

Do Large Language Models Know Folktales? A Case Study of Yokai in Japanese Folktales

Tsutsumi Ayuto
Tokyo Metropolitan University
tsutsumi-ayuto@ed.tmu.ac.jp

Yuu Jinnai
CyberAgent
jinnai_yu@cyberagent.co.jp

Abstract

Although Large Language Models (LLMs) have demonstrated strong language understanding and generation abilities across various languages, their cultural knowledge is often limited to English-speaking communities, which can marginalize the cultures of non-English communities. To address the problem, evaluation of the cultural awareness of the LLMs and the methods to develop culturally aware LLMs have been investigated. In this study, we focus on evaluating knowledge of folktales, a key medium for conveying and circulating culture. In particular, we focus on Japanese folktales, specifically on knowledge of *Yokai*. Yokai are supernatural creatures originating from Japanese folktales that continue to be popular motifs in art and entertainment today. Yokai have long served as a medium for cultural expression, making them an ideal subject for assessing the cultural awareness of LLMs. We introduce YokaiEval, a benchmark dataset consisting of 809 multiple-choice questions (each with four options) designed to probe knowledge about yokai. We evaluate the performance of 31 Japanese and multilingual LLMs on this dataset. The results show that models trained with Japanese language resources achieve higher accuracy than English-centric models, with those that underwent continued pretraining in Japanese, particularly those based on Llama-3, performing especially well. The code and dataset are available at <https://github.com/CyberAgentAILab/YokaiEval>.

1 Introduction

Large Language Models (LLM) have shown remarkable performance in language understanding and generation tasks (Ouyang et al., 2022; Touvron et al., 2023; OpenAI et al., 2024). Despite many LLMs being predominantly trained in English, their generalization capabilities allow them to transfer knowledge across languages, achieving



Figure 1: The Night Parade of One Hundred Demons (百鬼夜行) by Kawanabe Kyosai. It is said that a parade of supernatural creatures known as yokai march through the streets of Japan at night, and anyone who comes across would be spirited away. The folktale became one of the popular motifs in the Edo period portrayed in many media including ukiyo-e, toys, and picture scrolls.

decent performance even in resource-limited languages (Chen et al., 2023; Shaham et al., 2024; OpenAI et al., 2024).

While there is evidence for strong cross-lingual transfer capability, cross-cultural transfer is known to be challenging for LLMs (Arango Monnar et al., 2022; Hershcovich et al., 2022; Lee et al., 2023; Huang and Yang, 2023; Rao et al., 2024; Adilazuarda et al., 2024; Cao et al., 2024; Liu et al., 2024). Prior work shows that LLMs tend to be biased toward the values and opinions of certain communities, rather than representing the diversity of human values (Santurkar et al., 2023; Conitzer et al., 2024).

To address this issue, many studies have investigated methods to evaluate the cultural awareness of LLMs (Rao et al., 2024; Adilazuarda et al., 2024; Liu et al., 2024). Simultaneously, approaches to develop **culturally aware LLMs** using the language resources of target communities are being explored (Pires et al., 2023; Lin and Chen, 2023; Nguyen et al., 2023; Huang et al., 2024; Owen et al., 2024;

Tran et al., 2024; Etxaniz et al., 2024).

While many prior studies have investigated differences in the values and opinions of communities (Xu et al., 2024; Sorensen et al., 2024; Wang et al., 2024a; Naous et al., 2024; Durmus et al., 2024), the *medium* that conveys these values and opinions has received less attention. Specifically, there is less focus on **folktales**, which are important for creating, receiving, and circulating a community’s traditions, values, opinions, and culture (Abello et al., 2012).

The goal of the study is to investigate how much LLMs know about folktales. Specifically, we focus on *Yokai* (Figure 1), supernatural phenomena or entities believed to cause such phenomena, which originated from Japanese folktales (Komatsu et al., 2003; Foster, 2024). Yokai have been popular motifs in art and entertainment from the Edo period to the present, appearing in Ukiyo-e (a genre of Japanese paintings), Kabuki (a traditional Japanese theater featuring performance and dance), toys, manga, and anime (Kagawa, 2005, 2006).

In this paper, we present **YokaiEval**, a benchmark dataset designed to evaluate knowledge about yokai. YokaiEval is created using Wikipedia articles, with GPT-4o generating the question-answer pairs (QAs). These QAs are then curated using GPT-4omini to filter out inconsistent questions, followed by manual verification to ensure their correctness, appropriateness, and the presence of references. We use YokaiEval to evaluate 31 LLMs, including Japanese-centric, English-centric, and other multilingual models. The results show that Japanese-centric models achieve comparatively higher scores than the others, particularly those models that have been continually pretrained from highly capable English-centric models (e.g., Llama-3).

This study suggests that using language resources from the community is crucial for acquiring regional folktale knowledge and that continual pre-training is an effective method for training LLMs with these resources.

2 Yokai: From Traditional Folktales to Today’s Art and Entertainment

Yokai are supernatural phenomena or entities believed to cause such phenomena (Komatsu et al., 2003; Foster, 2024). Traditionally, yokai are described in oral lore and rumors as *superstitions*, but they have also become popular *characters* in contemporary art and entertainment (Kagawa, 2005;

Komatsu, 2009).

Yokai in folktales. Yokai in folktales are thought to have emerged from the human tendency to assign meaning to mysterious phenomena that defy everyday understanding (Kagawa, 2005). When faced with inexplicable occurrences, humans often experience fear and anxiety. The concept of yokai is believed to have arisen as a means to recognize and thereby alleviate these fears (Kagawa, 2005).

History of yokai in art and entertainment. Although they originate from folktales, yokai have also been recognized as motifs in art and entertainment. By the Edo period, many people in urban areas (e.g., Edo) already viewed the world from a scientific perspective and considered yokai to be fictional rather than real (Kagawa, 2005). As yokai became recognized as fictional superstitions, people began to enjoy these mysterious characters in the form of art and entertainment.

For example, ghost stories featuring yokai became popular in theater (Kagawa, 2005). One notable example is Yotsuya Kaidan (四谷怪談), a ghost story featuring yokai that remains one of the famous plays in Kabuki today (Tsuruya and Gunji, 2019). Yokai also became a popular theme in Ukiyo-e. In 1776, Sekien Toriyama painted *The Illustrated Night Parade of a Hundred Demons* (画図百鬼夜行), which portrayed each yokai individually, characterizing their unique features (Toriyama, 2005). His art became remarkably popular in the Edo era, and even today, many descriptions of yokai are based on his work (Takoshima, 2018). Toys featuring yokai were also popular; for instance, *Ghost Playing Cards* (おぼけかるた) were made in 1860 and gained popularity among children (Tada, 1998).

However, rural beliefs and folktales, including those about yokai, were suppressed by the Meiji government in its efforts to Westernize and modernize the country (Kagawa, 2022).

Yokai in today’s Japan. After the wars, manga artist Shigeru Mizuki revitalized the popularity of yokai through his picture stories and mangas, including GeGeGe no Kitaro (Mizuki, 1986). Mizuki researched Ukiyo-e and folktales and incorporated these elements into his art (Mizuki, 1974; Takoshima, 2018), preserving aspects of yokai culture.

Although most people consider yokai to be fictional, stories about them remain present and rel-

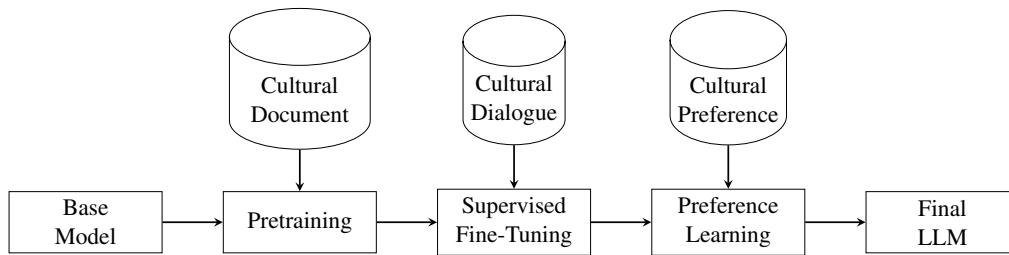


Figure 2: Common approach to training a culturally aware LLM is roughly divided into pretraining, supervised fine-tuning (SFT), and preference learning (PL). One of the questions of the study is to evaluate which step is the most critical to cultural knowledge.

evant in today’s Japan. For example, the transformation of a human in tragedy into an Oni (a kind of demon in Japanese folktales) is a common plot in these stories. An Ukiyo-e piece from 1902 by Tsukioka Yoshitoshi portrays a woman whose father was killed by a samurai, Omori Hikoshichi, and who then turns into an Oni due to her anger and sorrow (Yoshitoshi, 1902). Today, we see similar plots in some of the Oni characters in the series *Demon Slayer: Kimetsu no Yaiba* (Gotouge, 2016).

Summary. Yokai is a concept that originated from folktales and has flourished as a popular motif in arts and entertainment, preserved through the efforts of folklorists and artists. They remain popular in today’s Japan, both in folktales and in art. In this paper, **we consider both aspects of yokai to be cultural heritage** that we want LLMs to be knowledgeable about.

3 Related Work

We describe prior work on evaluating and training culturally aware LLMs.

