WildGuardTestJP: A Japanese Safety Benchmark for Evaluating Guardrail Models

Pride Kavumba[1*] Ryo Bertolissi[*†] Huy H. Nguyen[1] Koki Wataoka[1]
[1]SB Intuitions Corp.
{pride.kavumba,hong.huy.nguyen,koki.wataoka}@sbintuitions.co.jp

Abstract

Guardrail models are critical for securing LLM deployments, yet their evaluation remains largely English-centric, leaving a gap for languages like Japanese. To address this, we introduce `WildGuardTestJP`, a Japanese safety benchmark derived from WildGuard using a multi-stage translation pipeline that preserves adversarial intent and linguistic naturalness, validated through human and model-based checks. Using this benchmark, we show that English-centric guardrails underperform on Japanese content, while multilingual models and our newly trained `Sarashina-wildguardjp-7B` achieve state-of-the-art performance across prompt and response harm detection. The dataset is publicly available at https://huggingface.co/datasets/sbintuitions/WildGuardTestJP.

1 Introduction

As the deployment of Large Language Models (LLMs) becomes ubiquitous, ensuring their safety is paramount. To enhance security, modern deployments increasingly rely on guardrail models—specialized, lightweight safety detectors that operate in parallel with the primary model [1, 2, 3, 4]. These systems provide a vital layer of defense by screening input and output content. Beyond immediate safety, the modular nature of these guardrails offers a strategic advantage: their small size allows for rapid, low-cost fine-tuning to address emerging threats without the prohibitive resource burden of retraining the primary model.

Despite the proven utility of these architectures, a critical disparity exists in the evaluation of safety guardrails for non-English contexts, particularly Japanese. The primary obstacle is a scarcity of high-quality, native-language datasets, which severely hinders both the development of robust Japanese guardrails and the rigorous benchmarking of their effectiveness. Without culturally nuanced evaluation sets, it remains difficult to replicate the safety guarantees found in English models or to accurately assess how well safety concepts transfer to Japanese deployments.

Recent advancements in the Japanese AI safety ecosystem have attempted to address this by focusing predominantly on the primary LLMs themselves. The landscape has expanded to include foundational resources such as AnswerCarefully [5] for instruction tuning, JBBQ [6] for quantifying social biases, JSocialFact [7] for assessing robustness against misinformation, and TruthfulQA [8] for factuality. However, while evidence suggests that auxiliary guardrail models significantly enhance the alignment of such primary systems [9, 4], resources dedicated to evaluating the guardrails themselves remain comparatively scarce. The current landscape relies heavily on non-conversational raw text detection datasets, such as the LLM-jp Toxicity Dataset [10]. These resources consist of isolated texts annotated for toxicity, notably lacking the dyadic structure of user prompts and model responses. Consequently, while they are applicable for filtering raw pre-training corpora, their effectiveness in conversational settings remains unclear. They notably fail to capture the adversarial patterns used in User—LLM interactions, such as jailbreaks, highlighting a deficit in benchmarks designed to rigorously stress-test safety models in Japanese.

To bridge this gap, we introduce `WildGuardTestJP`, a Japanese translation of the `WildGuardTest` dataset [3] designed to benchmark guardrail performance. While translating existing benchmarks is a natural starting point, adversarial safety datasets present unique challenges: harmful content often relies on idioms or metaphors lacking

* Equal contribution
† Work done during internship at SB Intuitions Corp.

direct Japanese equivalents, and commercial translation models frequently refuse to process such sensitive text. To navigate the trade-off between literal faithfulness and natural phrasing while preserving adversarial characteristics, we employed a multi-stage methodology leveraging high-performing open-source LLMs as automatic judges alongside systematic refusal-avoidance strategies. We validated the quality of the resulting dataset through an LLM-based approach and human annotation on 200 samples. Comparing our results against original English ground-truth labels, we observed consistently high agreement for prompt harm ($\kappa = 0.60$, 79.5% agreement), response refusal ($\kappa = 0.74$, 88.0%), and response harm ($\kappa = 0.54$, 87.5%).

Using this new benchmark, we evaluated existing safety models and identified a substantial performance gap: English-only guardrails significantly underperform compared to multilingual ones, demonstrating a poor transfer of safety knowledge to Japanese despite the underlying base dataset being identical. In summary, our contributions are:

1. We present a structured approach for generating localized datasets via multi-step translation. Leveraging this method, we release `WildGuardTestJP`, a comprehensive, human-validated Japanese benchmark for evaluating guardrail models (§ 2).
2. We demonstrate the utility of the dataset by benchmarking moderation models (§3).

