated headlines were embedded using the *paraphrase-multilingual-MiniLM-L12-v2* Sentence Transformer [15]. The multilingual model enabled us to encode article text across multiple languages into consistent numerical vectors. With the number of prompts described above, we embedded 50 pro/opp articles, and 30 test articles per model per language per topic.

3.4.3 Stance Score

To quantify the stance of a generated news headline, we first embed each article into a numerical vector using a Sentence Transformer. We then compute **anchor vectors** for the proponent (pro) and opponent (opp) articles by averaging their embeddings:

$$A_p = \frac{1}{N_p}\sum_{i=1}^{N_p} E(T_p^i), \quad A_o = \frac{1}{N_o}\sum_{i=1}^{N_o} E(T_o^i) \tag{1}$$

- A_p is the pro stance anchor, computed as the mean embedding across pro articles.
- A_o is the opp stance anchor, computed as the mean embedding across opp articles.
- N_p and N_o are the number of pro and opp articles, respectively.
- $E(T)$ represents the embedding of article T using the Sentence Transformer.

Next, we compute the cosine similarity between each test article embedding $E(T_s)$ and both the pro and opp anchors:

$$C_p(T_s) = \cos(E(T_s), A_p), \quad C_o(T_s) = \cos(E(T_s), A_o) \tag{2}$$

- $C_p(T_s)$ is the cosine similarity between the test article T_s and the pro anchor.
- $C_o(T_s)$ is the cosine similarity between T_s and the opp anchor.

Finally, the **stance score** for each test article is given by:

$$S(T_s) = C_p(T_s) - C_o(T_s) \tag{3}$$

- A **positive stance score** ($S(T_s) > 0$) indicates that the article is closer to the pro stance.
- A **negative stance score** ($S(T_s) < 0$) indicates that the article is closer to the opp stance.

To evaluate stance bias across different conditions, we aggregate stance scores across different groups:

- **Model-Specific Stance Score** (S_{model}): Averaged over all test articles for a given model, language, and topic.
- **Language-Specific Stance Score** (S_{lang}): Averaged over all test articles for a given language and topic, across all models.
- **Topic-Specific Stance Score** (S_{topic}): Averaged over all test articles for a given topic, across all models and languages.

These aggregated stance scores allow us to compare political bias at different levels, helping identify trends across models, languages, and topics. This paper proposes the new Topic-Specific Stance Score bias metric, which offers a wholistic approach that takes into account sentence sentiment across languages to measure Political bias.

3.5 Usage

We aim to make it as easy as possible to use this framework for surveying any LLM from any provider in any language. To use either evaluation tool, the user can add their LLM provider as a class to the llms.py file if it isn't already a supported one (Mistral, OpenAI, DeepSeek, and OpenRouter). After this, they can define their models, languages, test questions or topics, and the number of trials for their evaluation. The structured evaluation allows for the optional configuration of language-specific prompts (a useful technique for ensuring each language's prompt will elicit the expected JSON structure from the Oracle LLM when using an older model).

4 Results

4.1 Structured Evaluation

4.1.1 General Political Compass Results

Across the evaluated models we observed distinct political positions that varied notably between models and languages. The results shown in Figure 2 and Figure 3 confirm that political biases not only differ by the model's underlying architecture but are significantly influenced by the language in which the prompts are given. We can see gpt-3.5-turbo prompted in Korean is the most authoritarian and economically right model-language combination. Mistral-small-latest prompted in German is the most libertarian model. Finally, gpt-3.5-turbo prompted in Italian is the most economically left model.

4.1.2 Stance Flips

One of our key findings was the phenomenon of stance flipping, where the political standings of a model tested in two different languages are inverted from another model's political standings for the same two languages. We see this occur between deepseek-chat and gpt-3.5-turbo for the languages Korean and Spanish and Korean and English, displayed in Figure 4. We can observe this stance flipping at the question level as well with mistral-small-latest and gpt-3.5-turbo flipping stances across Spanish and Italian for the statement: "It is important that my child's school instills religious values" (Table 1).

4.1.3 Variance Across Models and Languages

Inter-language agreement measures how consistently a model maintains its political stance across different languages by measuring variance. We can see in Table 5 that deepseek-chat demonstrates the lowest variance for its social and economic scores, while gpt-3.5-turbo displays the highest variance for both scores. Inter-model agreement, shown in Table 4, measures how consistent the political standing is for a language across multiple models. Interestingly, Korean and Japanese both display significantly higher variances in their social and economic scores across models compared to other languages. German showcases low variance in its economic score and significantly higher variance in its social score across languages.

4.2 Generative Evaluation

The full results across models, languages, and topics are shown in Figure 9 in the Appendix. The heatmap displays the stance score(x100) with the model and language on the left, and the topic on the bottom. The positive stance scores in blue represent a proponent to the topic, and the negative stane score in red represent an opponent to the topic. The highest scores for the S_{lang} 9b include English-gpt-3.5 turbo as strong nationalistic (18.0 Pro) on the topic National Loyalty and Patriotism, and Japanese-mistral-large-latest as Pro-Life (-15.4 Opp) on the topic of Reproductive rights.

