# Commonsense Reasoning and World Knowledge Benchmark Tests: A Comparison of ChatGPT and Humans on The Japanese Winograd Schema Challenge

A Thesis submitted to the faculty of
San Francisco State University
In partial fulfillment of
the requirements for
the Degree

Master of Arts

In

English: Linguistics

by

May Lynn Reese

San Francisco, California

May 2024

Copyright by
May Lynn Reese
2024

**Certification of Approval**

I certify that I have read Commonsense and World Knowledge Benchmark Tests: A Comparison of ChatGPT and Humans on the Japanese Winograd Schema Challenge by May Lynn Reese, and that in my opinion this work meets the criteria for approving a thesis submitted in partial fulfillment of the requirement for the degree Master of Arts in English: Linguistics at San Francisco State University.

Anastasia Smirnova, Ph.D.
Associate Professor,
Thesis Committee Chair

Jenny Lederer, Ph.D.
Associate Professor

**Abstract**

Modern Large Language Models (LLMs) are increasingly influencing the way people get information and make decisions. Recent advances have led to significant improvement on benchmark tests in commonsense reasoning and world knowledge and it is imperative to start focusing on comparison of LLMs with human baselines. Although most research is done in English, LLMs are developing rapidly in many languages. Due to a disparity in the amount of training data, performance in English does not necessarily translate to other languages. Using the Japanese translation of the Winograd Schema Challenge (WSC) as a case study, this work discusses benchmark tests in commonsense reasoning and world knowledge. It describes two experimental studies comparing humans and ChatGPT on the Japanese WSC. Results show that humans outperform ChatGPT on the WSC task, but differ in naturalness judgments. The study results contribute to a high-level analytical discussion of the WSC as a benchmark test for modern LLMs.

**Acknowledgements**

First, I would like to thank my advisor, Dr. Anastasia Smirnova, for being supportive, flexible and a wonderful guide through the process. I'd also like to thank Dr. Jenny Lederer for going out of her way to support me and make me feel like a part of the community despite me working remotely.

I'd also like to express my gratitude to Mika, Yuka and Mr. Rikuta for helping me with the project, Mie-san for giving me a place to study that felt like home, and my friends for keeping me going.

Finally, I'd like to thank my Dad for always encouraging me to learn, be creative and enjoy life, and my Mom for believing in me no matter what.

## Table of Contents


**Certification of Approval** 2

**Abstract** 3

**Acknowledgements** v

**Table of Contents** vi

**List of Tables** viii

**List of Figures** ix

**List of Appendices** x

**1 Introduction** 1

**2 Large Language Models and their Evaluation** 4
2.1 Large Language Models 4
2.2 LLM Benchmark Tests 8
2.2.1 Problems with Benchmark Tests 12
2.2.2 Non-English Benchmark Tests 14
2.2.3 Benchmark Test Translation Challenges 17

**3 The Winograd Schema Challenge** 19
3.1 The Japanese Winograd Schema Challenge 24
3.2 Zero Anaphora in Japanese 30
3.2.1 Interpreting Zero Anaphora 32
3.2.2 Zero Anaphora and the Winograd Schema Challenge 36

**4 Experimental Studies** 38
4.1 Study 1: Human Performance on the Japanese Winograd Schema Challenge 38
4.1.1 Participants 39
4.1.2 Design 39
4.1.3 Materials 39
4.1.4 Procedure 43
4.1.5 Results 46
4.2 Study 2: ChatGPT performance on the Japanese Winograd Schema Challenge 47
4.2.1 Procedure 47
4.2.2 Results 48
4.3 Comparing ChatGPT and Human Performance 49

**5 Experimental Studies Discussion** 51
5.1 Accuracy 52

5.2 Naturalness 59

**6 General Discussion** **68**
6.1 The Winograd Schema as a Modern LLM Evaluation Method 69
6.1.1 Contamination 70
6.1.2 Saturation 71
6.1.3 Suitability 72

**7 Conclusion** **79**

**References** **81**

**Appendices** **93**

## List of Tables

**Table 1. Examples of Excluded Items ..... 41**

**Table 2: Example of overt and null pronoun versions used. ..... 43**

**Table 3. Study 1 Results ..... 47**

**Table 4. Study 2 Results ..... 49**

**Table 5. List of Questions by Scenario ..... 53**

**Table 6. Items on which Humans Achieved Perfect Performance ..... 54**

**Table 7. Questions Rated Higher by Humans than ChatGPT ..... 64**

## List of Figures

**Figure 1. Principles of Interpretation for Zero Anaphora..... 33**

**Figure 2. Instructions for Task 1: Pronoun Resolution ..... 44**

**Figure 3. Example Question for Task 1 ..... 44**

**Figure 4. Instructions for Task 2: Naturalness Rating..... 45**

**Figure 5. Example Question for Task 2 ..... 45**

**Figure 6. Comparison of Human and ChatGPT Performance..... 49**

**Figure 7. Comparison of Human and ChatGPT on Naturalness Rating ..... 49**

**Figure 8. WinoWhy Knowledge Categories ..... 76**

## List of Appendices

**Appendix A: Full List of Stimuli Used in Task 1 and 2..... 93**

**Appendix B: ChatGPT Naturalness Explanations ..... 97**

# 1 Introduction

Commonsense reasoning and the use of world knowledge are skills that have historically differentiated humans from machines. However, recent developments in Large Language Model (LLM) technology have led to considerable improvements on a range of psycholinguistic competencies as evaluated by Natural Language Understanding (NLU) benchmark tests. New generation LLMs show increasingly high performance on tasks that require human-like reasoning. As LLMs become more widely used and accessible, they will have a great influence on how people get information and make decisions. Their contributions are already noticeable in a variety of sectors including education, journalism, science etc. However, to ensure their appropriate usage, we need a better understanding of how they work and how they compare to humans in their linguistic competencies and decision-making heuristics.

Currently used NLU benchmark tests, such as SuperGLUE (Wang et al. 2019), SQuAD 2 (Rajpurkar et al. 2018) and various Natural Language Inference (NLI) sets (e.g., Bowman et. al 2015; Williams et. al. 2018) aim to evaluate systems on a variety of tasks that require world knowledge and commonsense reasoning. The implicit conclusion is that a system which performs well on the test is demonstrating, or at least superbly imitating, linguistic performance and reasoning abilities on par with humans. The majority of studies focus on comparing one system to another with the goal of determining which performs best on the tasks. However, the difference between the capabilities of most modern models is small and it is becoming less informative to compare systems to each other. It is now more imperative to compare LLMs to human baselines on tasks that can deliver more specific, detailed data. In some human and LLM

comparisons (e.g., Nangia & Bowman 2019; He et al. 2021), the models already surpass human baseline data on certain tasks. However, most human and LLM comparison studies are on performance in English.

LLMs are progressing quickly in languages other than English as well and their usage is becoming widespread around the world. However, performance on benchmark tests in English does not necessarily correspond to performance in other languages. Due to linguistic differences, this is true for human data also. Because of an imbalance in LLM training data, language specific baselines are necessary. Data from 70 typologically diverse languages studied in MEGA, a model for Multilingual Evaluation of Generative AI by Ahuja et al. (2023), suggests that the amount of pre-training data for a specific language affects the performance of an LLM on tasks in that language. The researchers found that OpenAI models perform much better on paraphrase identification tasks in French (72.1% accuracy) than they do in Japanese (67% accuracy). The authors suggest that the disparity in training data explains the difference in accuracy (~3.5B tokens in French, ~240M in Japanese). These findings highlight the necessity for testing in different languages because success in one language does not translate to another. The difference between French and Japanese is striking, but the effects may be even more pronounced when considering the wide gap between English and other languages. As an example, GPT-3 is trained on 119 languages (Brown et al. 2020), but 93% of this data is in English (Ahuja et al. 2023). For LLMs to reach their full potential and avoid the disparity in distribution and usage of language technologies, more research in other languages is necessary (Blasi et al. 2022).

In this study, I will use the Japanese translation of the Winograd Schema Challenge as a case study and address the following research questions:

1. How does human and LLM performance compare on benchmark tests in languages other than English?
2. How effective are benchmark test translations?

First, I give a broad overview of modern LLMs and their benchmark tests with a focus on studies in human baseline data and translations. In section 3, I provide an in-depth explanation of the Winograd Schema Challenge and discuss translation difficulties due to typological differences between English and Japanese. In section 4, I present two experimental studies, one with human participants to establish a preliminary baseline, and the other with ChatGPT (GPT-3.5) to evaluate performance on the Japanese Winograd Schema Challenge. Results demonstrate that ChatGPT does not achieve the same level of accuracy as humans and differs from them in naturalness judgments. In the following sections, I discuss the merits and drawbacks of the Winograd Schema Challenge and its Japanese translation. Through this case study, I explore directions for the future of benchmarking and development of tests in non-English languages.

# 2 Large Language Models and their Evaluation

The demand for research in benchmark testing and multilingual approaches is rapidly growing with the most recent evolutions of language technology. The following sections attempt to give the rudimentary background information necessary to form the basis of the experimental studies. They will cover the basics of language modeling and give an overview of currently used benchmark tests.

## 2.1 Large Language Models

Language Models (LMs) are statistical representations of human language that compute probability distributions of words or sequences of words. Put simply, LMs predict what words will come next in a sequence based on the words that came prior. Language modeling is one of the core concepts of natural language processing (NLP) and the technology has progressed rapidly since the 1980s. It is an important part of virtually every NLP function. In this section, I will give a short, broad overview of significant developments in language modeling that have led to current state-of-the-art large language models (LLMs).

The first LMs were N-gram based, probabilistic models that relied on simple frequency data. They were usually trained on small sets of domain or task specific data and usage was very context-dependent. Given a starting word *I*, an N-gram based model would use the statistics learned in training data, and simply generate the word that most often comes after *I*, perhaps *am*. This seems reasonable, but N-gram models were very limited because they only encoded statistics about small windows of context (2 words for a 2-gram model, 3 words for a 3-gram model, etc).

Ground-breaking advances in language modeling occurred with the success of neural networks in computer science as well as innovations in word representation. In neural network LMs, words are represented by embeddings, which are dense vectors that encode information about the word's meaning. Two of the most widely used embedding methods were Word2Vec (Mikolov et al. 2013) and GLoVE (Pennington et al. 2014). These methods generate vectors for a word *x* by considering the immediate context of *x* and sampling words not in the immediate context to get co-occurrence frequencies. Similar vectors are an indication that their represented words are semantically similar. For example, *cake* and *pie* would have similar vectors because they are both nouns and appear in similar contexts. However, *car* and *sorrowful* would not as they would typically occur in very different contexts. Modern LLMs use even more complex neural networks for this process. During pre-training, a neural net takes training data and learns associations and dependencies in order to build embeddings for each word in the corpus. The size of each embedding depends on the method used and the amount of training data, but embeddings in modern models usually contain over 1,000 dimensions (Mars 2022). This means that LLMs have a vast amount of knowledge about word meaning that is built on co-occurrence statistics learned in pre-training.

N-gram based models were improved upon by Feedforward Networks (FNNs), some of the first neural net LMs. An FNN is a basic layered neural network that consists of an input layer, hidden layer(s) and an output layer. Natural language input gets translated into an embedding in the input layer. The hidden layer(s) contain a number of computational units, which take information from the input embeddings, apply weights and biases to calculate what to send to the output layer.

The next significant improvements in this architecture were Recurrent Neural Networks (RNNs) and long short-term memory networks (LSTMs). In these networks, the output of a hidden layer is not only dependent on its input, but also on its own output from a previous cycle. This addition allowed for networks to have a longer memory, extending the context which it could use as information for its output. This led to greater ability at many NLP tasks, including language modeling, sequence labeling, classification, and text generation.

RNNs also saw the introduction of encoder-decoder networks which allow a system to take an input of one length and output a sequence of a different length. This was especially significant for machine translation tasks and much of modern language modeling is a form of encoder-decoder or decoder-only networks.

Perhaps the biggest revolution in language modeling came in 2017 with the introduction of transformer technology (Vaswani et al. 2017). Transformer based systems benefit from increased parallel processing, leading to larger, faster models. They rely on self-attention, allowing models to utilize information about the word relations and their sequence positions, which results in more relevant and accurate output. It also gives models the ability to handle long-range dependencies more efficiently.

This evolution of LMs from FNNs to transformer based encoder-decoder models led to vast improvements in language technology. LMs became faster, more efficient and more effective on tasks such as machine translation, text summarization, sentiment analysis, information retrieval, etc. In human interaction with computers, LM output was more coherent and relevant, therefore more useful and practical for a variety of applications.

Modern Large Language Models (LLMs) are possible because of three developments: transformer technology, massive amounts of training data and new training processes.[1] As humans are increasingly interacting with LLMs on a daily basis in a variety of settings, including word processing, customer service, IT support, etc., more focus has been placed on making human interactions with LLMs go smoothly and naturally. An integral part of this is Reinforcement Learning from Human Feedback (RLHF), an important training process for modern LLMs.