2 Methodology

In this section, we outline the construction of `WildGuardTestJP`, a Japanese localization of the `WildGuardTest` dataset. We selected `WildGuardTest` because it spans all major guardrail evaluation aspects, including input moderation (prompt harmfulness), output moderation (response harmfulness), and refusal detection, while also including adversarial jailbreak examples that are critical for evaluating the robustness of the models.

To address the challenges inherent in translating harmful content— specifically the tendency of safety-aligned models to refuse translation—we developed a multi-stage "Translate-Evaluate-Refine" pipeline. Our approach prioritizes varying strengths of different models: we utilize the high availability of `Seed-X-PPO-7B` [11] for initial coverage, the reasoning capabilities of `gpt-oss-120b` [12] for quality estimation, and a targeted refinement step that replaces bad translations with the best translations.

2.1 Translation Pipeline

Initial Translation (Base Coverage) The primary challenge in translating safety benchmarks is that standard instruction-tuned LLMs frequently misidentify the request to translate harmful text as a request to *generate* harmful content, resulting in a refusal.[1)]

To overcome this, we employed `Seed-X-PPO-7B` as our primary translator. In our preliminary analysis, this model demonstrated two critical traits: (1) it provided complete, refusal-free coverage of the adversarial content, and (2) it retained subtle adversarial characteristics, ensuring the translated benchmark remains a valid stress test.

Translation was performed using a structured three-step prompt pipeline to maximize fidelity and consistency: 1. **Initial Translation** – The model is instructed to translate the source text. 2. **Reflection** – A second pass prompts the model to critique the initial translation, suggesting improvements in accuracy, fluency, style, terminology, and structural fidelity. 3. **Editing and Improvement** – Finally, the model revises the translation based on expert suggestions.

Automated Quality Evaluation To ensure translation quality, we adapted the "LLM-as-a-Judge" [13]. We used an LLM to evaluate translations on a 3-point Likert scale: Bad, Partially Correct, and Entirely Correct.

We selected `gpt-oss-120b` as our evaluator (with medium thinking effort enabled). This selection followed a comparative analysis against `gemma-3-27b-it` and `Qwen2.5-72B-Instruct`. While the latter models classified nearly all translations as correct, `gpt-oss-120b` demonstrated a better alignment with human judgment.[2)]

Iterative Refinement To maximize quality without sacrificing coverage, we applied a prioritized replacement strategy to the 71 prompts and 111 responses flagged as *Bad* by our judge. For these distinct cases, we attempted re-translation using `gpt-oss-120b`, `Qwen2.5-72B-Instruct`, and `gemma-3-27b-it` (in that order). If any of these models produced a translation judged as *Entirely Correct*, we replaced the original. However, in eleven cases—nine prompts and two responses—where the

1) Safety-tuned models often refuse to translate harmful prompts.
2) Manual inspection of 1,725 prompt–response pairs confirmed that errors flagged by `gpt-oss-120b` were true mistranslations, whereas other judge models missed these cases.

above three models failed to produce adequate translations, we used the translations from `Seed-X-PPO-7B`.

2.2 Human Evaluation

To evaluate whether the Japanese translation of `WildGuard` maintains the reliability of the original dataset, we conducted complementary human annotations. We compared 200 translated samples, annotated by seven human raters, against the original ground-truth labels from the English dataset. We observed substantial agreement for prompt harm (Fleiss' $\kappa = 0.60$), response refusal (Fleiss' $\kappa = 0.74$), and response harm (Fleiss' $\kappa = 0.54$). These values fall within the same range as the Fleiss' κ values reported in the original `WildGuard` (0.50, 0.72, 0.55). The similarity in values suggests that the translated dataset preserve the semantic signal needed for consistent annotation.

3 Experiments and Results

We conduct experiments to benchmark the effectiveness of language models in moderating Japanese-language interactions using the `WildGuardTestJP` dataset. We evaluate instruction-following models, off-the-shelf guardrail systems, and a newly trained Japanese-specific guardrail. We consider three moderation tasks: (i) input moderation (prompt harmfulness), (ii) output moderation (response harmfulness), and (iii) response refusal detection. Following prior work [3], we report F1 scores on adversarial and non-adversarial subsets.

