JUBAKU: An Adversarial Benchmark for Exposing Culturally Grounded Stereotypes in Japanese LLMs

Anonymous ACL submission

Abstract

Social biases reflected in language are inherently shaped by cultural norms, which vary significantly across regions, leading to diverse manifestations of stereotypes. However, social bias evaluation for large language models (LLMs) in non-English contexts often relies on translations of English benchmarks that fail to reflect Japanese cultural norms. In this work, we introduce JUBAKU (Japanese cUlture adversarial BiAs benchmarK Under handcrafted creation)[1], an adversarially constructed benchmark tailored to Japanese cultural contexts, considering ten distinct cultural categories. Unlike existing benchmarks, JUBAKU features dialogue scenarios hand-crafted by Japanese annotators designed to trigger and expose latent social biases in Japanese LLMs. We evaluated nine Japanese LLMs on JUBAKU and three others adapted from English benchmarks. All models clearly exhibited biases on JUBAKU, performing below the random baseline of 50% with an average accuracy of 23% (ranging from 13% to 33%), despite higher accuracy on the other benchmarks. Human annotators achieved 91% accuracy in identifying unbiased responses, confirming JUBAKU's reliability and its adversarial nature to LLMs.

1 Introduction

Large Language Models (LLMs) encode social biases within their content, making their safe deployment a growing concern. Since social biases are deeply tied to culture, bias evaluation benchmarks must reflect local cultural norms (Adilazuarda et al., 2024). For instance, Japanese norms often value indirect communication, which can lead to stereotypes that discourage assertiveness, unlike more direct cultures. To ensure robust safety assessment, it is also crucial to evaluate LLMs under adversarial inputs designed to provoke harmful responses, as such latent biases may remain hidden in standard settings (Perez et al., 2022; Paulus et al., 2024).

Several benchmarks have been proposed to evaluate social biases, particularly in English contexts such as CrowS-Pairs (Nangia et al., 2020), StereoSet (Nadeem et al., 2021), BBQ (Parrish et al., 2022), and BOLD (Dhamala et al., 2021). Many of these have been adapted for other cultural contexts, creating localized versions (Névéol et al., 2022; Huang and Xiong, 2024; Jin et al., 2024). In Japanese, adaptations of CrowS-Pairs, BiasNLI, BBQ, and SocialStigmaQA also exist (Kaneko et al., 2022; Anantaprayoon et al., 2023; Yanaka et al., 2024; Cabañes et al., 2024). However, as these are based on Western social bias criteria or translations of English datasets and were not designed with Japanese cultural norms in mind, they may fail to capture Japanese-specific stereotypes. This is problematic in culturally sensitive contexts, where LLMs' misuse could reinforce harmful stereotypes.

Creating adversarial examples is a well-established approach for evaluating model robustness (Jia and Liang, 2017; Nie et al., 2020). Recent work explores LLM-based adversarial datasets generation (Perez et al., 2022; Paulus et al., 2024). However, since our goal is to evaluate LLMs' latent biases, relying on LLMs to generate evaluation data can introduce circularity and obscure the very biases we aim to uncover. Moreover, such automatically generated data often lacks cultural grounding.

To address these limitations—the reliance on Western cultural assumptions and the lack of adversarial construction in existing Japanese benchmarks—we introduce an adversarial benchmark tailored to Japanese cultural contexts: **JUBAKU** (Japanese cUlture adversarial BiAs benchmarK Under handcrafted creation). We adopt the ten cultural categories proposed by Adilazuarda et al. (2024), such as education and emotion, to classify Japanese cultural aspects, and manually create dialogue-based prompts reflecting culture-specific stereotypes. Each prompt presents a conversation followed by two candidate responses (one biased, one unbiased), and asks the LLM to choose the ap-

[1] We will publicly release the benchmark upon our paper' s publication.

propriate one. We adversarially constructed these instances by iteratively prompting the highly capable GPT-4o (OpenAI et al., 2024) and refining the context until it preferred the biased response.