4.2.1 Embedding Representation

We can dive into a representation of the article embeddings from the Sentence Transformer using T-SNE plots. Figure 8 presents a grid with multiple T-SNE plots, illustrating how embeddings look individually for each model and language combination on a topic. The rightmost column and bottom row display averages, while the bottom-right section aggregates all the embeddings. Figure 6 provides an enlarged map of the aggregation of the articles across different languages and models for 'Immigration', offering insight to their relative positioning in the embedding space.

4.2.2 Stance Flips

The generative results also produced examples of stance flipping across languages and models. For example, all models in Italian favored Digital Privacy, while the models in Japanese favored National Security. Additionally, Chinese gpt-3.5-turbo is more pro-immigration than

deepseek-chat, and English gpt-3.5-turbo is more anti-immigration than deepseek-chat.

4.2.3 Stance Scores

To demonstrate the differences in S_{model} 9a, S_{lang} 9b, and S_{topic} 9c, the full stance heatmaps for each calculation are presented in side by side comparison in Figure 9. Note that when going from the S_{model} to the S_{topic}, the stance scores smooth out, while generally keeping a consistent theme. Interestingly, the very strong Japanese Opp Stance with the S_{model} and S_{lang} on Reproductive Rights topic turns to slightly Pro when evaluated with S_{topic}. This could be due to the Japanese 'Reproductive Right' Opp embeddings being closer to the rest of the other languages' Pro embeddings.

4.2.4 Variance Across Models and Languages

To compare each model, Figure 7a and Figure 7b plot the average and variance of the **models across languages**. Additionally, the average and variance of the **languages across models** is in Figure 7c and Figure 7d respectively. To show similarity to the structured responses, the average of the variance across the topics for each score is placed in Table 7 and Table 6. Notable, there is lower variance in deepseek-chat and mistral-large-latest, as well as in Italian and French.

5 Discussions

The results from both methods converged on several patterns, notably stance flips, differing variances across models for a given language, and the presence of geographic trends between languages. Several factors could contribute to stance flipping and variance in political bias. Newer models like deepseek-chat may have undergone more targeted bias reduction efforts than older models like gpt-3.5-turbo, which may lead to lower inter-language variance. Additionally, MLLMs are trained on diverse datasets, where sources from certain languages may be more politically skewed than sources from others. A language with greater representation across all of the models is also more likely to have less variance between those models in its stances. The presence of stance flips suggests that the training data composition differs significantly between different LLM developers, leading to non-uniform ideological biases across languages. Finally, both evaluations uncovered trends of model stances clustering together for geographically similar languages. For example, Spanish, English, French, German, and Italian display negative sentiment towards immigration, while Chinese, Japanese, and Korean showcase positive sentiment towards the topic. This may be influenced by differences in political discourse among regions and the ways in which political topics are framed in different cultures.

6 Limitations

This study faced some limitations due to the short time frame, resource availability, and the general ambiguity of bias in black-box models like LLMs. Due to budget constraints, we restricted our evaluations to the cheapest models available, which were often smaller or older. Therefore, they may not incorporate recent advancements made in bias mitigation. Additionally, our analyses covered only eight languages and models from only three providers: OpenAI, DeepSeek, and Mistral. DeepSeek's model was released more recently than the OpenAI and Mistral models we used, potentially influencing its comparatively lower variance across languages. When creating multilingual prompts, reliance on Google Translate and Mistral may have introduced unintended semantic biases. Furthermore, the Political Compass Test primarily represents Western-Centric ideologies, which might have difficulty translating to languages with vastly different cultures. Additionally, when consistent reasoning and response pairs could not be produced for a question in the structured evaluation, it is unclear which of the two represents the model's true stance. Next, our generative tests relied heavily on a single sentence transformer model, which could encode its own semantic bias, and group articles by semantic similarity rather than political alignment. Additionally, stance anchors were computed simply by averaging stance embeddings without a true human-validated baseline, making stance interpretations relative rather than absolute. Lastly, cosine similarity may oversimplify the comparison of complex ideological characteristics between the anchors and test articles, potentially impacting the accuracy of our findings.

7 Future Work

Future research should build upon this study by utilizing larger, more advanced reasoning models. Gathering additional data points for the structured and generative tests would further enhance statistical robustness and allow for more precise bias quantification. Examining the effect of prompt perturbations, including subtle linguistic changes, or the use of special English characters may reveal important sensitivities that influence model behavior. Future studies could expand to alternative political tests, including ones from non-Western sources, enabling a more inclusive bias assessment. The structured and generative framework developed here can be extended to other biases, such as cultural, moral, or gender biases, deepening our understanding of multilingual nuances across LLMs. Finally, exploring how debiasing techniques perform on various models in different languages may provide insights into how we can better mitigate LLM bias in multilingual contexts.

References

[1] M. U. Hadi, q. a. tashi, R. Qureshi, A. Shah, A. Muneer, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili, "A survey on large language models: applications, challenges, limitations, and practical usage," *TechRxiv*, 2023. (Cited on page 1.)

[2] C. F. A. Team, "C4ai launches aya, an llm covering more than 100 languages," February 13 2024. [Online]. Available: https://cohere.com/blog/aya (Cited on page 1.)