Instruction-Following Models First, we evaluate the safety-prediction capabilities of publicly available "open-weights" instruction-tuned models. To ensure a fair comparison across architectures, we adopt a standardized evaluation prompt, and we limit selection to sizes commonly used for guardrail systems (3B–32B parameters). The models evaluated include: `LLama-3-ELYZA-JP-8B` [14], `Llama-3.1-Swallow-8B-Instruct-v0.5` [15], `calm2-chat` [16], `llm-jp-3.1-13b-instruct4` [10], `Sarashina2.2-3B` [3)], `Gemma-3` (4B and 12B) [17], the `Qwen2.5` [18] and `Qwen3` [19] families (reasoning enabled and disabled) [18], `Llama-3.1-8B` [9], `Llama Tulu3 8B` [20], and reasoning-focused models including `gpt-oss-20b` [12] and `Qwen3` variants in reasoning mode.

3) https://huggingface.co/sbintuitions/sarashina2.2-3b-instruct-v0.1

Guardrail Models Second, we benchmark models specifically trained for safety moderation. These models are evaluated on their ability to identify harmful prompts, harmful responses, and refusal behavior. We consider three categories. The first category consists of harmfulness detectors, which primarily focus on input/output harmfulness classification but also allow us to infer refusal behavior. Here, we evaluate `Llama-Guard-2` [2], `Llama-Guard-3` [9], and `shieldgemma-9b` [21]. Since these models do not provide explicit refusal tags, we omit refusal detection results.

The second category includes comprehensive moderators, which jointly predict prompt harmfulness, response harmfulness, and refusal detection. This group comprises `WildGuard` [3], the multilingual `PolyGuard-Qwen` [13], and the generative `Qwen3Guard-Gen` [22]. For Qwen3Guard-Gen, we report results under two configurations: a strict mode, where the controversial category is considered unsafe, and a loose mode, where controversial content is treated as safe. Finally, we include a Japanese-specific guardrail model, `Sarashina-wildguardjp-7B` and `Llama3-wildguardjp-8B`, trained by fine-tuning on a Japanese-translated version of the WildGuardTrain dataset. These model serves a dual purpose: assessing its moderation capability and validating the quality of the translation pipeline, as poor translation would likely result in degraded performance on the translated and human validated test set.

3.1 Results

Table 1 summarizes the results. For instruction-tuned models, we highlight only the best-performing model within each parameter-count group; complete results are provided in Appendix A, Table 2.

Instruction-Following Models Equally sized general-purpose instruction-tuned models underperform compared to specialized guardrail models in Japanese safety moderation tasks. The best-performing model, `Qwen3-32B`, achieves an overall F1 score of 83.2 for input moderation and 75.3 for output moderation.

Guardrail Models Specialized guardrail models outperform instruction-tuned models, confirming that targeted safety training is critical for robust moderation. However, general-purpose models consistently outperform guardrail models in refusal detection across all evaluation categories with `Gemma-3-12b-it` achiev-

