Mind the Values: Assessing the Values of Large Language Models for Southeast Asian Languages

Spencer Hong
Department of Computer Science
National University of Singapore
spencer.hong@u.nus.edu

Hwee Tou Ng
Department of Computer Science
National University of Singapore
nght@nus.edu.sg

Abstract

Recent works have introduced large language models (LLMs) that were trained specifically for Southeast Asian (SEA) languages, which have demonstrated superior performance on SEA languages compared to generic multilingual models. Given that these SEA LLMs were trained on a focused set of languages, a natural question to ask is if these LLMs capture the unique values and opinions of human populations from this geographic location. In this work, we leverage the Global Attitudes Survey and the World Values Survey to assess the alignment between the responses of LLMs when prompted in five different languages (i.e., English, Indonesian, Malay, Thai, and Vietnamese) and human survey respondents. Our findings show that both SEA LLMs and non-SEA LLMs can capture the diverse values of SEA populations, and that all LLMs exhibit a high level of cross-lingual consistency when expressing subjective opinions.

1 Introduction

The advent of large language models (LLMs) in recent years has led to increasing interest in developing multilingual models in order to make this technology accessible to users in all parts of the world. In particular, research efforts for low-resource languages include (1) monolingual models, (2) LLMs trained on an abundance of languages (e.g., over 100) (Hernández-Cano et al., 2025), and, in between the two extremes, (3) models trained on smaller sets of languages (i.e., 10 to 20) that are connected by geography or linguistic similarities (Pava et al., 2025).

Recent efforts on building Southeast Asian (SEA) LLMs (Dou et al., 2025; Ng et al., 2025; Zhang et al., 2025) exemplify the third approach. In Southeast Asia, there are over 1,000 languages spoken by millions of people living in the region. However, these languages have low representation in common pre-training corpora of LLMs (Lovenia et al., 2024) and additionally, existing multilingual LLMs not trained specifically for the region experience performance degradation on pertinent tasks (Ng et al., 2025). This has motivated the release of several open-weight SEA LLMs as well as benchmarks that cover conversational, cultural, and linguistic evaluation specifically for Southeast Asian languages (Susanto et al., 2025; Liu et al., 2025; Wang et al., 2024).

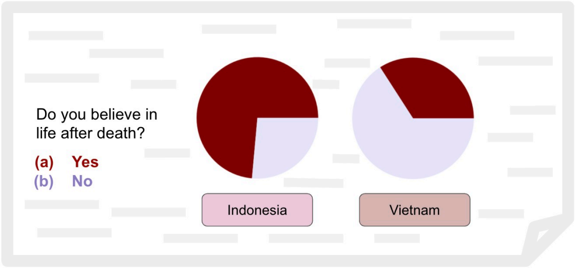

Figure 1: An example question from the World Values Survey and human respondent distributions (for brevity, the question was paraphrased and the list of answer options was shortened to include only those that were selected).

Given the recent interest in this new paradigm for multilingual LLM development, a natural question to ask is if these SEA LLMs can accurately capture the values and opinions of Southeast Asian populations and how they compare with LLMs that were exposed to a larger and more diverse set of languages during the training process. Prior works have investigated the values of multilingual LLMs with established human surveys (Aksoy, 2025; Arora et al., 2023), but to the best of our knowledge, there are no studies that specifically focus on the values of SEA LLMs.

In this work, we leverage survey questions from social science research to extensively evaluate four SEA LLMs and two non-SEA LLMs on a wide range of topics such as religion, politics, gender, and technology. We assess the alignment between the responses of these models and the opinions expressed by human survey respondents on five languages (i.e., English, Indonesian, Malay, Thai, and Vietnamese) and observe that the models that align best with humans can be SEA-specific or non-SEA-specific, which demonstrates that training solely on regional data may not be necessary for capturing the values of relevant human populations. We also find that all LLMs are slightly Anglocentric, despite the numerous languages present in their training data. While LLMs generally give consistent responses across languages, we observe that they can be inconsistent at times, especially on Thai and Vietnamese when compared to English.