RLHF models begin with a pre-trained LM, which takes multiple prompts and generates text outputs. Human annotators then evaluate the outputs and score them by preference, either choosing one of two options or ranking several options from best to worst. A reward model then uses the human responses in reinforcement learning to fine-tune the LM in order to optimize its use of language with high preference from human annotators (Lambert et al. 2022). RLHF models are now widely used because of their superior ability to follow user instructions in a way that is aligned with human preferences and values. OpenAI's InstructGPT are now the default models available through ChatGPT (Ouyang et al. 2022).

---

[1] To name a few, significant transformer based models include: BERT (Devlin et al., 2019), GPT (Radford et al. 2019, Brown et al. 2020), T5 (Raffel et al. 2020) and PaLM (Chowdhery et al. 2023). BERT was perhaps the best known transformer and there have been many descendants of the original model that either improved architecture, training methods or fine-tuned it for specific tasks. They are bidirectional, encoder-only models, which means that they have access to all positions of an input and can utilize bidirectional contexts to predict words. They're good for tasks that require the understanding of full sentences, such as text classification (Wei et al. 2023). PaLM and GPT-1, 2 and 3, are unidirectional, decoder-only models. They can only use prior tokens to predict the next and excel at text generation tasks, such as story generation. T5 is an encoder-decoder, sequence to sequence model, which can access tokens before and after the target token, and is good for translation and summarization tasks. Also particularly noteworthy and successful are XLM models (Conneau & Lample 2019), which are LLMs that have been pre-trained with cross-lingual data.

GPT-3 was OpenAI's first RLHF model. It was an autoregressive, decoder-only transformer based model, and was released in 2020. It has 175 billion parameters and is trained on 570 GB of text, including web text, books and code. This was the biggest model to be released by OpenAI at the time and it excelled at a number of evaluation metrics (Brown et al. 2020). The model used in this study is GPT-3.5, released in 2022. It is a larger and improved version of GPT-3. GPT-4 was released in 2023 and the number of parameters and size of training data has not been publicly released at the time of this writing.

With this deep architecture and vast amounts of training data, we can say that modern LLMs gain considerable statistical knowledge about human language and the world. This is evident in their ability to excel at a variety of tasks, including translation, text generation, text summarization, Q&A, etc. In the following section, I will discuss evaluation methods designed to measure LLMs' knowledge, reasoning and their ability to apply those to the relevant tasks.

**2.2 LLM Benchmark Tests**

To address this need, many evaluation methods have been developed to test natural language understanding (NLU). These consist of a variety of tasks including sentence and story completion, inference and logical reasoning questions, classification questions and reading comprehension. These benchmark tests are designed in such a way that high performance implies the ability to use commonsense reasoning, world knowledge and human-like decision heuristics.

Reading comprehension tasks challenge a system to respond to questions about a passage, sentence or dialogue. The questions can be multiple choice, as in RACE (Lai et al. 2017), or require a system to extract the answer span from the sample text as in SQuAD 2 (Rajpurkar et al. 2018). Questions in DROP (Dua et al. 2019) concern the whole passage and

require semantic parsing for discrete reasoning, such as addition, subtraction, comparison, etc. CoQA (Reddy et al. 2019) presents a passage, but the evaluation is done with free form answer conversational Q&A. QuAC (Choi et al. 2018) is a comprehension test of an information seeking dialogue.

Brown et al. found that GPT-3 performed well on CoQA (85.0%), coming within 3 points of the human baseline, and got a 69.8% accuracy on SQuAD 2 when using a few-shot setting. It was well below state-of-the-art and human data on the other datasets tested. Overall, GPT-3 has a lot of room for improvement on reading comprehension benchmark tests (Brown et al. 2020). However, Ye et al. (2023) found that GPT-3.5's best performance on SQuAD 2 was 81.76%.

Another common benchmark task is Natural Language Inference (NLI) which tests a system's ability to use logical reasoning to understand connections between events. A system is presented with two sentences or passages and is required to classify them into entailment, contradiction or independent/neutral. Widely used NLI datasets include SNLI (Bowman et al. 2015), MNLI (Williams et al. 2017) and Adversarial NLI (Nie et al. 2019). NLI sets have also been developed based on datasets from other tests, such as QNLI, which was derived from SQUAD (Rajpurkar et al. 2016), and WNLI, derived from Winograd Schema Challenge data (Levesque et al. 2012). RTE is an NLI test used in the original GLUE (Wang et al. 2018) and SuperGLUE (Wang et al. 2019) and is made up of datasets from the PASCAL Recognizing Textual Entailment Challenges (Dagan et al. 2005; Haim et al. 2006; Bentivogli et al. 2009).

BERT is shown to outperform humans on QNLI and comes within 2 points of human performance on MNLI (Nangia & Bowman 2019). Brown et al. (2023) finds that the largest GPT-3 model at the time performed at random chance at best on Adversarial NLI and slightly

better (56%) on RTE. GPT-3.5's best performance was 68.56% on SNLI and 68.13% on MNLI (Ye et al. 2023).

The tasks discussed thus far require a variety of linguistic abilities, but are solvable with only the knowledge given in the text. With the advances in LLM architecture, it is reasonable to assume that a pre-trained model also gains some general knowledge about the world in addition to linguistic knowledge. To evaluate this, there are also benchmark tasks designed specifically to test a system's commonsense reasoning and use of world knowledge.

The Winograd Schema Challenge (WSC) (Levesque et al. 2012) and WSC style tests, such as Winogrande (Sakaguchi et al. 2021) are multiple choice or open answer pronoun resolution tasks which require world knowledge to solve. COPA (Roemmele et al. 2011) is a binary choice task requiring commonsense reasoning about the cause or effect of a premise sentence. ReCoRD (Zhang et al. 2018) is a cloze style reading comprehension test that requires commonsense reasoning.

On the WNLI set, an NLI test based on the WSC, BERT performs well below the 96% human baseline (Nangia & Bowman 2019). Brown et al. (2023) finds that GPT-3 achieves 89% on the WSC and 78% on Winogrande.

One of the most widely-used benchmarks is SuperGLUE (Wang et al. 2019), an improvement on the previously published GLUE benchmark test (Wang et al. 2018). It consists of eight tasks: yes/no reading comprehension Q&A (BoolQ), three-class textual entailment classification (CB), cause and effect reasoning (COPA), multiple choice Q&A reading comprehension (MultiRC), cloze-style reading comprehension (ReCoRD), two-class textual

entailment classification (RTE), word disambiguation binary classification (WiC) and yes/no Q&A based coreference resolution (WSC).

Most current research using these tests focuses on comparing different models to each other to determine which has the best performance. However, as LLM technology progresses, performance has increased rapidly, and the difference between different models has decreased significantly. This calls for a greater focus on evaluating LLMs in relation to human baselines, but also on more direct comparison of LLM and human reasoning.

An increasing number of tests address this concern and focus on comparing LLM performance to humans on more complex tasks or problem-solving. Frohberg & Binder (2021) developed a test for counterfactual reasoning and found that LLMs do not achieve human accuracy level (98.18% for humans, 72.63% highest LLM performance by T0pp(Sahn et al. 2021)). They also found that difference in surface syntax features does not significantly affect performance. Collins et al. (2022) evaluated GPT-3 on its responses to prompts asking for a plan to accomplish a goal and explain a hypothetical situation. For each prompt, they created two extra conditions, which eliminated the most likely responses. The researchers found that GPT-3 underperforms on all three conditions compared to humans (see also Valmeekam et al. 2022). Cai et al. (2023) studied Vicuna's (Chiang et al. 2023), GPT-3.5's and GPT-4's responses to a range of psycholinguistic tasks in the categories of sounds, words, syntax, meaning, discourse, and interlocutor sensitivity. They found that ChatGPT performance is comparable to human results on 10 out of 12 tasks. Kauf et al. (2023) studied LLMs' generalized event knowledge, the ability to understand common events, which humans learn about through physical and linguistic experiences. They tested a variety of models on the ability to identify possible vs. impossible and

likely vs. unlikely events. The examples were shown in a number of different conditions created by altering surface-level linguistic features (voice, word frequency, sentence length, etc). They found that LLM performance on possible vs. impossible events was on par with human judgements, but slightly lower on likely vs. unlikely events. Unlike humans, LLM performance was found to be affected by the linguistic properties of the example sentences.

These types of evaluation offer informative insights into LLM reasoning and judgment in comparison to humans, but traditional benchmark tests, such as SuperGLUE, are still the industry standard for evaluating LLMs. Moving forward, it is imperative to refocus benchmark test research towards comparison with human data, which may also need to be updated and refined as the tests evolve. In the next section, I will discuss the drawbacks of modern evaluation methods which may lead to fundamental changes in the future of benchmarking.

### *2.2.1 Problems with Benchmark Tests*

Despite their widespread use and continuing development, there are a number of problems with modern benchmark tests. The problems fall into three broad categories: contamination, saturation and suitability. As viable alternatives do not yet exist, many researchers acknowledge these issues when using benchmark tests as a measure of performance. However, there is a growing literature beginning to offer solutions to these problems.

The first category, contamination, refers to systems being exposed to the benchmark data during their pre-training. Because most benchmark tests are publicly available on the internet, there is a high likelihood that LLMs encountered them in their training data. In their evaluation of GPT-3, Brown et al. (2020) did a contamination analysis of the tests they administered and found that six groups of benchmark tests were flagged for high probability of contamination.

>90% of reading comprehension sets, such as SQuAD2 and DROP were found to be potentially contaminated and 45% of Winograd schema examples were flagged. Ahuja et al. (2023) found similar results in their contamination verification with 12 of the 15 tests they used found to be potentially contaminated. This is a serious problem for benchmark tests as it implies that results may not be a good representation of the system's real ability.

The second category is saturation, which refers to the time it takes for models to reach or surpass human performance on benchmark tests (Kiela et al. 2021). Quick saturation of benchmark tests is becoming a more pressing problem with new models easily outperforming human baselines (Nangia & Bowman 2019; Bowman & Dahl 2021). For new models, supposedly hard to achieve milestones are reached within a few years, but because of issues of suitability, this does not necessarily indicate a model's ability of human-level thinking (Kiela et al. 2021). This is problematic, especially considering how much human effort usually goes into creating these tests (Chandran et al. 2024). Fixing the problem of saturation is going to require more careful test design and data collection methods, as well as larger and more difficult datasets (Bowman & Dahl 2021).

The third category, suitability, refers to how well the benchmark task delivers information on the target of evaluation. Ultimately, benchmark tests are proxy problems for other targets (Kocijan et al. 2023). For example, a test may claim that high performance indicates human-like reasoning or intelligence, but the actual task being performed is Q&A. So the test may provide data on a system's ability to perform Q&A, but it is difficult to make any judgment about a system's reasoning ability based on that data alone. There may be alternative ways to solve a benchmark task, such as using built-in biases or utilizing statistical irregularities in the

dataset. Poliak et al. (2018) tested this possibility by using hypothesis-only versions of NLI datasets. In an NLI test, two sentences are given: the premise, and the hypothesis. The system must determine, using background knowledge from the premise sentence, if the hypothesis is an entailment, neutral or a contradiction. However, in their experiment, researchers only gave the hypothesis sentence and found that the model they tested often performed better on hypothesis only tests, suggesting that there is some non-NLI related signal for systems to exploit. Additionally, researchers found that issues with benchmark tests can be traced back to their construction using crowdworkers. Parmar et al. (2023) investigated 14 benchmark tests and found "instruction bias" evident in about 73% of instruction sentences given to crowdworkers, leading to pattern propagation in the created datasets. These issues mean that certain patterns may be overrepresented in the test sets, allowing for the possibility that a system is using statistical shortcuts rather than the targeted psycholinguistic ability.

### *2.2.2 Non-English Benchmark Tests*

The vast majority of benchmark tests have been developed in English, but with the increased size of training data, which is not monolingual English, LLM performance in other languages has been steadily improving. GPT-3's training data was 93% English, but the other 7% were other languages (Brown et al. 2020). GPT-3's total training data was 570GB of text data, which theoretically means that it saw almost 40GB of non-English data in training. The numbers for GPT-3.5 and GPT-4 are not known, but we can assume that they are much larger. Other language models like BLOOM (Le Scao et al. 2022) and PaLM (Chowdhery et al. 2023) have an even higher percentage of non-English training data (Ahuja et al. 2023). Due to

typological differences between languages and unbalanced amounts of training data, LLM performance on English benchmark tests does not represent their performance in other languages. To address this, many non-English benchmark tests have been developed.

Some benchmark tasks have been translated from English into a target language. The Korean NLI sets KorNLI and KorSTS were developed by machine translating English NLI sets with human translators editing the results (Ham et al. 2020). Similarly, Japanese NLI sets have been translated using machine translation as in JSNLI (Yoshikoshi et al. 2020) and manual human translation as in JSICK (Yanaka & Mineshima 2022). XNLI (Conneau et a. 2018) is a cross-lingual NLI set that was based on an English dataset translated into 15 different languages by professional translators. To address translation problems, they had annotators re-annotate a sample of 100 translated sentences to generate new gold labels and they found that annotators recover the original English gold label 85% of the time and 83% on the French translation. From this, they conclude that there is no significant problem with preserving semantic relationships between sentences. This conclusion may be flawed, especially in languages with greater differences than English and French.