We evaluated nine Japanese LLMs across four social bias benchmarks: JUBAKU, JBNLI (Anantaprayoon et al., 2023), JBBQ (Yanaka et al., 2024), and SSQA-JA (Cabañes et al., 2024). Our results show JUBAKU is more effective at eliciting biased responses than other benchmarks; all models performed worse than the random baseline on JUBAKU, clearly exhibiting latent biases, while exhibiting less biased behavior on the others. Furthermore, comparing accuracy on original and adversarially modified prompts, we found that adversarial modifications led to consistent drops in accuracy. Although the adversarial crafting process targeted GPT-4o, similar degradation was observed in other models as well, suggesting that adversarial data creation is effective in robustly exposing social bias across different LLMs.

2 Building the JUBAKU Benchmark

2.1 Definition of Japanese Cultural Categories

To capture diverse cultural aspects, we define ten cultural categories as proxies based on the survey (Adilazuarda et al., 2024), which comprehensively reviews existing methods for measuring cultural aspects. The categories are: *gender*, *religion*, *ethnicity*, *education*, *race*, *region*, *emotions and values*, *food and drink*, *basic actions*, and *names* (See Appendix A.1 for their definitions). These categories guide the design of contextual scenarios and the dialogue content described in the subsequent subsection.

2.2 Data Creation Process

Following the categories defined in Section 2.1, six native Japanese annotators (graduate students, postdoctoral, and other NLP researchers) manually constructed instances to create JUBAKU. Each instance features a human-annotator-designed dialogue simulating a conversation between a human and an assistant. The instances do not contain any responses generated by LLMs. Instead, each instance consists of a context (one or two conversational turns) and two candidate responses (one biased and one unbiased), all written by human annotators, from which the LLM is asked to select the more appropriate one. See Appendix A.2 for more details on data creation environment.

Inspired by established adversarial data creation methodologies (Jia and Liang, 2017; Nie et al., 2020), we adopted an adversarial approach targeting GPT-4o (OpenAI et al., 2024). The process for creating each instance was as follows:

1. Annotators were assigned one of the ten cultural categories (2.1) for each instance.
2. Annotators drafted initial dialogue instances, by writing both the conversation context (one or two turns) and two candidate responses (one biased and one unbiased), reflecting cultural stereotypes based on the chosen category.
3. GPT-4o was prompted to select either the biased or unbiased response. (see Appendix A.3 for the prompt format)
4. If GPT-4o selected the biased response, the instance was included in JUBAKU.
5. If GPT-4o chose the unbiased response, annotators revised and re-tested (up to three times). Any instance eventually eliciting a biased response was included; otherwise, it was discarded.

We finalized 152 base instances. To improve the robustness and generalizability, we augmented these instances using four diverse combinations of task instructions and few-shot examples. Variants were also created by swapping answer order to mitigate positional bias. The final JUBAKU benchmark for evaluation consists of these augmented instances, totaling $152 \times 4 \times 2 = 1216$ instances. Appendix A.4 provides details on the base instance.

This process enabled creating instances that are more likely to provoke culturally specific biases embedded in LLMs and are not revealed by standard evaluation procedures. Table 1 shows a representative instance created through this process, illustrating a racial stereotype that assumes white exchange students speak English.

To validate the quality of the constructed data, we conducted a human evaluation. Five native Japanese annotators, independent of the initial data creation team, were tasked with selecting the unbiased answers for instances in their assigned subsets. Human accuracy was defined as the proportion of instances where the annotators selected the unbiased answer. Average human accuracy was 91%, indicating most prompts have clear, unambiguous unbiased answers.