2 Values of Southeast Asian LLMs

2.1 Data

Survey Questions. To measure alignment between LLMs and humans, we leverage questions from the Global Attitudes Survey[1] and the World Values Survey (Haerpfer et al., 2022), as consolidated in the GlobalOpinionQA dataset from Durmus et al. (2024). These surveys were developed by social science researchers to capture the values and beliefs from human respondents located in countries from all over the world. Each question is multiple-choice and comes with statistics describing the opinions of humans at the country level. More specifically, if a question q with o_q answer options was administered in a country C_i, it will come with a vector $h_i \in \mathbb{R}^{o_q}$ where the jth item of h_i is the proportion of people from C_i who selected the jth answer option (see Figure 1 for an example). Note that the number of country-level annotations varies across the questions. For this study, we are interested in the questions that have been answered by respondents located in Southeast Asian countries, including Indonesia, Malaysia, Thailand, and Vietnam. Table 1 shows the number of survey questions for each country. The United States is also included for additional comparison.

Translations. In order to determine the behavior of SEA LLMs on non-English languages, machine translation was done with GPT-5 to translate the survey questions into Indonesian, Malay, Thai, and Vietnamese[2]. By prompting language models with non-English inputs, we can study the cross-lingual consistency of the models when expressing opinions and determine the effect that different languages have on the degree of human value alignment. The authors confirmed the quality of the translations by backtranslating into English and checking for semantic consistency on a small subset of the examples.

Country	**GAS**	**WVS**	**Total**
Indonesia	724	212	936
Malaysia	297	219	516
Thailand	107	212	319
Vietnam	246	210	456
USA	893	211	1104

Table 1: Number of survey questions with human statistics for each country.

2.2 Quantifying Alignment

Let $D = \{(q, \{h_i\}_{i=1}^{c_q})\}$ be our dataset of survey questions where c_q is the number of countries that question q was administered in. In order to represent an LLM's opinions for question q, we construct a vector $t \in \mathbb{R}^{o_q}$ by sampling $m > 1$ responses from the model non-deterministically and setting t_j to the fraction of the m responses that map to answer option j. We can then quantify the alignment between any pair of probability distributions $t, h \in \mathbb{R}^{o_q}$ using the Jensen Shannon distance, as proposed by Durmus et al. (2024):

$$JS(t,h) = \sqrt{\frac{KL(t||\frac{t+h}{2}) + KL(h||\frac{t+h}{2})}{2}} \quad (1)$$

where $KL(\cdot||\cdot)$ represents the Kullback-Leibler divergence with a logarithm base of 2 and $\frac{t+h}{2}$ represents the pointwise mean between the vectors. Larger values of $JS(\cdot,\cdot)$ correspond to higher degrees of dissimilarity. Note that $JS(\cdot,\cdot)$ is symmetric and bounded between 0 and 1, where a value of 0 is achieved when the two input distributions are the same.

Equipped with this metric, we can now measure the similarity between LLM-human pairs as well as pairs of LLM distributions. Appendix B shows the distances between human survey respondents from different countries.

3 Experimental Setup

Open-Source Multilingual LLMs. Four Southeast Asian LLMs are used in this study:

[1] https://www.pewresearch.org/

[2] Indonesian, Malay, Thai, and Vietnamese are the official languages of Indonesia, Malaysia, Thailand, and Vietnam respectively. The official language of the United States is English.

Survey	Model	Indonesia		Malaysia		Thailand		Vietnam		USA
		id	en	ms	en	th	en	vi	en	en
GAS	Apertus-8B-Instruct-2509	44.74	44.86	44.03	43.25	**40.06**	**41.04**	47.68	49.09	42.45
	gemma-3-12b-it	62.29	63.29	62.71	63.66	61.66	62.11	61.48	62.74	58.52
	SeaLLMs-v3-7B-Chat	**38.69**	**38.83**	**40.66**	**41.56**	41.63	43.0	**39.21**	**37.56**	**33.05**
	Llama-SEA-LION-v3-8B-IT	48.3	48.3	49.36	51.13	48.72	51.35	48.69	48.71	45.59
	Gemma-SEA-LION-v3-9B-IT	57.01	58.92	59.52	60.28	58.08	58.15	54.9	57.68	55.46
	Sailor2-8B-Chat	60.08	58.96	62.11	60.72	58.46	59.97	55.0	54.69	54.69
WVS	Apertus-8B-Instruct-2509	**44.78**	**47.16**	**43.58**	45.99	46.93	46.77	**50.19**	49.99	**44.69**
	gemma-3-12b-it	65.78	67.06	65.0	67.34	68.87	69.5	68.03	66.72	63.68
	SeaLLMs-v3-7B-Chat	50.76	**47.16**	48.8	**45.7**	**44.67**	**40.69**	55.46	**49.47**	44.77
	Llama-SEA-LION-v3-8B-IT	54.63	56.27	54.04	56.01	52.94	55.54	54.88	54.42	51.68
	Gemma-SEA-LION-v3-9B-IT	64.58	63.39	64.13	63.31	64.45	64.56	64.06	64.51	59.46
	Sailor2-8B-Chat	61.39	60.59	64.38	62.45	65.64	65.51	66.7	64.11	60.36