Other tests have been developed directly in that language with tasks modeled on GLUE (Wang et al. 2018). The Chinese Language Understanding Evaluation Benchmark (CLUE) (Xu et al. 2020) is one example. It consists of nine tasks including single-sentence and sentence pair classification, Chinese WSC, NLI, and reading comprehension tasks. The tasks resemble those used in GLUE for LLM evaluation in English, but were constructed using native Chinese corpuses. The authors also established human baselines for tasks which did not already have them by averaging annotators' responses to a randomly selected set of 100 items. FLUE, the

French version of GLUE was constructed in a similar way, using previously constructed evaluation tasks in French from cross-lingual datasets or monolingual sets (Le et al. 2020). Wilie et al. (2020) developed an Indonesian benchmark test, IndoNLU, with various sentence classification and sequence labeling tasks using Indonesian datasets. The Korean language version KLUE (Park et al. 2021) was developed by constructing a Korean corpus with rigorous selectional requirements. It includes eight tasks which are structured similarly to the original GLUE paper and include classification, NLI, named entity recognition (NER), dependency parsing and reading comprehension. BasqueGLUE (Urbizu et al. 2022) followed a similar pattern, collecting Basque datasets either from cross-lingual sets or mono-lingual Basque corpuses, which were used to create ten NLU tasks. The Russian SuperGLUE benchmark test (Shavrina et al. 2020) used existing Russian corpuses to construct test items for nine different tasks. The WSC is also included in the nine tasks, but is a translation of the English WSC. They also provided human baselines for all tasks, which they collected by using the majority vote of three to five crowd workers' responses to a sample of items. Similarly, researchers created a Polish version of GLUE, KLEJ (Rybak et al. 2020) using native Polish corpuses. A Japanese GLUE, JGLUE (Kurihara et al. 2022) was developed using no translation and only Japanese corpuses to test three main skills: text classification, sentence classification and Q&A.

Some benchmark tests have also been developed with the goal of testing a model's cross-lingual abilities. XGLUE (Liang et al. 2020) is a set of 11 cross-lingual tasks which cover single-input understanding, pair-input understanding and generation tasks. The corpuses used for XGLUE cover 100 languages. Xtreme (Hu et al. 2020) covers 40 languages and is made up of nine tasks. The authors indicate that they estimated human performance using the baselines

established for the originally sourced tasks. However, it is important to note that the original human baselines were established on monolingual English benchmark tasks.

### *2.2.3 Benchmark Test Translation Challenges*

Creating non-English benchmark tests has a number of challenges, including gathering usable data, especially for low resource languages. Using existing corpuses is time efficient, but their size and level of diversity may be much smaller compared to high resource languages like English and Mandarin. For certain benchmark tasks that require a specific structure, researchers have to construct their own corpuses.

There are even more drawbacks and challenges when it comes to translation methods. Translation takes a considerable amount of time and labor and the product, while still useful and necessary, often has significant issues that could have an effect on performance results. The creators of Mandarinograd (Bernard & Han 2020), the Mandarin translation of the WSC, discuss translation difficulties. Due to linguistic differences between languages, appropriate translations for some things may not exist. For items where there was no direct translation possible, they opted to create a new WSC problem that is thematically similar to the eliminated one. The researchers who translated NLI tests into Korean acknowledge that despite every effort to provide quality translations, the semantic relationships between sentences may not be preserved in the Korean translation (Ham et al. 2020). Additionally, due to translation errors and differences in conceptual meaning, the gold label taken from the original English may not be applicable to the Korean version. This is a significant drawback and may have a high impact on the quality of the dataset and its usefulness as an accurate benchmark test. However, researchers

would probably argue that inconsistency in the quality of the dataset is mitigated by the large number of examples.

While this may be true to some extent, I think that it depends on the percentage of poor quality and high quality items, which can only be evaluated if there are clear selection criteria established and multiple human evaluators. This is very difficult, if not impossible in many cases due to time and labor costs, but at the very least, more scrutiny should be applied to the quality of benchmark datasets and their individual limitations should be clearly stated when used in evaluating LLMs.

# 3 The Winograd Schema Challenge

As introduced in section 2.2, one benchmark task currently used to evaluate commonsense reasoning and world knowledge in LLMs is the Winograd Schema Challenge (WSC). It was originally presented as an alternative to the Turing test. The Turing test, developed in 1950 by Alan Turing, involved text based conversations between a computer system and a human tester. After a long exchange, if the human tester could not reliably identify the conversation partner as a computer system, the system was said to have passed the test. Passing the Turing test was said to demonstrate an ability to think comparable to human intelligence. However, the dialogue format and goal of the test encouraged deceptive techniques. The creators of the WSC set out to make an evaluation method that showed whether or not a system exhibited a human-like ability to think, but did not rely on deception (Levesque et al. 2012).

The result was a short reading comprehension test. A Winograd Schema consists of one sentence and one question. The sentence introduces two entities and then uses a pronoun or possessive article to refer back to one of the entities. The question requires the system to identify the referent of the pronoun or possessive article. High performance on the WSC is supposed to indicate human-level ability to use commonsense reasoning and world knowledge.

The development of Winograd Schemas is guided by three criteria. First is that Winograd Schemas should not be too difficult that a native English speaking adult cannot answer the question. The second is that they must not be so simple that they could be answered by pragmatic or syntactic clues alone. An example of a pragmatic clue could be presenting a dog and a human

as the two entities, with the question involving the ability to speak. Examples of syntactic clues include agreement in gender, number or verbal inflection. The third criteria is that Winograd Schemas should be *google-proof*, unanswerable by word co-occurrence patterns. An example of a Winograd Schema that violates the third criteria would be a sentence involving a mechanic and a baker, and the question concerned with fixing cars. Sentence (1a), shown below, is an example of a valid Winograd Schema that adheres to the format of the test and fulfills all three criteria.

(1) a. The trophy would not fit into the brown suitcase because it was too big.

What was too big?

Answer: the trophy

(Davis 2018)

The first sentence of (1a) introduces two entities: *the trophy* and *the brown suitcase*, and the ambiguous pronoun *it* refers to one of them. In order to answer the question correctly, a system would have to use spatial reasoning as well as lexical and syntactic understanding of the concept *too big* to arrive at the answer, *the trophy*.

Despite the ambiguity, the question is not so difficult that a native English speaking adult would be unable to answer, which fulfills the first criteria. The two entities introduced are both referenceable by the pronoun *it*, so syntactic patterns alone will not provide an answer to the question. As a rule of thumb, Winograd Schemas in English must introduce a male-male, female-female or neutral-neutral pairing. *Too big* is also an appropriate descriptor for both of the objects, so there are no pragmatic clues available either, therefore fulfilling the second criteria. The word

*big* does not have a preferred association with *trophy* or *suitcase*, so word collocation statistics are also insufficient to arrive at the correct answer, therefore fulfilling the third criteria.

In addition to following this basic structure and restrictions, the sentence and question must also have a single word or phrase that can be changed to flip the correct answer. This is to ensure that the answer of the question is not determinable by pragmatics or frequency statistics. (1b), shown below, is the alternate pair of (1a).

Special word: big/small

(1) a. The trophy would not fit into the brown suitcase because it was too big.

What was too big?

Answer: the trophy

b. The trophy would not fit into the brown suitcase because it was too small.

What was too small?

Answer: the brown suitcase

(Davis 2018)

The original paper that introduced the WSC included 18 hand-crafted examples. These were expanded upon by Ernest Davis, Ray Jackendoff, David Bender and other scholars. Their collection of 150 examples were published online in 2011. An expanded dataset, also available on the WSC website, consists of 273 Winograd Schemas, including the original 150. It is known as the WSC273 set and it is the standard set applied as a benchmark test. The questions in the

WSC273 dataset adhere to the basic format and the three development criteria. The authors also intentionally tried to cover a wide range of knowledge and linguistic features (Davis 2018).

Since that original set, many scholars have developed other datasets. However, as it is very difficult to handcraft examples that fit the original requirements, many subsequent datasets adopt less rigid rules for viable test items. Winogrande (Sakaguchi et al. 2021) is a set of 44,000 crowdsourced problems. The Definite Pronoun Resolution set (DPR, also referred to as the WSCR) (Rahman, Ng 2012) is a set of 941 sentence pairs constructed with relaxed restrictions by volunteer university students. The Pronoun Disambiguation Problem Dataset (PDP) (Morgenstern et al. 2016) is a set of Winograd Schema-like questions which were sourced and modified from published material or hand-constructed. The Winograd Natural Language Inference set (WNLI) (Wang, et al. 2019) appears in the GLUE and SuperGLUE benchmark tests, but rather than presenting a choice of two entities, it is a binary classification task in which the system must determine if the pronoun refers to the indicated entity or not. WinoGender (Rudinger et al. 2018) and WinoBias (Zhao et al. 2018) were WSC style tests created to investigate gender bias in computer systems.

Because of the format and criteria of the test, human performance on the WSC is expected to be close to 100 percent. A computer system that performs better than chance can be said to demonstrate some commonsense thinking, but in order to be considered human-level, it would have to match or exceed human performance (Kocijan 2023). In 2015, four years after the test's introduction, David Bender used 140 of the original Winograd Schemas and 407 human participants answered a randomly selected 40 questions. The questions were divided into easy and hard items. Easy items were defined as questions which could be answered using pragmatic

clues or statistical correlations. Notably, things that qualified as “easy” items would not be valid Winograd Schemas according to the original stipulations. The participants answered approximately 36 hard items and 4 easy items. The overall mean score was 92.1%, but for just easy items, it was 98.6% (Bender 2015). During the creation of Winogrande, human answers were used as a baseline for questions that remained in the final dataset (Sakaguchi 2021). Each question was answered by three independent annotators and only questions which were answered correctly by two out of three people were included in the final set. This eliminated 32 percent of the proposed questions. Abdou et al. (2020) created a set of “perturbed” Winograd Schemas which were sourced from the WSC273 dataset. To create their set of 2,330 examples, they took the original questions and changed the tense, number, gender or voice, added a relative clause, adverbial qualification or substituted names with synonyms. Following conventions from Winogrande, they tested three human annotators and measured agreement between them. The three annotators agreed on the most correct answer in 82-83% of the questions.

The WSC has been translated into many languages and translations are available on the WSC website. The French translation was adapted from the English version and totals 214 questions (Amsili & Seminck 2022). The Portuguese translation adapted some questions to match Portuguese gender agreement and other grammatical functions and the result totals 285 questions (de Melo et al. 2019). The available Mandarin translation is of the original 150 questions (Bernard & Han 2020). The Chinese WSC used in CLUE (Xu et al. 2020) was constructed using examples from contemporary Chinese literature, but selection criteria are not explained. The Hungarian WSC was created using machine translation and checks by linguists and totals 122 questions (Vadász & Ligeti-Nagy 2022). The Russian translation of the WSC was

used as one of the tests in the Russian SuperGLUE benchmark (Shavrina et al. 2020). Also listed on the WSC website is a Japanese translation of 142 questions from the original 150 (Tanaka et al. 2014). It was recently updated with corrections and additions (Hashimoto et al. 2023).

### 3.1 The Japanese Winograd Schema Challenge

There are two Japanese translations of WSC datasets. The one available on the WSC website is of the original 150 questions. Hashimoto et al. 2023 made corrections and additions to it and the updated dataset is referred to as the WSC273-ja (Hashimoto et al. 2023).

The other is a translation of the Rahman & Ng (2012) Definite Pronoun Resolution set (DPR, also referred to as the WSCR). The WSCR set follows the standard format but with relaxed restrictions on the *google-proof* rule, allowing for items that might be answerable by correlation statistics or pragmatic reasoning (Rahman & Ng 2012). The Japanese version is called the WSCR-ja and it consists of 941 question pairs translated into Japanese (Shibata et al. 2015).

There are some significant translation differences between the WSC372-ja and the WSCR-ja. One difference is that the items in the WSCR-ja set use two sentences rather than one as in the English WSC. Thus, the Japanese WSCR-ja tests inter-sentential analysis rather than intra-sentential (Takahashi et al. 2023). The reason for this translation decision may be due to Japanese sentence structure. In the English version of the brown suitcase (1a-b) example: *The trophy doesn't fit into the brown suitcase because it's too big*, the subordinate clause *because it's too big* comes after the main clause. In English, this order could be switched: *Because it's too big, the trophy doesn't fit into the brown suitcase*. However, in Japanese, if the goal is to have

one sentence, the subordinate clause must always come before the main clause. See examples (2a-b).

(2) a. **torofi ga chairoi sutsukesu ni osamara-na-katta amari ni ookii node.*

Trophy TOP brown suitcase in fit-NEG PST too big because

'The trophy didn't fit into the brown suitcase because it was too big.'

b. *amari ni ookii node, torofi ga chairoi sutsukesu ni osamara-na-katta.*

too big because trophy TOP brown suitcase in fit-NEG-PST

'Because the trophy is too big, it didn't fit into the brown suitcase.'

