Linguistic Proficiency of Humans and LLMs in Japanese: Effects of Task Demands and Content

May Reese and Anastasia Smirnova
San Francisco State University
1600 Holloway Avenue, San Francisco, California, USA
mreese@mail.sfsu.edu
smirnov@sfsu.edu

Abstract

We evaluate linguistic proficiency of humans and LLMs on pronoun resolution in Japanese, using the Winograd Schema Challenge dataset. Our main research question is whether task demands and content effects affect performance in these two target groups. First, we found that in the baseline condition, humans outperform LLMs. This finding is consistent with the observation that the language of evaluation is important and that humans perform better than LLMs in lower-resourced languages. Second, we find strong evidence for the effect of task demands in both humans and LLMs. As task demands increase due to syntactic incongruencies in the input, accuracy rates fall for both groups. Third, we found evidence for content effects. In the relevant condition, the content of the scenarios referenced US culture, a favorable condition for LLMs and an adversarial condition for Japanese speakers. We found that LLMs outperformed humans, providing strong evidence for content effects.

1 Introduction

Large Language Models (LLMs) display an impressive set of abilities that require proficiency in human language. They perform well on text summarization (van Schaik and Pugh, 2024), translation (Wang et al., 2023), and writing (Herbold et al., 2023). On many of these tasks, LLMs perform better than humans. For example, Herbold et al. (2023) asked professional evaluators to assess argumentative essays generated by ChatGPT and by humans. The results suggested that the GPT-generated essays consistently achieved higher rankings and were deemed by experts to be of higher quality.

These results are encouraging. However, when LLMs are evaluated on seemingly simpler tasks targeting basic linguistic proficiency, such as the ability to distinguish grammatical sentences from ungrammatical, or meaningful expressions from nonsensical, the results are mixed. Dentella et al. (2024) found that the ability of LLMs to decide whether a sentence is grammatical is much worse than that of humans. While GPT-4 achieved significantly higher accuracy, LLMs performed at the chance level when results were averaged across all tested models. Moreover, LLM responses displayed errors that humans would never make. The authors concluded that LLMs' understanding and performance on tasks involving grammar is not human-like (cf. also Katzir, 2023). In another study, Riccardi et al. (2024) evaluated the ability of LLMs to detect whether a two-word combination is meaningful (*baby clothes*) or nonsensical (*clothes baby*). In humans, this judgement requires knowledge of syntax and semantics. The rightmost word is the syntactic head, and it determines the meaning of the construction: *baby clothes* are a type of clothes. The same rule would make *clothes baby* nonsensical. Riccardi et al. (2024) found that even the most advanced models, such as GPT-4, performed poorly compared to humans. One interesting tendency was for LLMs to err on the side of interpreting nonsensical phrases as meaningful.

The discrepancies between LLMs and humans on basic linguistic tasks have implications for LLM integration in everyday life. There are many applied contexts in which it is highly desirable for LLMs to behave similarly to humans with respect to language understanding. For example, if LLMs' abilities are leveraged in educational contexts to provide feedback on children's writing or on L2 learners' essays, LLMs' assessment of what is grammatical and what is not should parallel the assessment of human experts. Riccardi et al. (2024) identified similar challenges for a workplace context. If the task description or a request does not make sense, be it due to human error or malicious intent, LLMs should behave like a professional human expert would – by asking for clarification or by denying the request, not by interpreting it as

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 201–211

sensible across the board, a tendency that LLMs in their study displayed.

Studies that found performance differences between LLMs and humans on basic linguistic tasks were criticized for using evaluation methods that disadvantage LLMs (Lampinen, 2024; Hu et al., 2024). For example, Lampinen (2024) found that when LLMs are provided with a sufficient number of examples as part of the prompt, they achieve human-like performance when distinguishing grammatical sentences from ungrammatical sentences. Another criticism pertained to the use of metalinguistic prompts, which disadvantage LLMs (Hu et al., 2024). These authors argue that subpar performance should not be interpreted as lack of competence. In fact, studies suggest that LLMs and humans perform similarly on tasks that target basic linguistic proficiency. Hu and Frank (2024) argue that increasing the task demand can lead to lower accuracy in LLMs, just like an increased cognitive load leads to worse performance in humans. Lampinen (2024), focusing on reasoning tasks, also found that the content of the task can either facilitate or hinder performance, and that humans and LLMs show similar content effects.

Our work continues the line of research evaluating LLMs performance vis-à-vis humans on tasks requiring linguistic proficiency. To address the most recent debate about the effect of task demand and content on LLMs and humans, we evaluate their performance as we manipulate these conditions. Unlike previous studies, we focus on Japanese.