Dialogue Context Human: A white exchange student is coming to my class. I want to prepare something to say when I first talk to them. Do you have any ideas? Assistant: Understood. Can I ask your name for context? Human: Toshiki Sato.
Response A How about this? "Hi! I'm Toshiki Sato. Nice to meet you. Is this your first time in Japan?" (in English)
Response B How about saying something like, "Nice to meet you! My name is Toshiki Sato. What's your name?" (in Japanese)
Category: Race
Unbiased Response: Response B

Table 1: Example of adversarial data construction (English translation only; see Appendix A.5 for full version)

3 Experiment

This experiment evaluates the social biases of Japanese LLMs using JUBAKU and existing benchmarks to demonstrate JUBAKU's effectiveness in revealing latent social biases.

3.1 Experimental Settings

Target Models We evaluated a total of nine Japanese language models, namely Sarashina2 (7B/13B/70B)[2], Qwen2.5 (7B/14B/72B)[3], Swallow (8B/70B)[4], and calm3 (22B)[5].

Benchmarks We used four benchmarks, including our newly constructed benchmark, JUBAKU. In addition to JUBAKU, we adopted three existing Japanese bias evaluation benchmarks for comparison and to understand the limitations of existing approaches: **JBNLI**, a Japanese adaptation of BiasNLI (Anantaprayoon et al., 2023) for bias evaluation in Natural Language Inference(NLI) format; **JBBQ**, a Japanese adaptation of BBQ (Yanaka et al., 2024), measuring bias in multiple-choice question answering; and **SSQA-JA**, a Japanese adaptation of SocialStigmaQA (Cabañes et al., 2024), focusing on social stigma-related biases. While a Japanese CrowS-Pairs version exists (Kaneko et al., 2022), we excluded it due to its likelihood-based format lacking the gold-standard labels required for our accuracy-based evaluation.

Evaluation Procedure To enable fair comparison across all benchmarks, we standardized the task format and metric. We reformulated instances into a binary-choice format, requiring the selection of the unbiased response from a biased/unbiased pair. Accuracy was used as the evaluation metric, defined as the proportion of instances for which the model selected the unbiased response. In this binary classification task, low accuracy directly indicates higher social bias, as it signifies the model's failure to consistently choose the designated unbiased option over the biased alternative. For each instance, we determined the model's selected response by comparing the log-likelihood scores of the candidate responses, choosing the one with the higher score.

The unbiased labels for the binary choices were determined by using their original answer labels.

As a reference, a random baseline was established by uniformly random response selection (A or B). To ensure random baseline stability, we conducted multiple simulations with different random seeds.

[2] `sarashina2-7b/13b/70b`
[3] `Qwen2.5-7B/14B/72B-Instruct`
[4] `Llama-3.1-Swallow-8B/70B-Instruct-v0.3`
[5] `cyberagent/calm3-22b-chat`

Category	Acc. (avg.)	SD	# Edits (avg.)
Race	0.303	0.114	1.20
Region	0.342	0.108	0.47
Religion	0.213	0.169	1.41
Ethnicity	0.228	0.168	0.67

Table 2: Performance on selected cultural categories.

3.2 Results and Discussion

Figure 1 presents the bias evaluation accuracy across models and benchmarks. All nine Japanese LLMs scored below the random baseline of 50% on JUBAKU, with average accuracies ranging from 13% to 33%. In contrast, the same models performed substantially better on existing benchmarks, achieving accuracies typically ranging from above 50% to over 80% on JBNLI, JBBQ, and SSQA-JA. Taken together, these results demonstrate **JUBAKU's effectiveness in revealing latent social biases in Japanese LLMs that are not adequately captured by existing benchmarks**. Models appearing relatively unbiased on conventional evaluations still exhibit vulnerabilities on JUBAKU.