3.1 Evaluating Cultural Awareness of LLM

While LLMs have shown strong capability in cross-lingual transfer, cross-cultural transfer remains challenging (Rao et al., 2024; Adilazuarda et al., 2024; Liu et al., 2024). As a result, LLMs often exhibit a bias towards Western culture, where many language resources are more readily available on the internet (Santurkar et al., 2023; Conitzer et al., 2024).

Therefore, efforts are being made to evaluate various cultural aspects of LLMs, including values (Xu et al., 2024; Sorensen et al., 2024; Wang et al., 2024a), opinions (Naous et al., 2024; Durmus et al., 2024), social norms (Yu et al., 2024; Rao et al., 2024; Yuan et al., 2024; Agarwal et al., 2024), commonsense knowledge (Wang et al., 2024b; Shen

et al., 2024; Myung et al., 2024), dietary preferences (Palta and Rudinger, 2023; Cao et al., 2024), and offensive languages (Zhou et al., 2023; Lee et al., 2023).

Several existing studies have investigated computational approaches to understand folktales (Declerck et al., 2012; Lestari and Manurung, 2015; Declerck et al., 2016; Schraagen, 2016; Meaney et al., 2024). Burda-Lassen (2022, 2023) investigate the performance of existing models in translating folktales and highlight the need for a larger folktale corpus with a broader range of cultural terms. Hobson et al. (2024) shows that GPT-4o’s identification of the values and lessons conveyed in folktales has a high agreement with what human annotators in Western countries identify. Benkler et al. (2022); Wu et al. (2023a) investigate how human values, morals, and gender biases are expressed in folktales across cultures using an international corpus of folktales.

While many existing studies seek to compare folktales from multiple cultures and compare their differences, we focus on folktales of a single country (Japan). We argue that establishing a thorough analysis of the folktales within each community is important as it will enhance the quality and depth of multi-cultural comparisons, which we discuss in Appendix J.

3.2 Training Culturally Aware LLM

There are plenty of efforts to train culturally aware LLM. The training procedures for LLMs can be broadly categorized into three approaches: continual pretraining, supervised fine-tuning, and preference learning.

Continual pretraining. Given the huge gap in available language resources between English and other languages, several studies have investigated an approach to train a model starting from an al-

ready highly capable English-centric model rather than from scratch. Continual pretraining has been reported to be effective for learning cultural knowledge across many communities (Pires et al., 2023; Lin and Chen, 2023; Nguyen et al., 2023; Huang et al., 2024; Owen et al., 2024; Tran et al., 2024; Etxaniz et al., 2024).

Supervised fine-tuning (SFT). SFT is a more computationally efficient process than continual pretraining. Many studies have shown that SFT can improve the cultural awareness of LLMs (Lin and Chen, 2023; Owen et al., 2024; Yoo et al., 2024; Huang et al., 2024; Cahyawijaya et al., 2024; Zhang et al., 2024). Choenni et al. (2024) show that cultural values can be influenced by the language used during the fine-tuning process. On the other hand, Moore et al. (2024) indicate that fine-tuned models tend to be more inconsistent on controversial topics.

Preference learning (PL). The effect of preference learning on cultural awareness of LLM has been studied less extensively than pretraining and SFT. Jinnai (2024) evaluate the effect of intra and cross-cultural alignment, showing that LLMs can learn the commonsense morality of a community with higher accuracy using the dataset annotated by community members. Chakraborty et al. (2024) propose a method to train a model that reflects the diversity of human preferences. Yao et al. (2025) show that a naive approach to preference learning can be dominated by the votes of the majority, marginalizing the preference of minorities.

4 Constructing YokaiEval

We first describe the procedure for generating YokaiEval in Section 4.1. Then, in Section 4.2, we analyze the characteristics of the generated dataset to evaluate if it is suitable for measuring the knowledge of yokai in LLMs.

4.1 Dataset Construction

YokaiEval is constructed using the process described in Figure 3.

Question generation. We generate QAs using 1,054 Wikipedia articles listed in the “List of Japanese yokai.”¹ For each yokai article, we provide the article to GPT-4o and ask it to generate a four-option QA that can be answered using the

¹<https://ja.wikipedia.org/wiki/日本の妖怪一覧>

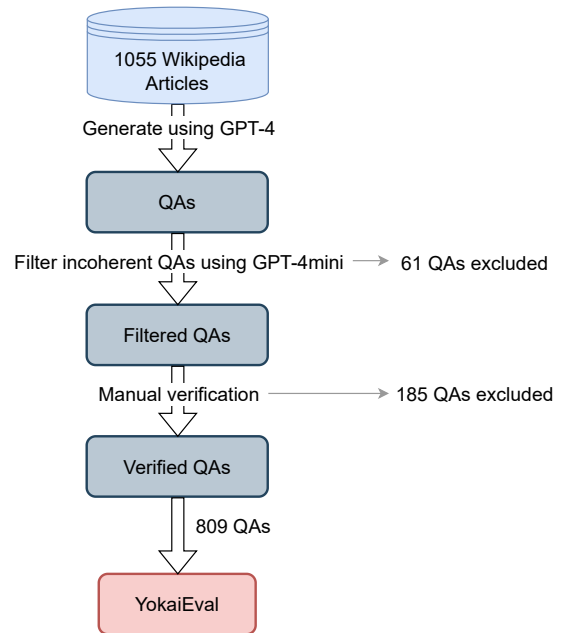


Figure 3: The procedure for generating YokaiEval involves several steps. First, we use articles from Wikipedia as the sources of information and generate QA pairs using GPT-4o. Next, we use GPT-4omini to remove QA pairs that are not coherent with the articles. Finally, the first author manually reviews all the data entries and their references to verify the validity of the QA pairs.

information within the article. See Appendix A.1 for the prompt used for question generation. We manually generate five questions and use them as 5-shot examples.

Since folktales have variations and lack scientific ground truth, we prompt GPT-4o to generate questions that ask which option is often considered true in Japanese folktales, rather than what is scientifically correct.

Automated filtering. To ensure that the questions can be solved by someone with knowledge of the yokai, we use GPT-4omini to answer the generated questions using the corresponding articles as prompts (Appendix A.2). Out of 1,055 QAs, GPT-4omini correctly answers 994 QAs. We manually check some of the QAs that GPT-4omini fails to answer correctly and observe that most of them either lack sufficient clues in the article or are not correctly formed as four-option QAs.

Manual verification. For the 994 QAs that GPT-4o answers correctly, we manually check the quality of all the QAs. We evaluate whether each QA is unique, relevant, appropriate, and associated with references. QAs without associated references are

Question	Choices	Answer
In the case of the Japanese yokai <i>Oitekebori</i> , what is the content of the voice that is often heard when trying to leave?	Go back, Leave it, Run away, Go	Leave it
When is the day considered particularly important for the appearance of the <i>Ippondatara</i> ?	January 1, May 5, December 20, August 8	December 20
What is often considered necessary to perceive the existence of the Japanese yokai <i>Enraenra</i> ?	Having a relaxed mind, Chanting a specific spell, Waiting at a specific place, Watching at a specific time	Having a relaxed mind

Table 1: Examples of data entries of the dataset translated to English. The original texts in the dataset are in Japanese.

excluded to ensure that the questions are verifiable via the references and not based solely on Wikipedia articles written by anonymous contributors. We manually check all the references of the QAs, as we find that GPT-4o often fails to identify the correct references.

Initially, the first author annotated the data, and two additional annotators subsequently verified these annotations. The verification process ensured that each QA pair met the following criteria:

1. The answer to the question is verifiable using the corresponding Wikipedia article.
2. The statement in the article, which corresponds to the QA and is associated by reference, was written by a domain expert.
3. The QA contains no inappropriate content.
4. The QA is not duplicated.

The annotation process was conducted interactively. Whenever an annotator encountered such cases, we held meetings or discussions to decide on the appropriate annotations. Due to this procedure, we do not have a record of inter-annotator agreement. We exclude 99 QAs for overlapping content, 41 QAs for irrelevant content, 3 for inappropriate content, and 42 QAs for lacking references, resulting in 809 QAs.

Note that the QAs in the dataset do not concern cultural preferences, values, or opinions. The questions assess factual accuracy based on the studies of folklorists and Japanese literary researchers. Each question presents four answer choices, with one correct answer and three incorrect ones. Although folktales themselves exhibit variations, the documented information can be uniquely defined.

Although the initial data entries come from Wikipedia, the QAs in the resulting dataset (YokaiEval) are all manually verified to be supported by references. Therefore, all QAs are associated with references.

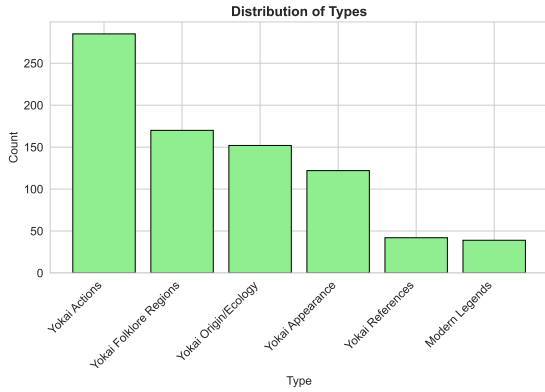
4.2 Analysis of the Generated Dataset

We analyze the diversity and the quality of the dataset to evaluate if it is valid for measuring knowledge of yokai.