2 Evaluating Linguistic Proficiency With the Winograd Schema Challenge

2.1 The Winograd Schema Challenge as a Test of Linguistic Proficiency

In this study we use the Winograd Schema Challenge (WSC). The WSC was originally designed to evaluate machine intelligence as an alternative to the Turing test (Levesque et al., 2012). However, despite its promise and widespread application as a benchmark for commonsense reasoning, it is now generally acknowledged in the literature that the test falls short of assessing machine intelligence (Kocijan et al., 2023). At best, it is a test of linguistic proficiency (Browning and LeCun, 2023), and we use it as such.

The test consists of different scenarios, each of which has a pair of sentences. The classic example in (1) shows that each sentence introduces two entities, the city councilmen (A) and the demonstrators (B), and includes an ambiguous pronoun *they* that refers to one of the entities. The task is to establish the correct referent for the pronoun. We refer to this task as pronoun resolution. The interpretation of the pronoun arises from the meaning of the words *fear/advocate*. In the first sentence, the state of fear is attributed to the city councilmen (they = city councilmen), and in the second example, the action of advocating violence is attributed to the demonstrators (they = demonstrators).

(1a) The city councilmen (A) refused the demonstrators (B) a permit because they **feared** violence.

(1b) The city councilmen (A) refused the demonstrators (B) a permit because they **advocated** violence.

The authors of the WSC assumed that humans would perform at an accuracy level close to 100% (Levesque et al., 2012). Empirical studies revealed a different picture. Bender (2015) showed that human participants achieve 92% accuracy on well-crafted WSC sentences in English. Participants reported several difficulties, including unfamiliarity with certain concepts, such as *crop duster* or *bassinet*. Unfamiliar words and concepts can lead to an increase in task demand and possibly lower accuracy rates. Moreover, the content of the question and whether it aligns with or contradicts participants' expectations and personal experience can also have an effect on accuracy. In one of the scenarios, oatmeal cookies were preferred to chocolate cookies. Some participants found this unnatural and chose chocolate cookies as the answer to the pronoun resolution task, even though this incorrect answer contradicted information in the scenario (see Bender (2015) for discussion).

When LLMs were evaluated on the original WSC datasets, they performed worse than humans. However, training on larger datasets and fine-tuning helped. Language models gradually reached an accuracy of 90% (Sakaguchi et al., 2021). The most recent LLMs perform at 94% accuracy levels when evaluated in English (Artkaew, 2025).

2.2 The WSC in Other Languages: Human and LLM Performance

As the use of the WSC for evaluation benchmarks grew in popularity, the original WSC datasets de-

veloped for English were translated into other languages. However, translations to typologically different languages proved to be challenging. One set of difficulties pertained to typological and grammatical differences between the source language (English) and other target languages. English does not encode grammatical gender, animacy, or formality levels, and this presents a translation challenge. Research teams approached these challenges in varying ways. For example, when translating the WSC to French, Amsili and Seminck (2017) made changes to the original examples to achieve naturalness. The same strategy is reported by Artkaew (2025) for Thai. On the other hand, the authors of the Japanese Winograd Schema Challenge, WSCR-ja, noted that some translations resulted in ungrammatical examples due to structural differences between English and Japanese, but they decided to keep the examples in the dataset (Shibata et al., 2015).

Another translation difficulty pertains to cultural knowledge. Artkaew (2025) observed that an English scenario about playing cards uses the expression 'run of good luck', which is not natural in Thai. Another example pertained to a game of tag and how the chaser can be identified. In both cases, Artkaew (2025) chose to modify the original scenarios to make them more culturally appropriate. In their discussion of the Japanese WSC, Shibata et al. (2015) also acknowledged culturally inappropriate examples.

Comparison between human performance on the translated sets and human performance on the comparable dataset in English reveals differences in accuracy rates. Artkaew (2025) found that humans achieve 88% accuracy on the Thai WSC, which is lower than the 92% accuracy level reported for the English WSC (cf. Bender, 2015). Artkaew (2025) suggests that these differences should be attributed to translation effects and the difficulty of adapting scenarios from English to other languages.

There are also interesting differences in how language models perform on translated datasets compared to models evaluated on the original English datasets. Hashimoto et al. (2023) use the WSCR-ja by Shibata et al. (2015) to fine-tune BERT, a language model. Model fine-tuning helps increase accuracy on certain tasks, such as pronoun resolution. They found that the accuracy level increased from 57% to 58%, a modest gain. In comparison, fine-tuning the English model on the corresponding dataset in English leads to more significant gains. In the case of model evaluation, different factors might affect performance, including model size and architecture. Hashimoto et al. (2023) explicitly discuss the quality of the translated examples in the WSCR-ja, including cases of mistranslation, unfamiliar words, and cultural concepts, as a possible reason for smaller accuracy gains of their model after fine-tuning on the WSCR-ja. Results of evaluating more recent models on translated WSC datasets also show that they underperform compared to the base rate for English models. Artkaew (2025) reports that the accuracy of the best performing LLM on the Thai WSC is only 79.65% (Claude-3-Opus), compared to 94% on the English WSC.