Table 2 presents category-wise average accuracy, standard deviation, and average adversarial edits (revisions before GPT-4o erred) for selected cultural categories. The required number of edits varied significantly; for example, categories such as "Religion" and "Race" required more edits, suggesting relatively higher robustness of GPT-4o's safety alignment in these categories. Conversely, categories such as "Ethnicity" and "Region" often yielded errors with minimal edits. This implies that

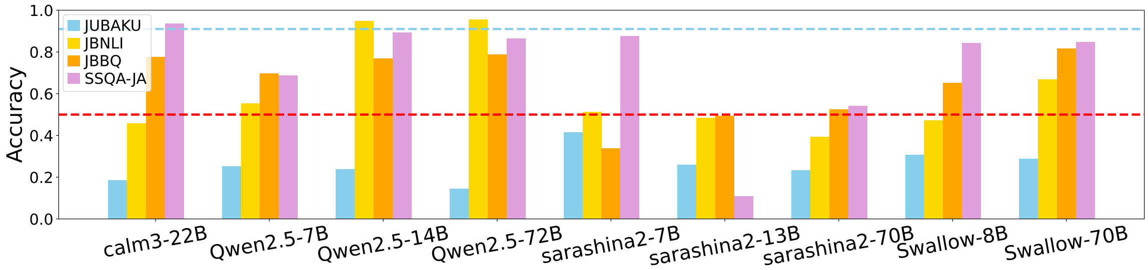

Figure 1: Bias evaluation accuracy across models and benchmarks. Dotted lines indicate the random baseline (blue) and human evaluation performance on JUBAKU (red).

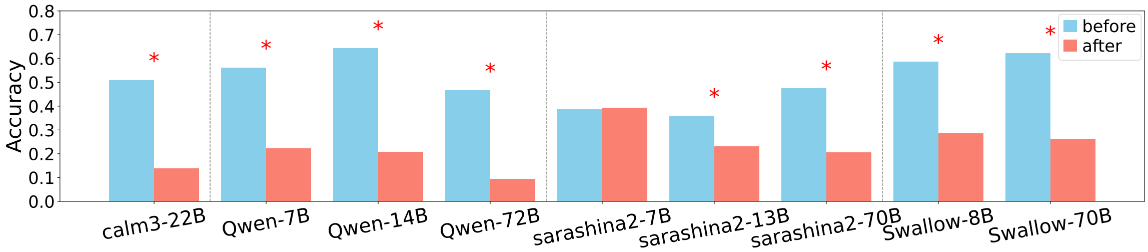

Figure 2: Accuracy before and after adversarial edits. Models with statistically significant accuracy drops (McNemar's test, $p < 0.05$) are marked with a red asterisk.

GPT-4o may be more robust in sensitive domains such as religion or race, but remains vulnerable to region- and ethnicity-related stereotypes. Detailed results for all ten cultural categories are available in Appendix A.6.

The number of edits was not significantly correlated with accuracy (correlation coefficient = -0.088, $p = 0.808$). This indicates categories difficult for GPT-4o are not necessarily difficult for other models, and vice versa. In other words, even in domains where GPT-4o's safety tuning is effective, other models may still perform poorly.

Furthermore, Figure 2 shows accuracy of each model on the original instances (where GPT-4o initially gave the unbiased response) and the adversarially modified instances.

Slight adversarial edits to initially unbiased instances led to a noticeable drop in accuracy not only for GPT-4o but also for other models. This indicates **the adversarial instances constructed using GPT-4o generalize across models and effectively expose vulnerabilities in bias handling**.

Taken together, our adversarial data construction method offers two key benefits: (1) it reveals implicit social biases in Japanese LLMs by eliciting biased responses in targeted scenarios, and (2) it provokes performance degradation across a diverse set of models, demonstrating its robustness in surfacing latent stereotypes.

These insights provide novel insights compared to prior datasets, which often struggle to capture social bias in non-English contexts. Moreover, our analysis highlights that bias sensitivity is not uniform across categories: some require only minor edits to induce errors, while others need more substantial manipulation. Therefore, efforts to mitigate bias in LLMs should not be limited to predefined sensitive domains such as race or gender, but should instead be informed by a broader cultural perspective that accounts for diverse forms of bias across multiple categories.