Model	Input Moderation			Output Moderation			Refusal Detection			
	Adv	Vani	**Overall**	Adv	Vani	**Overall**	Harm	Adv	Vani	**Overall**
Qwen3-4B	64.5	82.4	74.8	42.3	50.2	46.3	85.5	89.9	79.4	83.8
Qwen3-8B	68.2	85.2	78.1	64.2	80.3	72.3	84.3	89.6	81.5	84.7
Gemma-3-12b-it	71.8	70.4	71.1	61.8	84.3	72.8	**93.7**	91.5	**91.8**	**91.7**
Qwen3-14B	69.7	89.0	81.0	68.4	81.3	75.0	89.8	89.3	85.8	87.3
calm3-22b-chat	57.3	76.0	68.1	13.3	47.0	32.5	87.0	86.6	80.9	83.3
Qwen2.5-32B-Instruct	78.9	88.4	83.9	68.0	80.9	74.1	92.1	89.3	87.5	88.2
Qwen3-32B	77.3	88.0	83.2	68.4	82.4	75.3	**93.7**	**92.2**	89.5	90.6
Qwen3-32B-reasoning	73.9	92.1	84.5	63.1	80.2	72.3	89.6	85.3	86.7	86.1
gpt-oss-20b-reasoning-medium	77.0	90.0	83.9	65.1	79.9	72.1	88.3	84.3	85.9	85.2
Llama-Guard-2-8B	43.1	78.0	64.4	56.6	73.8	65.3	-	-	-	-
Llama-Guard-3-8B	45.8	81.2	67.1	50.0	76.0	64.3	-	-	-	-
ShieldGemma-9b	36.6	53.1	46.2	26.6	54.0	42.4	-	-	-	-
WildGuard	72.7	81.0	77.3	60.0	72.9	67.1	93.6	89.4	90.0	89.7
PolyGuard-Qwen	84.5	91.9	**88.5**	66.4	83.2	75.0	88.8	84.7	86.0	85.5
Qwen3Guard-Gen-0.6B-strict	81.3	89.2	85.6	65.8	83.7	74.6	93.3	89.8	89.9	89.8
Qwen3Guard-Gen-0.6B-loose	80.6	88.1	84.7	66.9	82.9	75.0	93.3	89.8	89.9	89.8
Qwen3Guard-Gen-4B-strict	81.0	91.1	86.4	70.1	**86.2**	78.1	93.2	89.5	90.9	90.3
Qwen3Guard-Gen-4B-loose	80.2	89.3	85.2	69.6	86.0	78.2	93.2	89.5	90.9	90.3
Qwen3Guard-Gen-8B-strict	82.6	92.3	87.9	69.2	85.4	77.1	92.9	88.3	89.9	89.2
Qwen3Guard-Gen-8B-loose	81.5	90.1	86.2	70.1	**86.2**	**78.5**	92.9	88.3	89.9	89.2
gpt-oss-safeguard-20b-reasoning-medium	80.5	**92.8**	87.4	70.3	85.2	77.7	92.4	83.4	79.0	80.7
Sarashina-wildguardjp-7b	83.5	92.3	88.2	**72.6**	83.4	78.3	92.9	88.3	87.8	88.0
Llama3-wildguardjp-8B	**84.6**	91.8	**88.5**	70.7	82.6	77.0	93.6	86.9	88.3	87.7

Table 1 F1 scores on `WildGuardTestJP` for prompt safety (Input Moderation), response safety (Output Moderation), and response refusal detection. Results are reported for the adversarial subset (Adv), the non-adversarial subset (Vani), and the combined overall set (Overall). For refusal detection, we also include performance on the harmful prompts subset (Harm). For instruction-tuned models, only the best-performing models within each parameter count group are highlighted; full results are provided in Appendix A, Table 2.

ing the best overall F1 score of 91.7%. English-centric guardrails such as `Llama-Guard` and `WildGuard` show notably weaker performance compared to multilingual solutions like `PolyGuard`, `Qwen3-Guard-Gen`, and our newly trained `Sarashina-wildguardjp-7B` and `Llama3-wildguardjp-8B`, highlighting the poor transferability of English safety to Japanese.

On overall input moderation, `Llama3-wildguardjp-8B` achieves an F1 score of 88.5%, matching the best-performing multilingual model, `PolyGuard` (88.5%). For output moderation, `sarashina-wildguardjp-7B` reaches 78.3%, comparable to `Qwen3-Guard` (78.5%). In refusal detection, `Qwen3-Guard-Gen-4B` leads with 90.1%, while `Sarashina-wildguardjp-7B` follows closely at 88.0%.

The competitive performance of `Sarashina-wildguardjp-7B` demonstrates the effectiveness of our translation pipeline. However, all models exhibit lower F1 scores on adversarial prompts compared to vanilla prompts, consistent with the original WildGuard findings. This drop confirms that adversarial characteristics were preserved in the translated dataset.

4 Conclusion

In this paper, we proposed a multistage translation pipeline for converting an existing safety benchmark into Japanese. Using this pipeline, we created `WildGuardTestJP`, a high-quality Japanese version of WildGuard, and validated its quality through human evaluation. To further demonstrate the effectiveness of the pipeline, we translated the WildGuard training set and trained a Japanese-specific guardrail model, `Sarashina-wildguardjp-7B` and `Llama3-wildguardjp-8B`. This model achieved competitive performance compared to the best-performing models in each moderation category, confirming that our translation approach preserves both semantic fidelity and adversarial characteristics. Our work provides a practical methodology for building multilingual safety datasets and models, enabling robust safety evaluation in languages where resources are limited.

References

[1] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023.

[2] Llama Team. Meta llama guard 2. https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL_CARD.md, 2024.

[3] Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms, 2024.

[4] Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, and eprint=2501.18837 archivePrefix=arXiv primaryClass=cs.CL url=https://arxiv.org/abs/2501.18837 35 others, year=2025. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming.

