Synthetic Data in NLP: Evaluating GPT-4's Impact on Japanese Linguistic Acceptability Judgment Models

Masaki Kurita
Uppsala University
masaki.kurita.5792@student.uu.se

Abstract

This study investigates the use of GPT-4 generated synthetic data in training models for Japanese linguistic acceptability judgments. We investigate the impact of synthetic data across different syntactic categories and its implications for Natural Language Processing (NLP). The models, trained with different combinations of synthetic data and human-made data, are evaluated using the Matthews Correlation Coefficient. Results indicate that while synthetic data enhances performance in certain categories, its effectiveness is not uniform across all linguistic phenomena. This highlights the importance of careful synthetic data generation strategies for each linguistic phenomena. Our findings also suggest that models could overfit to synthetic data and emphasize the need for its careful and balanced integration into human-made data.

1 Introduction

Natural Language Processing (NLP) encompasses a wide variety of complex tasks, one of which is linguistic acceptability judgment. In this task, individuals intuitively differentiate grammatically correct from incorrect sentences (Someya et al., 2023). This provides empirical insights that help uncover the underlying rules and structures of language (Someya et al., 2023). It is fundamental to human linguistic competence and serves as a vital measure of grammatical knowledge and understanding (Chomsky, 1957). In light of this, Warstadt et al. (2019) took neural network approaches to automating acceptability judgments and also developed the English-based Corpus of Linguistic Acceptability (CoLA). This initiative paved the way for subsequent efforts in other languages, such as the Italian CoLA corpus and the Japanese Corpus of Linguistic Acceptability (JCoLA) (Trotta et al., 2021; Someya et al., 2023). However, these valuable datasets often face challenges, including limited representation of complex syntactic phenomena and imbalanced distributions between acceptable and unacceptable sentences.

The advancement of large language models like GPT langauge models offers new methods to address these challenges. GPT models have shown promise in generating synthetic datasets that are comparable to human-annotated data, especially in linguistically low-resource settings (Mercity, 2023; Ubani et al., 2023; Møller et al., 2023; Sashida et al., 2023; Xin et al., 2024). GPT-4 is the latest GPT model. Its potential to enhance the richness and diversity of linguistic datasets, which is highly expected but is still not well explored particularly in Japanese, forms the core of our research.

This study explores the utility of synthetic data generated by GPT-4 in enhancing models for Japanese linguistic acceptability judgments. Central to our research are two pivotal questions: Firstly, how does the performance of models vary across different syntactic categories and depending on various combinations of synthetic and human-made training data. Secondly, what insights can be gained about the strengths and limitations of synthetic data in augmenting NLP tasks from the variability in model performance across these categories and different training datasets?

Our investigation leads to several key findings. We observe that the effectiveness of synthetic data varies notably across different syntactic categories, highlighting the importance of data quality and representativeness. Certain categories see significant improvements with the inclusion of synthetic data, while others do not, underscoring the need for careful and tailored generation strategies to align with specific task requirements. We also find that the models' performance varies depending on the combinations of synthetic and human-made training data, indicating the need for the strategic and balanced integration of synthetic data into human-made data.This study sheds light on the capabilities and constraints of using synthetic data in NLP, of-

fering valuable insights for future research in the field.

2 Related Work

2.1 GPT-4

GPT-4 is a large language model which accepts text inputs and produces text outputs (OpenAI, 2023). GPT-4 performs comparably to humans on various benchmarks, such as Massive Multitask Language Understanding (MMLU) (OpenAI, 2023). It also passed a simulated bar exam with a score around the top 10% of test takers (OpenAI, 2023).

GPT-4 excels in text generation, which was demonstrated in standard similarity metrics like Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and bilingual evaluation understudy (BLEU), where its responses closely align with reference answers (Bubeck et al., 2023). Furthermore, GPT-4 excels in creating complex and creative text, blending skills such as mathematical reasoning and poetic expression with natural language generation (Bubeck et al., 2023).

GPT-4, like its predecessors such as GPT-3.5, is trained on a dataset comprising a vast range of internet text in various languages, which enables the model to learn and predict text in those languages (OpenAI, 2023; Brown et al., 2020).

The ability of GPT models in Japanese is acknowledged by several researches. For example, GPT-4 passed several Japanese exams, such as medical licensing examinations (Toyama et al., 2023; Kasai et al., 2023; Kaneda et al., 2023). Furthermore, GPT-4 and its predecessors were also effective in text data augmentation in Japanese (Sashida et al., 2023; Xin et al., 2024)

These strengths make GPT-4 an ideal candidate for generating synthetic Japanese text data.