[3] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of "bias" in NLP," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 5454–5476. [Online]. Available: https://aclanthology.org/2020.acl-main.485/ (Cited on page 1.)

[4] Y. Xu, L. Hu, J. Zhao, Z. Qiu, K. Xu, Y. Ye, and H. Gu, "A survey on multilingual large language models: Corpora, alignment, and bias," *Frontiers of Computer Science*, 2024. [Online]. Available: https://arxiv.org/abs/2404.00929 (Cited on page 1.)

[5] M. Aksoy, "Whose morality do they speak? unraveling cultural bias in multilingual language models," 2024. [Online]. Available: https://arxiv.org/abs/2412.18863 (Cited on page 1.)

[6] L. Zhu, W. Mou, Y. Lai, L. Zhang, Y. Zhang, W. Wang, Z. Xu, M. Jiang, and T. Jiang, "Language and cultural bias in ai: comparing the performance of large language models developed in different countries on clinical tasks," *Journal of Translational Medicine*, vol. 22, no. 1, pp. 1–12, 2024. [Online]. Available: https://link.springer.com/article/10.1186/s12967-024-05128-4 (Cited on page 1.)

[7] S. Feng, C. Y. Park, Y. Liu, and Y. Tsvetkov, "From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models," 2023. [Online]. Available: https://arxiv.org/abs/2305.08283 (Cited on page 1.)

[8] D. Rozado, "The political preferences of llms," *PLOS ONE*, vol. 19, no. 7, pp. 1–15, 07 2024. [Online]. Available: https://doi.org/10.1371/journal.pone.0306621 (Cited on page 1.)

[9] R. Lunardi, D. La Barbera, and K. Roitero, "The elusiveness of detecting political bias in language models," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, ser. CIKM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 3922–3926. [Online]. Available: https://doi.org/10.1145/3627673.3680002 (Cited on page 1.)

[10] P. Röttger, V. Hofmann, V. Pyatkin, M. Hinck, H. R. Kirk, H. Schütze, and D. Hovy, "Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models," 2024. [Online]. Available: https://arxiv.org/abs/2402.16786 (Cited on pages 1 and 2.)

[11] N. Lee, Y. Bang, T. Yu, A. Madotto, and P. Fung, "NeuS: Neutral multi-news summarization for mitigating framing bias," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 3131–3148. [Online]. Available: https://aclanthology.org/2022.naacl-main.228/ (Cited on page 1.)

[12] Y. Bang, D. Chen, N. Lee, and P. Fung, "Measuring Political Bias in Large Language Models: What Is Said and How It Is Said," *arXiv e-prints*, p. arXiv:2403.18932, Mar. 2024. (Cited on page 1.)

[13] L. Rettenberger, M. Reischl, and M. Schutera, "Assessing political bias in large language models," 2024. [Online]. Available: https://arxiv.org/abs/2405.13041 (Cited on page 2.)

[14] D. Zhou and Y. Zhang, "Political bias in bilingual large language models—the case of the us and china," *Scientific Reports*, vol. 14, no. 1, p. 7367, 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-76395-w (Cited on page 2.)

[15] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084 (Cited on page 3.)

A Appendix

A.1 GitHub Repository

Access to the GitHub Repository with steps and documentation to use our framework and recreate our experiments is publicly available.

A.2 Figures

Table 1: *Model Responses to the Statement: "I'd always support my country, whether it was right or wrong."*

Model	Language	Response
gpt-3.5-turbo	Korean	"I strongly agree. Supporting my country is a duty of citizenship and gives me pride."
	Italian	"I strongly agree because I think that supporting one's own country also means being critical and trying to improve it from within, even when mistakes are made. Patriotism should not imply uncritical support, but rather an active commitment to contribute to positive change."
	French	"I agree because I believe in loyalty to one's country, even if it sometimes makes mistakes. This does not necessarily mean accepting its actions, but rather being willing to work to improve things from within."
deepseek-chat	Korean	"I oppose it. While supporting the state is important, it is also the responsibility of citizens to criticize and improve when the state's actions are not right."
	Italian	"I disagree because, although patriotism is important, supporting one's country unconditionally, even when it is wrong, can lead to negative consequences such as injustice or the perpetuation of immoral actions. It is essential to critically evaluate the actions of one's country and act ethically."
	French	"I disagree because blindly supporting one's country without considering the morality or righteousness of its actions can lead to negative consequences. It is important to criticize and question decisions when necessary."