Table 2: Average JS distances between LLM and human distributions on Global Attitudes Survey (GAS) and World Values Survey (WVS). Each column represents a set of questions with human distributions for that country. Each subcolumn represents the language used to prompt the model (i.e., **en** for English, **id** for Indonesian, **ms** for Malay, **th** for Thai, and **vi** for Vietnamese). For each survey, the lowest value in each subcolumn is in **bold** and the second lowest is underlined.

(1) SeaLLMs-v3-7B-Chat, (2) Llama-SEA-LION-v3-8B-IT, (3) Gemma-SEA-LION-v3-9B-IT, and (4) Sailor2-8B-Chat. The two SEA-LION models (Ng et al., 2025) and Sailor2-8B-Chat (Dou et al., 2025) were trained via continued pre-training and subsequent post-training, while SeaLLMs-v3-7B-Chat (Zhang et al., 2025) was instead the result of first merging together several monolingual models and then supervised fine-tuning.

We also consider two multilingual models that were not developed specifically for the SEA region: (1) Apertus-8B-Instruct-2509 and (2) gemma-3-12b-it. Apertus-8B-Instruct-2509 (Hernández-Cano et al., 2025) was trained on 1,811 languages and comes with a completely transparent development pipeline. gemma-3-12b-it supports over 140 languages and can handle both image and text inputs.

Implementation. For each question, an LLM is prompted $m = 30$ times with a temperature of 1.0 and these predictions are then converted into a probability distribution over the answer options as described in Section 2.2. Note that though all LLMs are told to format their answers in a certain way in the prompts given to them so that the outputs can be deterministically mapped to one of the answer options, the models may not always follow instructions[3]. To ensure that the valid responses are representative of the LLM's output distributions, questions are excluded from reported calculations when the number of valid predictions is less than 15. All reported results are presented in %.

4 Results and Discussion

Table 2 presents the average JS distances between LLM and human distributions on the Global Attitudes Survey and the World Values Survey[4]. For each SEA country, each original question corresponds to two prompts, one in English and another in the official language of that country. We describe our observations below.

SEA LLMs and non-SEA LLMs can capture the opinions of SEA populations. As we would expect, the LLM that achieves the best overall alignment with SEA populations is a SEA LLM (i.e., SeaLLMs-v3-7B-Chat). However, a counterintuitive observation is that the second best SEA LLM (Llama-SEA-LION-v3-8B-IT) is consistently less aligned when compared to a non-SEA LLM (Apertus-8B-Instruct-2509), which usually achieves the second lowest scores on both surveys. We also observe that Gemma-SEA-LION-v3-9B-IT always has lower alignment when compared to Llama-SEA-LION-v3-8B-IT despite coming from the same model family. Additionally, the results show that gemma-3-12b-it has the worst alignment with humans.

Scores on GAS are generally lower than WVS scores. The range of scores on GAS per model is usually lower than the range of scores on WVS. To

[3] See Appendix A for more details.

[4] We also calculate the average normalized entropies of the LLMs' response distributions along with average F1 scores between the LLMs' and humans' sets of answer options with nonzero probability mass. See Appendix C for discussion. See Appendix E for the distances between LLMs.