2.2 GPT for Data Augmentation

As large language models become more powerful, GPT language models have been shown as effective for text data augmentation in recent years (Mercity, 2023; Ubani et al., 2023; Møller et al., 2023; Sashida et al., 2023; Xin et al., 2024).

Møller et al. (2023) explored the efficacy of GPT-4 in creating synthetic data for tasks like sentiment analysis, hate speech detection, and social dimensions classification. They found that models trained on synthetic data generated by GPT-4 can perform comparably to those trained on human-annotated data, when detailed prompts were used to generate synthetic data (Møller et al., 2023). However, in their experiments, models trained on human-annotated data still held a slight edge in predictive power in certain tasks. They underscored the continued importance of human oversight and the need for effective prompt engineering to maximize the benefits of synthetic data generated by language models like GPT-4.

Dai et al. (2023) also performed data augmentation by GPT-3, which is GPT-4's predecessor. They gave samples from all various classes of their existing data as inputs to GPT-3 and prompted it to generate additional samples that maintain semantic consistency with the existing data. Specifically, they prompted GPT-3 to rephrase each input sentence into six new sentences, thereby enriching and diversifying the existing dataset. In text classification tasks, their BERT model trained with synthetic data showed the superior performance over models trained on data augmented by conventional methods, such as exchanging random characters, in terms of testing accuracy.

Additionally, Xin et al. (2024) proved GPT-3's Japanese text generation ability through its use in creating a Japanese emotional speech corpus. GPT-3 was employed for automatic script generation, producing dialogues that effectively conveyed six basic emotions (Xin et al., 2024). The effectiveness of GPT-3 in this context was evidenced by the improved phoneme coverage and emotion recognizability in the corpus compared to the previous corpora (Xin et al., 2024).

GPT-4 outperforms its predecessors on various tasks and benchmarks (OpenAI, 2023). Based on its high performance across various tasks and benchmarks as detailed earlier, GPT-4's potential for text data augmentation, while not as extensively investigated as GPT-3, is anticipated to be more robust and effective.

2.3 JCoLA

Japanese Corpus of Linguistic Acceptability (JCoLA) is the first Japanese large-scale acceptability judgment task dataset, which consists of 10,020 sentences manually extracted from linguistics textbooks,handbooks and journal articles (Someya et al., 2023). The data is split into in-domain data and out-of-domain data. In-domain data is relatively simple sentences and acceptability judgments extracted from textbooks and handbooks, without annotation of syntactic categories. Out-of-

domain data is theoretically significant acceptability judgments and sentences extracted from journal articles. It is categorized by 12 linguistic phenomena and has annotations of syntactic categories.

The 12 syntactic phenomena are "simple", "Arg. Str.", "ellipsis", "filler-gap", "control/raising", "island effects", "NPI/NCI", "verbal agr.", "binding", "morphology", "nominal structure", "quantifier". To clarify the ambiguous categories, the 'Arg. Str.' (Argument Structure) category includes sentences evaluated for the correct ordering of words around a verb and their corresponding grammatical roles. The 'Simple' category contains sentences that do not have marked syntactic structures and do not belong to other categories. For instance, it includes a simple transitive sentence.

Someya et al. (2023) trained various Japanese language models on in-domain data of JCoLA and tested them on out-of-domain data for acceptability judgment task. Their language models trained on in-domain data performed well in simpler categories such as "Simple", indicating their proficiency in basic syntactic recognition. However, there was a notable disparity in performance across more complex linguistic categories. Models tended to show relative strength in categories like "Binding" and "Argument Structure" due to the presence of clear syntactic patterns and a higher proportion of positive examples. Conversely, they struggled with categories like "NPI/NCI" and "Verbal Agreement", which involve more abstract and nuanced rules, including long-distance dependencies. This variation in model performance across different syntactic phenomena underscored the current limitations of language models in fully grasping the breadth and depth of complex linguistic structures.

3 Methodology

In our experiment, we trained models on various combinations of Japanese synthetic data generated by GPT-4 and JCoLA's in-domain data for linguistic acceptability judgements and compared their performance on JCoLA's out-of-domain data. This section explains the methods used in the experiment.

3.1 Data Extraction

As the first step, we extracted sentences from JCoLA to provide GPT-4 with them as examples to generate synthetic data. Regarding out-of-domain data of JCoLA, some syntactic categories only have a few acceptable and/or unacceptable sentences. Therefore, we extracted two to three acceptable and unacceptable sentences from every syntactic category for an equal number of examples to be extracted from all the categories.