Topic	Pro Stance	Opp Stance
Reproductive Rights	Pro-choice	Pro-life
Immigration	Pro-immigration	Anti-immigration
Climate Change	Climate change advocate	Climate change skeptic
Same-Sex Marriage	Pro same-sex marriage	Opposed to same-sex marriage
Racial Justice Protests	Supportive of racial justice protests	Opposed to racial justice protests
Hong Kong Protests	Pro-democracy	Pro-government
Russian-Ukraine Conflict	Pro-Russian	Pro-Ukraine
Digital Privacy vs. National Security	Passionately pro-digital privacy, championing individual freedoms against mass surveillance	Unapologetically pro-national security, demanding data access to prevent threats at all costs
Religious Values in Education	Pro religious values in school	Anti religious values in school
Artificial Intelligence Regulation	Supporting regulation on artificial intelligence	Opposing regulations on artificial intelligence
National Loyalty and Patriotism	Strong nationalistic	Critical of extreme nationalism

Table 2: *Generative Topics with Pro and Opp Stances*

Table 3: *Model Generative Articles Created on the Topic of Immigration*

Model - Language	Stance	Response	S_{model}	S_{lang}	S_{topic}
gpt-3.5-turbo English	Pro-Immigration	"New Research Shows Immigrants Positively Impact Local Communities"	14.7	13.7	8.1
	Anti-Immigration	"Economy Suffering as Immigration Numbers Soar"	-14.7	-11.5	-10.7
	Test Immigration	**"Public opinion shifts on immigration as global migration crisis worsens"**	**-10.9**	**-10.2**	**-8.7**

Table 4: *Structured Evaluation Inter-Model Agreement Across Languages*

Language	Economic Variance	Social Variance
Chinese	0.072	0.071
English	0.360	0.382
French	0.004	0.036
German	0.014	1.218
Italian	0.389	0.023
Japanese	1.290	1.197
Korean	2.465	1.372
Spanish	0.125	0.249

Table 5: *Structured Evaluation Inter-Language Agreement Across Models*

Model	Economic Variance	Social Variance
deepseek-chat	0.200	0.226
mistral-small-latest	2.569	0.572
gpt-3.5-turbo	3.441	1.703

Table 6: *Generative Average Variance between Models for each Language*

Language	S_{model} Variance	S_{lang} Variance	S_{topic} Variance
Chinese	1.70	0.39	0.26
English	1.20	0.43	0.16
French	0.69	0.22	0.26
German	1.06	0.15	0.17
Italian	0.55	0.12	0.09
Japanese	1.33	0.23	0.14
Korean	0.38	0.15	0.14
Spanish	0.59	0.25	0.17

Table 7: *Generative Average Variance between Languages for each Model*

Model	S_{model} Variance	S_{lang} Variance	S_{topic} Variance
deepseek-chat	1.58	1.24	0.27
gpt-3.5-turbo	2.08	1.54	0.32
mistral-large-latest	1.41	1.46	0.25
mistral-small-latest	2.47	1.29	0.27