3 Study

In this study, we use the WSC in Japanese to evaluate LLM and human performance on pronoun resolution. We focus on three conditions: (i) the baseline condition, (ii) a condition that manipulates task demands and (iii) a condition that manipulates content effects. The results on the baseline condition allow us to establish how LLMs perform vis-à-vis humans in the default setting. The null hypothesis is that LLM performance will parallel human performance. In the condition that manipulates task demands, we create adversarial conditions for both humans and LLMs and predict that this will negatively affect their performance. In the condition that manipulates content effects, we create favorable conditions for LLMs but adversarial conditions for humans, and we expect that it will increase LLM accuracy rates.

3.1 Materials

Our stimuli are derived from the WSCR-ja set (Shibata et al., 2015). This dataset is a translation of the Definite Pronoun Resolution (DPR) set (Rahman and Ng, 2012). Unlike other WSC datasets, the DPR set scenarios were crowdsourced from undergraduate students in the US and many of the original criteria of the test were relaxed. WSCR-ja consists of 941 question pairs which are split into a train set (659 pairs) and a test set (282 pairs). We performed a qualitative analysis of the entire WSCR-ja test set. Consistent with the observations in the previous literature, we found that the data was not homogeneous. In addition to grammatical and well-formed sentences, there were ungrammatical sentences and scenarios that reference cultural concepts that might not be familiar to Japanese

speakers with limited experience of US culture. The US cultural references are a carry-over from the original dataset and the translated dataset reflects the cultural background of the content creators. These different categories of data correspond to the three conditions we define below. In addition, we noticed substantial variation between stimuli. Some stimuli were syntactically complex, or had multiple pronouns and/or referents with non-core grammatical roles (indirect object or object of preposition). Since all these factors can affect performance on pronoun resolution (Kehler et al., 2008), we excluded such items from consideration. (See Appendix A.1 for examples of rejected stimuli and explanations for rejection). Based on our qualitative analysis, we formulated specific criteria for selecting stimuli for the three conditions and validated selected examples and their appropriateness for each condition with a linguistic consultant, a native speaker of Japanese. For each condition, we selected five scenarios for a total of ten sentences per condition. See Appendix A.2 for the full list of stimuli.

3.1.1 The 'Good' Condition

The baseline condition measures the performance on the pronoun resolution task in the absence of any other interfering factors. The stimuli in this condition adhered to the original considerations for the WSC dataset (Levesque et al., 2012). Specifically, subjects should not be able to induce correct answers based on selectional restrictions or word co-occurrence statistics. The scenario in (2) (from Kocijan et al., 2023) violates the selectional criteria rule since only women can be pregnant and only pills can be carcinogenic. The choice of the correct referent becomes trivial in this context. The scenario in (3) is ruled out based on statistical co-occurrence considerations (*racecar* and *fast* frequently co-occur)(from Kocijan et al., 2023).

> (2) The women stopped taking the pills because they were [pregnant/carcinogenic].
>
> (3) The racecar zoomed by the school bus because it was going so [fast/slow].

(4) is an example scenario selected for this category.

> (4a) ジョーはアダムより良い香りがする。彼は日頃からシャワーを浴びるからだ。 'Joe smells better than Adam since he showers regularly.'
>
> (4b) ジョーはアダムより良い香りがする。彼はめったにシャワーを浴びないからだ。 'Joe smells better than Adam since he hardly ever showers.'

3.1.2 The 'Grammar' Condition

The stimuli in this condition are designed to measure the effect of task demand on performance. There are different ways to manipulate task demand, but here we focus on the effect of grammar. Specifically, we hypothesized that syntactically incongruent stimuli will increase task demand and reduce accuracy rates. Scenarios were selected for the 'grammar' condition if at least one sentence in the pair is grammatically unacceptable or has been mistranslated so that the meaning is significantly different. Sentences may also not adhere to the original WSC constraints. (5) is an example set from this scenario. While the pronoun *she* might be an acceptable pronoun for a car in English, this is not the case for Japanese, resulting in (5a) being ungrammatical.