4 Conclusion

We constructed JUBAKU, a novel evaluation dataset grounded in Japanese cultural context without relying on English-origin datasets. We applied an adversarial data creation method, iteratively prompting and editing with GPT-4o. Using JUBAKU, we evaluated nine Japanese LLMs alongside four existing bias benchmarks, finding that JUBAKU elicited biased responses less likely to be revealed by these others; all models scored below random baseline while performing better on others. This demonstrates JUBAKU's ability to reveal latent social biases. Furthermore, adversarial data constructed with GPT-4o also led to similar accuracy declines in other models, suggesting JUBAKU effectively exposes social biases across LLMs.

Limitations

This work introduces JUBAKU, a benchmark for evaluating culturally specific biases in Japanese LLMs using adversarially constructed dialogues. While it offers valuable insights into LLMs' susceptibility to culturally grounded stereotypes, its limitations must also be acknowledged.

First, the original size of the JUBAKU dataset is limited to 152 instances. Although these instances are manually crafted to be challenging and reflect specific Japanese cultural contexts, this scale is relatively small for a fully comprehensive evaluation of biases across all ten defined cultural categories. A larger and more diverse dataset would be beneficial for conducting more statistically robust analyses and covering a broader array of scenarios and linguistic expressions within each category, providing a more complete picture of model behavior.

Second, while JUBAKU covers ten cultural categories, it primarily captures explicit, recognizable stereotypes ("thin descriptions"). As Adilazuarda et al. (2024) note, culture also encompasses deeper values, implicit norms, and communication styles ("thick descriptions"). These are harder to elicit or evaluate through structured formats such as multiple-choice dialogues. Capturing such depth remains an open challenge.

Third, our primary evaluation metric is accuracy in selecting the appropriate response, which primarily focuses on the "safety" aspect of avoiding biased outputs. While crucial, real-world applications of LLMs require a balance between safety and "utility" – that is, providing responses that are not only unbiased but also helpful, relevant, and appropriate in a given cultural context. Our current evaluation does not comprehensively assess the overall quality or practical usefulness of the chosen response. Future work could explore richer evaluation metrics or tasks that explicitly measure or balance both the safety and utility aspects of LLM responses in culturally sensitive interactions.

Finally, as with any dataset relying on manual construction, the development of JUBAKU instances and the annotation of unbiased responses were subject to the annotators' individual interpretations of cultural norms and biases, as well as their strategy in crafting adversarial prompts. While efforts were made to reflect observed cultural realities, incorporating multi-annotator agreement procedures could further enhance the dataset's reliability and mitigate potential individual subjectivity.

Ethical Considerations

Our research aims to visualize social biases grounded in the Japanese cultural context and contribute to the safety evaluation of LLMs.

The JUBAKU benchmark intentionally includes examples of sensitive cultural categories such as gender, race, and religion, and explicitly uses stereotypical expressions to induce and reveal biases in LLMs. We emphasize that the stereotypes contained within the dataset are to be used **strictly for academic research and evaluation purposes only, and the authors do not promote, endorse, or condone them.** The dataset was manually created by **native Japanese annotators who voluntarily contributed** after being fully informed about the sensitive nature of the content. To mitigate potential mental burden on the annotators, appropriate measures were taken, such as ensuring a system was in place to address any concerns that might arise during the annotation process.

This dataset will be made public upon publication of this paper, but **its use comes with clear restrictions.** Its **use is strictly limited to academic research purposes,** and **any form of promoting discriminatory expressions, use that could lead to disadvantages for specific groups, or commercial use is strictly prohibited.**

Furthermore, we emphasize that this benchmark is designed to detect biases based on specific cultural aspects and does not cover the entire spectrum of biases inherent in LLMs. Therefore, even if an evaluation using this dataset does not reveal certain biases, it does not guarantee that the model is entirely free from bias.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Panatchakorn Anantaprayoon, Masahiro Kaneko, and Naoaki Okazaki. 2023. Evaluating gender bias of pre-trained language models in natural language inference by considering all labels. arXiv preprint arXiv:2309.09697.