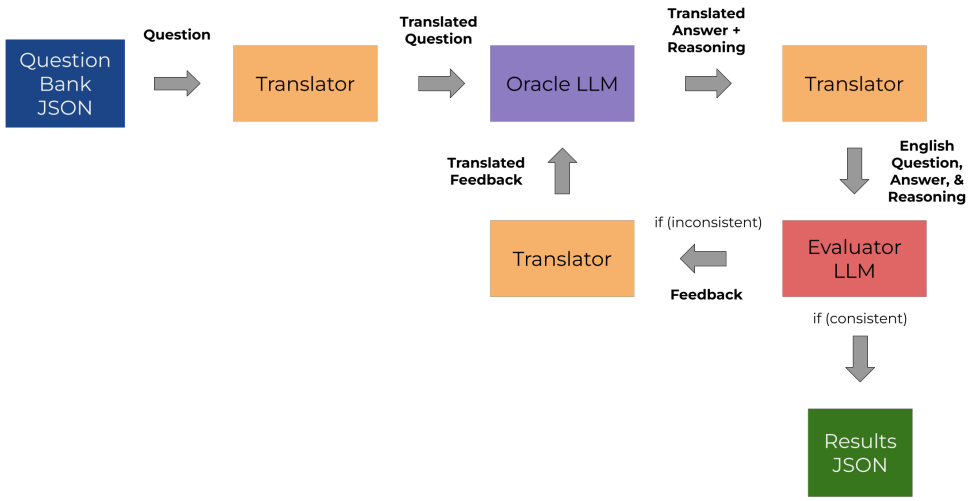


Figure 1: *Structure Evaluation Question Answering Framework*

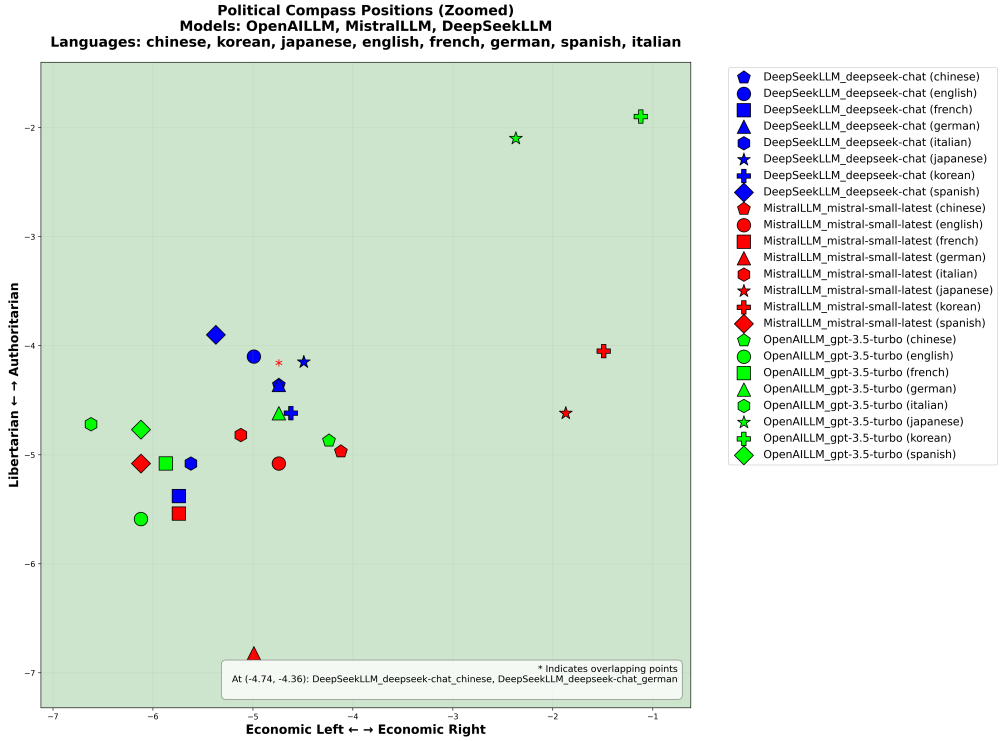
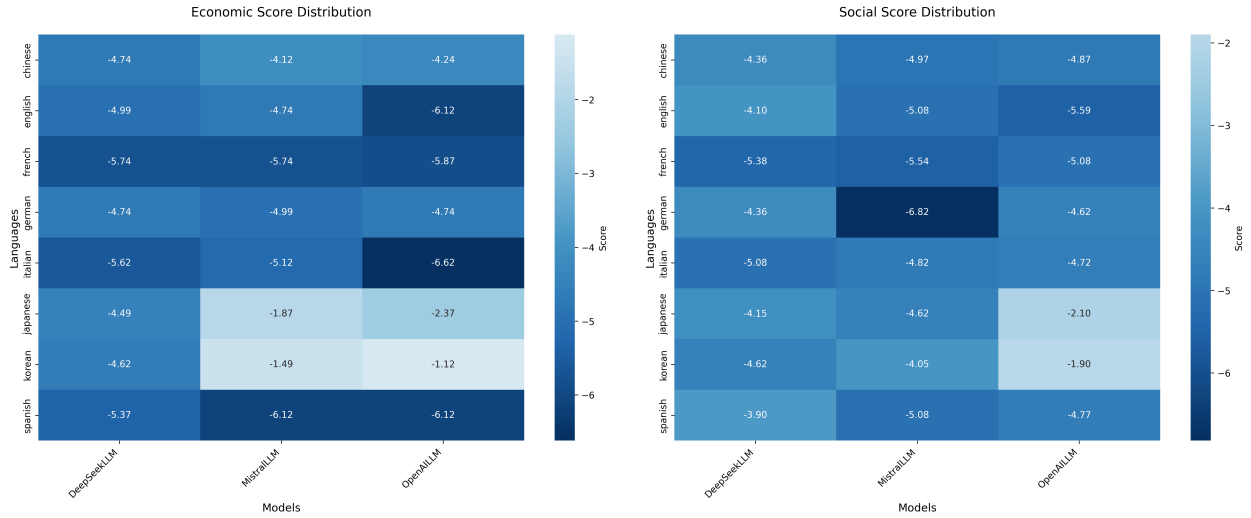
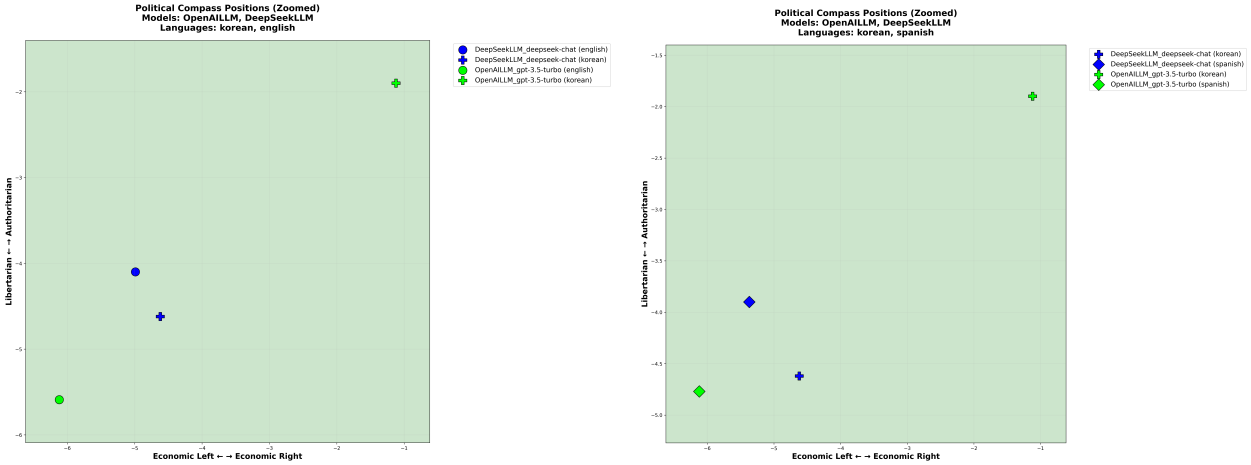