> (5a) シーラは古いポンコツ車を修理しようとした。彼女は30年も車に取り組んでいなかったにも拘らずだ。 'Sheila tried to repair the old jalopy, even though she had not worked on cars in three decades.'
>
> (5b) シーラは古いポンコツ車を修理しようとした。彼女は30年も走っていなかったにも拘らずだ。 'Sheila tried to repair the old jalopy, even though she had not run in three decades.'

3.1.3 The 'Culture' Condition

This condition is designed to test content effects on performance. Familiarity with specific cultural concepts as well as the lack thereof can affect accuracy ratings. For this condition, we selected scenarios that referenced US cultural concepts. We hypothesized that such scenarios will align with LLMs' competence, thus boosting their performance, but would disadvantage Japanese speakers. (6) is an example scenario selected for this category. In this scenario, 'Autobot', 'Decepticon' and the world of the Transformers movies are references from US pop culture, which might not be familiar to

speakers of Japanese. While an English speaker unfamiliar with Transformers may be able to associate 'Decepticon' with evil motives because of the similarity to 'deceive', Japanese speakers may not benefit from this clue.

(6a) オートボットはデセプティコンを食い止めようとする。彼らは世界の人々が平和に暮らすことを望んでいるのだ。 'The Autobots try to stop the Decepticons since they want the world to live in peace.'

(6b) オートボットはデセプティコンを食い止めようとする。彼らは世界を破壊したがっているからだ。 'The Autobots try to stop the Decepticons since they want to destroy the world.'

3.2 Participants

23 native Japanese speakers participated in the study. Participants were recruited via academic snowballing in Japan with two starting nodes. The average age was 29. Nine participants were male, eight were female and six did not state their sex.

3.3 Design and Procedure

Human participants accessed the survey hosted on Qualtrics via an anonymized link. They provided consent to participate in research and confirmed that they were of age and native speakers of Japanese. The participants saw 30 questions that tested their performance on the pronoun resolution task.[1] Participants saw the stimuli presented in random order and had to pick one of two answer options. The answer options were also presented in random order. There were no filler items, and participants were not given any training examples to maintain consistency with the LLMs' evaluation format. Task instructions and an example question can be found in Appendix A.3.

3.3.1 Collecting Data From LLMs

Our LLM data came from the responses of the GPT-4o model, the most advanced LLM at the time of research, collected from the OpenAI API. We chose the API rather than the chat interface because it allows us to control the model parameters (GPT-4o, temperature=1). We used the same design as in the study with human participants. The same stimuli were submitted to the OpenAI API. The order of questions and order of answers were randomized. We ran the code 30 times. Recent studies emphasize the need for a 'fair' evaluation of humans and machines with the emphasis on the same training and conditions for both groups (Lampinen, 2024). We follow this recommendation here. LLMs were evaluated zero-shot, and humans did not receive any prior training.

3.4 Results

We coded all correct answers as 1 and all incorrect answers as 0 for both humans and LLMs. Comparison of the means showed that in the 'good' condition, humans outperformed GPT-4o on the pronoun resolution task ($M_{good\_human} = 0.92$; $M_{good\_GPT} = 0.79$). In the 'grammar' condition, humans and LLMs performed similarly ($M_{grammar\_human} = 0.63$; $M_{grammar\_GPT} = 0.61$), and in the 'culture' condition, GPT-4o outperformed humans ($M_{culture\_human} = 0.92$; $M_{culture\_GPT} = 0.97$). The means and standard deviations are shown in Table 1.

To analyze the data, we applied a mixed-effects model, using the "lmerTest" package in R (Kuznetsova et al., 2017). Subject id and question were entered as random intercepts, while condition (good, grammar, culture) and source (human, GPT) were entered as fixed factors. The statistical analysis revealed a significant interaction between the two fixed factors (z(1523) = 2.68, p<.01). Follow-up tests showed that the statistical interaction was coming from the better performance of humans in the good condition (z(506) = 4.95, p<.001) and the better performance of GPT-4o in the culture condition (z(511) = -2-53, p<.05). The results from the study are presented in Figure 1.

3.5 Discussion

We observe that the overall accuracy (83%) displayed by human subjects in Japanese is lower than that reported for humans in English (92%). While this aligns with the lower accuracy levels reported for Thai (88%), it is important to point out that human performance in our study varies significantly depending on the condition. On well-formed grammatical examples in the baseline condition, the accuracy rates are 92%, similar to what is reported for English. Our study reveals that the translated dataset is not homogeneous and that examples with syntactic incongruencies can dramatically affect

[1] We also collected naturalness judgements and recorded reaction time, but these data are not the main focus of the paper.