Question types. We classified the question types into six categories using GPT-4o. See Appendix C for the prompt. The number of questions in each category is shown in Figure 4a. The results indicate that answering multiple aspects of yokai is necessary for an LLM to achieve a high score in YokaiEval.

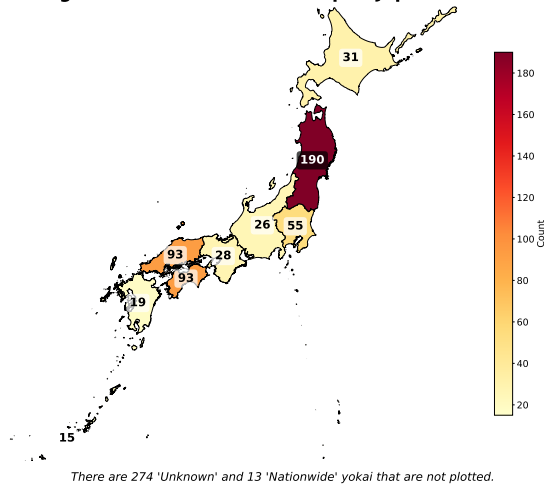
Regions. The region of each yokai is extracted from the corresponding Wikipedia article using GPT-4o (see prompt in Appendix C). Figure 4b shows the statistics of the regions of the yokai in YokaiEval. Many yokai are difficult to assign to a specific origin because the folktales are spread across multiple regions, resulting in some being labeled as Unknown. Still, the results indicate that the dataset provides good coverage of the regions in Japan.

References. Table 2 shows the authors whose work is cited ten or more times in YokaiEval. We observe a good balance among contemporary artists (e.g., Shigeru Mizuki and Natsuhiko Kyogoku), artists from the Middle Ages (e.g., Sekien Toriyama), and scholars of folklore and Japanese literature (e.g., Kunio Yanagita, Atsunobu Inada, Yasaburo Ikeda, Tokihiko Oto, and Iwao Hino). A



(a) Types of questions in YokaiEval.

Regional Distribution Heat Map of Japan



There are 274 'Unknown' and 13 'Nationwide' yokai that are not plotted.

(b) Regions of Yokais in YokaiEval.

list of frequently referenced materials is provided in Appendix D.

Questions solvable with reasoning. We observe that some questions can be guessed if one knows the tendencies of Japanese folktales and the meanings of the names of the yokai. For example, *Oitekebori* (置行堀; Table 1) literally means "left behind in the ditch." Thus, one can guess that "Leave it" is likely to be the correct answer. On the other hand, some questions are difficult to guess without specific knowledge of the particular yokai. For example, it is extremely difficult to infer that *Ippondatarara* (一本だたら; Table 1) appears on December 20th without specific knowledge of this yokai, making it extremely difficult to deduce from the text alone.

Therefore, while some questions are solvable with inductive knowledge of Japanese folktales, others require specific knowledge of the particular yokai. We consider this to be a desirable feature of the dataset, as our goal is to evaluate general knowledge of yokai.

Author	#refs
村上健司 [Kenji Murakami]	183
水木しげる [Shigeru Mizuki]	123
多田克己 [Katsumi Tada]	127
柳田國男 [Kunio Yanagita]	67
高田衛 [Mamoru Takada]	65
鳥山石燕 [Sekien Toriyama]	60
稲田篤信 [Atsunobu Inada]	50
田中直 [Nao Tanaka]	50
草野巧 [Takumi Kusano]	23
京極夏彦 [Natsuhiko Kyogoku]	19
池田彌三郎 [Yasaburo Ikeda]	14
柴田宵曲 [Shokyoku Shibata]	13
大藤時彦 [Tokihiko Oto]	13
寺島良安 [Ryoan Terashima]	12
日野巖 [Iwao Hino]	11
藤沢衛彦 [Morihiro Fujisawa]	11
千葉幹夫 [Mikio Chiba]	11
島田勇雄 [Isao Shimada]	10
田野理夫 [Morio Tano]	10
戸部民夫 [Tamio Tobe]	10

Table 2: The list of authors with more than or equal to ten references to their work in the QAs of YokaiEval.

Offensive and biased contents. By nature, some folktales contain stories or views that can be harmful (e.g., Kappa in Yanagida, 1976). We find that most of the QAs generated by GPT-4o do not include such content. Only three QAs were excluded during manual verification due to potentially inappropriate content. We hypothesize that this is due to using GPT-4o via the Azure OpenAI API. GPT-4o is aligned to avoid generating potentially harmful content, and the Azure OpenAI API includes a content moderation filter that prevents the generation of harmful content.

Although offensive and biased content in folktales is also important for folktale studies, our research questions do not necessarily require the inclusion of such content for assessment.

5 Evaluating Japanese and Multilingual LLMs on YokaiEval

With YokaiEval generated, we run experiments to evaluate the performance of 31 publicly available LLMs, including Japanese-centric, English-centric, and Chinese-centric models. We then further look into the performance of Japanese-centric models to investigate which training procedure (pretraining, SFT, and PL) is important to acquiring the knowledge of yokai.

5.1 Experimental Setup

We use the prompt in Table 5 to generate outputs

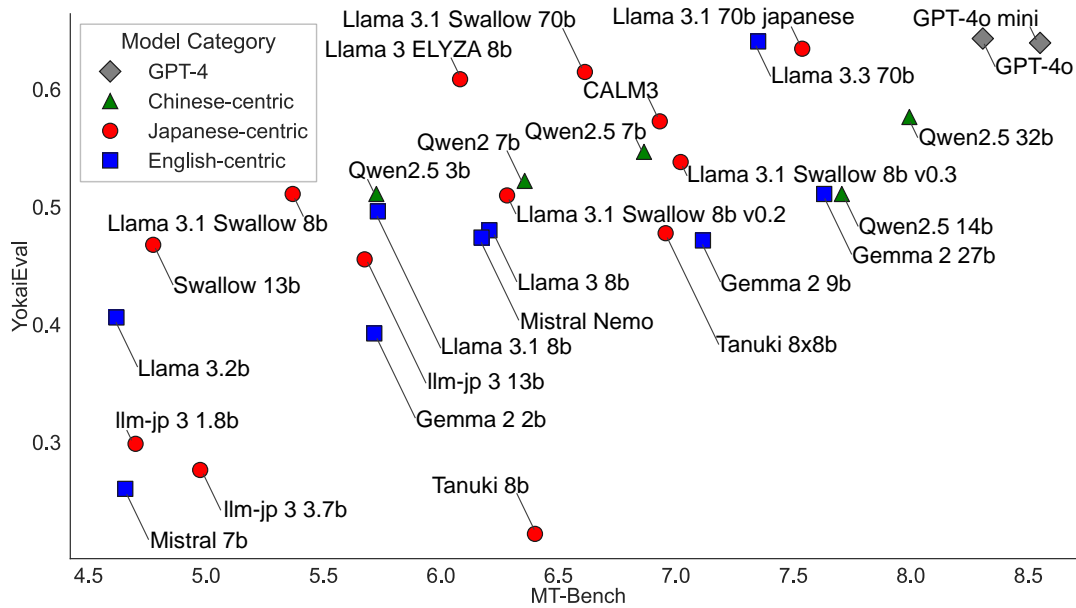


Figure 4: Scores of YokaiEval and JMT-Bench. Red: Japanese-centric models (including models continually pretrained on Japanese), Blue: English-centric languages models, Green: Chinese-centric models. Gray: GPT-4o models which we use for the dataset construction, which may have biases. The scores are included as references. See Appendix E for the exact values of the scores.

以下に、日本の妖怪に関する質問をする指示があります。質問に対する回答を記述してください。(Translation: Below are instructions for asking questions about Japanese yokai. Please provide answers to the questions.)

```
{question}
- {option1}
- {option2}
- {option3}
- {option4}
```

Figure 5: The prompt for generating the response from LLMs for YokaiEval.

from the LLMs. See Appendix H for the generation hyperparameters. We use GPT-4o to determine if the responses are correct, rather than relying on lexical matching, because the notation of yokai inherently has a lot of fluctuations. See Appendix F for the prompt used for evaluation. As far as we are aware, there were no errors in the GPT-4o evaluation.²

²We implemented a lexical matching parser capable of extracting answers from approximately 80% of the LLMs. For those answers extracted through lexical matching, the same answers were identified by GPT-4o. Additionally, we

We use Japanese MT-Bench (JMT-Bench) as a reference for the generic capability of the LLMs in Japanese.³ See Appendix G for details.

5.2 Results

Figure 4 shows the scores on YokaiEval and JMT-Bench. The YokaiEval score is computed as the number of correct answers divided by the total number of questions. Overall, Japanese-centric models achieve higher scores on YokaiEval compared to models with similar performance on the JMT-Bench. Notably, models that undergo continual pretraining from the Llama-3 family tend to perform particularly well on YokaiEval.

The study suggests that training LLMs with Japanese language resources improves performance on YokaiEval, suggesting a greater knowledge of Japanese folktales. In particular, continually pretraining a highly capable English-centric model (e.g., the Llama-3 family) on Japanese documents appears to be important for acquiring knowledge of Japanese folktales.