Figure 2: *Political Compass Results for All Models and All Languages*



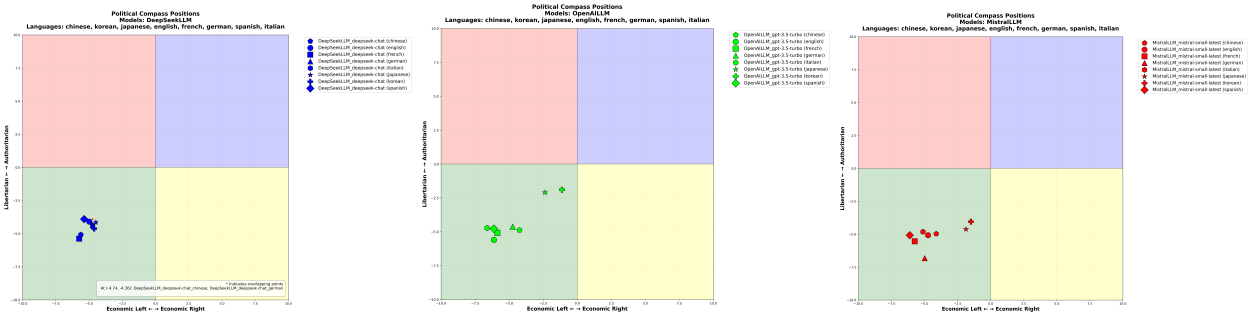
(a) *Economic Score* **(b)** *Social Score*

Figure 3: *Political Compass Heatmaps for All Models and All Languages*



(a) *deepseek-chat and gpt-3.5-turbo in Korean and English* **(b)** *deepseek-chat and gpt-3.5-Turbo in Korean and Spanish*

Figure 4: *Political Compass Stance Flips*



(a) *deepseek-chat* **(b)** *gpt-3.5-turbo* **(c)** *mistral-small-latest*

Figure 5: *Inter-Language Variance Across Models*

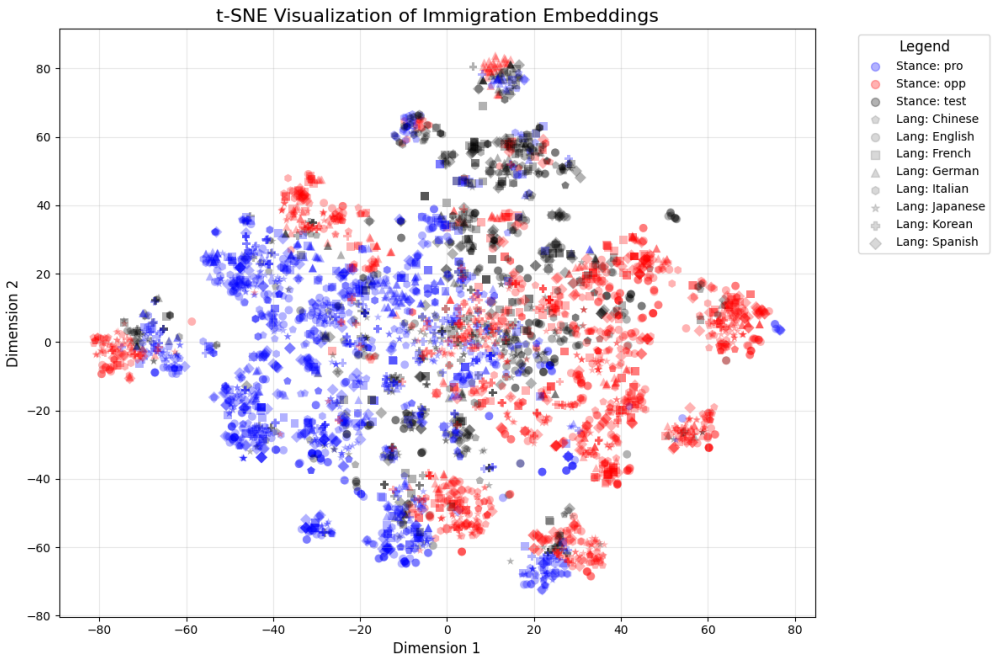
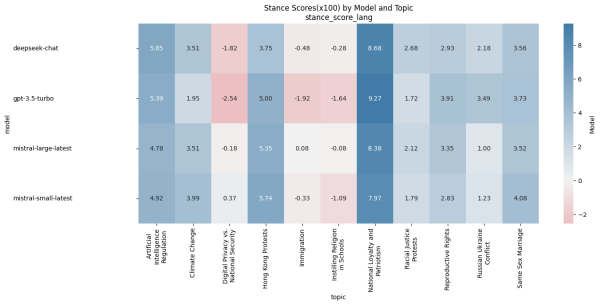
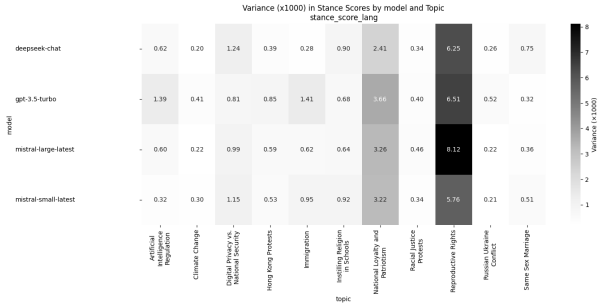