(Tanaka et al. 2014)

In the WSC273-ja set, the translators chose to flip the order of clauses as in (2b), therefore keeping it to one sentence and remaining grammatical in Japanese. However, the WSCR-ja splits the items into two sentences, maintaining the order of information presented and overall structure of the statement as shown in (3a).

(3) a. 教授は学生に低い成績をつけた。彼は厳格だったからだ。

*kyoju wa gakusei ni hikui seiseki o tsuketa. kare wa genkakudatta kara da.*

The professor gave the student a bad grade because he was strict.

(Shibata et al. 2015)

Another significant difference in translation is the decision to keep or omit the pronoun in Japanese. Japanese is a pro-drop language, which means that pronouns can be omitted from the sentence (Shibatani 1990). The WSCR-ja set keeps an overt pronoun as seen in (3a) with *kare* 'he', but the WSC273-ja set omits the pronoun as seen in (2b) where there is no Japanese morpheme that corresponds to *it*. This phenomenon and its implications for the WSC will be explored in more depth in section 3.2.

To my knowledge, most studies on the Japanese WSC have used the WSCR-ja set. It was used in the development of the Japanese Feature Change Knowledge Base (JFCKB) (Nakamura & Kawahara 2018) and in the development of XWINO (Tikhonov & Ryabinin 2021), a multilingual WSC benchmark test made up of datasets in seven different languages. Takahashi et al. (2023) uses the WSCR-ja in their comparison of pre-trained language models. They compare BERT, RoBERTa, XLM, T5, GPT and GPT-2 and find that a version of XLM using a siamese network achieves the highest performance when tested on the dataset. Hashimoto et al. (2023) used both the WSCR-ja and WSC273-ja datasets to test BERT. Following practices from Kocijan et al. (2019), a study on the English WSC, they used the WSCR-ja set to fine-tune BERT and found a significant increase in performance on the WSC273 set by the fine-tuned model. Both Kocijan et al. (2019) and Hashimoto et al. (2023) find an increase in performance on WSC273/WSC273-ja when the model has been fine-tuned on the WSCR/WSCR-ja.

In addition to the translation choices discussed above, there are other significant difficulties in translating English to Japanese. Particularly relevant to the WSC is the difference between the pronoun systems of English and Japanese. During the translation of the WSCR, researchers found some instances in which the two entities of the sentence could not be referred

to with the same pronoun in Japanese due to gender or number agreement. In those sentences, the researchers decided to change the meaning of the sentence or use an unnatural pronoun in order to avoid giving the system a syntactic hint. See (4a-b) for the example they present.

(4) a. シボレーは最近彼らの車をリコールした。彼らは自社製品のブレーキに関する苦情を受けたからだ。

*shibore wa saikin karera no kuruma o rikorushita. karera wa jisha seihin no bureki ni kansuru kujo o uke-ta kara da.*

'Chevrolet had to recently recall their cars because they received complaints about the brakes in their product.'

b. シボレーは最近彼らの車をリコールした。だから彼らは車のブレーキの修理を受けることができた。*shibore wa saikin karera no kuruma o rikoru-shita. dakara karera wa kuruma no bureki no shuri o ukeru-koto ga dekita.*

'Chevrolet had to recently recall their cars so they could receive a repair in their brakes.'

(Shibata et al. 2015)

This pair of Winograd Schemas is acceptable in English because the pronoun *they* can refer to both animate and inanimate referents. In (4a), the underlined pronoun *they* in the English version can refer to either *Chevrolet* or *their cars* but a human is able to reason that the entity receiving complaints is *Chevrolet* rather than *their cars*. Likewise, in (4b), *they* could refer to

either entity, but based on commonsense reasoning, we can determine that the most likely referent is *their cars*.

This is not the case for Japanese. The third person plural pronoun in Japanese differs on animacy and gender. A group of animate female entities takes the pronoun *kanojora* 'they' (plural, animate, female), and a group with male and female entities or only male entities is referred to with *karera* 'they' (plural, animate, male). However, a group of inanimate objects needs one of the following three pronouns: *sorera* 'those', *korera* 'these' or *asokora* 'those over there' (plural, inanimate, neuter).

The Japanese translation of (4a) has no problem because *karera* 'they' (plural, animate, male), matches the gender and animacy of its antecedent *shibore* 'Chevrolet'. However, this is not the case in (4b) as the grammatically acceptable pronoun to use for *karera no kuruma* 'their cars' would be *sorera* 'those' (plural, inanimate, neuter). To avoid giving the system this syntactic clue, the translators kept the pronoun *karera* 'they' (Shibata 2015). However, due to the animacy requirement of *karera* 'they' and syntax of the sentence, the current Japanese translation will cause readers to resolve it as *the car owners* instead of the cars themselves.

These are just a few examples of the difficulties of translating WSC questions. Kurihara et al. (2022) discuss other linguistic differences between English and Japanese which make translation challenging, including the complex Japanese alphabet, lack of spaces between words and relatively free word order. They argue that as a result of these differences, findings from research on English may not be applicable to Japanese and that one drawback of contemporary studies in Japanese is that most are machine or human translation of English datasets. Translations can result in text that is considerably different from native text and there may be

cultural or social features that require more specification when translated. As a result, despite efforts to translate text into natural sounding language, translated corpuses may still sound English-like (Clark et al. 2020). To address these issues, Kurihara et al. (2022) built the JGLUE benchmark test from scratch without translation, but their collection does not include a WSC task.

Because of these problems, linguistic differences and a number of other issues, the English WSC human baseline established by Bender (2015) does not translate to the Japanese WSC datasets. The human baseline, established on the original 150 questions, may be a good estimate of performance in English, but not for the Japanese sets which were built on WSC273 and WSCR. In the creation of the WSCR set, researchers had multiple annotators evaluate created sentences but they did not establish a human baseline and instead relied on the assumption that human performance is close to 100% (Rahman & Ng 2012). The researchers that developed the Winogrande set calculated a human baseline for the WSCR set by taking the majority vote of three crowdworkers and found performance to be at 95.2% (Sakaguchi et al. 2021). These estimates and assumptions do not apply for Japanese because language specific features can have a significant effect on the ability to respond correctly to a test item. As noted earlier, in the creation of the Korean NLI translations, some of the gold standards even changed in the translated set (Ham et al. 2020). While this particular problem may or may not be an issue for the Japanese WSC, it highlights the degree to which translations can change semantic relationships. Therefore, a usable human baseline for Japanese WSC datasets does not yet exist.

In this study, I will use the WSCR-ja dataset to establish a preliminary baseline and investigate the efficacy of translated benchmark tests. To examine naturalness in the Japanese

WSC and also evaluate human and LLM sensitivity to surface linguistic features, two conditions will be tested. The conditions will be based on a significant typological difference between English and Japanese: the use of ellipsis and zero anaphora.

**3.2 Zero Anaphora in Japanese**

In English, all arguments of a sentence are usually morphologically realized. This is not the case for Japanese. In Japanese, arguments that are recoverable through contextual or linguistic clues can be omitted. The omission of arguments is referred to as ellipsis. This is most pronounced in Japanese with nominal forms, and specifically with first person and second person pronouns (Shibatani 1990). Shibatani (1990) uses the following examples (5a-b) to illustrate this difference between English and Japanese.

(5) a. *Tokorode, moo kono hon yonda?*
By the way already this book read
'By the way, (have you) read this book already?'
b. *Un, moo yonda.*
Yeah, already read
'Yeah, (I've) already read (it).'

(Shibatani 1990)

In the English translation of (5a), the word *you* is a required argument and the sentence would not be grammatical if it were missing. The auxiliary verb *have* is also in parentheses, but

this is due to a difference in verb conjugation rather than ellipsis. Likewise, in the (5b) English translation, the subject *I* and the object *it* are required elements.

In contrast, the Japanese text of (5a) omits the pronoun *anata*, 'you' and (5b) omits both the first person pronoun and the object. Despite the elided arguments, these sentences are acceptable and natural in Japanese. Considering the short verbal form and lexical choices, this exchange is in an informal register and there is more ellipsis present that is not being highlighted. The particles *wa* and *o*, used to mark subjects and objects respectively, are also omitted.

In the (5b) English translation, the relation of the word *it* to *this book* is called anaphora. In the Japanese text of (5b), the pronoun is elided and may therefore be called a null pronoun. The relation of that null pronoun to *kono hon* 'this book' in the first sentence is called zero anaphora. Unlike English, where ellipsis is only acceptable in very specific settings, zero anaphora and ellipsis are very widespread in Japanese in both informal and formal contexts.

Kameyama (1985) describes how languages differ typologically in their requirement or non-requirement for overtly expressed grammatical functions. She identifies two ways for languages to encode pronominal information: free pronouns, such as *he*, and bound morphemes, such as *-s* as in *He walk-s*.

In English, with the exception of specific dialogue contexts such as informal question and answer settings, overt free pronouns are always required for major grammatical functions and there is some amount of agreement inflection between subjects and verbs.

However, in Japanese, there is no agreement between subject-verb or verb-object and no obligatory pronoun. Both the overt pronoun and null pronoun are acceptable, but an overt pronoun is not *required*. In fact, Japanese shows a preference for the null pronoun, which is

equivalent to an unstressed pronoun in English. Just as repetitions of nouns with the same referent are avoided in English for style reasons, Japanese avoids using an overt pronoun when it is not necessary for understanding (Kuroda 1965).

### *3.2.1 Interpreting Zero Anaphora*

Kameyama (1985) argues that the use or omission of an overt pronoun is pragmatically decided, so zero anaphora interpretation must be informed by discourse-level information. She proposes that one significant result of this is the increased semantic specification of arguments as per the following examples.

1. Strict animacy constraint: Constructions like *A traffic accident killed a teenager* are not acceptable in Japanese because the subject of the verb *korosu* 'kill' must be a higher animal capable of thought and action.
2. Human/nonhuman distinction: Kameyama uses an example from Chinese, of the differences in the verbs for 'kill' and 'give birth to' based on the level of humanness of the direct object. In Japanese, there is some distinction between human and nonhuman in certain verbs, but these may be related more to hierarchy than the human/nonhuman distinction. For example, the verb *ageru* 'give' may be used to refer to an act of giving when the indirect object is a human of the same or lower social status as the speaker. The verb *yaru* 'give', however, is reserved for things perceived as lower beings, including plants and animals, but occasionally children or lower social-status humans.
3. More differentiated semantic field: In English, the phrasal verb *put on* may be used for many different kinds of clothing, while in Japanese, there are different verbs depending on what part of the body the clothing is intended for and the manner in which it covers

the body. For example *boushi wo kaburu* ‘to put on (one’s head) a hat’ or *kutsu wo haku* ‘to put on (one’s feet) shoes’.

4. Honorificity agreement: In Japanese, there is subject-verb agreement based on politeness, formality and social status. For example, *irassharu* ‘come/go’ describes the actions of a respected person while *mairu* ‘come/go’ is a humbling reference to the speaker or someone within their group or family.

Shibatani (1990) provides a similar analysis of how semantic phenomena in Japanese serve to help interlocutors identify the referents of elided elements. He argues that the complex honorific system in Japanese helps indicate who actions are ascribed to, which partially compensates for the lack of a verbal agreement system. He also draws attention to deictic expressions, which encode information about the direction of an action or location of an object, and the evidential system, which indicates the source of information and level of confidence.

The increased semantic specification, honorific system and phrases with deictic encoding pose many challenges for translators. The strict animacy constraint and the human/nonhuman distinction are particularly of note for translating WSC examples as demonstrated in section 3.1. This is a significant typological difference in the context of the WSC as the test hinges on a subject’s ability to resolve a pronoun.

Okamoto (1985) comes to the same conclusion that zero anaphora interpretation depends on context and semantic properties. She proposes four interpretation strategies for nominal ellipsis in Japanese.

**Figure 1. Principles of Interpretation for Zero Anaphora**

1. **Principle of Role Assignment for the Referent : Assign the semantic role of the unexpressed item by making reference to the frame of the situation described by the sentence as well as examining the relevancy/importance of the item to the situation.**

2 **Principle of Local Interpretation : Regard items in the immediate context as the potential "referents."**

3 **Principle of the Use of Syntactic Clues : Use the syntactic clues when available and when necessary.**

4 **Principle of Pragmatic Interpretation : On the basis of the knowledge of the world, choose from the potential "referents" an item which is pragmatically most appropriate/normal as the "referent."**

*Figure 1. Four principles of interpretation by Okamoto (1985)*

Using principle 1, speakers identify ellipsis based on semantic frames and use their knowledge of relevant frame elements to interpret the discourse. Principle 2 describes speakers narrowing down the possible referents by only considering only the immediate context. Principle 3 deals with syntactic features such as the honorific system, deictic expressions, evidentials, etc. Finally, principle 4 concerns the speaker's use of pragmatic knowledge, world knowledge and common sense.