The effect of SFT and PL. As shown in Figure 4, models continually pretrained on Japanese datasets achieve high scores on YokaiEval. These models

manually reviewed the evaluation results of GPT-4o provided by two LLMs and found no errors.

³<https://github.com/Stability-AI/FastChat/tree/jp-stable>

Model	YokaiEval
(meta-llama-3.1-8b-instruct)	0.496
llama-3.1-swallow-8b-instruct-v0.1	0.511
llama-3.1-swallow-8b-instruct-v0.2	0.510
llama-3.1-swallow-8b-instruct-v0.3	0.538

Table 3: The effect of SFT on the YokaiEval scores.

also undergo SFT and PL in addition to continual pretraining. This raises the question of whether fine-tuning alone is sufficient for acquiring cultural knowledge. To investigate this, we conduct an observational ablation study to assess whether SFT and PL alone effectively improve YokaiEval scores. Note that our focus is on the effects of standard fine-tuning processes rather than fine-tuning specifically designed to enhance knowledge of yokai.

Table 3 presents the YokaiEval scores for the llama-3.1-swallow-8b-instruct series. These models are continually pretrained on llama-3.1 and llama-3.1-instruct. Additional SFT processes are applied between versions v0.1, v0.2, and v0.3. While SFT leads to a significant improvement in JMT-Bench scores, its impact on YokaiEval is marginal, suggesting that SFT may have a limited effect on acquiring knowledge of Japanese folktales.

Since only a few models are available both before and after the PL process, we trained models ourselves to evaluate the effect of PL. Specifically, we used ChatbotArena-Ja (Jinnai, 2024) to train three LLMs with direct preference optimization (DPO) (Rafailov et al., 2023) using LoRA (Hu et al., 2022).

ChatbotArena-Ja is a dataset consisting of English instructions from Chatbot Arena (Zheng et al., 2023), translated into Japanese, along with responses generated by Japanese LLMs. This dataset has been shown to improve Japanese LLM performance on the JMT-Bench. Since the instructions originate from English sources, it is unlikely to contain knowledge relevant to yokai. By using this dataset, we can isolate the effect of the PL procedure itself, independent of any additional knowledge introduced through training data. See Appendix H for the hyperparameters used in the DPO process.

Table 4 presents the scores of models before and after DPO. Overall, the improvements are inconsistent across models, suggesting that DPO on

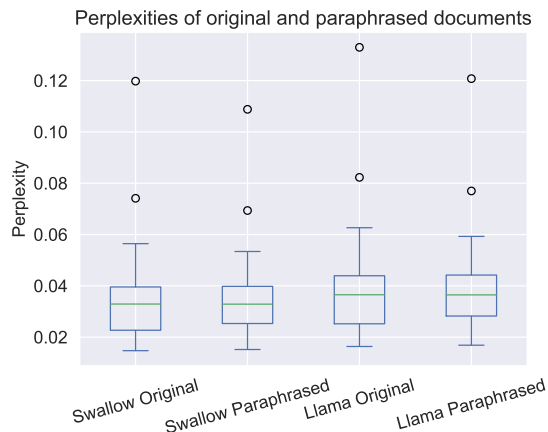


Figure 6: The distribution of perplexities of the original and paraphrased articles to assess the potential of data leakage.

ChatbotArena-Ja may not be effective in improving performance on YokaiEval.

Note that these results do not imply that SFT and PL on a dataset specifically designed for yokai knowledge would fail to improve YokaiEval scores. Rather, our findings indicate that fine-tuning on a standard dataset does not significantly impact YokaiEval performance. The purpose of the experiment is to separate out the effect of SFT and PL so that we can evaluate the effect of continual pertaining more precisely.

The result shows that the post-training processes are less likely to be effective for improving performance on YokaiEval compared to JMT-Bench, which implies that continual pertaining is likely to be the key step in acquiring knowledge of Japanese folktales.

Analysis of potential data leakage. A potential issue with using Wikipedia articles is that they may have been included in the training data of the LLMs, raising concerns about the suitability of YokaiEval for evaluation.

To investigate this, we follow the method of Deng et al. (2024) by generating paraphrased versions of 20 randomly selected articles using GPT-4o and comparing the perplexity of the LLMs. If an LLM has “memorized” an article, it should exhibit significantly lower perplexity for the original article compared to its paraphrased version.

We evaluate Swallow (llama-3.1-swallow-8b-instruct-v0.1) and Llama (meta-llama3.1-8b-instruct), as these models and their derivatives achieve high scores on YokaiEval. Figure 6 shows the perplexity of both models on the original

	SFT model	DPO Model	Improvement
calm2-7b-chat	0.242	0.199	-0.043
llm-jp-13b-instruct-full-jaster-v1.0	0.154	0.119	-0.036
swallow-7b-instruct-v0.1	0.340	0.424	0.084

Table 4: The effect of DPO on the YokaiEval scores.

Types of questions	#Questions	Accuracy
Regions of Yokai	170	59.41%
Yokai’s References	42	69.05%
Yokai’s Appearance	122	66.39%
Yokai’s Origin and Ecology	153	54.25%
Actions of Yokai	285	64.21%
Modern Legends	38	55.26%

Table 5: Accuracy of Swallow-70b in YokaiEval for each question type.

Regions of Yokai	#Questions	Accuracy
Unknown	274	67.88%
Chugoku / Shikoku	93	54.84%
Chubu	26	61.54%
Kyushu	19	78.95%
Hokkaido	31	58.06%
Nationwide (Japan)	13	61.54%
Tohoku	190	60.00%
Okinawa	15	40.00%
Kinki	28	64.29%
Kanto	55	56.36%

Table 6: Accuracy of Swallow-70b in YokaiEval for each region of Yokai.

and paraphrased articles. Overall, no significant difference is observed between the two groups of articles for either model. This suggests no clear evidence of data leakage in either case. We speculate that articles about yokai are often overlooked by developers as they are unlikely to be relevant for existing benchmark datasets.

Error analysis. Table 5 shows the accuracy of the Swallow-70b (llama-3.1-swallow-70b-instruct-v0.1) model on each type of question in YokaiEval. Overall, the accuracy is balanced over the types. Table 6 shows the accuracy of the model on each region of Yokai. Overall, the accuracy is balanced except for Okinawa, where the accuracy is only 40%. It may be because Okinawa has a historically distinct background compared to the other regions,

as it was once the Ryukyu Kingdom and the United States. In-depth analysis is required to analyze the effect of the regions to the knowledge of the LLMs.

6 Conclusions

In this paper, we introduce YokaiEval, a benchmark dataset designed to evaluate LLMs’ knowledge of Japanese folktales. The dataset is generated by extracting QAs from Wikipedia articles using GPT-4o, followed by curation with GPT-4omini and human verification.

We evaluate 31 LLMs using YokaiEval, showing that Japanese LLMs outperform the others, with models continually pretrained on Japanese language resources achieving the highest scores. The result implies the importance of using community-sourced resources to develop LLMs knowledgeable of folktales.

7 Limitations

Our study focuses on Yokai in Japanese folktales, and the findings may differ when applied to other folktales, communities, or cultural domains. Our methodology may be extended to other cultural contexts in future research.

The dataset used in this study only includes yokai recorded in digital media (i.e., Wikipedia) up to the present day. Thus, it may be biased and contain underrepresented regions or eras. It may not represent the folktales in their original forms. Additionally, the QAs were extracted using GPT-4o, which may introduce bias into the dataset.

YokaiEval is not intended to serve as a standalone benchmark for assessing the general cultural understanding of LLMs. Instead, we expect it to be most useful when combined with other benchmarks (e.g., Japanese MT-Bench). Since YokaiEval addresses an area not covered by existing benchmarks, it serves as a valuable complementary evaluation dataset.

In 2025, the National Diet Library of Japan holds approximately 5,590 digital records, 10,800 books

and articles not available in digital form, 2,313 research articles, and 2,130 multimedia records related to Yokai.⁴ Using non-digital resources to further enhance the dataset is future work.

The analysis in Section 5 is observational, relying on publicly available models rather than models trained specifically for this analysis. Therefore, several factors, beyond the cultural training procedures we discuss, could influence the YokaiEval scores. For instance, some Japanese LLMs use tokenizers specifically designed for Japanese, which may contribute to their higher performance on YokaiEval (Fujii et al., 2024; Aizawa et al., 2024).

Our analysis is limited to evaluating the knowledge embedded in the LLMs themselves. However, it may not be the only way to achieve culturally aware LLMs. One potential approach for acquiring cultural knowledge is Retrieval-Augmented Generation (RAG), which generates output by using a search engine or vector database to retrieve relevant information during inference (Lewis et al., 2020).

8 Ethical Considerations

The dataset is constructed from publicly available Wikipedia articles. Analysis of folktales should be done carefully, as it may promote biased or polarized views of the community. We encourage users of YokaiEval to be mindful of the potential harm when using the dataset.

Acknowledgments

We thank the anonymous reviewers for their insightful comments on the manuscript. We would like to thank our colleagues at CyberAgent AI Lab for the helpful discussions.

References

James Abello, Peter Broadwell, and Timothy R Tangherlini. 2012. Computational folkloristics. *Communications of the ACM*, 55(7):60–70.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhymna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. *arXiv preprint arXiv:2403.15412*.