Model	Indonesian	Malay	Thai	Vietnamese
Apertus-8B-Instruct-2509	29.72	32.1	**39.9**	34.24
gemma-3-12b-it	25.2	28.44	**35.35**	27.89
SeaLLMs-v3-7B-Chat	29.46	31.87	**34.58**	34.12
Llama-SEA-LION-v3-8B-IT	28.77	31.52	**40.25**	31.89
Gemma-SEA-LION-v3-9B-IT	25.56	27.45	**35.2**	31.24
Sailor2-8B-Chat	25.05	24.31	**36.55**	26.5

Table 3: Average JS distances between English distributions and SEA language distributions per question. The highest value in each row is in **bold** and the second highest is underlined.

explain this, we automatically classify each question into one of four topical categories: (1) *Society & Culture*, (2) *Politics & World*, (3) *Economy & Work*, and (4) *Technology & Knowledge* (see Appendix D) and notice that for all countries, *Society & Culture* and *Politics & World* make up the majority of the questions. However, these categories are usually balanced in WVS, while GAS usually has more questions in *Politics & World*. Therefore, the difference in ranges may be explained by the lower alignment of the models on *Society & Culture*.

All LLMs are slightly Anglocentric. This observation aligns with previous findings (Durmus et al., 2024; AlKhamissi et al., 2024). On GAS, all LLMs achieved lower scores for USA compared to the scores for other countries; notably, for gemma-3-12b-it, SeaLLMs-v3-7B-Chat, Llama-SEA-LION-v3-8B-IT, and Sailor2-8B-Chat, the USA scores were the lowest in the model-specific ranges. On WVS, we see a similar trend, with gemma-3-12b-it, Llama-SEA-LION-v3-8B-IT, Gemma-SEA-LION-v3-9B-IT, and Sailor2-8B-Chat achieving the lowest scores on USA compared to other countries.

Prompting in official SEA languages does not improve human alignment. We observe that the average distances are always consistent across languages for SEA countries. To see if this means that there is consistency between the languages for each question, the Jensen Shannon distance between the distribution from prompting in the official language and the distribution from prompting in English is computed and then the average over each dataset is reported in Table 3. We can see that all LLMs demonstrate good levels of consistency, but the distances are relatively higher for Thai and Vietnamese.

5 Related Work

SEA-Centric Evaluation. Several works have released benchmarks focusing on evaluating LLM performance on Southeast Asian languages. SEACrowd (Lovenia et al., 2024) is a standardized collection of approximately 500 corpora in about 1,000 SEA languages and three modalities: text, image, and audio. Liu et al. (2025) introduced two benchmarks consisting of multiple-choice questions drawn from regional educational exams and open-ended tasks that resemble interactions in SEA communities. SEA-HELM (Susanto et al., 2025) provides a holistic suite of cultural and linguistic evaluation tasks in Filipino, Indonesian, Tamil, Thai, and Vietnamese. BHASA (Leong et al., 2023) and SeaEval (Wang et al., 2024) aim to evaluate LLMs on natural language understanding, generation, reasoning, linguistics, cultural representation, and cross-lingual consistency.

Subjective Opinions of LLMs. Previous works have attempted to study the morals and values of LLMs by utilizing social science surveys and frameworks from psychology such as the World Values Survey (Rystrøm et al., 2025; Durmus et al., 2024; Tao et al., 2024; AlKhamissi et al., 2024; Arora et al., 2023), Hofstede's cultural dimensions (Masoud et al., 2025; Cao et al., 2023; Arora et al., 2023), and the Moral Foundations Questionnaire (Aksoy, 2025; Ji et al., 2025; Abdulhai et al., 2024). Additionally, some works have focused on evaluating multilingual LLMs on non-English languages, like we do in this study (Aksoy, 2025; Agarwal et al., 2024; Arora et al., 2023; Cao et al., 2023). However, we are the first to perform value assessment on Southeast Asian language models. Aside from these efforts, there is also a line of research focusing specifically on evaluating the political leanings of LLMs (Rettenberger et al., 2025; Röttger et al., 2024; Fulay et al., 2024; Chen et al., 2024; Feng et al., 2023).

6 Conclusion

In this work, we assessed the values of Southeast Asian language models by leveraging questions from Global Attitudes Survey and the World Values Survey. By evaluating the LLMs on five different languages (i.e., English, Indonesian, Malay, Thai, and Vietnamese), we quantified the alignment between the responses of LLMs and the opinions expressed by human survey respondents, and analyzed the cross-lingual consistency of the models. Our findings show that both SEA and non-SEA LLMs can achieve moderately high alignment with humans, and that the language that we prompt them in does not have an impact on the results. We also observe that the models in this study are slightly Anglocentric and less consistent with English on Thai and Vietnamese languages.