Modern natural language processing has machine equivalents for principles 1-3 but whether or not current LLMs such as ChatGPT can use principle 4 is the target of evaluation for many benchmark tests, including the Winograd Schema Challenge.

Kameyama (1985) draws on Centering Theory (Grosz et al. 1983; Sidner 1983) to examine usage patterns of zero anaphora. The "center" is defined as a non-linguistic entity centrally being talked about or in the center of attention at the time the utterance is produced or understood in discourse. See the following English and Japanese examples.

(6) a. Who is Max waiting for?

b. He is waiting for Rosa.

c He invited (Rosa/her) to dinner.

(7) a. *Max wa dare o matte-iru no.*

'Who is Max waiting for?'

b. *Ø Rosa o matte-iru n-da yo*.

'Ø is waiting for Rosa.'

c *Ø { Rosa o / kanojo o / Ø } yushoku ni sasot-ta n-da*.

' Ø invited {Rosa/her/} to dinner.'

(Kameyama 1985)

In English, the center is usually referred to with an unstressed pronoun as seen in sentences (6b-c) with the use of *he*. In Japanese, a center most often uses a null pronoun as seen in sentences (7b-c). The center may change throughout the discourse, but certain thematic roles and grammatical functions are given preference. In English discourse, the entity that fulfills the theme role is the most preferred center and the agent is the least preferred (Sidner 1979). In Japanese, there is a clear preference for the subject.

Kameyama uses this theory to describe the usage of zero anaphora in Japanese, but it also illuminates some questions of interpretation as well. We can assume that examples that follow the expected pattern will be easily interpretable and that those which don't will either be perceived as unnatural or be more difficult to interpret.

Kameyama and Okamoto agree that their analysis of ellipsis is mostly applicable to spoken discourse. In spoken language, the high frequency of information exchange provides the necessary context for interpretation of elided elements and both speaker and hearer are understood to be possible referents (Shibatani 1990). Therefore, some of the conventions discussed above may not be true or may be less pronounced for written or formal language. However, ellipsis in writing is still more widespread in Japanese than it is in English.

### *3.2.2 Zero Anaphora and the Winograd Schema Challenge*

In the context of the WSC, the difference between English and Japanese regarding ellipsis and zero anaphora is significant. The WSC requires a system to resolve an ambiguous pronoun. In Japanese, a pronoun may be used, but as shown, the null pronoun would also be grammatical and in some cases, the default. As discussed in section 3.1, the majority of items from the WSC273-ja set use a null pronoun. For items which use a null pronoun, the task would be zero anaphora resolution rather than pronoun resolution.

While native speakers would presumably not have a problem interpreting the overt or null pronoun version of a sentence, zero anaphora resolution is a significantly different linguistic processing task than pronoun resolution for an LLM, and particularly for Japanese NLP, it is still an ongoing area of research (e.g. Konno et al. 2021; Sasano & Kurohashi 2011; Ueda et al. 2020; Ueyama & Kano 2023).

In contrast, the WSCR-ja dataset almost always uses overt pronouns with the exception of 11 sentences in the training set for which both overt and null versions exist. It preserves the pronoun resolution task of the WSC while using a dataset that adheres to less stringent rules than the original collection. While its format is closer to the original, the relaxed restrictions and

easier task of pronoun resolution over zero anaphora resolution means that it may be a simpler test for LLMs than the WSC273-ja.

Accordingly, we would also expect humans to achieve high performance on this pronoun resolution task, but perhaps see discrepancies in judgments of naturalness based on the widespread usage of ellipsis in Japanese. If LLM training data in Japanese accurately represents its usage, then we would also expect LLMs to show no difference in performance for overt or null pronoun examples, but perhaps show some preference for null pronoun items in terms of naturalness.

# 4 Experimental Studies

This section describes two experimental studies. The first study was done with human participants and the second tested ChatGPT. Both studies were structured the same way with two tasks and the stimuli presented in the two tasks were the same. In Task 1, the participant or ChatGPT resolved the pronoun in a Japanese WSC problem. In task 2, the participant or ChatGPT rated the WSC stimuli for naturalness in Japanese.

## 4.1 Study 1: Human Performance on the Japanese Winograd Schema Challenge

Currently, the WSCR-ja dataset does not have an established human baseline, nor does the DPR set from which it was translated. When using the WSC to evaluate LLMs, it is necessary for baseline data because while human performance might be assumed to be close to 100%, studies may show lower accuracy as in Bender (2015), which revealed 92% accuracy by human participants. The goal of study 1 was to establish a human baseline for the WSCR-ja dataset to use in comparison with LLM performance.

As discussed in sections 2.2.3 and 3.1, translation challenges often lead to unnatural sounding language. As the researchers who translated the WSCR-ja prioritized faithfulness to the original syntax and goals of the text, the results were often unnatural in Japanese. In evaluation of LLMs, the effect of some unnatural examples might be mitigated by the large size of the dataset. However, linguistic naturalness may significantly affect human performance. When establishing a human baseline for a translated test, It is necessary to investigate not only accuracy, but also how natural the examples sound to native speakers and whether that might affect their performance. Therefore, the experiment consisted of two tasks:

1. Pronoun resolution task: the original WSC task which tests world knowledge and commonsense reasoning.
2. Naturalness rating task: rating how natural the examples sounded in Japanese.

### *4.1.1 Participants*

There were 33 native Japanese speaking participants. The youngest participant was 36, the oldest was 90 and the average age was 54.7. Participants consisted of 10 men, 20 women and 3 who selected "other" or did not state their gender. Recruitment was done via snowball sampling. There were three starting nodes all connected to different regions of Japan, so it is likely that a wide range of regional dialects are represented in the data.

### *4.1.2 Design*

The stimuli were presented in two conditions. Condition 1 was unmodified Japanese WSC problems. This was the overt pronoun condition. Condition 2 was the null pronoun condition in which the target pronouns were omitted. Table 2 shows example sentences for both conditions. Previous studies showed that LLMs are sensitive to surface sentence features, which suggests that more research is needed on the effect of structural variations on LLM performance (Kauf et al. 2023). The use of the overt pronoun condition and the null pronoun condition in this study takes this into consideration and the choice of variation is informed by the prevalence of zero anaphora in Japanese.

### *4.1.3 Materials*

The stimuli in this study are taken from the WSCR-ja dataset (Shibata et al. 2015). It has 941 question pairs and 10 of them were selected for this study. Prior to selecting items, there was a detailed study of the dataset. Many sentences were direct translations from English and

required specific cultural knowledge to resolve (e.g. American football terminology). Also, there were a number of structural inconsistencies. For example, some sentences had complex clausal structure, while others only had one simple clause. Sentences had varying numbers of unresolved pronouns and many had a mixture of active and passive voice. These inconsistencies are all features of natural language, so presumably they would not have affected the ability to answer the question correctly. However, including them in the study would have added many extra variables that are not feasible to analyze sufficiently with a small dataset. Therefore, there was a careful selection process which was informed by cultural appropriateness, structural factors and semantic considerations. In order for an item to be used in the study, it had to fulfill the following criteria.

1. The problem must consist of two simple sentences with no subordinate clauses.
2. The problem must not contain any long foreign names in Katakana (Japanese alphabet used to write foreign words) or culturally specific knowledge that is not familiar to most Japanese people.
3. The target referent of the pronoun must either be the subject, object, or indirect object of the first sentence.
4. There must be only one pronoun targeted for resolution.
5. The voice must be consistent between the first sentence and the second sentence. (active-active, or passive-passive)

Items were selected by generating a list of 100 random question pairs. From that set, the pairs that met all of the requirements were chosen for the study. Table 1 shows examples of items which were not selected because they violated the selection criteria.

**Table 1. Examples of Excluded Items**

| Item | Reason for Exclusion |
|---|---|
| ペイパルは不満を持つ客のために毎年数百万件の争議を解決しなくてはならない。彼らは取引保護策をとっているからだ。<br>Paypal has to settle millions of disputes every year by disgruntled users because they provide a transaction protection policy. | Comprised of more than just two simple clauses. |
| エイドリアン・ピーターソンはレジー・ブッシュよりタッチダウンの数が多い。彼はボールを前に運ぶのがあまり上手くないからだ。<br>Adrian Peterson has more touchdowns than Reggie Bush because he cannot run the ball as well. | Contains long, probably unfamiliar foreign names and requires cultural knowledge of the rules of American football. |
| ダンはハリーの車を改造した。彼が丁寧に頼んだからだ。<br>Dan modified Harry's car because he asked politely. | The referent of the target pronoun for resolution is “Harry”, who is neither the subject, object or indirect object of the first sentence. |
| 父は息子にいつか騎士になれると説得した。彼は彼の夢を打ち砕きたくなかったからだ。<br>The father convinced his son that it is possible for him to one day become a knight, because he did not want to crush his dreams. | There is an ambiguous pronoun (“he”) and an ambiguous possessive article (his). |
| クーガーはウサギと友達になろうとした。それが皆に愛されていたからだ。<br>The cougar tried to befriend the rabbit because it was loved by everybody. | The first sentence is active voice, and the second sentence is passive. |

Next, the WSCR-ja dataset does not include a question prompting for pronoun resolution, so this was created for each problem. The WSC273-ja set was referenced when deciding how to formulate the questions. It uses the two patterns shown below:

Pattern 1:

*dare/nani ga X ka?*

who/what SUB X Q

'Who/what X(s/ed)?'

Pattern 2:

*X (no) wa dare/nani ka?*

X (GN) SUB who/what Q

'As for X, who/what is it?'

Both patterns are natural expressions in Japanese, but depending on the syntax of the sentence, pattern 1 is not always applicable, making pattern 2 more versatile. To match the consistency of the questions in the English WSC, the same pattern was used for all questions in this study and the copula *desu* was added. See the following structure and example.

*X (no) wa dare/nan desu ka?*

X (GN) SUB who/what COP Q

'As for X, who/what is it?'

*neko wa nezumi o tsukama-e-ta. sore wa kashikoka-tta-kara*.

cat SUB mouse OB catch-can-PST it SUB clever-PST-because

'The cat caught the mouse because it was clever.'

*kashikoka-tta  no wa nan desu ka*

clever-PST GN SUB what COP Q

'What was clever?' lit. 'As for clever, what was it?'

Finally, a second version of each problem was created to serve as the null pronoun condition. To do this, the target pronoun and its related particle were omitted from the second sentence of each problem. Table 2 shows a complete question pair used in the experiments. See Appendix 1 for a full list of stimuli.

**Table 2. Example of Complete Question Pair.**

| Overt Pronoun (Original) | Null Pronoun |
|---|---|
| 教授は学生に低い成績をつけた。彼は厳格だったからだ。厳格だったのは誰ですか？<br>The professor gave the student a bad grade because he was strict. Who was strict?<br>Answer: the professor | 教授は学生に低い成績をつけた。∅ 厳格だったからだ。厳格だったのは誰ですか？<br>The professor gave the student a bad grade because Ø was strict. Who was strict?<br>Answer: the professor |
| 教授は学生に低い成績をつけた。彼が勉強しなかったからだ。勉強しなかったのは誰ですか？<br>The professor gave the student a bad grade because he did not study. Who did not study?<br>Answer: the student | 教授は学生に低い成績をつけた。∅ 勉強しなかったからだ。勉強しなかったのは誰ですか？<br>The professor gave the student a bad grade because Ø did not study. Who did not study?<br>Answer: the student |

### *4.1.4 Procedure*

The study was conducted via the online survey platform Qualtrics and participants accessed it with an anonymous link distributed by email, instant message or QR code. The first section of the survey was an implied consent form in which participants verified that they were over 18 years old and spoke Japanese as their native language. Next, they completed task 1,

testing accuracy in pronoun resolution, and task 2, rating the sentences for naturalness in Japanese. Both tasks had 40 questions and they were presented in a random order. There were no filler items and the human participants were not given any training questions in the interest of emulating the zero-shot prompting approach that was used with ChatGPT in study 2. The response time for each question was also recorded.

The instructions for task 1 informed the participants about the format of the questions and asked them to select the most appropriate answer. See figure 2 for the instructions shown to participants for task 1. Each problem had two Japanese sentences, one question prompt and two answer selections. Participants had to select an answer to the question, then click to progress the survey before they were shown a new problem. Both the order of problems and the order of answer selections were randomized. See figure 3 for an example. The English translation is provided in the figures, but participants only saw the Japanese text.