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.

Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, et al. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *arXiv preprint arXiv:2407.03963*.

Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. [Resources for multilingual hate speech detection](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Noam Benkler, Scott Friedman, Sonja Schmergalunder, Drisana Mosaphir, Vasanth Sarathy, Pavan Kantharaju, Matthew D. McLure, and Robert P. Goldman. 2022. Cultural value resonance in folktales: A transformer-based analysis with the world value corpus. In *Social, Cultural, and Behavioral Modeling*, pages 209–218, Cham. Springer International Publishing.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. [GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus](#). In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47, Denver, Colorado, USA. Association for Computational Linguistics.

Olena Burda-Lassen. 2022. [Ukrainian-to-English folk-tale corpus: Parallel corpus creation and augmentation for machine translation in low-resource languages](#). In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Workshop 2: Corpus Generation and Corpus Augmentation for Machine Translation)*, pages 28–31. Association for Machine Translation in the Americas.

Olena Burda-Lassen. 2023. [Machine translation of folktales: small-data-driven and LLM-based approaches](#). In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 68–71, Gothenburg, Sweden. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Afina Putri, Emmanuel Dave, Jhonson Lee, Nuur Shadieq, Wawan Cenggoro, Salsabil Maulana Akbar, Muhammad Ihza Mahendra, et al. 2024. Cendol: Open instruction-tuned generative large language models for indonesian languages. *arXiv preprint arXiv:2404.06138*.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. [Cultural](#)

⁴<https://www.ndl.go.jp/>

- adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.
- Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Bedi, and Mengdi Wang. 2024. **Maxmin-RLHF: Alignment with diverse human preferences**. In *Forty-first International Conference on Machine Learning*.
- Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. Phoenix: Democratizing ChatGPT across languages. *arXiv preprint arXiv:2304.10453*.
- Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. **The echoes of multilinguality: Tracing cultural value shifts during language model fine-tuning**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15042–15058, Bangkok, Thailand. Association for Computational Linguistics.
- Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailley Schoelkopf, Emanuel Tewolde, and William S. Zwicker. 2024. **Position: Social choice should guide AI alignment in dealing with diverse human feedback**. In *Forty-first International Conference on Machine Learning*.
- Thierry Declerck, Tyler Klement, and Antonia Kostova. 2016. **Towards a WordNet based classification of actors in folktales**. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 83–87, Bucharest, Romania. Global Wordnet Association.
- Thierry Declerck, Nikolina Koleva, and Hans-Ulrich Krieger. 2012. **Ontology-based incremental annotation of characters in folktales**. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 30–34, Avignon, France. Association for Computational Linguistics.
- Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gestein, and Arman Cohan. 2024. **Benchmark probing: Investigating data leakage in large language models**. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning - The Good, the Bad, and the Ugly*.
- Esin Durmus, Karina Nguyen, Thomas I. 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. *arXiv preprint arXiv:2306.16388*.
- Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. **BertaQA: How much do language models know about local culture?** In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Larisa Fialkova and Maria N Yelenevskaya. 2001. Ghosts in the cyber world. an analysis of folklore sites on the internet. *Fabula*.
- Michael Dylan Foster. 2024. *The Book of Yokai, Expanded Second Edition: Mysterious Creatures of Japanese Folklore*. Univ of California Press.
- Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. **Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities**. In *First Conference on Language Modeling*.
- Koyoharu Gotouge. 2016. *鬼滅の刃 [Demon Slayer: Kimetsu no Yaiba]*. Shueisha.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarelo, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. **Challenges and strategies in cross-cultural NLP**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.
- David G Hobson, Haiqi Zhou, Derek Ruths, and Andrew Piper. 2024. **Story morals: Surfacing value-driven narrative schemas using large language models**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12998–13032, Miami, Florida, USA. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. **LoRA: Low-rank adaptation of large language models**. In *International Conference on Learning Representations*.
- Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. **AceGPT, localizing large language models in Arabic**. 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 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.
- Jing Huang and Diyi Yang. 2023. **Culturally aware natural language inference**. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

- Yuu Jinnai. 2024. Does cross-cultural alignment change the commonsense morality of language models? In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 48–64, Bangkok, Thailand. Association for Computational Linguistics.
- Masanobu Kagawa. 2005. 江戸の妖怪革命 [*The Yokai Revolution in Edo*]. Kawade Shobo Shinsha.
- Masanobu Kagawa. 2006. 日本人の妖怪観の変遷に関する研究：近世後期の「妖怪娯楽」を中心に [*A Study of Changes in the Japanese View of Specters : Focusing on “Yokai Amusements” in the Late Modern Period*]. Ph.D. thesis, The Graduate University for Advanced Studies, Sokendai.
- Masanobu Kagawa. 2022. 図説日本妖怪史 [*Illustrated history of Japanese Yokai*]. Kawade Shobo Shinsha Publishers Inc.
- Kazuhiko Komatsu. 2009. 妖怪文化研究の最前線 [*The Frontiers of Yokai Culture Research*]. Serica Syobo.
- Kazuhiko Komatsu, Toru Tsunemitsu, Shoji Yamada, and Kazuhisa Nakayama. 2003. 異界へのいざない：怪異・妖怪伝承データベースの試み [invitation to the otherworld: An attempt at a database of oddity and yokai traditions]. *Sokendai journal*, pages 42–43.
- John Laudun and Jonathan Goodwin. 2013. Computing folklore studies: Mapping over a century of scholarly production through topics. *Journal of American Folklore*, 126(502):455–475.
- Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.
- Victoria Anugrah Lestari and Ruli Manurung. 2015. Measuring the structural and conceptual similarity of folktales using plot graphs. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 25–33, Beijing, China. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487*.
- Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *arXiv preprint arXiv:2406.03930*.
- J. A. Meaney, Beatrice Alex, and William Lamb. 2024. Testing and adapting the representational abilities of large language models on folktales in low-resource languages. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 319–324, Miami, USA. Association for Computational Linguistics.
- Shigeru Mizuki. 1974. 妖怪なんでも入門 [*Introduction to Yokai Anything*]. Shogakukan.
- Shigeru Mizuki. 1986. ゲゲゲの鬼太郎 [*Gegege no Kitaro*]. Chikumashobo.
- Jared Moore, Tanvi Deshpande, and Diyi Yang. 2024. Are large language models consistent over value-laden questions? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15185–15221, Miami, Florida, USA. Association for Computational Linguistics.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Victor Gutierrez Basulto, Yazmin Ibanez-Garcia, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzaev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. BLEnd: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.
- Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023. Seallms—large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.
- OpenAI et al. 2024. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Louis Owen, Vishesh Tripathi, Abhay Kumar, and Bidwan Ahmed. 2024. Komodo: A linguistic expedition into indonesia’s regional languages. *arXiv preprint arXiv:2403.09362*.
- Yei Theodora Ozaki. 2024. *Japanese Fairy Tales*. E-Kitap Projesi & Cheapest Books.

- Shramay Palta and Rachel Rudinger. 2023. [FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.
- Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*, pages 226–240. Springer.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. [Whose opinions do language models reflect?](#) In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.
- Marijn Schraagen. 2016. [Folktale similarity based on ontological abstraction](#). In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 355–363, Bucharest, Romania. Global Wordnet Association.
- Uri Shaham, Jonathan Herzig, Roei Aharoni, Idan Szepkator, Reut Tsarfaty, and Matan Eyal. 2024. [Multilingual instruction tuning with just a pinch of multilinguality](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.
- Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. [Understanding the capabilities and limitations of large language models for cultural commonsense](#). 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 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.
- Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. 2024. [Value kaleidoscope: Engaging AI with pluralistic human values, rights, and duties](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19937–19947.
- Katsumi Tada. 1998. *江戸妖怪かるた [Edo Yokai Karuta]*. Kokushokankoukai.
- Sunao Takoshima. 2018. [Folkloristic elements and their embellishments among the works of shigeru mizuki](#). *Transactions of the Institute for Cultural Studies, Aichi Gakuin University NINGEN BUNKA*, pages 194–164.
- Sekien Toriyama. 2005. *鳥山石燕画図百鬼夜行全画集 [Toriyama Sekien Painting Collection of Hyakki Yagyō]*. Kadokawa Shoten.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Khanh-Tung Tran, Barry O’Sullivan, and Hoang D Nguyen. 2024. Uccix: Irish-excellence large language model. *arXiv preprint arXiv:2405.13010*.
- Nanboku Tsuruya and Masakatsu Gunji. 2019. *Tokaido Yotsuya Kaidan*. Shinchosha.
- Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024a. [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 Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024b. [Not all countries celebrate thanksgiving: On the cultural dominance in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Winston Wu, Lu Wang, and Rada Mihalcea. 2023a. [Cross-cultural analysis of human values, morals, and biases in folk tales](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5113–5125, Singapore. Association for Computational Linguistics.
- Winston Wu, Lu Wang, and Rada Mihalcea. 2023b. [Word category arcs in literature across languages and genres](#). In *Proceedings of the 5th Workshop*

on *Narrative Understanding*, pages 36–47, Toronto, Canada. Association for Computational Linguistics.

Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, and Deyi Xiong. 2024. Exploring multilingual concepts of human value in large language models: Is value alignment consistent, transferable and controllable across languages? *arXiv preprint arXiv:2402.18120*.