Regarding in-domain data of JCoLA. since they are not categorized, we randomly extracted 100 unacceptable sentences and 100 acceptable sentences from it.

3.2 Prompt Design

Møller et al. (2023) stated that when detailed prompts explaining relevant concepts and tasks were given to GPT-4, it produced better synthetic data. Therefore, we crafted prompts that contain example sentences and their syntactic categories. We also included the explanations about both the categories and the acceptability judgement task. In addition, we included why the examples are unacceptable in the prompts when we generated unacceptable sentences.

To generate in-domain synthetic data, as it does not have annotation of syntactic categories, we did not include explanations of syntactic phenomena in the prompts.

We also used the responses generated by GPT-4 in the following prompts. Specifically, we responded to GPT-4 by explaining which synthetic sentences in their responses were desirable or not and the reasons. We also included previously generated sentences to avoid repetitive sentence generation.

Li et al. (2023) found that various large language models, such as GPT-4, improve their output when the prompts contain emotional stimuli, such as "Believe in yourself and push your limits". Therefore, we also contained emotional stimuli in our prompts.[1]

3.3 Synthetic Data Generation

We gave the crafted prompts to GPT-4 to generate synthetic data. Regarding out-of-domain synthetic data, we generated 100 acceptable sentences and 100 unacceptable sentences for every syntactic phenomena category. Some generated sentences belonged to several categories. As a result, we generated 1090 acceptable sentences and 1060 unacceptable sentences for out-of-domain synthetic

[1]Basic prompts that we used can be found at: https://github.com/masauppsala/Synthetic-Japanese-Data-by-GPT-4/blob/main/Example%20Prompts

data.

To keep the amount of training data equal, we also generated 1090 acceptable sentences and 1060 unacceptable sentences for in-domain synthetic data.

Initially, a Japanese native speaker evaluated a small sample of synthetic sentences (about ten) to assess quality, examining if acceptable sentences were indeed acceptable and aligned with the syntactic phenomena of the examples, and if unacceptable sentences maintained their unacceptability while preserving the basic syntax of their acceptable counterparts. Sentences with context-dependent acceptability were excluded, in line with Someya et al. (2023). Upon confirming the validity of these initial synthetic sentences, we increased the volume of data generation. Thereafter, for each set of sentences generated by GPT-4 to a prompt, we reviewed each sentence to ensure they accurately represented the intended acceptability and syntactic categories. Duplicates were removed. [2]

3.4 Model Training

We trained Waseda RoBERTa-seq128LARGE (Kawahara Lab, Waseda University, 2021) for acceptability judgement. We chose this model, as it exhibited the highest performance in acceptability judgements when it was trained on JCoLA by Someya et al. (2023).

As Table 1 shows, we trained 7 different Waseda RoBERTa models in total on various combinations of Japanese synthetic data generated by GPT-4 and JCoLA's in-domain data, which is randomly extracted 1090 acceptable and 1060 unacceptable in-domain sentences. We did not use any of JCoLA's out-of-domain data for training and only used it for testing and extracting examples to generate out-of-domain synthetic data. This is because its amount is limited and its syntax categorization enables us to evaluate the models performances across different syntax phenomena when it is used for testing.

Specifically, we trained three models: JCoLAID on JCoLA's in-domain data, SynOD on out-of-domain synthetic data, and SynID on in-domain synthetic data, each with 1090 acceptable and 1060 unacceptable sentences.

Additionally, we trained three more models: SynOD+JCoLAID on a combination of out-of-domain synthetic data and JCoLA's in-domain data, SynID+JCoLAID on a combination of in-domain synthetic data and JCoLA's in-domain data, and SynOD+SynID on both out-of-domain and in-domain synthetic data, each with 2180 acceptable and 2120 unacceptable sentences. Lastly, we trained the AllDataMix model on all three types of data - out-of-domain synthetic data, in-domain synthetic data, and JCoLA's in-domain data, with a total of 3270 acceptable and 3180 unacceptable sentences.

We used these combinations of datasets to assess the impact of different types of synthetic and human-made data on model performance.

We split each combination of data into 80% training and 20% evaluation data.

We used the Trainer class from the transformers library, a learning rate of 5e-5, the number of training epochs of 10, and batch sizes of 16 and 64 for training and evaluation, respectively. The setup also includes warmup steps of 300 and an weight decay of 0.02 for regularization, a linear learning rate scheduler was used , and the training was conducted with mixed-precision (FP16) for enhanced performance.