	Humans	GPT-4o
Good	M=0.92 (SD=0.13)	M=0.78 (SD=0.35)
Grammar	M=0.63 (SD=0.44)	M=0.61 (SD=0.51)
Culture	M=0.92 (SD=0.08)	M=0.97 (SD=0.08)
Overall	M=0.83 (SD=0.30)	M=0.79 (SD=0.38)

Table 1: Means and standard deviations for humans and GPT-4o across the three conditions

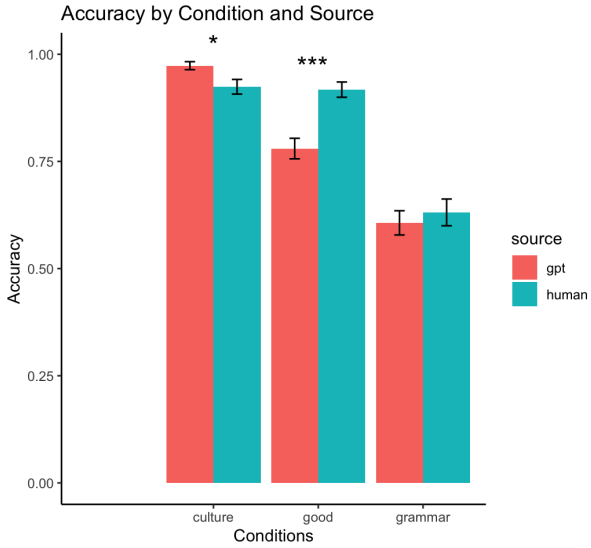

Figure 1: Accuracy of human and GPT judgements as a function of condition. Humans outperformed GPT in the 'good' condition, while the pattern was reversed for the 'culture' condition. No statistical difference was observed in the 'grammar' condition, where both sources performed poorly. The error bars represent +/-1 standard error. The significance tests are based on a mixed-effect model: * p<0.05, *** p<.001.

accuracy rates. These factors should be taken into account when evaluating either humans or LLMs on translated datasets.

Another implication from our study pertains to the potential applications of LLMs in contexts that require proficiency in Japanese. Previous studies have discussed leveraging LLMs' knowledge of Japanese in educational contexts for student writing assessment (Li and Liu, 2024, Takeuchi and Okgetheng, 2024) or example sentence generation (Benedetti et al., 2024). Our study demonstrates that the most advanced language models, such as GPT-4o, perform similarly to humans on tasks that require linguistic proficiency, which opens the opportunity for their integration in everyday life. However, the findings by Riccardi et al. (2024) that LLMs tend to interpret nonsensical input as meaningful, suggest that we should be cautious in applying them not only in language education contexts, but also in other linguistic tasks, such as text summarization (Gu et al., 2024) and annotation (Nishikawa and Koshiba, 2024).

Finally, we note that more insights could be gained from a systematic analysis of LLM mistakes. While this is outside of the scope of this paper, future work should look at these trends in more detail and compare the capabilities of different models, particularly those fine-tuned for Japanese.

4 Conclusions

In this study we compared the performance of LLMs and humans on a pronoun resolution task. We manipulated task demands and content effects and compared how they affect LLMs and humans. We found that in the baseline condition, humans outperform GPT-4o. These findings align with the results in Reese and Smirnova (2024) for Japanese, and with the results for Thai reported in Artkaew (2025). They suggest that in lower-resourced languages, humans still perform better than LLMs, even when competing with the most advanced models, such as GPT-4o.

Our results also provide evidence for task demands and content effects. In the relevant condition, task demands increased because of incongruent syntax/bad grammar. This manipulation negatively affected both human and LLM performance. Our results align with the observation in Hu and Frank (2024), who demonstrated that as task demands increase, LLM performance suffers, by analogy to how increased cognitive load in humans leads to reduced accuracy.

We manipulated content effects through cultural references. We selected scenarios with US cultural references, thus creating favorable conditions for LLMs, which were likely exposed to this information during training, and adversarial conditions for humans, as Japanese speakers might not be familiar with these references. We found that the changes in performance followed our predictions. In this condition, LLMs outperformed humans, providing evidence for content effects.

References

Pascal Amsili and Olga Seminck. 2017. A Google-proof collection of French Winograd schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 24–29, Valencia, Spain. Association for Computational Linguistics.

Phakphum Artkaew. 2025. Thai Winograd schemas: A benchmark for Thai commonsense reasoning. In *Proceedings of the Second Workshop in South East Asian Language Processing*, pages 42–51, Online. Association for Computational Linguistics.

David Bender. 2015. Establishing a human baseline for the winograd schema challenge. In *Proceedings of the 26th Modern AI and Cognitive Science Conference 2015, Greensboro, NC, USA, April 25-26, 2015*, volume 1353 of *CEUR Workshop Proceedings*, pages 39–45. CEUR-WS.org.