Kunio Yanagida. 1976. *遠野物語・山の人生 [Tales of Tono, Life in the Mountains]*. Iwanami Shoten.

Binwei Yao, Zefan Cai, Yun-Shiuan Chuang, Shanglin Yang, Ming Jiang, Diyi Yang, and Junjie Hu. 2025. No preference left behind: Group distributional preference optimization. In *The Thirteenth International Conference on Learning Representations*.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. 2024. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*.

Tsukioka Yoshitoshi. 1902. *新形三十六怪撰 大森彦七道に怪異に逢ふ図 [Omori Hikoshichi encountering a demon, new form of thirty-six-ghosts]*.

Linhao Yu, Yongqi Leng, Yufei Huang, Shang Wu, Haixin Liu, Xinmeng Ji, Jiahui Zhao, Jinwang Song, Tingting Cui, Xiaoqing Cheng, Liutao Liutao, and Deyi Xiong. 2024. CMoralEval: A moral evaluation benchmark for Chinese large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11817–11837, Bangkok, Thailand. Association for Computational Linguistics.

Ye Yuan, Kexin Tang, Jianhao Shen, Ming Zhang, and Chenguang Wang. 2024. Measuring social norms of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 650–699, Mexico City, Mexico. Association for Computational Linguistics.

Wenjing Zhang, Siqi Xiao, Xuejiao Lei, Ning Wang, Huazheng Zhang, Meijuan An, Bikun Yang, Zhaoxiang Liu, Kai Wang, and Shiguo Lian. 2024. Methodology of adapting large english language models for specific cultural contexts. *arXiv preprint arXiv:2406.18192*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023. Cultural compass: Predicting transfer learning success in offensive language detection with cultural features. In *Findings of the Association for Computational Linguistics*:

EMNLP 2023, pages 12684–12702, Singapore. Association for Computational Linguistics.

A Prompts for Constructing YokaiEval (Section 4.1)

YokaiEval is constructed using GPT-4o in various ways. This Appendix lists the prompts used for each of the processes on the dataset construction.

A.1 Prompt for Generating QAs

Figure 7 shows the 5-shot prompt used to generate a QA from a Wikipedia article using GPT-4o.

A.2 Prompt for Filtering QAs

Figure 9 shows the prompt used for the filtering process. The filtering process tests if the QA is generated in a way that can be solved given the information of the corresponding article.

A.3 Prompt for Extracting the Reference of the QA

Figure 11 is the prompt we use to identify the reference that corresponds to the QA using GPT-4o (Section 4.1). We find that many of the references identified by GPT-4o are incorrectly attributed, so we correct them manually in the manual verification process.

B Annotation Tool for Manual Verification

We develop a lightweight annotation tool to manually verify the data entries. Figure 13 is the snapshot of the annotation tool. The tool is designed to show the Wikipedia article, the generated QA, and the question reference. The task is to manually check if the QA is appropriate and has a reference. The interface is designed to edit the reference as it has the highest frequency or errors.

C Prompts for Analyzing YokaiEval (Section 4.2)

Figure 14 shows the prompt used to categorize the type of question. Figure 16 shows the prompt used to extract the region of the yokai from the Wikipedia article.

D List of Frequently Referred Materials

Table 7 is a list of frequently referred materials used in the QAs of YokaiEval.

```

###Instruction###
日本の妖怪{yokai_name}の知識を確認する質問を1つ作成しなさい。
###conditions###
1. 説明文から検証可能な質問にしなさい
2. 4択形式の質問にしなさい
3. その妖怪についての知識がないと正解出来ない質問にしなさい
4. 妖怪の名前から推測不可能な質問にしなさい
5. 妖怪に関する記述は文献によって異なる場合が多々あるため、「とされていることが多いでしょうか？」のように質問し、回答の存在を保証できるような質問にしなさい
6. コードブロックのないjson形式にしなさい
7. 必ず予想では正解出来ない質問を作りなさい
###{yokai_name}についての説明文###
{detail}
###Example###

{{
  "question": "日本の妖怪である「赤い」の大きさはどの程度とされていることが多いでしょうか？以下の4つから回答を1つ選び出力しなさい。",
  "choices": [
    "10メートル",
    "100メートル",
    "1キロメートル",
    "10キロメートル"
  ],
  "answer": "10キロメートル"
}}

{{
  "question": "日本の妖怪である「鍛冶壺」の正体とされている動物はなんとされていることが多いでしょうか？以下の4つから回答を1つ選び出力しなさい。",
  "choices": [
    "牛",
    "鼠",
    "狼",
    "馬"
  ],
  "answer": "狼"
}}

{{
  "question": "日本の妖怪である「片耳豚」にあることをされると人間は魂を抜かれてしまうとされています。何をされると魂を抜かれてしまうとされていることが多いでしょうか？以下の4つから回答を1つ選び出力しなさい。",
  "choices": [
    "影を踏まれる",
    "股の下をくぐられる",
    "見つめられる",
    "前を横切る"
  ],
  "answer": "股の下をくぐられる"
}}

{{
  "question": "日本の妖怪である「木心坊」はある木を材料にすりこぎを作るとその木が変化して生まれるといわれています。何の木を材料にすると生まれるとされていることが多いでしょうか？以下の4つから回答を1つ選び出力しなさい。",
  "choices": [
    "桜",
    "椿",
    "樅",
    "榲"
  ],
  "answer": "椿"
}}

{{
  "question": "日本の妖怪「馬魔」は馬に何かするとされています。何をするとされていることが多いでしょうか？以下の4つから回答を1つ選び出力しなさい。",
  "choices": [
    "転ばす",
    "殺す",
    "乗る",
    "吹き飛ばす"
  ],
  "answer": "殺す"
}}

```

Figure 7: Prompt for generating a QA from a Wikipedia article using GPT-4o. **The English translation is available in Figure 8.**

```

###Instruction###
Create a question to test knowledge about the Japanese yokai {yokai_name}.
###Conditions###
1. Make the question verifiable from the description.
2. Make the question a multiple-choice question with four options.
3. Make the question one that cannot be answered correctly without knowledge of the yokai.
4. Make the question one that cannot be guessed from the yokai's name.
5. Since descriptions of yokai often vary by source, frame the question in a way that ensures the existence of an answer, such as "What is often said to be...?"
6. Format the question in JSON without any code blocks.
7. Ensure the question cannot be answered correctly by guessing.
###Example###
{{
  "question": "What is often said to be the size of the Japanese yokai 'Akaei'? Choose one answer from the following four options.",
  "choices": [
    "10 meters",
    "100 meters",
    "1 kilometer",
    "10 kilometers"
  ],
  "answer": "10 kilometers"
}}
{{
  "question": "What animal is often said to be the true form of the Japanese yokai 'Kajibokuro'? Choose one answer from the following four options.",
  "choices": [
    "Cow",
    "Rat",
    "Wolf",
    "Horse"
  ],
  "answer": "Wolf"
}}
{{
  "question": "What action is often said to cause a human to lose their soul when performed by the Japanese yokai 'Kataeributa'? Choose one answer from the following four options.",
  "choices": [
    "Stepping on their shadow",
    "Passing under their legs",
    "Staring at them",
    "Crossing in front of them"
  ],
  "answer": "Passing under their legs"
}}
{{
  "question": "What type of tree is often said to give birth to the Japanese yokai 'Kishinbo' when used to make a pestle? Choose one answer from the following four options.",
  "choices": [
    "Cherry",
    "Camellia",
    "Fir",
    "Zelkova"
  ],
  "answer": "Camellia"
}}
{{
  "question": "What is often said to happen to horses when the Japanese yokai 'Umama' is around? Choose one answer from the following four options.",
  "choices": [
    "They are tripped",
    "They are killed",
    "They are ridden",
    "They are blown away"
  ],
  "answer": "They are killed"
}}

```

Figure 8: Prompt for generating a QA from a Wikipedia article using GPT-4o (Figure 7) translated to English.

```

###Task###
これは{yokai_name}の説明です。説明を参考にして、質問に教えてください。
###説明###
{yokai_detail}。
###質問###
{question}

```

Figure 9: Prompt for evaluating the correctness of QAs using GPT-4o. **The English translation is available at Figure 10.**


```

###Task###
This is a description of {yokai_name}. Please refer to the description and answer the question.
###Description###
{yokai_detail}.
###Question###
{question}

```

Figure 10: Prompt for evaluating the correctness of QAs using GPT-4o (Figure 9), translated to English.

```

[[Task]]
以下に示す質問は[detail]の情報を元に作成されました。
質問と解答に基づいている情報源・文献を、detailの中の番号で出力してください。
複数存在する場合は複数出力してください。例えば、[番号]
質問が文献に基づいていない場合は[0]と出力してください。
[[質問]]
{question}
{choices}
[[解答]]
{answer}
[[detail]]
{yokai_name}
{yokai_detail}

```

Figure 11: Prompt for identifying the reference of the QA using GPT-4o. The English translation is available in Figure 12.

```

[[Task]]
The following question was created based on the information in [detail].
Please output the source(s) or document(s) that the question and answer are based on, using the
number(s) in the detail. If there are multiple sources, output multiple numbers, for example,
[number]. If the question is not based on any document, output [0].
[[Question]]
{question}
{choices}
[[Answer]]
{answer}
[[Detail]]
{yokai_name}
{yokai_detail}

```

Figure 12: Prompt for identifying the reference of the QA using GPT-4o (Figure 11), translated into English.