3.5 Evaluation

JCoLA's out-of-domain data was used as test data, which includes the sentences that were given as examples to GPT-4 to generate synthetic out-of-domain data.

Following the methodology adopted by Someya et al. (2023), we utilized the Matthews Correlation Coefficient(MCC) as our evaluation metric.

MCC is particularly suitable for datasets with imbalanced classifications. MCC is a correlation coefficient between the actual and predicted binary classifications; it returns a value between -1 and 1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction, and -1 an inverse prediction.

The choice of MCC is especially valid given the imbalanced nature of our test dataset(JCoLA's out-of-domain data), where the distribution of acceptable and unacceptable sentences is skewed. In such cases, metrics like accuracy can be misleading, as they don't adequately reflect the model's performance on the minority class. MCC, on the other hand, takes into account all four confusion matrix categories (true positives, false positives, true negatives, and false negatives), offering a more balanced

[2]The synthetic data generated for this study can be found at: https://github.com/masauppsala/Synthetic-Japanese-Data-by-GPT-4

measure that is less prone to being skewed by the class imbalance.

4 Results and Discussion

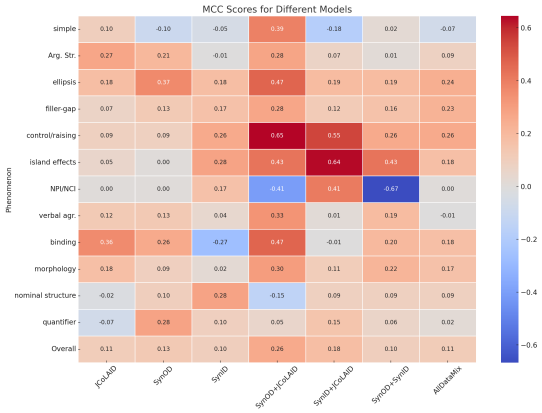

Figure 1: Performance of each language model on JCoLA out-of-domain test set by phenomenon.

As Figure 1 shows, the models trained on different combinations of synthetic data and JCoLA's in-domain data showed varied performance across different syntactic phenomena.No single model consistently outperformed others across all categories.

Model	Overall MCC
JCoLAID	0.110809343
SynOD	0.129790085
SynID	0.09801437823
SynOD+JCoLAID	0.2571400201
SynID+JCoLAID	0.1794172079
SynOD+SynID	0.09751318873
AllDataMix	0.1141004299

Table 1: Overall Average MCC Scores for Different Models

The results of our experiment provide valuable insights into the potential and limitations of using synthetic data generated by GPT-4 for training NLP models, specifically in the context of Japanese linguistic acceptability judgment.

4.1 Overall Performance

Firstly, in terms of the overall performances, among JCoLAID , SynOD, and SynID models, the SynOD model has the highest MCC, which could be because out-of-Domain synthetic data was generated based on extracted examples from test data(JCoLA's out-of-domain data) and could be the most similar to it. Surprisingly, the JCoLAID

Phenomenon	Average MCC Score
simple	0.0166
Arg. Str.	0.131
ellipsis	0.2614
filler-gap	0.1641
control/raising	0.3072
island effects	0.2861
NPI/NCI	-0.0714
verbal agr.	0.1183
binding	0.1691
morphology	0.1562
nominal structure	0.0688
quantifier	0.0843
Overall	0.1360

Table 2: Average MCC Scores for Each Syntax Category

model performed comparably to the SynOD model, even though it was not trained on sentences generated from examples of the test data. We assume this could be because of the rich diversity in the sentence features of JCoLA's in-domain data, such as sentence structures, word order, and vocabulary, as it was extracted from various textbooks and handbooks about Japanese syntax. Furthermore, it is also possible that JCoLA's in-domain data and JCoLA's out-of-domain data are not significantly different in terms of sentence features because according to Someya et al. (2023), the difference is that the former was extracted from textbooks and hand-books on Japanese syntax and the latter was extracted from research journals, instead of the difference being based on sentence features' differences. On the other hand, the relatively lower MCC shown by the SynID model suggests that the data does not well reflect the features of the JCoLA's in-domain data. This could be because JCoLA's in-domain data does not have syntax categories and lacks unique features that GPT-4 could use for generating synthetic data. We also included all the examples from JCoLA's in-domain data in a single prompt and kept using it to generate in-domain synthetic data. When GPT-4 was generating in-domain synthetic data, it tended to generate sentences that are only similar to a subset of the examples and ignore other examples. Therefore, we had to include "pay equal attention to all the examples" in the following prompts, which probably was not effective given the experimental results. Segmenting these examples and using them in multiple prompts

might yield more balanced synthetic data.