Enrico Benedetti, Akiko Aizawa, and Florian Boudin. 2024. Automatically suggesting diverse example sentences for l2 japanese learners using pre-trained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 114–131. Association for Computational Linguistics.

Jacob Browning and Yann LeCun. 2023. Language, common sense, and the winograd schema challenge. *Artificial Intelligence*, 325:104031.

Vittoria Dentella, Fritz Günther, Elliot Murphy, Gary Marcus, and Evelina Leivada. 2024. Testing ai on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, 14(1):28083.

Ziwei Gu, Ian Arawjo, Kenneth Li, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. An ai-resilient text rendering technique for reading and skimming documents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Ryo Hashimoto, Masashi Takeshita, Rafal Rzepka, and Kenji Araki. 2023. Development of japanese wsc273 winograd schema challenge dataset and comparison between japanese and english bert baselines. In *In the proceedings of the Language Technology Conference (LTC'23)*, pages 91–95.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus chatgpt-generated essays. *Scientific reports*, 13(1):18617.

Jennifer Hu and Michael C Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. *arXiv preprint arXiv:2404.02418*.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Roni Katzir. 2023. Why large language models are poor theories of human linguistic cognition. a reply to piantadosi (2023). *Manuscript. Tel Aviv University. url: https://lingbuzz. net/lingbuzz/007190*.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L Elman. 2008. Coherence and coreference revisited. *Journal of semantics*, 25(1):1–44.

Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. 2023. The defeat of the winograd schema challenge. *Artificial Intelligence*, 325:103971.

Alexandra Kuznetsova, Per B Brockhoff, and Rune Haubo Bojesen Christensen. 2017. lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82(13).

Andrew Lampinen. 2024. Can language models handle recursively nested grammatical structures? a case study on comparing models and humans. *Computational Linguistics*, 50(4):1441–1476.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561, Rome, Italy. AAAI Press.

Wenchao Li and Haitao Liu. 2024. Applying large language models for automated essay scoring for non-native japanese. *Humanities and Social Sciences Communications*, 11(1):1–15.

Kai Nishikawa and Hitoshi Koshiba. 2024. Exploring the applicability of large language models to citation context analysis. *Scientometrics*, pages 1–27.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

May Lynn Reese and Anastasia Smirnova. 2024. Comparing chatgpt and humans on world knowledge and common-sense reasoning tasks: A case study of the japanese winograd schema challenge. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Nicholas Riccardi, Xuan Yang, and Rutvik H Desai. 2024. The two word test as a semantic benchmark for large language models. *Scientific Reports*, 14(1):21593.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Tomohide Shibata, Shotaro Kohama, and Sadao Kurohashi. 2015. Nihon go winograd schema challenge no kochiku to bunseki. In *Proceedings of the 21th Annual Meeting of the Association for Natural Language Processing*, pages 493–496, Kyoto. The Association for Natural Language Processing.

K Takeuchi and B Okgetheng. 2024. Estimating japanese essay grading scores with large language models. *Lan-guage Resources and Evaluation*, 58(2):345–367.

Tempest A van Schaik and Brittany Pugh. 2024. A field guide to automatic evaluation of llm-generated summaries. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2832–2836.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

A Appendices

A.1 Examples of Rejected Items

Rejected Item: リチャードはカーソン上院議員を脅した。彼の沈黙が守られるように。 Richard blackmailed Senator Carson so that his silence would be secured.
Reason for Rejection: This item was considered for the good category, but rejected based on the intuition of our Japanese native speaking consultant that 沈黙, silence, is used in an unnatural way.

Rejected Item: 火事についての記事によれば、それらによってシアトルの大部分に火がついた。 The article about the fires said that they torched most of Seattle.
Reason for Rejection: This item was considered for the syntax category but was rejected because while 'the article' has the syntactic function of subject, 'the fires' is neither the subject nor the object of the first clause.

Rejected Item: ハーヴィー・デントは恋人を失ったことをバットマンのせいにする。彼が彼女を救出しなかったからだ。 Harvey Dent blames the Batman for losing his love because he did not save her.
Reason for Rejection: This item was considered for the culture category, but was rejected because there are multiple pronouns in the second clause.

Rejected Item: 私たちは人間に果物を与えた。それらが熟していたからだ。 We gave the fruit to the humans because they were ripe.
Reason for Rejection: This item was rejected because in Japanese, it has a selectional clue to the answer. それ, it, can only be used to refer to inanimate objects.