**Figure 2. Instructions for Task 1: Pronoun Resolution**

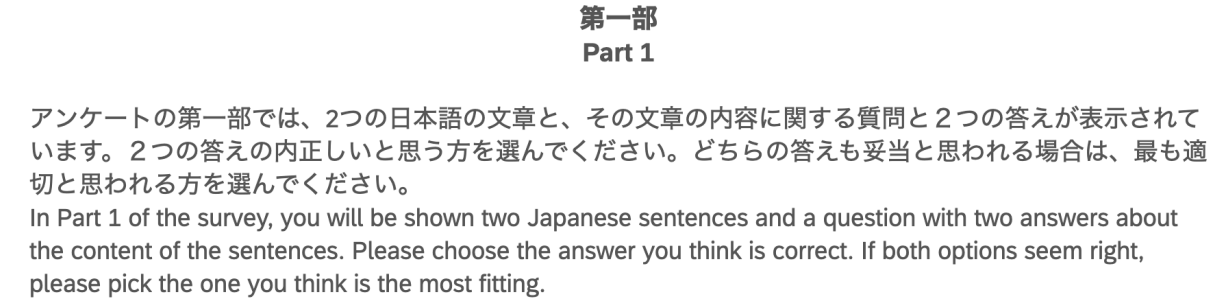


*Figure 2. Instructions given in Qualtrics to human participants for Task 1. The English text was not shown to participants.*

**Figure 3. Example Question for Task 1**

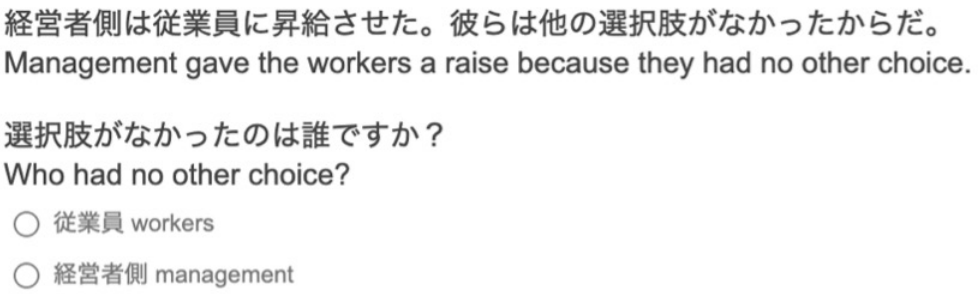


*Figure 3. Example pronoun resolution task question as presented to participants in Qualtrics. The English text was not shown to participants.*

Next, participants completed task 2. See figure 4 for the instructions given to participants describing the format of the task. The task 2 items consisted of the two sentences from a WSC problem, a prompt asking for a naturalness rating, and a five point likert scale. 1 was labeled "very unnatural" and 5 was "very natural". The items were shown in random order and as with task 1, participants had to select a response, then click to progress the survey to the next question. See figure 5 for an example item from task 2. The English translation is provided, but the participants only saw the Japanese text.

**Figure 4. Instructions for Task 2: Naturalness Rating**

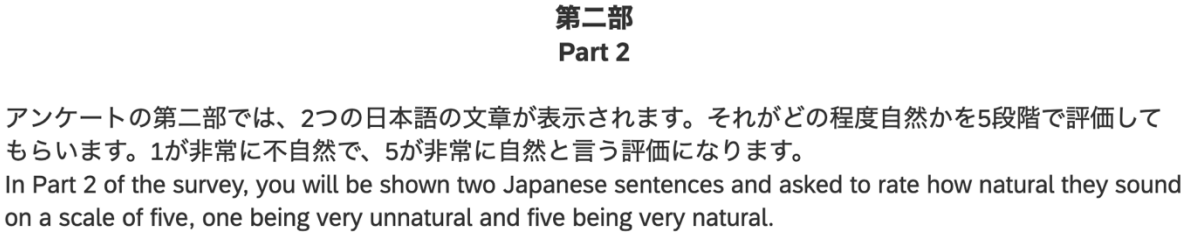


*Figure 4.Instructions given in Qualtrics to human participants for Task 2. The English text was not shown to participants.*

**Figure 5. Example Question for Task 2**

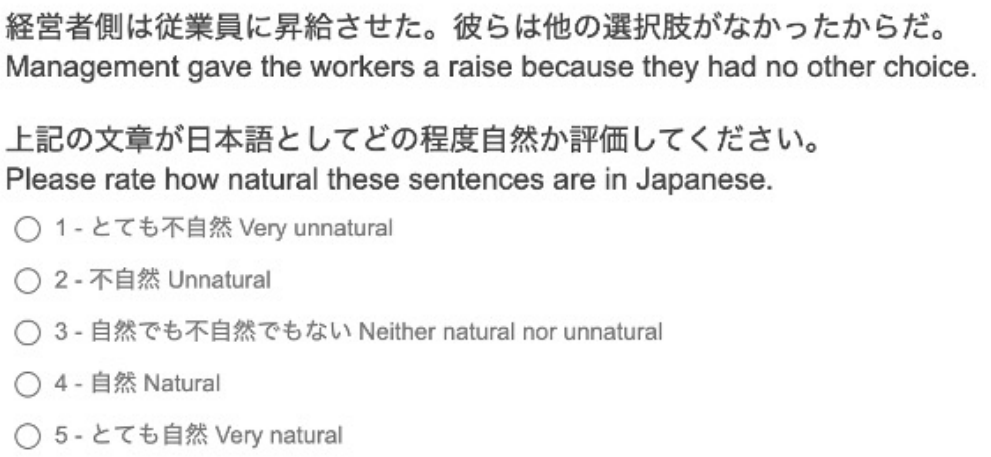


*Figure 5. Example Naturalness Rating task question as presented to participants in Qualtrics. The English text was not shown to participants.*

### *4.1.5 Results*

For task 1, answers were coded as 1 if the pronoun was resolved correctly and 0 if incorrect. Unanswered questions were not included in the results. The answer values were averaged across participants to get an accuracy number for each problem. Participants showed high accuracy levels for both the overt pronoun condition ($M_{human_overt}$ = 0.97, SD =0.05) and the null pronoun condition ($M_{human_null}$ = 0.96, SD = 0.05). The difference between the two conditions was not statistically significant on a paired t-test ($t(32) = 1$, $p = 0.28$, two-tailed). This is an indication that interpretation is not affected by the overt/null pronoun variation in the surface form. Considering how widespread zero anaphora is in Japanese, this result is to be expected.

Naturalness ratings were also averaged across participants for each question. On average, participants assigned higher naturalness ratings to sentences with overt pronouns ($M_{human_overt}$=3.28, SD = 0.53), than sentences with null pronouns ($M_{human_null}$ = 3.08, SD = 0.56). The differences

between the two conditions were statistically significant on a paired t-test ($t(32) = 4.27$, $p < 0.01$, two-tailed). These results are summarized below in table 3.

**Table 3. Study 1 Results**

| | **Accuracy** | **Naturalness** |
|---|---|---|
| **Overt Condition** | 0.97 (SD 0.05) | 3.28 (SD 0.53) |
| **Null Condition** | 0.96 (SD 0.05) | 3.08 (SD 0.56) |

## 4.2 Study 2: ChatGPT performance on the Japanese Winograd Schema Challenge

Study 2 collected ChatGPT responses on the same tasks in order to compare GPT accuracy on the Japanese WSC problems and naturalness judgements to human data. The study was performed on GPT-3.5, but the newer model GPT-4 is also currently available. Research on the Japanese WSC using GPT-4 is a direction for future work.

### *4.2.1 Procedure*

The experiment design and materials used in study 2 are the same as in study 1. The only difference is the order in which items were presented to ChatGPT. Since the overt pronoun condition of the sentences is openly available for download, it is possible that they were used in pre-training data for GPT-3.5. This means that performance on the null pronoun items is less likely to be influenced by pre-training data, and may more accurately reflect ChatGPT's true ability. Therefore, the stimuli in the null pronoun condition (pronoun resolution and naturalness rating tasks) were presented first, then all stimuli in the overt pronoun condition.

In preparation for the study, different prompting strategies were tested using WSC items that are not included in the test data. There were no significant differences between zero-shot and few-shot prompts, so a zero-shot prompt format was used in the experiment.

The experiment was run twice. Each task for each condition was run in different sessions, totalling 8 sessions. The order of sessions was as follows: Task 1 null condition, Task 2 null condition, Task 1 overt condition, Task 2 overt condition.

In task 1 sessions, instructions were first provided to ChatGPT in Japanese: “Please answer the following questions with A or B.” The problems were structured to mirror the format used with human participants as much as possible. Each problem had the two WSC sentences, the prompting question, and two answer selections which were labeled A and B. ChatGPT was shown the problems one at a time, in random order with the order of answer selections also randomized.

Task 2 also tried to create the same testing conditions as were used with human participants. ChatGPT was shown the two WSC sentences, then the same naturalness rating prompt and likert scale as in the human survey. The items were shown one at a time in random order. It was not requested to provide an explanation, but for each answer, ChatGPT generated a short passage describing the reasoning it used when selecting its response.

### *4.2.2 Results*

The results of the two runs were averaged. ChatGPT demonstrated slightly lower accuracy on the overt pronoun condition (MChatGPT_overt=0.80) than on the null pronoun condition (MChatGPT_null= 0.83). In naturalness ratings, it gave sentences in the overt pronoun condition (MChatGPT_overt = 3.48)  lower ratings than in the null pronoun condition (M ChatGPT_null = 3.58). These means were used to compare ChatGPT’s performance and ratings with those from the human baseline. See table 4 for a summary of the study 2 results from ChatGPT.

**Table 4. Study 2 Results**

| | Accuracy | Naturalness |
|---|---|---|
| **Overt Condition** | 0.80 | 3.48 |
| **Null Condition** | 0.83 | 3.58 |

### 4.3 Comparing ChatGPT and Human Performance

**Figure 6. Comparison of Human and ChatGPT Performance**

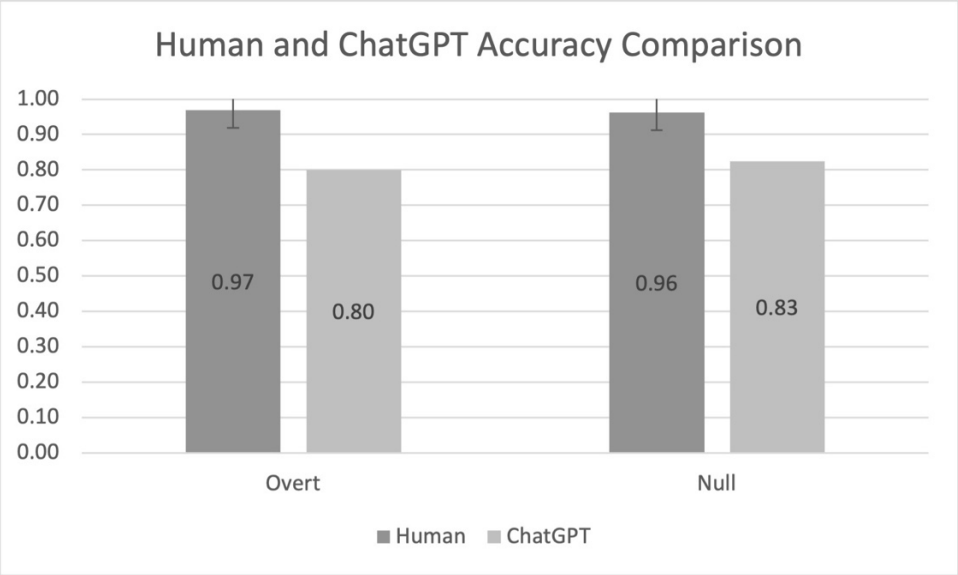


*Figure 6 A graph comparing human and ChatGPT performance on Task 1- Pronoun Resolution.*

Comparing ChatGPT and humans on the same pronoun resolution tasks, we can see that humans achieve better accuracy in both conditions ($M_{human_overt} = 0.97$, $M_{human_null} = 0.96$, $M_{ChatGPT_overt} = 0.8$; $M_{ChatGPT_null} = 0.83$). The differences are statistically significant on a one sample t-test for both the overt pronoun condition ($t(32) = 19.2$, $p < 0.01$, two-tailed) and the null pronoun condition ($t(32) = 15.6$, $p < 0.01$, two-tailed).

**Figure 7. Comparison of Human and ChatGPT on Naturalness Rating**

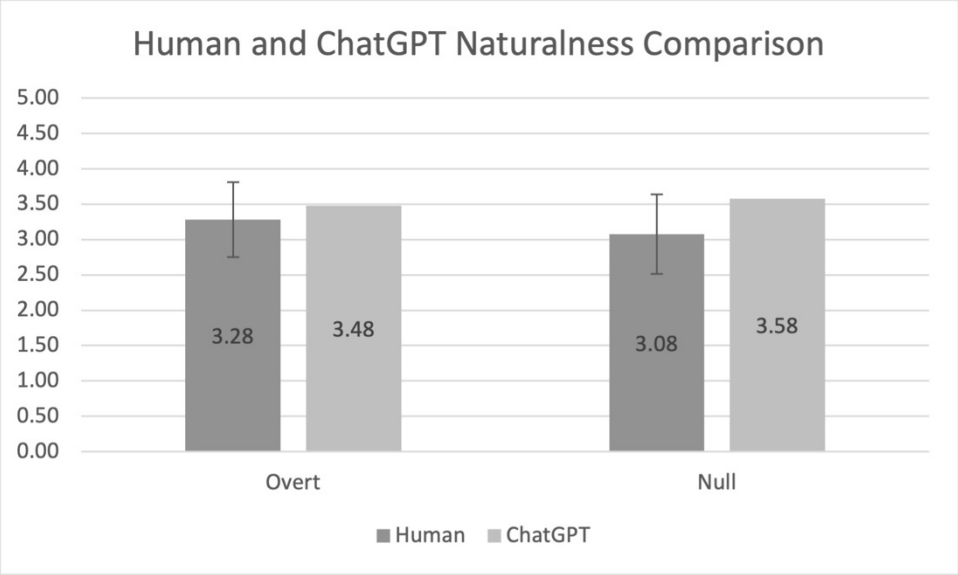


*Figure 7 A graph comparing human and ChatGPT responses on Task 2- Naturalness Rating.*

Humans and ChatGPT also demonstrate differences on the naturalness judgment task. On average, humans judge Japanese WSC sentences as less natural ($M_{human_overt} = 3.28$; $M_{human_null} = 3.08$) compared to ChatGPT ($M_{ChatGPT_overt} = 3.46$; $M_{ChatGPT_null} = 3.58$). The differences are statistically significant on a one sample t-test in both the overt pronoun condition ($t(32) = -2.09$, $p = 0.04$, two-tailed) and in the null pronoun condition ($t(32) = -5.06$, $p < 0.01$, two-tailed).