Limitations

In our work, we leveraged questions from two surveys that were previously administered to human respondents in order to compare LLM responses with the opinions of human populations. However, this comes with a few limitations; (1) all questions are multiple-choice, but the values expressed by LLMs may be more varied in open-ended settings; (2) these survey questions were originally designed for humans, and not for evaluating LLMs; and (3) the survey data may not be perfectly representative of the values of the underlying human populations (e.g., the questions for each country may not be comprehensive).

Additionally, each survey question in our study corresponded to two LLM prompts: the original English question and the machine-translated non-English question. However, LLMs may be sensitive to the wording of the prompt (i.e., using questions that are semantically the same but lexically different may affect the results). We did not perform any experiments to evaluate the robustness of the models.

Lastly, we aimed to assess alignment with human populations from Southeast Asia, but only selected four countries and prompted LLMs using only four SEA languages. Future work can extend this evaluation to other locations in the region with other languages.

References

Marwa Abdulhai, Gregory Serapio-García, Clement Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752, Miami, Florida, USA. Association for Computational Linguistics.

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia. ELRA and ICCL.

Meltem Aksoy. 2025. Whose morality do they speak? unraveling cultural bias in multilingual language models. *Natural Language Processing Journal*, page 100172.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Kai Chen, Zihao He, Jun Yan, Taiwei Shi, and Kristina Lerman. 2024. How susceptible are large language models to ideological manipulation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17140–17161, Miami, Florida, USA. Association for Computational Linguistics.

Longxu Dou, Qian Liu, Fan Zhou, Changyu Chen, Zili Wang, Ziqi Jin, Zichen Liu, Tongyao Zhu, Cunxiao Du, Penghui Yang, and 1 others. 2025. Sailor2: Sailing in South-East Asia with inclusive multilingual LLMs. *arXiv preprint arXiv:2502.12982*.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Suyash Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2024. On the relationship between truth and political bias in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9018, Miami, Florida, USA. Association for Computational Linguistics.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bjorn Puranen, and 1 others. 2022. World values survey: Round seven–country-pooled datafile version 5.0.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, and 1 others. 2025. Apertus: Democratizing open and compliant LLMs for global language environments. *arXiv preprint arXiv:2509.14233*.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2025. Moralbench: Moral evaluation of llms. *SIGKDD Explor. Newsl.*, 27(1):62–71.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. BHASA: A holistic Southeast Asian linguistic and cultural evaluation suite for large language models. *arXiv preprint arXiv:2309.06085*.

Chaoqun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, Anh Tuan Luu, and Lidong Bing. 2025. SeaExam and SeaBench: Benchmarking LLMs with local multilingual questions in Southeast Asia. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6119–6136, Albuquerque, New Mexico. Association for Computational Linguistics.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on Hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.

Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, and 1 others. 2025. SEA-LION: Southeast Asian languages in one network. *arXiv preprint arXiv:2504.05747*.

J Pava, C Meinhardt, HB Uz Zaman, T Friedman, ST Truong, D Zhang, V Marivate, and S Koyejo. 2025. Mind the (language) gap: Mapping the challenges of LLM development in low-resource language contexts.

Luca Rettenberger, Markus Reischl, and Mark Schutera. 2025. Assessing political bias in large language models. *Journal of Computational Social Science*, 8(2):1–17.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

Jonathan Rystrøm, Hannah Rose Kirk, and Scott Hale. 2025. Multilingual!= multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs. *arXiv preprint arXiv:2502.16534*.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xianbin Yong, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. SEA-HELM: Southeast Asian holistic evaluation of language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12308–12336, Vienna, Austria. Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2025. SeaLLMs 3: Open foundation and chat multilingual large language models for Southeast Asian languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 96–105, Albuquerque, New Mexico. Association for Computational Linguistics.

A Mapping LLM Outputs to Answer Choices

For each survey question, in addition to presenting each LLM with the question and corresponding answer options, the prompt also begins with the following instructions: "*You are provided with a multiple-choice question. Answer the question by selecting exactly one of the answer options. Just output the letter corresponding to your answer and nothing else.*"

While LLMs can usually select answer options, they do not always follow the formatting instructions. To handle this, for each question q with answer options $a_1, \cdots, a_{o_q}$, there is a deterministic mapping defined to map free-form outputs to each answer option that involves looking for both exact and inexact matches. For example, if q was "*Do you believe in life after death?*" with option a_1 corresponding to "*(A) Yes*", then LLM outputs like "*A*", "*(A)*", "*A.*", "*A)*", "*(A) Yes*", "*A. Yes*", and "*A) Yes*" will all map to a_1. Additionally, if the LLM output starts with one of the last three aforementioned phrases, it will also map to a_1. If we cannot find any matches for an LLM output according to these mappings, it is considered invalid.