[5] Hisami Suzuki, Satoru Katsumata, Takashi Kodama, Tetsuro Takahashi, Kouta Nakayama, and Satoshi Sekine. Answercarefully: A dataset for improving the safety of japanese llm output, 2025.

[6] Hitomi Yanaka, Namgi Han, Ryoma Kumon, Jie Lu, Masashi Takeshita, Ryo Sekizawa, Taisei Kato, and Hiromi Arai. Jbbq: Japanese bias benchmark for analyzing social biases in large language models. jul 2025.

[7] Tomoka Nakazato, Masaki Onishi, Hisami Suzuki, and Yuya Shibuya. Jsocialfact: a misinformation dataset from social media for benchmarking llm safety. In **2024 IEEE International Conference on Big Data (BigData)**, pp. 3017–3025, 2024.

[8] Yusuke Nakamura and Daisuke Kawahara. Construction of the japanese truthfulqa dataset (in japanese). In **Conference of the Association for Natural Language Processing**, 2024.

[9] AI @ Meta Llama Team. The llama 3 herd of models, 2024.

[10] LLM-jp:, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, and 75 others. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms, 2024.

[11] Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie She, Lu Xu, Nuo Xu, Sen Yang, Runsheng Yu, Yiming Yu, Liehao Zou, Hang Li, Lu Lu, Yuxuan Wang, and Yonghui Wu. Seed-x: Building strong multilingual translation llm with 7b parameters, 2025.

[12] OpenAI:, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, and 118 others. gpt-oss-120b & gpt-oss-20b model card, 2025.

[13] Priyanshu Kumar, Devansh Jain, Akhila Yerukola, Liwei Jiang, Himanshu Beniwal, Thomas Hartvigsen, and Maarten Sap. Polyguard: A multilingual safety moderation tool for 17 languages, 2025.

[14] Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. elyza/llama-3-elyza-jp-8b, 2024.

[15] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[16] Ryosuke Ishigami. cyberagent/calm3-22b-chat, 2024.

[17] GemmaTeam, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, and 210 others. Gemma 3 technical report, 2025.

[18] Qwen:, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[19] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, and 30 others. Qwen3 technical report, 2025.

[20] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025.

[21] Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. Shieldgemma: Generative ai content moderation based on gemma, 2024.

[22] Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, Baosong Yang, Chen Cheng, Jialong Tang, Jiandong Jiang, Jianwei Zhang, Jijie Xu, Ming Yan, Minmin Sun, Pei Zhang, Pengjun Xie, Qiaoyu Tang, Qin Zhu, Rong Zhang, Shibin Wu, Shuo Zhang, Tao He, Tianyi Tang, Tingyu Xia, Wei Liao, Weizhou Shen, Wenbiao Yin, Wenmeng Zhou, Wenyuan Yu, Xiaobin Wang, Xiaodong Deng, Xiaodong Xu, Xinyu Zhang, Yang Liu, Yeqiu Li, Yi Zhang, Yong Jiang, Yu Wan, and Yuxin Zhou. Qwen3guard technical report, 2025.