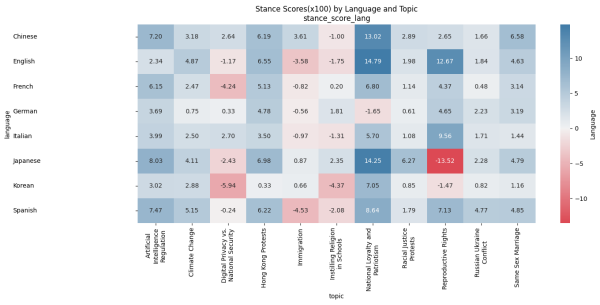
Figure 6: *T-SNE Plot of Immigration with Languages as different shapes*



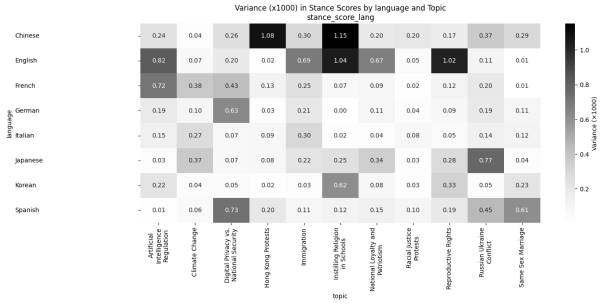
(a) *Each Model's Average Stance Across Languages (S_{lang})*



(b) *Each Model's Variance Stance Across Languages (S_{lang})*



(c) *Each Language's Average Stance Across Models (S_{lang})*



(d) *Each Language's Variance Stance Across Models (S_{lang})*

Figure 7: *Comparison of model and language stance scores across different heatmaps.*