We recognize that our mappings may not cover all valid responses; however, we observed that they are comprehensive enough for our setting.

B Alignment Between Survey Respondents from Different Countries

Figure 2 shows the average JS distance per question between human groups from different countries[5]. According to the figure, we can see that Malaysia has higher alignment with Indonesia and Thailand, and the USA is most dissimilar with Indonesia and Vietnam[6].

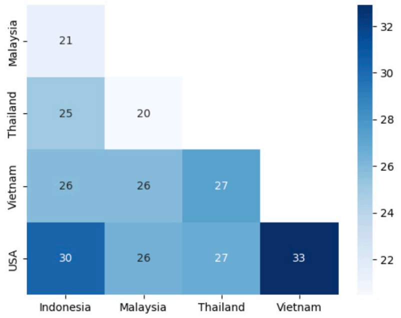

Figure 2: Average JS distances per question between human groups

C Normalized Entropies & F1 Scores

Table 4 shows the average normalized entropy of the LLM response distributions per question, defined as:

$$NE(p) = \frac{-\sum_{i=1}^{n} p_i \cdot log(p_i)}{log(n)} \qquad (2)$$

where p is a discrete probability distribution over n items. The maximum value of $NE(\cdot)$ is 1, which is achieved with a uniform distribution. The minimum value is 0, which is achieved by setting the probability of one item to 1 and everything else to 0.

Table 5 shows the average F1 scores between the LLM's and humans' sets of answer options with nonzero probabilities. The humans' set is treated as the ground truth. The F1 score is computed for each question and then averaged over all questions.

From the tables, we can see that both metrics correlate with the Jensen Shannon distances from Table 2, which means that lower JS scores correspond to LLMs not only increasing the entropy of response distributions to match human diversity, but also achieving high F1 scores by placing probability mass on the answer options that are actually selected by human respondents.

D Topical Distribution of Survey Questions

Using GPT-5, we assign each survey question one out of 17 topic labels, and then group the 17 topics into 4 coarse-grained categories[7]. One of the authors sampled 202 questions and their corresponding *fine-grained* topic labels, and observed that approximately 95% were correct. Here is the list of topics/categories:

[5] For each country pair, the questions considered must have statistics for both countries in the pair.

[6] Please keep in mind that these statements may not be perfectly representative of human populations in the real world. As stated in Durmus et al. (2024), relying on social science surveys comes with limitations.

[7] We used ChatGPT to assist with deriving the coarse-grained categories.

Survey	Model	Indonesia		Malaysia		Thailand		Vietnam		USA
		id	en	ms	en	th	en	vi	en	en
GAS	Apertus-8B-Instruct-2509	51.63	51.76	55.33	54.32	57.29	61.92	43.96	50.31	49.25
	gemma-3-12b-it	3.51	3.76	4.09	3.43	6.22	5.14	2.98	4.81	3.34
	SeaLLMs-v3-7B-Chat	70.19	66.82	73.56	66.61	75.14	72.41	68.74	66.3	67.0
	Llama-SEA-LION-v3-8B-IT	38.13	36.66	40.51	34.91	47.74	39.49	37.28	36.16	34.42
	Gemma-SEA-LION-v3-9B-IT	15.54	13.38	15.92	12.62	19.1	14.45	15.69	13.43	11.69
	Sailor2-8B-Chat	12.07	13.78	9.9	13.16	19.63	14.6	11.99	12.7	12.6
WVS	Apertus-8B-Instruct-2509	46.91	54.43	51.38	52.78	46.75	52.21	50.15	52.81	53.55
	gemma-3-12b-it	3.85	3.88	5.43	3.78	3.62	3.84	2.66	4.07	3.79
	SeaLLMs-v3-7B-Chat	70.53	72.86	70.7	71.56	71.67	72.89	73.03	71.27	71.34
	Llama-SEA-LION-v3-8B-IT	40.86	36.19	43.41	35.17	47.71	35.27	44.45	35.18	37.13
	Gemma-SEA-LION-v3-9B-IT	14.38	15.25	14.23	15.18	14.03	15.72	13.27	14.42	15.08
	Sailor2-8B-Chat	12.07	12.94	12.99	12.5	17.04	13.19	15.65	12.46	12.97