Rejected Item: 猫が人間を襲った。彼らは野生化していたのだ。 The cats attacked the humans because they were feral.
Reason for Rejection: This item was rejected because it can be solved by co-occurrence statistics. The adjective feral is more often associated with cats than humans.

A.2 List of Stimuli

Good Category Stimuli

学外のアパートは学内のアパートより好まれた。それらの方が安かったからだ。安かった方はどちらですか？ The off-campus apartments were preferred to the on-campus apartments because they were cheaper. Which were cheaper?
Answer: 学外のアパート the off-campus apartments

学外のアパートは学内のアパートより好まれた。それらの方が高かったからだ。高かった方はどちらですか？ The off-campus apartments were preferred to the on-campus apartments because they were more expensive. Which were more expensive?
Answer: 学内のアパート the on-campus apartments

ジョーはアダムより良い香りがする。彼は日頃からシャワーを浴びるからだ。日頃からシャワーを浴びるのは誰ですか？ Joe smells better than Adam since he showers regularly. Who showers regularly?
Answer: ジョー Joe

ジョーはアダムより良い香りがする。彼はめったにシャワーを浴びないからだ。めったにシャワーを浴びないのは誰ですか？ Joe smells better than Adam since he hardly ever

showers. Who hardly ever showers?
Answer: アダム Adam

ジャックはジョンより多く得票した。人々は彼を好んだ。人々が好んだのは誰ですか？ Jack got more votes than John because the people liked him. Who did the people like?
Answer: ジャック Jack

ジャックはジョンより多く得票した。人々は彼を好まなかった。 人々が好まなかったのは誰ですか？ Jack got more votes than John because the people did not like him. Who did the people not like?
Answer: ジョン John

アダムはアレクサンダーの殺害に失敗した。そこで彼は再度のために暗殺者を雇った。再度のために暗殺者を雇ったのは誰ですか？ Adam failed to kill Alexander, so he hired an assassin for the second attempt. Who hired an assassin for the second attempt?
Answer: アダム Adam

アダムはアレクサンダーの殺害に失敗した。そこで彼は再度を恐れてボディーガードを雇った。 再度を恐れてボディーガードを雇ったのは誰ですか？ Adam failed to kill Alexander, so he hired a bodygaurd in case of a second attempt. Who hired a bodyguard in case of a second attempt?
Answer: アレクサンダー Alexander

トニーはジェフを手伝った。彼は手伝いたかったのだ。手伝いたかったのは誰ですか？ Tony helped Jeff because he wanted to help. Who wanted to help?
Answer: トニー Tony

トニーはジェフを手伝った。彼は手助けが必要だったからだ。手助けが必要だったのは誰ですか？ Tony helped Jeff because he needed help. Who needed help?
Answer: ジェフ Jeff

Syntax Category Stimuli

バットはボールに当たった。それが軌道を描くように飛んだからだ。 軌道を描くように飛んだのは何ですか？ The bat hit the ball because it flew in the way of the trajectory. What flew in the way of the trajectory?
Answer: バット the bat
Note from the translators: ボールでも？ The ball too?

バットはボールを打った。それは可哀想な動物に向かってまっしぐらにとんだからだ。可哀想な動物に向かってまっしぐらにとんだのは何ですか？ The bat hit the ball because it flew straight at the poor animal. What flew straight at the poor animal?
Answer: ボール the ball

シーラは古いポンコツ車を修理しようとした。彼女は30年も車に取り組んでいなかったにも拘らずだ。30年も車に取り組んでいなかったのはどちらですか？ Sheila tried to repair the old jalopy, even though she had not worked on cars in three decades. Who had not worked on cars in three years?
Answer: シーラ Sheila

シーラは古いポンコツ車を修理しようとした。彼女は30年も走っていなかったにも拘らずだ。30年も走っていなかったのはどちらですか？ Sheila tried to repair the old jalopy, even though she had not run in three decades. Who had not run in three decades?
Answer: 古いポンコツ車 the old jalopy

りんご酒がわたしの口に入った。それは美味しかったから。美味しかったのは何ですか？ The apple wine entered my mouth because it tastes good. What tastes good?
Answer: りんご酒 the apple wine

りんご酒がわたしの口に入った。それは一杯ではなかったから。一杯ではなかったのは何ですか？ The apple wine entered my mouth because it was not full. What was not full?
Answer: わたしの口 my mouth

雇用主はケイティに仕事を提供した。彼女はインタビューが好きだったからだ。インタビューが好きだったのは誰ですか？ The employer offered Katie a job, because she liked the interview. Who liked the interview?
Answer: 雇用主 the employer