## 5 Experimental Studies Discussion

The WSC as a benchmark test can give us valuable insights into how closely ChatGPT resembles humans in commonsense reasoning and world knowledge. The results of the two studies presented show that humans outperform ChatGPT (GPT-3.5) on the Japanese WSC. While ChatGPT clearly performed better than chance, it would have to match or exceed human performance in order to be considered on par with human thinking (Kocijan et al. 2023). It seems that for now, GPT-3.5 is not capable of human-level commonsense reasoning and use of world knowledge in Japanese. This could have a significant influence on how we use information and decisions processed and delivered by ChatGPT.

The results also show differences in naturalness ratings of the Japanese WSC test sentences. Ratings from human participants were surprisingly low, which is reflective of translation difficulties as discussed in section 2.2.3 (Amsili & Seminch 2017; Bernard & Han 2020; de Melo et al. 2020; Shibata et al. 2015). ChatGPT ratings were more moderate, but the difference in naturalness ratings calls into question the use of this translated test for natural language processing in Japanese.

This is the first attempt to gather human data on the Japanese WSC, so there are several limitations to address. First, as participants did not receive any compensation, it was necessary to keep the survey short. The WSC items used in this study were only a small, curated set of the original WSCR-ja dataset. Items were chosen after rigorous screening for culturally specific knowledge and structural inconsistencies. Human and ChatGPT performance on the full, unfiltered dataset may be lower than shown here. Second, the sample size was small at 33

participants. Both the limited dataset and small sample size make it difficult to draw more detailed conclusions about the differences between LLM and human reasoning, but the analysis of results offer some suggestions for the direction of future research.

**5.1 Accuracy**

Previous research on human baseline data for the original WSC set in English showed a 92% accuracy (Bender 2015). Participants answered a selection of 40 random questions taken from a collection of 160 questions and all 160 questions were answered at least 50 times. Data from this study shows higher accuracy on the Japanese WSC at 96% overall. However, the test consisted of a smaller set of questions and the WSCR-ja set is much easier than the original WSC set because many items are resolvable by selectional restrictions or statistical correlation. The human baseline for the English WSCR, as reported by Sakaguchi et al. (2021), was 95.2%, slightly lower than the Japanese data presented here. However, other than stating that performance was calculated by averaging answers from three crowdworkers, no further elaboration on the human data collection method was given. For these reasons, it is still unclear whether the human performance on the English set corresponds well to the Japanese.

As expected, humans showed no difference in performance between overt and null conditions with high performance for both (96% accuracy and 97% respectively). This is reflective of widespread zero anaphora in Japanese. ChatGPT also showed no significant difference in performance between the two conditions, suggesting that there were enough examples of zero anaphora in its Japanese training data for it to learn how to interpret null pronouns.

20 scenarios were used in this study. Each scenario consisted of two questions, one in which the answer was the subject of the sentence, and one in which the answer was either the object or direct object. Each question also had two versions: overt pronoun condition and null pronoun condition. Please refer to Appendix ? for the list of stimuli used. Throughout the rest of this discussion section, I will be referencing scenarios and question numbers, so the question numbers, scenarios and answers are reproduced in table 5 for convenience. Odd numbered questions are overt conditions and even numbered questions are null conditions.

**Table 5. List of Questions by Scenario**

| Question Numbers | Scenario | Answer |
|---|---|---|
| 1, 2 | Raise | Management (S) |
| 3, 4 | Raise | Workers (IO) |
| 5, 6 | Ask | John (S) |
| 7, 8 | Ask | Bill (IO) |
| 9, 10 | Protest | The people (S) |
| 11, 12 | Protest | The government (IO) |
| 13, 14 | Candy | The man (S) |
| 15, 16 | Candy | The boy (IO) |
| 17, 18 | Grades | The professor (S) |
| 19, 20 | Grades | The student (IO) |
| 21, 22 | Window | The ball (S) |
| 23, 24 | Window | The window (O) |
| 25, 26 | Catch | The cat (S) |
| 27, 28 | Catch | The mouse (O) |

| Question Numbers | Scenario | Answer |
|---|---|---|
| 29, 30 | Visit | John (S) |
| 31, 32 | Visit | Luke (O) |
| 33, 34 | Hit | Shawn (S) |
| 35, 36 | Hit | Joe (O) |
| 37, 38 | Firing | John (S) |
| 39, 40 | Firing | Dave (O) |

S= Subject, IO= Indirect Object, O= object

Out of the 20 null/overt condition pairs, human participants achieved perfect performance on six pairs, shown below in table 6. The only scenario in which all four questions were answered correctly was the protest scenario (Questions 9-12).

**Table 6. Items on which Humans Achieved Perfect Performance**

| Numbers | Question & Answer |
|---|---|
| 9 Overt<br>10 Null | 民衆は政府に抗議した。(彼らは)飢えていたからだ。飢えていたのは誰ですか？<br>The people protested the government because (they) were hungry. Who was hungry?<br>Answer: The people |
| 11 Overt<br>12 Null | 民衆は政府に抗議した。(彼らは)増税しようとしていたからだ。増税しようとしていたのは誰ですか？<br>The people protested the government because (they) were raising taxes. Who was raising taxes?<br>Answer: The government |
| 15 Overt<br>16 Null | 男は少年にキャンディをあげると言った。しかし(彼は)キャンディをもらわなかった。キャンディをもらわなかったのは誰ですか？<br>The man offered the boy candy, but (he) got no candy. Who got no candy?<br>Answer: The boy |

| Numbers | Question & Answer |
|---|---|
| 19 Overt<br>20 Null | 教授は学生に低い成績をつけた。彼が勉強しなかったからだ。勉強しなかったのは誰ですか？<br>The professor gave the student a bad grade because he did not study. Who did not study?<br>Answer: The student |
| 23 Overt<br>24 Null | ボールは窓ガラスを割った。それが壊れやすかったからだ。壊れやすかったのは何ですか？<br>The ball broke the window because it was fragile. What was fragile?<br>Answer: The window |
| 39 Overt<br>40 Null | ジョンはデイブを首にした。彼は欠勤が多すぎた。欠勤が多すぎたのは誰ですか？<br>John fired Dave because he had too many absences. Who had too many absences?<br>Answer: Dave |

The protest scenario questions (Q9-Q12) are easily answerable by selectional restrictions. The word *hungry* describes the state of a living organism, therefore *the government* is not a viable answer. Also, *raising taxes* is usually an action taken by *the government* and not by *the people*. Q15 and Q16, Q19 and Q20, Q23 and Q24 can be answered with co-occurrence statistics. The act of *getting candy* is more frequently associated with children than adults. *Studying* is more often associated with students than professors and windows are more often described as *fragile* than are balls. Q39 and Q40 are harder to explain, but as we will see in section 5.2, the phrase used in this sentence *kubi ni suru*, ‘to fire’, is particularly strong in Japanese and because of that, the situation may be more salient and easy to parse.

The two questions shown below had considerably lower accuracy in comparison to the overall performance of 0.96. Notably, both are instances where the target referent was in the subject position.

Human Performance: 0.88
Q1: Raise_Subject_Overt
経営者側は従業員に昇給させた。彼らは他の選択肢がなかったからだ。選択肢がなかったのは誰ですか？
Management gave the workers a raise because they had no other choice. Who had no other choice?

Human Performance: 0.85
Q34: Hit_Subject_Null
ショーンはジョーを叩いた。しかし⦸謝らなかった。謝らなかったのは誰ですか？
Shawn hit Joe but (he) did not apologize. Who did not apologize?

While the most likely answer is clearly *management* for Q1 and *Shawn* for Q34, both questions are susceptible to over-thinking or strong associations. In Q1 there may be a strong association between *workers* and *no other choice*, causing participants to mistakenly select *the workers*. Q34 can easily be overthought to imagine a scenario in which the answer is *Joe*.

In general, items in which the target referent was the subject had lower accuracy than the target referents of object or indirect object. As discussed in section ?, Japanese exhibits a strong preference for the subject as the center, which means that the subject of a sentence is usually expressed with a null pronoun, assuming its referent is clear from context. As a result, an overt subject may have encouraged participants to overthink and arrive at the wrong answer. However, with the current data, it is not possible to draw further conclusions about this topic. Deeper analysis of the performance between different syntactic positions is a possible direction for future studies.

ChatGPT had perfect answers on over half of the questions, but there were two questions which it never answered correctly. They were the overt pronoun conditions of one scenario, shown below.

Q37: Firing_Subject_Overt
ジョンはデイブを首にした。彼はあまりに多くの従業員を抱えていた。あまりに多くの従業員を抱えていたのは誰ですか？
John fired Dave because he had too many employees. Who had too many employees?

Q39: Firing_Object_Overt
ジョンはデイブを首にした。彼は欠勤が多すぎた。欠勤が多すぎたのは誰ですか？
John fired Dave because he had too many absences. Who had too many absences?

Strangely, it answered the null pronoun versions of those same two sentences perfectly. Without more data and examples similar to this, it is impossible to determine what may have contributed to these results. Human participants did not show the same pattern and question 39 was one of the questions all participants did perfectly.

Opposite to the human data, ChatGPT had higher performance on pronoun resolution in both conditions when the referent had the grammatical function of subject, than when it was an object or indirect object. These opposite tendencies point to a difference in language understanding or decision making heuristics between humans and ChatGPT. Humans might be overthinking questions in which the subject was the target of pronoun resolution, perhaps due to unnecessary redundancy in the sentence. However, ChatGPT seems to be over-using the heuristic that the most likely referent of the pronoun is the subject of the previous sentence. Take the following set of questions as an illustration of this:

Q21 & 22: Window_Subject
ボールは窓ガラスを割った。(それが)堅かったからだ。堅かったのは何ですか？
The ball broke the window because (it) was hard. What was hard?
Human Accuracy: 0.92
GPT Accuracy: 0.75

Q23 & 24: Window_Object
ボールは窓ガラスを割った。(それが)壊れやすかったからだ。壊れやすかったのは何ですか？
The ball broke the window because (it) was fragile. What was fragile?
Human Accuracy: 1
GPT Accuracy: 0.50

Humans were more accurate on questions 23 and 24 where the answer was *the window*, the object of the previous sentence. However, ChatGPT was more accurate on questions 21 and 22, where the answer was *the ball*, the subject of the sentence.

Both of these sentences have a problem with redundancy. There is no reason to state that the ball was hard after we already know it broke a window. Likewise there is no reason to state that the window was fragile after we already know it broke. As a result, humans may be looking for a reason for the redundancy, causing them to overthink and answer incorrectly. In questions 21 and 22, both windows and balls could theoretically be hard, perhaps leading a person to mistakenly pick *the window*. In questions 23 and 24, however, a fragile ball would not be able to break a window, therefore the answer is easy to determine despite the awkward redundancy.

ChatGPT, on the other hand, may be using a simpler decision heuristic: when there is an ambiguous pronoun, the most likely referent is the subject of the previous sentence. It mistakenly picked the object or indirect object only five times throughout the two runs, but it mistakenly picked the subject ten times.

There wasn't a discernible correlation between the questions that human participants answered perfectly and the ones that ChatGPT answered perfectly. The only questions in both null and overt conditions that humans and ChatGPT got 100 percent accuracy on were questions 9-12. As discussed earlier, these items may have been particularly easy due to selectional restrictions or co-occurrence statistics. However, because there are many other questions in the dataset that have the same flaws, it does not seem to be a predicting factor of ease for ChatGPT.

**5.2 Naturalness**

The overall low rating of naturalness by humans could be due to a number of factors, including redundancy, lack of clarity and rigid syntax. Because the translation process focused mainly on preserving the aspects of the sentences necessary for the WSC, it may have resulted in a lot of syntactic or semantic awkwardness. The translators acknowledge this drawback directly and emphasize that they prioritized whether or not the sentence required the same use of commonsense knowledge in Japanese rather than what the natural Japanese wording would be (Shibata et al. 2015).