Table 4: Average **normalized entropies** on Global Attitudes Survey (GAS) and World Values Survey (WVS). Each column represents a set of questions with human distributions for that country. Each subcolumn represents the language used to prompt the model (i.e., **en** for English, **id** for Indonesian, **ms** for Malay, **th** for Thai, and **vi** for Vietnamese).

Survey	Model	Indonesia		Malaysia		Thailand		Vietnam		USA
		id	en	ms	en	th	en	vi	en	en
GAS	Apertus-8B-Instruct-2509	79.77	79.81	79.45	80.19	80.88	82.73	74.71	78.26	78.99
	gemma-3-12b-it	46.49	46.4	44.79	45.21	43.12	42.19	44.98	45.88	46.86
	SeaLLMs-v3-7B-Chat	90.53	91.85	89.16	90.45	88.64	89.89	88.61	89.33	92.08
	Llama-SEA-LION-v3-8B-IT	71.86	71.23	71.04	69.05	72.68	68.91	68.6	71.19	69.64
	Gemma-SEA-LION-v3-9B-IT	59.3	56.66	56.88	55.47	52.77	54.23	59.09	55.31	55.17
	Sailor2-8B-Chat	52.63	53.88	51.54	52.81	54.25	50.33	52.51	53.31	53.39
WVS	Apertus-8B-Instruct-2509	76.73	77.73	81.89	78.27	72.2	72.16	77.67	76.03	72.86
	gemma-3-12b-it	37.2	37.0	42.7	39.86	33.28	32.95	37.65	39.24	32.64
	SeaLLMs-v3-7B-Chat	79.1	82.63	74.97	79.63	79.38	81.91	70.98	77.03	82.08
	Llama-SEA-LION-v3-8B-IT	66.49	61.03	65.56	60.61	66.13	58.73	62.57	61.1	59.64
	Gemma-SEA-LION-v3-9B-IT	47.37	48.55	49.34	50.37	43.06	44.18	48.11	48.33	45.81
	Sailor2-8B-Chat	43.44	44.45	48.26	49.05	42.42	39.47	48.02	46.98	40.8

Table 5: Average **F1 scores** on Global Attitudes Survey (GAS) and World Values Survey (WVS). Each column represents a set of questions with human distributions for that country. Each subcolumn represents the language used to prompt the model (i.e., **en** for English, **id** for Indonesian, **ms** for Malay, **th** for Thai, and **vi** for Vietnamese).

A. Society & Culture: (1) Social values and attitudes, (2) Religion and spirituality, (3) Gender and LGBTQ, (4) Family and relationships, (5) Race and ethnicity, (6) Generations and age, (7) Demographics

B. Politics & World: (8) Politics and policy, (9) Security, (10) Immigration and migration, (11) International affairs, (12) Regions and countries

C. Economy & Work: (13) Economy and work

D. Technology & Knowledge: (14) Science and technology, (15) Internet and technology, (16) News habits and media, (17) Methodological research

Table 6 shows the proportion of questions that fall in each of the four topical categories for each survey.

E Alignment Between LLMs

To examine the similarity between the six LLMs considered in this study, we compute the JS distances between distributions of each pair of LLMs for each question and again aggregate via averaging. This is done for each Southeast Asian country using the official languages only. The heatmap for Indonesia is shown in Figure 3, for Malaysia in Figure 4, for Thailand in Figure 5, and for Vietnam in Figure 6.

From the figures, we can see that how the different model pairs compare with each other is very similar across all countries/languages. In more detail, we make the following observations: (1) The SEA-LION models consistently achieve lower distances compared to other pairs, most likely because they are from the same family; (2) gemma-3-12b-it and SeaLLMs-v3-7B-Chat are far-

Survey	Topic	Indonesia	Malaysia	Thailand	Vietnam	USA
GAS	Society & Culture	21.13	19.53	14.02	15.85	20.27
	Politics & World	60.5	59.93	45.79	51.22	64.5
	Economy & Work	9.94	11.45	26.17	17.48	10.19
	Technology & Knowledge	8.43	9.09	14.02	15.45	5.04
WVS	Society & Culture	41.04	40.64	41.04	41.9	42.18
	Politics & World	46.23	44.29	46.7	44.29	44.55
	Economy & Work	7.55	10.05	7.08	8.57	8.06
	Technology & Knowledge	5.19	5.02	5.19	5.24	5.21

Table 6: Proportion of questions for each topic. Note that *for each survey*, the values in a column sum to 100%.