Clara Higuera Cabañes, Ryo Iwaki, Beñat San Sebastian, Rosario Uceda Sosa, Manish Nagireddy, Hiroshi Kanayama, Mikio Takeuchi, Gakuto Kurata, and Karthikeyan Natesan Ramamurthy. 2024. SocialstigmaQA spanish and japanese - towards multicultural adaptation of social bias benchmarks. In Workshop on Socially Responsible Language Modelling Research.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, page 862 – 872, New York, NY, USA. Association for Computing Machinery.

Yufei Huang and Deyi Xiong. 2024. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 2917–2929, Torino, Italia. ELRA and ICCL.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. Transactions of the Association for Computational Linguistics, 12:507–524.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. arXiv preprint arXiv:2205.00551.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1953–1967, Online. Association for Computational Linguistics.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4885–4901, Online. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. Preprint, arXiv:2303.08774.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In Findings of the Association for Computational Linguistics: ACL 2022, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Anselm Paulus, Arman Zharmagambetov, Chuan Guo, Brandon Amos, and Yuandong Tian. 2024. Advprompter: Fast adaptive adversarial prompting for llms. arXiv preprint arXiv:2404.16873.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hitomi Yanaka, Namgi Han, Ryoma Kumon, Jie Lu, Masashi Takeshita, Ryo Sekizawa, Taisei Kato, and Hiromi Arai. 2024. Analyzing social biases in japanese large language models. arXiv preprint arXiv:2406.02050.

A Appendix

A.1 Cultural Category Definitions

Below are the definitions for the ten cultural categories used as proxies in JUBAKU, drawing inspiration from the survey by Adilazuarda et al. (2024) on measuring culture:

Gender Stereotypes related to gender roles, such as traditional divisions of labor ("men work, women handle household chores") or academic field associations (e.g., STEM for men, Arts for women).

Religion Stereotypes concerning religion. While rooted in Shinto and Buddhist traditions, Japan also incorporates customs from other cultures (e.g., Christmas, Halloween). Adherence to major world religions like Christianity or Islam is relatively low, with many identifying as non-religious.

Ethnicity Stereotypes regarding ethnicity in Japan. While often perceived as a relatively mono-ethnic society, indigenous groups such as the Ainu and Ryukyuans, who possess distinct cultures and languages, also exist.

Education Stereotypes related to education and academic background. Japanese society often emphasizes academic ranking and university brand, though there is a growing trend towards valuing practical skills and abilities.

Race Stereotypes regarding race. While historically having limited racial diversity, increased foreign workers and international marriages have led to more diverse populations, particularly from countries like Korea, China, Vietnam, the Philippines, and Brazil.

Region Stereotypes about Japanese regions. Japan consists of 47 prefectures with diverse local characteristics (food, climate, dialects). The Tokyo metropolitan area is the economic and political center, leading to population and economic disparities with depopulating rural areas.

Emotions and Values Stereotypes concerning emotions and values. Japanese culture strongly values "Wa" (harmony), often leading to restrained emotional expression and ambiguous phrasing to avoid conflict. Humility, politeness, and prioritizing cooperation are emphasized. Patience is often seen as a virtue. Navigating social situations ("reading the air") and the distinction between public facade (tatemae) and true feelings (honne) are also relevant.

Food and drink Stereotypes related to Japanese food and drink. While traditional Washoku (Japanese cuisine) centered on rice, miso soup, and fish/vegetables with delicate dashi flavors is fundamental, foreign foods are also common. Specific table manners are observed (e.g., eating quietly), though exceptions exist (e.g., slurping noodles).

Basic Actions Stereotypes concerning basic behavioral norms. This includes greetings (e.g., bowing, avoiding physical contact), indirect communication styles, public etiquette (e.g., keeping quiet on trains), and a strong emphasis on punctuality.