Secondly, among SynOD+JCoLAID, SynID+JCoLAID, and SynOD+SynID models, the highest MCC was achieved by the SynOD+JCoLAID model, suggesting an effective synthesis of the two datasets without introducing bias. While the SynID+JCoLAID model has better MCC when the two datasets are jointly used, SynOD+SynID have lower MCC than when each of them is used alone despite their larger dataset sizes. This suggests that using a large amount of synthetic data without human-made data has a negative influence, possibly due to the model overfitting to the synthetic data, straying from the human-made test data(JCoLA out-of-domain data).

Finally, despite its largest training dataset size, the AllDataMix model has lower MCC than several other models. This again suggests that using a large amount of synthetic data with a small amount of human-made data has a negative influence causing bias.

These findings highlight the importance of a balanced approach to data augmentation in model training. While synthetic data can enhance the diversity and specificity of training material, it is crucial to maintain a proportionate inclusion of human-made data to ensure that the models remain grounded in the authentic structures and complexities of the real-world language. This balance is key to avoiding overfitting to synthetic patterns that might not accurately reflect real-world language use.

In our experiment, the size of the training dataset did not correlate with MCC. This reinforces the notion that training data quality, in terms of alignment with the test set and absence of bias-inducing noise, is crucial over mere dataset size.

4.2 Performance by Syntax Phenomena

Our results indicate that synthetic data can contribute to capturing certain complex linguistic structures. This is evident, as in certain categories like 'control/raising' and 'island effects' models trained with synthetic data are showing relatively high performance. This could be attributed to several reasons.

Firstly, it could be attributed to Japanese language characteristics. For instance, control/raising and island effects feature distinct syntactic patterns in Japanese.In a control structure, a subject in the main clause also often serves as the subject in the embedded clause. In island effects, there are constraints on elements, elements within certain clauses unable to be arbitrarily rearranged. These distinct patterns might help models to more easily distinguish acceptable from unacceptable sentences.

Secondly, when models perform well on some syntax categories, it could be that those syntax categories were represented by relatively many sentences in the training data. JCoLA's in-domain data or the in-domain synthetic data might have more sentences belonging to those well-performed categories than others, although this is speculative due to the lack of syntactic category annotations in the in-domain data.

However, the effectiveness of synthetic data was not uniform across all syntactic categories. In 'NPI/NCI' and 'nominal structure' categories, models trained with out-of-domain synthetic data exhibited negative MCC scores. This could be due to unacceptable sentences' features in these categories of the test data(JCoLA's out-of-domain data). In 'NPI/NCI', unacceptability in synthetic data and test data tended to be due to negative elements being used in inappropriate syntactic and semantic contexts. That means understanding the overall structure and meaning of the sentence is important and might have been difficult for models to learn. In 'nominal structure', unacceptability in synthetic data and the test data tended to be due to incorrect use of particles, which could be smaller and more difficult to capture compared with sentence and word level mistakes. This is how when we were generating unacceptable sentences of synthetic out-of-domain data, there tended to be a pattern in how to violate syntactic rules in each syntactic phenomena category. It is likely that some types of unacceptability are more distinct and easier to capture in than other types, leading to different performances across various syntactic phenomena.

The low accuracy in the NPI/NCI category, observed in both our study and Someya et al. (2023), might be attributed to the same underlying challenge they proposed: the complexity of capturing long-distance dependencies, which appears to be a consistent barrier impacting model performance in this syntactic phenomenon.

In the 'Simple' category, some models also showed negative MCC. This result contrasts with the high accuracy observed in the 'Simple' cate-

gory by Someya et al. (2023), suggesting that while our synthetic data approach holds promise, it may not yet fully capture the subtle features of such linguistically straightforward categories.

Both studies observed various performance across all syntactic categories. This aligns with the idea that certain linguistic constructs are inherently easier for models to grasp.

4.3 Synthetic Data Generation by GPT-4

Efficient synthetic data generation was achieved using GPT-4, which produced 50 Japanese sentences per prompt. Initially, GPT-4 faced difficulties in generating syntactically unacceptable sentences, often mistakenly generating acceptable ones. This issue likely stemmed from GPT-4 not being extensively trained on incorrect sentence structures (OpenAI, 2023). However, by specifying examples of incorrect sentences in their output and explaining their flaws, we effectively guided the model to produce more desired outputs. Evaluating whether the generated sentences accurately represented their intended syntactic categories was time-consuming, emphasizing the user's expertise in the relevant field as a key factor in the quality and efficiency of synthetic data generation.