Figure 13: Snapshot of the annotation tool developed for the manual verification step.

以下は妖怪に関する質問と回答です。この質問は以下のどの質問に分類されますか？
 [[回答]]という形式で答えてください。例えば [[A]]と答えてください。
 これら以外の分類である場合は自由記述で答えてください。

- A. 妖怪の行う行動
- B. 妖怪の形状・見た目
- C. 妖怪の生まれ・生態
- D. 妖怪の典拠
- E. 妖怪の伝承のある地域
- F. 現代における伝承

{question}

Figure 14: Prompt for categorizing the type of the QA using GPT-4o. **The English translation is available in Figure 15.**

The following is a question and answer about a yokai. Which category does this question fall into? Please answer in the format [[Answer]]. For example, answer [[A]]. If it falls into a category other than those listed, please answer in free text.

- A. Actions performed by the yokai
- B. Appearance of the yokai
- C. Origin or ecology of the yokai
- D. References to the yokai
- E. Region where the yokai is traditionally known
- F. Modern-day traditions

{question}

Figure 15: Prompt for categorizing the type of the QA using GPT-4o (Figure 14), translated to English.

以下は日本の妖怪「{yokai_name}」に関する説明です。{yokai_name}が伝わる地域はどこであると記述されていますか？
 [[回答]]という形式で答えてください。例えば [[A]]と答えてください。
 日本全国でなく、かつ複数にまたがる場合は自由記述で答えてください。
 これら以外の分類である場合は自由記述で答えてください。

- A. 日本全国
- B. 北海道
- C. 東北
- D. 関東
- E. 中部地方
- F. 近畿地方
- G. 中国地方
- H. 四国地方
- I. 九州
- J. 沖縄
- K. 不明

{article}

Figure 16: Prompt for extracting the region of the yokai from the Wikipedia article using GPT-4o. **The English translation is available in Figure 17.**

Book	Count
村上健司編著 『妖怪事典』 [Encyclopedia of Yokai by Kenji Murakami], 毎日新聞社 [Mainichi Shinbunsha], 2000. ISBN 978-4-620-31428-0	142
稲田篤信、田中直日 編 『鳥山石燕 画図百鬼夜行』 [Gazu Hyakki Yagyō by Sekien Toriyama, Edited by Atsunobu Inada and Nao Tanaka], 高田衛 監修 国書刊行会 [Supervised by Mamoru Takada, Kokushokankokai], 1992. ISBN 978-4-336-03386-4	55
民俗学研究所、柳田國男監修 編 『総合日本民俗語彙』 [Comprehensive Japanese Folklore Glossary, supervised by Kunio Yanagita], 平凡社 [Heibonsha], 1955. NCID BN05729787	41
多田克己 『幻想世界の住人たち』 IV [Residents of the Fantasy World IV by Katsumi Tada], 新紀元社 [Shinkigensha], 1990. ISBN 978-4-915146-44-2	56
村上健司編著 『日本妖怪大事典』 [Great Encyclopedia of Japanese Yokai by Kenji Murakami], 角川書店 [Kadokawa Shoten], 2000. ISBN 978-4-04-883926-6	31
水木しげる 『妖鬼化』 [Yokai Transformation by Shigeru Mizuki], Softgarage, 2004. ISBN 978-4-86133-006-3	29
水木しげる 『図説 日本妖怪大全』 [Illustrated Great Encyclopedia of Japanese Yokai by Shigeru Mizuki], 講談社 [Kodansha], 1994. ISBN 978-4-06-256049-8	25
多田克己 著、京極夏彦、多田克己 編 『妖怪図巻』 [Yokai Illustrated Book by Katsumi Tada, edited by Natsuhiko Kyogoku and Katsumi Tada], 国書刊行会 [Kokushokankokai], 2000年 [2000]. ISBN 978-4-336-04187-6	16
草野巧 『幻想動物事典』 [Encyclopedia of Fantasy Animals by Takumi Kusano], 新紀元社 [Shinkigensha], 1997年 [1997]. ISBN 978-4-88317-283-2	14
柳田國男 『妖怪談義』 [Discussions on Yokai by Kunio Yanagita], 講談社 [Kodansha], 1977年 [1977]. ISBN 978-4-06-158135-7	14
水木しげる 『決定版 日本妖怪大全 妖怪・あの世・神様』 [Definitive Edition of the Great Encyclopedia of Japanese Yokai, Spirits, and Gods by Shigeru Mizuki], 講談社 [Kodansha]	14
多田克己 編 『竹原春泉 絵本百物語 -桃山人夜話-』 [Picture Book of a Hundred Stories by Haruzen Takehara, edited by Katsumi Tada], 国書刊行会 [Kokushokankokai], 1997. ISBN 978-4-336-03948-4	12
日野巖 『動物妖怪譚』 [Tales of Animal Yokai by Iwao Hino], 中央公論新社〈中公文庫〉 [Chuokoron-Shinsha]	11
多田克己 編 『絵本百物語 桃山人夜話』 [Picture Book of a Hundred Ghost Stories: Night Tales of the Momoyama Man by Katsuki Tada] 国書刊行会 [Kokushokankokai], 1997. ISBN 978-4-336-03948-4	11

Table 7: List of references cited more than ten times in YokaiEval.

```

The following is an explanation about the Japanese yokai "{yokai_name}". Where is the region
mentioned where {yokai_name} is traditionally known?
Please answer in the format [[Answer]]. For example, answer [[A]].
If it is not nationwide in Japan and spans multiple regions, please answer in free text.
If it falls into a category other than those listed, please answer in free text.
A. Nationwide in Japan
B. Hokkaido
C. Tohoku
D. Kanto
E. Chubu region
F. Kinki region
G. Chugoku region
H. Shikoku region
I. Kyushu
J. Okinawa
K. Unknown
{article}

```

Figure 17: Prompt for extracting the region of the yokai from the Wikipedia article using GPT-4o (Figure 16), translated to English.

Model name	YokaiEval	JMT-Bench
gpt4o-mini	0.643	8.31
gpt4o	0.640	8.55
llama-3.1-70b-japanese-instruct-2407	0.635	7.54
llama-3.1-swallow-70b-instruct-v0.1	0.615	6.61
llama-3-elyza-jp-8b	0.609	6.08
llama-3.1-swallow-8b-instruct-v0.3	0.538	7.02*
calm3-22b-chat	0.573	6.93
llama-3.1-swallow-8b-instruct-v0.1	0.511	5.37
llama-3.1-swallow-8b-instruct-v0.2	0.510	6.28
tanuki-8x8b-dpo-v1.0	0.478	6.96
swallow-13b-instruct-v0.1	0.468	4.78
llm-jp-3-13b-instruct	0.456	5.68
llama-3.2-3b-instruct	0.406	4.62
llm-jp-3-1.8b-instruct	0.299	4.70
llm-jp-3-3.7b-instruct	0.277	4.98
tanuki-8b-dpo-v1.0	0.222	6.40
qwen2.5-32b-instruct	0.577	7.99
qwen2.5-7b-instruct	0.547	6.86
qwen2-7b-instruction	0.522	6.36
qwen2.5-14b-instruct	0.511	7.71
qwen2.5-3b-instruct	0.511	5.73
meta-llama-3.3-70b-instruct	0.641	7.35
eurollm-9b-instruct	0.535	-
gemma-2-27b-it	0.511	7.63
meta-llama-3.1-8b-instruct	0.496	5.73
meta-llama-3-8b-instruct	0.480	6.21
mistral-nemo-instruct-2407	0.474	6.17*
gemma-2-9b-it	0.472	7.12
gemma-2-2b-it	0.393	5.72*
mistral-7b-instruct-v0.2	0.260	4.66
eurollm-1.7b-instruct	0.086	-

Table 8: YokaiEval and JMT-Bench scores. * indicates that the score is estimated by Eq. 1 in Appendix G. The subgroups correspond to GPT-4, Japanese-centric, Chinese-centric, and English-centric models.

E JMT-Bench and YokaiEval Scores

Table 8 shows the scores of the JMT-Bench and YokaiEval in Figure 4.

F Prompt for GPT-4o Evaluation

Figure 18 shows the prompt we use for the evaluation of the output by the model for YokaiEval. We use a 1-shot prompt with an example of a response text which correctly answers the question but it tells the user that it is an educated guess. For the purpose of the experiment, we want to judge this to be a correct answer, so we place this example in the 1-shot prompt.