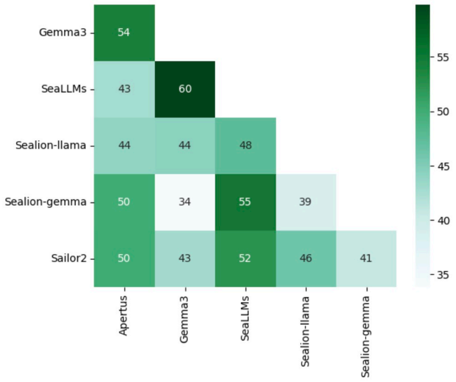

Figure 3: LLM-LLM alignment in the Indonesian language. Model names are abbreviated for better readability.

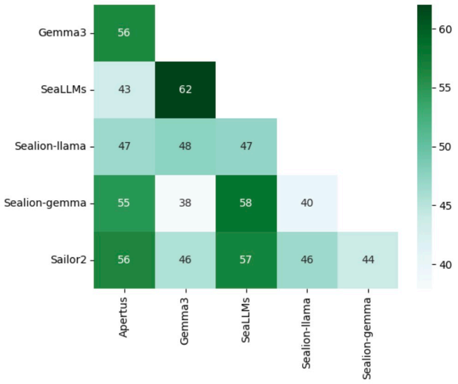

Figure 4: LLM-LLM alignment in the Malay language. Model names are abbreviated for better readability.

thest apart from each other, which is expected since SeaLLMs-v3-7B-Chat achieves the best human alignment while gemma-3-12b-it achieves the worst; and (3) Gemma-SEA-LION-v3-9B-IT and gemma-3-12b-it always achieve the lowest average distance.

F Use of Scientific Artifacts

For this work, we leveraged the GlobalOpinionQA dataset from Durmus et al. (2024), which is released under a CC-BY-NC-SA-4.0 license. All models used in the experiments are open-weight and available on HuggingFace[8]. Each model's page on HuggingFace comes with information about its license; Apertus-8B-Instruct-2509 and Sailor2-8B-Chat are released under apache-2.0, gemma-3-12b-it and Gemma-SEA-LION-v3-9B-IT are released under gemma, Llama-SEA-LION-v3-8B-IT is released under llama3.1, and SeaLLMs-v3-7B-Chat is released under seallms. To access the models, version 4.56.1 of the transformers package was employed.

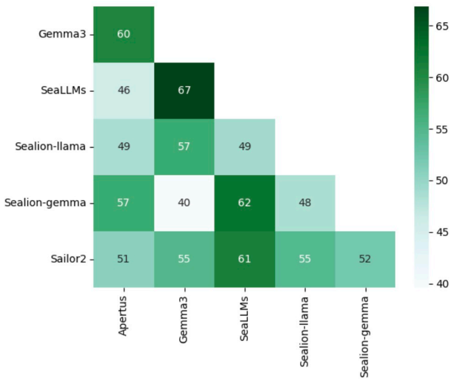

Figure 5: LLM-LLM alignment in the Thai language. Model names are abbreviated for better readability.

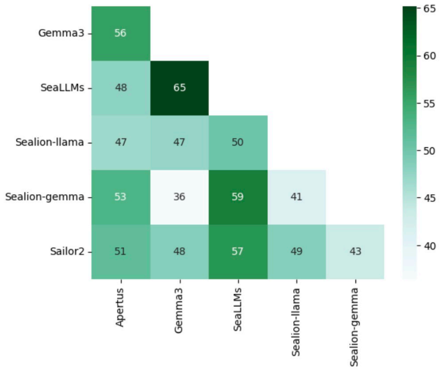

Figure 6: LLM-LLM alignment in the Vietnamese language. Model names are abbreviated for better readability.

[8] https://huggingface.co/

Usage of all artifacts is consistent with their intended use.