4.4 Limitations

This study encountered several limitations. Some syntactic categories of the test data(JCoLA's out-of-domain) only contained a few acceptable and/or unacceptable sentences. For example, the binding category has 46 acceptable but only 2 unacceptable sentences. The simple category has 28 acceptable but only 4 unacceptable sentences. Therefore, the models performances on the test data may not be generalizable.

Another limitation is the inconsistent size of training datasets, potentially affecting model performance. Variations in size make it difficult to attribute performance changes to data quality or quantity. Future research should standardize data size to isolate its impact.

Additionally, the reliance on MCC as the sole evaluation metric may not comprehensively capture all dimensions of model performance, particularly in the context of nuanced linguistic phenomena.

Finally, when we extracted examples from in-domain and out-of-domain data of JCoLA, we extracted more in-domain examples(200 sentences) than out-of-domain examples. This might have introduced unfairness in generating synthetic data.

4.5 Future Research Direction

Future research could benefit from employing a dual-classifier approach, training and utilizing both an acceptability classifier and a syntax category classifier. This would ensure a more refined synthetic data corpus, balancing acceptability status and syntactic representation. Exploring ensemble methods, such as weighted voting or stacking, could also be advantageous. Combining different models may enhance overall accuracy and address the varied performance across linguistic phenomena observed in this study.

5 Conclusion

This study investigated the utility of synthetic data generated by GPT-4 for training models in the context of Japanese linguistic acceptability judgment. Our findings show both the potential and constraints of employing synthetic data in this domain. The application of synthetic data showed promise in capturing complex syntactic structures, such as 'control/raising' and 'island effects', demonstrating the model's capability to capture intricate linguistic patterns. However, performance varied significantly across different syntactic categories, with some, including 'NPI/NCI' and 'nominal structure', displaying negative MCC scores. This variation underscores the need for the efforts to make a synthetic dataset diverse and representative, which could be achieved by elaborating on the synthetic data generation strategies for each syntax category. This study also emphasizes the importance of a balanced approach in the integration of synthetic data for training models in Japanese linguistic acceptability judgments. Our results demonstrate that while GPT-4 generated synthetic data can help models perform well in certain situations, it is essential to combine it with a proportionate amount of human-made data. This combination ensures that models are not distant from the authentic complexities and structures of real-world language, without overfitting to synthetic language patterns that may not accurately mirror actual language usage. This research contributes to the understanding of large language models' capabilities in data augmentation and sets the stage for future explorations into more sophisticated model training strategies.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation.

Yudai Kaneda, Ryo Takahashi, Uiri Kaneda, Shiori Akashima, Haruna Okita, Sadaya Misaki, Akimi Yamashiro, Akihiko Ozaki, and Tetsuya Tanimoto. 2023. Assessing the performance of gpt-3.5 and gpt-4 on the 2023 japanese nursing examination. *Cureus*, 15:e42924.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations.

Kawahara Lab, Waseda University. 2021. Waseda roberta-seq128large. `https://huggingface.co/nlp-waseda/roberta-large-japanese`. Accessed: 2023-11-01.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli.

Mercity. 2023. Using chatgpt to build synthetic datasets.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks.

OpenAI. 2023. Gpt-4 technical report.

Masaki Sashida, Kiyoshi Izumi, and Hiroki Sakaji. 2023. Extraction of sdgs-related sentences from sustainability reports using bert and chatgpt. *Jinkou Chinou Gakkai Daini Shu Kenkyuu Kai Shiryou [Materials of the Second Kind Research Society of the Japanese Society for Artificial Intelligence]*, 2023(FIN-031):55–60.

Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2023. Jcola: Japanese corpus of linguistic acceptability.

Y. Toyama, A. Harigai, M. Abe, et al. 2023. Performance evaluation of chatgpt, gpt-4, and bard on the official board examination of the japan radiology society. *Japanese Journal of Radiology*. Jpn J Radiol.

Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929–2940, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Solomon Ubani, Suleyman Olcay Polat, and Rodney Nielsen. 2023. Zeroshotdataaug: Generating and augmenting training data with chatgpt.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Detai Xin, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2024. Jnv corpus: A corpus of japanese nonverbal vocalizations with diverse phrases and emotions. *Speech Communication*, 156:103004.