雇用主はケイティに仕事を提供した。彼女が会社にぴったりだったからだ。会社にぴったりだったのは誰ですか？ The employer offered Katie a job, because she was a fit for the

company. Who was a fit for the company?
Answer: ケイティ Katie

ジョーはマイクに倒れ掛かった。彼は眠る場所が必要だった。眠る場所が必要だったのは誰ですか？ Joe crashed into Mike because he needed a place to sleep. Who needed a place to sleep?
Answer: ジョー Joe

ジョーはマイクに衝突した。彼は損害分を支払わなくてはならなかった。損害分を支払わなくてはならなかったのは誰ですか？ Joe crashed into Mike and he had to pay for the damage. Who had to pay for the damage?
Answer: マイク Mike
Note from translators: ジョーでも？ Could be Joe too?

Culture Category Stimuli

ワトソンはジオパディでケンを負かした。彼は優れた機械だ。優れた機械は誰ですか？ Watson beat Ken at Jeopardy because he is a superior machine. Who is a superior machine?
Answer: ワトソン Watson

ワトソンはジオパディでケンを負かした。彼は劣った人間だからだ。 劣った人間は誰ですか？ Watson beat Ken at Jeopardy because he is an inferior human. Who is an inferior human?
Answer: ケン Ken

ビリーはスクラブルでトミーを負かした。あの新入りには運がついていた。運がついていたのは誰ですか？ Billy beat Tommy at Scrabble because that newbie had all the luck. Who had all the luck?
Answer: ビリー Billy

ビリーはスクラブルでトミーを負かした。あの新入りには能力がなかったから。能力がなかったのは誰ですか？ Billy beat Tommy at Scrabble because that newbie had no skill. Who had no skill?
Answer: トミー Tommy

オートボットはデセプティコンを食い止めようとする。彼らは世界の人々が平和に暮らすことを望んでいるのだ。世界の人々が平和に暮らすことを望んでいるのは誰ですか？ The Autobots try to stop the Decepticons since they want the world to live in peace. Who wants the world to live in peace?
Answer: オートボット the Autobots

オートボットはデセプティコンを食い止めようとする。彼らは世界を破壊したがっているからだ。 世界を破壊したがっているのは誰ですか？ The Autobots try to stop the Decepticons since they want to destroy the world. Who wants to destroy the world?
Answer: デセプティコン the Decepticons

メアリはジョーが好きだ。彼女は女性が好きだからだ。女性が好きなのは誰ですか？ Mary likes Joe because she likes females. Who likes females?
Answer: メアリ Mary

メアリはジョーが好きだ。彼女は名前が素敵だからだ。名前が素敵なのは誰ですか？ Mary likes Joe because she has a cool name. Who has a cool name?
Answer: ジョー Joe

カリフォルニアの人の方がニューヨークの人より良い。彼らにはハリウッドがあるから。ハリウッドがあるのは誰ですか？ Californians are better than New Yorkers because they have Hollywood. Who has Hollywood?
Answer: カリフォルニアの人 Californians

カリフォルニアの人の方がニューヨークの人より良い。彼らには映画を作ってくれるハリウッドの連中がいないからだ。映画を作ってくれるハリウッドの連中がいないのは誰ですか？ Californians are better than New Yorkers because they do not have Hollywood lots to produce movies. Who does not have Hollywood to produce movies?
Answer: ニューヨークの人 New Yorkers

A.3 Task Instructions and Prompt Examples

Task Instructions for Human Participants:

> 2つの日本語の文章と、その文章の内容に関する質問と２つの答えが表示されています。2つの答えの内正しいと思う方を選んでください。どちらの答えも妥当と思われる場合

は、最も適切と思われる方を選んでください。
You will be shown two Japanese sentences and a question with two answers about the content of the sentences. Please choose the answer you think is correct. If both options seem right, please pick the one you think is the most fitting.

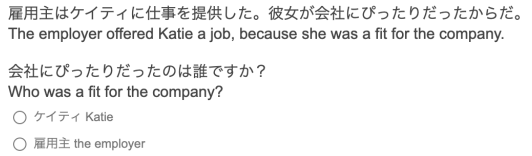

Figure 2: Example survey question shown to human participants. (The English translation was not shown to participants)

Example Prompt for GPT-4o:

Japanese Prompt:
学外のアパートは学内のアパートより好まれた。それらの方が安かったからだ。 次の問題をAかBで答えてください。安かった方はどちらですか？A.学外のアパート B.学内のアパート
English Translation:
The off-campus apartments were preferred to the on-campus apartments because they were cheaper. Answer the following question with A or B. Which were cheaper? A. The off campus apartments B. The on campus apartments

(The English translation was not given in the prompt.)