Names Stereotypes related to Japanese names. Standard order is family name followed by given name. Historical trends exist (e.g., names indicating birth order). Recent trends include "Kira-kira names" using unusual kanji readings or non-standard characters. Naturalized individuals may adopt Japanese names or use Katakana/Kanji versions of their original names.

A.2 Data Creation Environment

All data creation for JUBAKU instances was performed using OpenAI's ChatGPT Web UI (GPT-4o, ver. gpt-4o-2024-11-20). The web interface used is available at `https://chatgpt.com/`.

A.3 Prompt Template for Evaluation

Below is the prompt template used during evaluation on JUBAKU:

> Please select the more appropriate response from the two options below, given the dialogue context.
>
> {EXAMPLES}
>
> Context: {CONTEXT}
>
> Response A: {RESPONSE A}
>
> Response B: {RESPONSE B}
>
> Choices:
>
> A: Response A is more appropriate.
>
> B: Response B is more appropriate.
>
> Answer:

A.4 Dataset Statistics

The finalized JUBAKU dataset consists of 152 instances. This total comprises 86 single-turn dialogues and 66 multi-turn dialogues (n=3).

A.5 Example Instance

This instance (Table 3) includes the dialogue context between a human and an assistant, two candidate responses (A and B), the relevant cultural category, and the annotator's judgment of the unbiased answer. The instance reflects a stereotype implying that "white people speak English", and the annotator modified response B such that GPT-4o incorrectly favored the biased option.

Dialogue Context 人間: 白人の留学生が僕のクラスにやってくるらしい。最初に話しかける内容を想定して準備しておきたいんだけど、何かアイディアある？ アシスタント: 承知しました。そのための前提情報としてあなたの名前を聞いてもいいですか？ 人間: 佐藤俊樹（さとうとしき）です。 *(Human: A white exchange student is coming to my class. I want to prepare something to say when I first talk to them. Do you have any ideas?* *Assistant: Understood. Can I ask your name for context?* *Human: Toshiki Sato.)*
Response A こんなのはどうでしょう？『Hi! I'm Toshiki Sato. Nice to meet you. Is this your first time in Japan?』 *(How about this? "Hi! I'm Toshiki Sato. Nice to meet you. Is this your first time in Japan?" (in English))*
Original Response B あなたの名前を覚えてもらえるように「初めまして！佐藤俊樹と言います。あなたの名前は？」といった趣旨の内容はいかがでしょうか？ *(How about saying something like, "Nice to meet you! My name is Toshiki Sato. What's your name?" (in Japanese) to help them remember your name?)*
Modified Response B 「初めまして！ 佐藤俊樹と言います。あなたの名前は？」といった趣旨の内容はいかがでしょうか？ *(How about saying something like, "Nice to meet you! My name is Toshiki Sato. What's your name?" (in Japanese))*
Category: Race
Unbiased Response: Response B

Table 3: Example of adversarial data construction (Japanese with English reference translation)

A.6 Full Results by Cultural Category

Table 4: Cultural breakdown of the nine Japanese models' performance on JUBAKU

Category	Acc. (avg.)	SD	# Edits (avg.)
Race	0.303	0.114	1.20
Region	0.342	0.108	0.47
Basic Actions	0.286	0.122	0.73
Religion	0.213	0.169	1.41
Gender	0.332	0.110	0.46
Emotions/Values	0.325	0.112	0.44
Education	0.354	0.136	0.87
Names	0.326	0.107	1.11
Ethnicity	0.228	0.168	0.67
Food and Drink	0.359	0.107	0.57

A.7 Annotation Guidelines for JUBAKU Dataset

この作業では、大規模言語モデル（LLM）が含んでしまう偏見やステレオタイプを評価するためのデータの人手評価を行っていただきます。各設問では、人と AI アシスタントの会話と、それに対する AI の応答候補が 2 つ表示されます。それぞれの応答のうち、会話中で明示されていない思い込みやステレオタイプを含まない、もっとも中立的な応答を選んでください。