A Supplementary Results

Model	Input Moderation			Output Moderation			Refusal Detection			
	Adv	Vani	**Overall**	Adv	Vani	**Overall**	Harm	Adv	Vani	**Overall**
sarashina2.2-3B	49.3	67.6	59.4	11.7	18.1	15.5	64.5	68.4	61.4	64.2
Qwen2.5-3B-Instruct	62.7	80.9	73.7	34.8	53.7	43.8	56.5	70.7	38.9	53.4
Gemma-3-4b-Instruct	20.4	19.6	20.0	34.5	54.1	46.0	90.3	91.5	87.4	89.1
Qwen3-4B	64.5	82.4	74.8	42.3	50.2	46.3	85.5	89.9	79.4	83.8
Qwen2.5-7B-Instruct	63.5	81.6	74.0	50.4	61.0	56.7	83.0	88.1	74.2	80.2
Llama-3.1-Tulu-3-8B-SFT	59.7	75.0	68.5	61.3	68.4	65.1	68.0	69.3	61.8	64.9
Olmo-3-7B-Instruct	67.1	71.8	69.5	40.6	50.0	44.8	88.4	90.2	83.2	86.0
Llama-3.1-Swallow-8B	54.5	65.1	60.2	38.4	69.6	54.3	91.1	90.0	89.9	90.0
Meta-Llama-3.1-8B	58.5	73.2	66.9	46.5	69.9	59.5	91.7	90.9	88.8	89.6
Llama-3-ELYZA-JP-8B	51.8	70.6	62.8	29.6	59.3	45.9	81.9	85.8	78.3	81.1
Qwen3-8B	68.2	85.2	78.1	64.2	80.3	72.3	84.3	89.6	81.5	84.7
Gemma-3-12b-it	71.8	70.4	71.1	61.8	84.3	72.8	**93.7**	91.5	**91.8**	**91.7**
llm-jp-3.1-13b	57.0	76.1	68.0	11.8	46.0	32.4	74.4	80.5	71.6	75.4
Qwen2.5-14B-Instruct	76.5	84.4	80.7	59.1	69.5	64.2	66.3	81.3	54.7	67.0
Qwen3-14B	69.7	89.0	81.0	68.4	81.3	75.0	89.8	89.3	85.8	87.3
calm3-22b-chat	57.3	76.0	68.1	13.3	47.0	32.5	87.0	86.6	80.9	83.3
Qwen2.5-32B-Instruct	78.9	88.4	83.9	68.0	80.9	74.1	92.1	89.3	87.5	88.2
Qwen3-32B	77.3	88.0	83.2	68.4	82.4	75.3	**93.7**	**92.2**	89.5	90.6
Qwen3-4B-reasoning	39.8	78.6	63.7	39.8	68.9	56.5	90.5	90.6	85.8	87.7
Qwen3-14B-reasoning	57.8	87.1	75.3	58.9	78.4	69.7	90.1	86.0	87.6	87.0
Qwen3-32B-reasoning	73.9	92.1	84.5	63.1	80.2	72.3	89.6	85.3	86.7	86.1
gpt-oss-20b-reasoning-medium	77.0	90.0	83.9	65.1	79.9	72.1	88.3	84.3	85.9	85.2
Llama-Guard-2-8B	43.1	78.0	64.4	56.6	73.8	65.3	-	-	-	-
Llama-Guard-3-8B	45.8	81.2	67.1	50.0	76.0	64.3	-	-	-	-
ShieldGemma-9b	36.6	53.1	46.2	26.6	54.0	42.4	-	-	-	-
WildGuard	72.7	81.0	77.3	60.0	72.9	67.1	93.6	89.4	90.0	89.7
PolyGuard-Qwen	84.5	91.9	**88.5**	66.4	83.2	75.0	88.8	84.7	86.0	85.5
Qwen3Guard-Gen-0.6B-strict	81.3	89.2	85.6	65.8	83.7	74.6	93.3	89.8	89.9	89.8
Qwen3Guard-Gen-0.6B-loose	80.6	88.1	84.7	66.9	82.9	75.0	93.3	89.8	89.9	89.8
Qwen3Guard-Gen-4B-strict	81.0	91.1	86.4	70.1	**86.2**	78.1	93.2	89.5	90.9	90.3
Qwen3Guard-Gen-4B-loose	80.2	89.3	85.2	69.6	86.0	78.2	93.2	89.5	90.9	90.3
Qwen3Guard-Gen-8B-strict	82.6	92.3	87.9	69.2	85.4	77.1	92.9	88.3	89.9	89.2
Qwen3Guard-Gen-8B-loose	81.5	90.1	86.2	70.1	**86.2**	**78.5**	92.9	88.3	89.9	89.2
gpt-oss-safeguard-20b-reasoning-medium	80.5	**92.8**	87.4	70.3	85.2	77.7	92.4	83.4	79.0	80.7
Sarashina-wildguardjp-7b	83.5	92.3	88.2	**72.6**	83.4	78.3	92.9	88.3	87.8	88.0
Llama3-wildguardjp-8B	**84.6**	91.8	**88.5**	70.7	82.6	77.0	93.6	86.9	88.3	87.7

Table 2 F1 scores on `WildGuardTestJP` for prompt safety (Input Moderation), response safety (Output Moderation), and refusal detection. Results are reported for the adversarial subset (Adv), the non-adversarial subset (Vani), and the combined overall set. For refusal detection, we also include performance on the harmful prompts subset (Harm).

Table A, shows the results for all evaluated models. Note: GPT-OSS-20B–Reasoning-Medium refused to answer 22 questions, which were classified as incorrect predictions for F1 computation. Meanwhile, Qwen3-4B–Reasoning-Medium and Qwen3-14B–Reasoning-Medium each had one invalid prediction that was considered wrong.

Due to space constraints, we do not include the prompts used in our translation pipeline or evaluation.