G Calculating JMT-Bench Scores

We rely on publicly available leaderboards for the scores of JMT-Bench scores. In particular, we use the scores on Nejumi LLM Leaderboard 3 by Weights-and-Biases.⁵ Because some of the LLMs we evaluate are not available on this leaderboard, we use the leaderboard from the Swallow project.⁶ Because the JMT-Bench scores on the two leaderboards have slightly different scores, we adjust the scores using a linear regression model to we compute a linear regression model to predict the score on Nejumi LLM Leaderboard 3 given the score on the Swallow project. The linear regression model is computed using the LLMs that are evaluated on

⁵<https://wandb.ai/wandb-japan/llm-leaderboard-3/reports/Nejumi-LLM-3--Vmlldzo30Tg2NjM2>

⁶<https://swallow-llm.github.io/evaluation/index.ja.html>

Model name	JP MT-Bench	YokaiEval	Base Model	(Continual) Pretraining	SFT	PL
calm3-22b-chat	6.93	0.573	from scratch	✓	✓	✓
llama-3.1-swallow-8b-instruct-v0.2	6.28	0.510	Llama-3.1-8B-Instruct	✓	✓	
llm-jp-3-13b-instruct	5.68	0.456	from scratch	✓	✓	
swallow-13b-instruct-v0.1	4.78	0.468	Llama-2-13b-hf	✓	✓	
llm-jp-3-3.7b-instruct	4.98	0.277	from scratch	✓	✓	
llm-jp-3-1.8b-instruct	4.70	0.299	from scratch	✓	✓	
tanuki-8x8b-dpo-v1.0	6.96	0.478	from scratch	✓	✓	✓
llama-3.1-70b-japanese-instruct-2407	7.54	0.635	Meta-Llama-3.1-70B-Instruct	✓	?	?
llama-3.1-swallow-70b-instruct-v0.1	6.61	0.615	Meta-Llama-3.1-70B	✓	✓	
llama-3-elyza-jp-8b	6.08	0.609	Meta-Llama-3-8B-Instruct	✓	✓	?
llama-3.1-swallow-8b-instruct-v0.1	5.37	0.511	Llama-3.1-8B	✓	✓	
tanuki-8b-dpo-v1.0	6.40	0.222	from scratch	✓	✓	✓
llama-3.1-swallow-8b-instruct-v0.3	7.02*	0.538	Meta-Llama-3.1-70B-Instruct	✓	✓	

Table 9: List of Japanese LLMs and their training processes for learning Japanese. ✓ shows that the training process is conducted using Japanese datasets. ? indicates that there is no information available.

```

###指示###
以下に表示されるユーザーの4択問題に対するAIアシスタントの解答の正誤を判定してください。
出力例のフォーマットに厳密に従ってtrueかfalseかnullの3値で応答を評価する必要があります。
出力が誤った選択肢を回答としている場合にはfalseと出力しなさい。
出力が正解の選択肢を回答としている場合にはtrueと出力しなさい。
出力が何も回答していない場合はnullと出力しなさい。
出力が正解の選択肢を意味している場合は出力の形式によらず正解と判定しなさい。
ただし、複数の選択肢を解答している場合にはnullと出力しなさい。

###例###

###4択問題###
日本の妖怪である「あすこここ」は、どのような特徴を持つ妖怪とされていることが多いでしょうか？以下の4つから回答を1つ選び、回答のみ出力しなさい。
- 人間の姿をしている
- 水中に住む
- 足が三本ある
- 夜にだけ現れ
###正解の選択肢###
足が三本ある
###AIアシスタントの応答###
「あすこここ」という妖怪についての具体的な情報は、一般的な妖怪図鑑や文献には記載されていないようです。そのため、この妖怪の特徴については明確な答えを提供することができません。しかし、もし仮に「あすこここ」の特徴を仮定するとすれば、選択肢の中から「足が三本ある」というものが最も一般的かもしれません。
###あなたの出力###
true

(中略)

###Task###

###4択問題###
{question}
###正解の選択肢###
{correct}
###AIアシスタントの応答###
{response}
###あなたの出力###

```

Figure 18: Prompt for GPT-4o Evaluation of the answer of the LLMs (Section 5). The English translation is available in Figure 19.

```

### Instructions ###
Please determine the correctness of the AI assistant's answer to the user's multiple-choice question displayed below.
You need to evaluate the response strictly in the format of true, false, or null as shown in the output example.
Output false if the response contains an incorrect choice.
Output true if the response contains the correct choice.
Output null if the response does not provide any answer.
If the response implies the correct choice, consider it correct regardless of the format.
However, if the response contains multiple choices, output null.
### Example ###
### Multiple-Choice Question ###
What is the characteristic that the Japanese yokai "Asukokoko" is often said to have? Choose one answer from the following four
options and output only the answer.
- Human appearance
- Lives in water
- Has three legs
- Appears only at night
### Correct Choice ###
Has three legs
### AI Assistant's Response ###
There doesn't seem to be specific information about the yokai "Asukokoko" in general yokai encyclopedias or literature.
Therefore, I cannot provide a definitive answer about the characteristics of this yokai. However, if we were to speculate on the
characteristics of "Asukokoko," the most common one among the options might be "Has three legs."
### Your Output ###
true
(omitted)
### Task ###
### Multiple-Choice Question ###
{question}
### Correct Choice ###
{correct}
### AI Assistant's Response ###
{response}
### Your Output ###

```

Figure 19: Prompt for GPT-4o Evaluation of the answer of the LLMs (Figure 18), translated to English.

Parameter	Value
temperature	0.7
top_p	1.0
max_new_tokens	200
Version (GPT-4o)	2024-05-13
Version (GPT-4o-mini)	2024-07-18
Azure Content filter	All set to High

Table 10: Hyperparameters used for GPT-4o and GPT-4o-mini via Azure OpenAI API.

Parameter	Value
temperature	0.1
top_p	1.0
max_new_tokens	128

Table 11: Hyperparameters for generating responses of LLMs to YokaiEval QAs.

Parameter	Value
Epochs	3
Learning rate	1e-5
Optimizer	AdamW
Batch size	4
Regularization factor (β)	0.1
LoRA r	64
LoRA α	16

Table 12: Hyperparameters used for DPO training.

both of the leaderboards. The model is as follows:

$$y = 9.576 \cdot x + 0.868, \quad (1)$$

where x is the score on the Swallow project and y is the estimated score if it were evaluated on the Nejumi leaderboard. We use these estimated scores in Section 5.

H Hyperparameters

Table 10 shows the hyperparameters used for GPT-4o and GPT-4omini in Section 4.1. We use the same hyperparameters for all the tasks. Table 11 shows the hyperparameters used for generating the answer to the YokaiEval from LLMs in Section 5. Table 11 is the list of hyperparameters for DPO training in

Section 5.2. We use the same hyperparameters for all three models.

I Reproducibility Statements

The code is available at <https://github.com/CyberAgentAILab/YokaiEval>. The dataset is available at <https://huggingface.co/datasets/cyberagent/YokaiEval>.

Our experiments use a closed model (GPT-4o) for data generation, analysis, and evaluation. The outputs of the closed model are published alongside the code at the GitHub repository.

The code is based on Python 3.10. The Cuda version is 12.1.0. Huggingface’s Transformers library (Wolf et al., 2020) is version 4.47.1. OpenAI library is version 1.58.1.

Evaluation of the LLMs (Section 5) is conducted using an NVIDIA A100 GPU with 80GB VRAM. The total time required for the experiments is estimated to be less than 60 hours.

J On the Necessity of Mono-Cultural Computational Folktale Research

One of the limitations of the study is that it only evaluates Japanese folktales and not of other communities. We argue that the analysis of multiple communities is extremely difficult, and we need to first establish the analyses of folktales of each community in order to conduct a well-established multi-cultural analysis.

To make sure that the methodology is valid and unbiased, one needs to carefully investigate the data entries manually to ensure that they are valid resources for analyzing the community. Otherwise, the analysis may result in biased or calligraphed analysis. For example, *Japanese Fairy Tales* by Yei Theodora Ozaki (Ozaki, 2024) is one of the books in the Project Gutenberg,⁷ which is a digital archive of cultural works used in many analyses (Fialkova and Yelenevskaya, 2001; Laudun and Goodwin, 2013; Brooke et al., 2015; Wu et al., 2023a,b).

However, there are several problems with using this book. First, the stories are literally translated to be accessible to people in Western countries, thus it is not a firsthand source of Japanese folktales. In the preface of the book, it is described that the translator intended to literally translate the stories:

This collection of Japanese fairy tales is the outcome of a suggestion made to

me indirectly through a friend by Mr. Andrew Lang. They have been translated from the modern version written by Sadanami Sanjin. These stories are not literal translations, and though the Japanese story and all quaint Japanese expressions have been faithfully preserved, they have been told more with the view to interest young readers of the West than the technical student of folk-lore.

For example, in “My lord bag of rice”, the warrior of the story, Tawara Toda, is referred to as a “knight”. Knight does not exist in Japan. This shows that the translation is not written for the understanding of Japanese culture or folktales.

Note that the stories by Sadanami Sanjin referred to as “the technical student of folk-lore” in the preface are also not folktales. Sadanami Sanjin is an author of children’s books writing them as novels, and he is not a folklorist.

Second, using the book for folktale study is against the will of the writer. In the preface of the book it is stated below:

...they have been told more with the view to interest young readers of the West than the technical student of folk-lore.

Thus, the translator does not intend the book to be used for folklore studies.

These problems are identifiable if one reads the preface (written in English) of the material. However, qualitative analysis of the materials of the study is often missing in multi-cultural evaluation as it needs a lot of materials to be investigated. We claim that we should indeed investigate each material if we are to do cultural or folklore studies even when we use computational methods, as the materials may not be the representatives of what they are intended to be in the study. The result of the analysis of such materials may not reflect the communities accurately enough which can spread inappropriate views of the communities.

⁷www.gutenberg.org