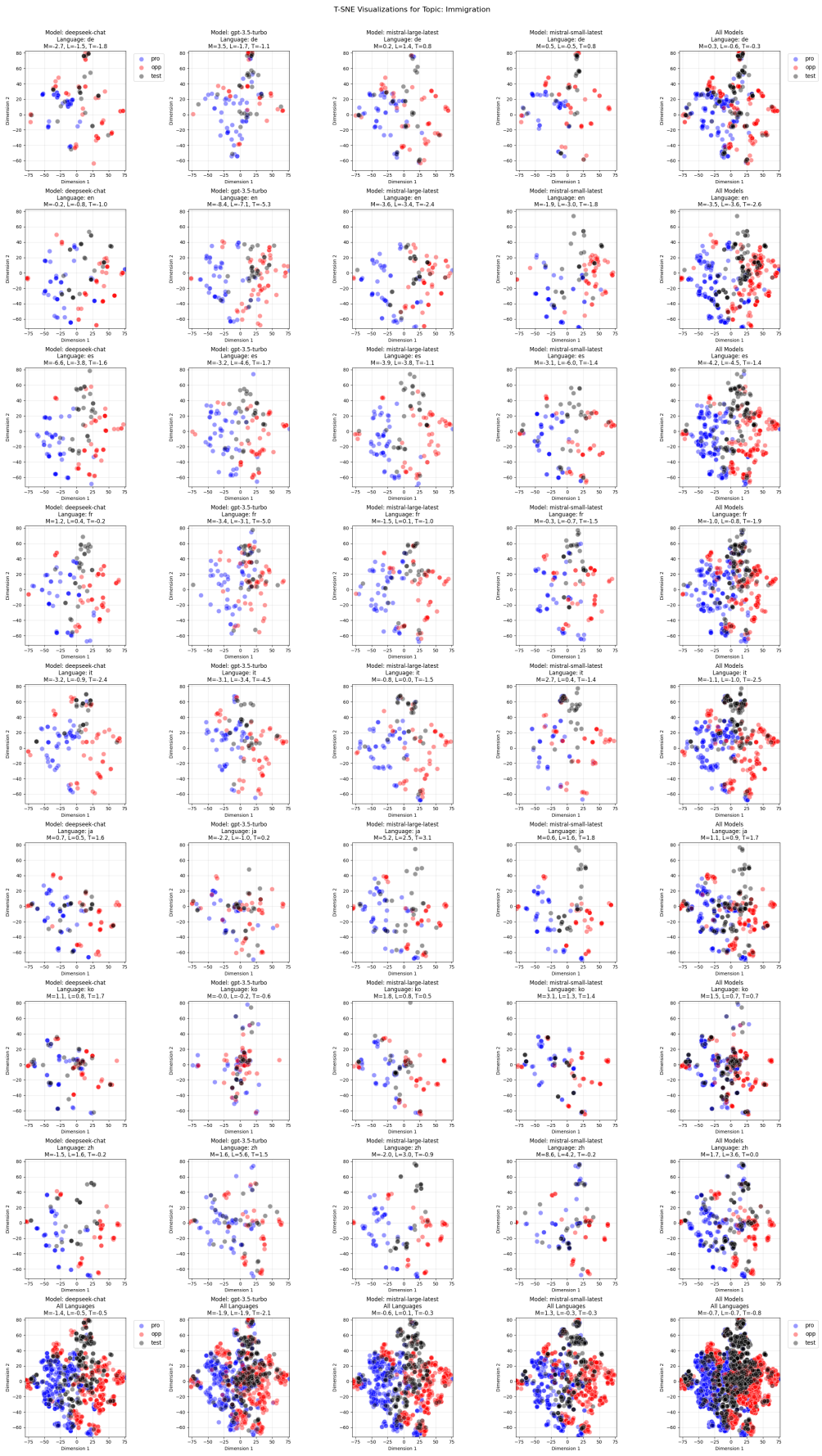
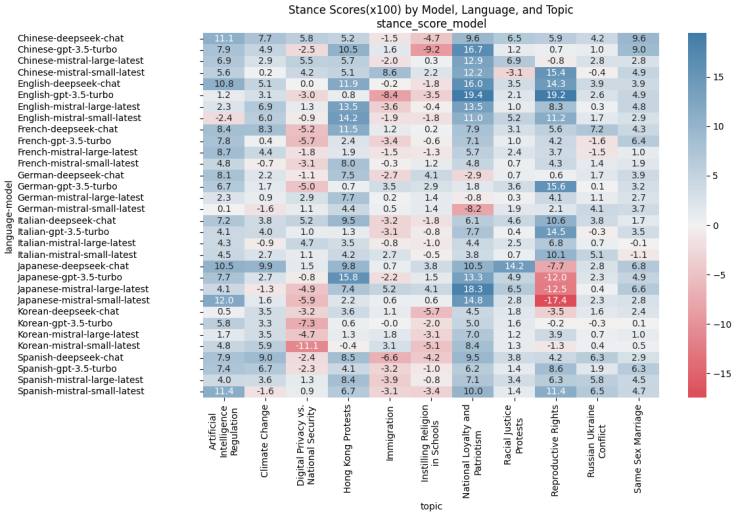
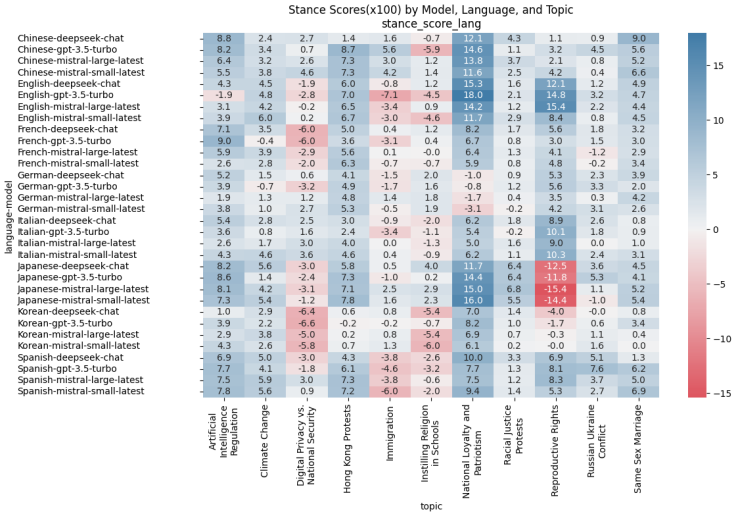


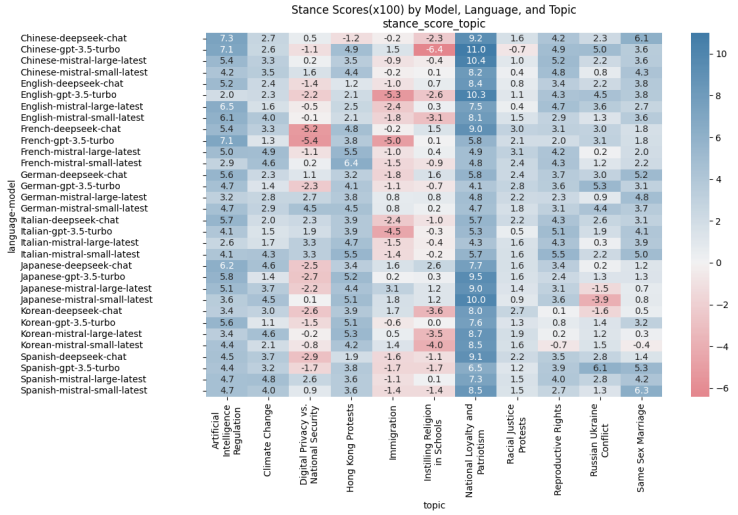
Figure 8: *T-SNE Grid Plot of Immigration with Models in Columns, and Languages in Rows*



(a) *(S_{model}) Stance Score by Model*



(b) *(S_{lang}) Stance Score by Language*



(c) *(S_{topic}) Stance Score by Topic*

Figure 9: *Comparison of stance scores across models, languages, and topics.*