My initial hypothesis was that null pronouns would be rated as more natural than overt pronouns due to the frequency of zero anaphora in Japanese. However, the data reveals the opposite: that human participants found overt pronouns to be more natural. The most likely reason for this finding is due to differences in register. While ellipsis may be very common in colloquial language, particularly in speech, it is less common in written language and formal settings (Shibatani 1990). This experiment was conducted in writing and due to the formal survey context, null pronoun sentences may have seemed too informal or out of place. Also, in most cases, null pronouns are used when there is very little or no ambiguity about their referents

(except in cases of intended ambiguity) (Kameyama 1985; Okamoto 1985; Shibatani 1990) Therefore, WSC sentences may not have fulfilled the criteria necessary for a null pronoun environment as they are inherently ambiguous. As a result, the null pronoun condition sentences may have seemed unnatural.

There was one scenario in which all four questions were rated considerably lower than others: Questions 21-24, the broken window scenario, reproduced below for convenience.

> Q21 & 22: Window_Subject
> ボールは窓ガラスを割った。(それが)堅かったからだ。堅かったのは何ですか？
> The ball broke the window because (it) was hard. What was hard?
> Human Naturalness Rating: 2.58 (Overt), 2.48 (Null)
>
> Q23 & 24: Window_Object
> ボールは窓ガラスを割った。(それが)壊れやすかったからだ。壊れやすかったのは何ですか？
> The ball broke the window because (it) was fragile. What was fragile?
> Human Naturalness Rating: 2.38 (Overt), 2.25 (Null)

As previously discussed, these sentences are awkward due to redundancy, but they are also the only questions in which the target referent is an inanimate object. Unlike most of the other questions, which use personal pronouns (*kare*, 'he', *kanojo*, 'she', *karera*, 'they'), this one uses *sore*, 'it'. While personal pronouns may be repeated in formal registers, such as writing, *sore*, 'it', encodes deictic information and is often used as a demonstrative, which means it may be unnatural as there is no physical context. While it can be used in some anaphoric contexts, previous analysis suggests that it is usually used conceptually or in very specific conditions (Hasegawa 2012).

This could explain why the overt conditions of this scenario received such low ratings, and the low ratings of the null pronoun versions can be explained by awkwardness and redundancy.

There may also be a syntactic choice contributing to awkwardness. In these sentences, presumably to stay as faithful to the English as possible, the verb *waru*, 'to break' is in the active voice. However, due to increased animacy constraints discussed in section 3.2.1, the active form of *waru*, may be reserved for animate, perhaps even sentient entities. Of the many verbs that exist for 'break' in Japanese, *waru* often implies intentionality or accident on the part of a breaker (Fujii 1996). In her semantic analysis of five verbs that mean 'to break' in Japanese, Fujii uses the following examples of *waru* to illustrate their central usage:

(8) a. *osara-o otoshite wat-te shimat-ta*
plate-ACC drop break-GRD end.up-PST
'(I) dropped the plate and broke it.'

b. *take -no -bou -de suika -o waru*
bamboo-GEN-stick-INST watermelon-ACC break
'break a watermelon with a bamboo stick'

(8a) is an example of an accidental breaking event and the subject is elided, but it is clear that there is an entity performing the action. Similarly, in (8b), although there is no subject, because the bamboo stick is indicated as an instrument, it is implied that someone must intentionally break the watermelon. In Q21-24, the ball is not described as an instrument but as the subject causing the breaking event, and thus may sound unnatural because it is being given agency when it has none.

While there are some interesting discussions and insights to gain from the analysis of human naturalness ratings, it is difficult to do more than speculate. In the interest of keeping the survey to a reasonable time limit, participants were not asked to provide reasoning for their answers.

In contrast to human judgment, ChatGPT's ratings of naturalness were significantly higher. While not explicitly requested to explain its naturalness decisions, ChatGPT provided short explanations with each naturalness judgment. So, unlike with the human data, we have greater insight into what it prioritized in its decision making and why it assigned the ratings it did. See appendix ? for all of ChatGPT's explanations.

For each question, ChatGPT gave an explanation of whether or not the sentences met certain criteria that it considered when deciding if the sentences were natural. The most often referenced criteria were (1) whether or not the sentences were grammatically correct (grammaticality), (2) easy to understand (ease of understanding) and (3) whether or not the expressions in the second sentences were commonly used as explanations for the situation established in the first sentence (context applicability). Although ChatGPT used these three basic criteria, its evaluation or explanation for each criteria did not always match what a human might think. This was particularly pronounced for criteria 2 and 3. See the following example of a high naturalness rating:

Stimuli:
教授は学生に低い成績をつけた。勉強しなかったからだ。
The professor gave the student a bad grade because (he) did not study.

GPT Response:

この文は文法的にも文脈的にも自然で、一般的な表現です。教授が学生に低い成績をつけた理由として「勉強しなかったからだ」というのは、一般的な説明です。したがって、４ー自然 と評価できるでしょう。
These sentences are grammatically and contextually natural and are common expressions. “Because he didn’t study” is a common explanation for why a professor would give a student a bad grade. Therefore, it can be evaluated as a 4 - natural.

In this example, ChatGPT assigns a high rating and references the sentences’ grammaticality, context and whether or not the second sentence is a commonly expected explanation for the situation. Its explanation is coherent and reasonable. For reference, human participants gave this sentence a naturalness judgment of 3.3.

In the following example, ChatGPT uses the same three criteria to determine naturalness, but gives the stimuli a low rating.

Stimuli:
ショーンはジョーを叩いた。しかし叫び声を上げなかった。
Shawn hit Joe, but (he) did not scream.

GPT Response:
この文は文法的には正しいですが、文脈的にはやや奇妙に感じられるかもしれません。通常、人が叩かれた場合には叫び声を上げることが一般的であるため、その点が驚きや違和感を生む可能性があります。ただし、文全体としては理解可能で、２ー不自然 というよりも、やや奇妙な表現といえるでしょう。
This sentence is grammatically correct, but it may seem a little strange in context. Normally, when a person is hit, it is common for them to scream, so that part may be surprising or unusual. However, the sentence as a whole is understandable. 2-unnatural, but it is an odd expression rather than an unnatural one.

In this example, ChatGPT again touches upon the same three criteria, but assigns a lower rating because the second sentence is an unexpected follow-up to the situation expressed in the first sentence. However, it does acknowledge that this doesn’t necessarily make the sentence “unnatural”, but rather, unusual. In this example, ChatGPT’s explanation for criteria 3 may be

flawed. Humans may not come to the same conclusion that it is common to scream after being hit. For reference, human participants gave this sentence a naturalness judgment of 2.8.

By generally using the same three criteria, ChatGPT seemed to use a methodical approach when assigning naturalness judgments. Items that fulfilled all three were rated as more natural. Notably, ChatGPT never assigned a judgment of 5 (very natural) and only once gave a rating of 1 (very unnatural).

Human participants most likely used a less systematic approach when assigning naturalness judgments. The term "naturalness" was not defined for the participants, so they were either using their own set of criteria or intuition. It seems likely that humans would also consider grammaticality and ease of understanding. However, ChatGPT's third criteria, whether or not the explanation or follow-up sentence is an often used expression for that situation, may be less related to human intuition about naturalness. Out of a total of 1,302 naturalness ratings by human participants, 164 ratings were a 5 (very natural), and 85 ratings were a 1 (very unnatural).

Based on this, we can see that ChatGPT was more moderate and more lenient than human participants and in most cases, humans rated sentences lower than ChatGPT. There were only 9 out of 40 questions in which humans assigned a higher naturalness judgment than ChatGPT and the largest difference was seen in the questions from the same scenario, laid out in the following table.

**Table 7. Questions Rated Higher by Humans than ChatGPT**

| Question Label | Question | Human Rating | ChatGPT Rating |
|---|---|---|---|
| Q37<br>Firing<br>Sub<br>Overt | ジョンはデイブを首にした。彼はあまりに多くの従業員を抱えていた。あまりに多くの従業員を抱えていたのは誰ですか？<br>John fired Dave because he had too many employees. Who had too many employees? | 3.0 | 2.0 |
| Q38<br>Firing<br>Sub<br>Null | ジョンはデイブを首にした。あまりに多くの従業員を抱えていた。あまりに多くの従業員を抱えていたのは誰ですか？<br>John fired Dave because (he) had too many employees. Who had too many employees? | 2.5 | 2.0 |
| Q39<br>Firing<br>Ob<br>Overt | ジョンはデイブを首にした。彼は欠勤が多すぎた。欠勤が多すぎたのは誰ですか？<br>John fired Dave because he had too many absences. Who had too many absences? | 3.8 | 1.5 |
| Q40<br>Firing<br>Ob<br>Null | ジョンはデイブを首にした。欠勤が多すぎた。欠勤が多すぎたのは誰ですか？<br>John fired Dave because (he) had too many absences. Who had too many absences? | 3.4 | 2.5 |

In both runs of the experiment, ChatGPT gave the same reason each time it encountered one of the sentences from this scenario. That reason was the choice of expression: *kubi ni suru* ‘to fire’. In each instance, it pointed out that the expression was too informal, inappropriate or too direct. The term *kubi ni suru* is very colloquial and the word *kubi* literally means ‘neck’. It is theorized that the phrase *kubi ni suru* meaning ‘to fire’ comes from the practice of beheading (Okada 2019). In some cases, ChatGPT suggested replacing it with the more neutral term *kaiko suru* ‘to fire’, which is the more formal, polite term for ‘to fire’.

In the firing scenario case, ChatGPT was very consistent in its judgments and reasoning, and in most cases, it delivered reasonably coherent explanations. However, there were some cases in which it rated the same question very differently and delivered confused, illogical reasoning. See the following two answers it gave when asked to rate Q21 for naturalness.

> **Q21:** ボールは窓ガラスを割った。それが堅かったからだ。堅かったのは何ですか？
> The ball broke the window because it was hard. What was hard?
>
> **Answer 1:** この文章は ２ー不自然、と評価される可能性があります。一般的に、「ボールは窓ガラスを割った。それが堅かったからだ。」という表現は不自然です。ガラスは硬い材料であるため、割れやすい性質を持っています。文脈や意図によっては、適切な表現に修正すると良いでしょう。例えば、「ボールは窓ガラスを割った。それが強化ガラスだったからだ。」などが考えられます。
> This sentence could be evaluated as 2 - unnatural. 'The ball broke the window because it was hard.' is an unnatural expression. Glass is a hard material, so it is easily broken. Depending on the context and intent, it may be good to modify the expression to something more appropriate. For example: "The ball broke the window because it was tempered glass."
>
> **Answer 2:** この文章は比較的自然な表現です。ボールが窓ガラスを割るという状況と、その原因として「それが堅かったからだ」という説明が一貫しています。しかし、より自然にするためには、「それが堅かった」の代わりに「それが頑丈だったからだ」とか「それが割れにくかったからだ」と表現すると、より適切かもしれません。
> 評価: ３ー自然でも不自然でもない
> This sentence is relatively natural. The situation of the ball breaking the window and the explanation "because it was hard" are consistent. However, to make it sound more natural, instead of "it was hard", it might be more fitting to say "because it was sturdy" or " because it was hard to break". Evaluation: 3 - Neither natural nor unnatural.

In both answers, ChatGPT misunderstands very basic information and this is reflected in its low accuracy for the questions in this scenario. In Answer 1, it seems to have a faulty understanding about the properties of glass and the connection between *hard* and *easy to break*.

Answer 2 is more coherent until it suggests the correction of “because it was hard to break”. In both instances, it delivers a naturalness rating, but it either does not resolve the pronoun correctly or is confused by the syntax of the sentence.

Without explanations from human participants on their choice of naturalness rating, we can only speculate about the differences between human and ChatGP reasoning. However, it is clear from the data collected that humans did not find the sentences very natural overall and that ChatGPT does not have human-like intuition about what constitutes natural Japanese.

# 6 General Discussion

This case study of the Japanese WSC shows us that ChatGPT (GPT-3.5) does not achieve human levels on the resolution task, but performs similarly to the XLM models which achieved highest accuracy in Takahashi et al. (2023) (80.85% and 83.51%). If the WSC is indeed an accurate metric of commonsense reasoning and world knowledge, we can conclude that ChatGPT is not capable of human-level thinking. In terms of linguistic processing, ChatGPT shows some similarities to humans and some differences. Like humans, its ability to process a sentence does not seem to be affected by surface features, such as the presence or absence of a pronoun.

The differences between humans and ChatGPT are more pronounced in the naturalness data. Human evaluation revealed that the sentences are not particularly natural Japanese and that overt pronoun condition sentences were rated more natural. As discussed earlier, this could be due to the formal context of the survey. However, ChatGPT's naturalness ratings were higher, and null pronoun versions of sentences were rated more natural. Without information about why human participants rated things the way they did, it is difficult to compare human reasoning directly to ChatGPT.

Despite the small scale of the experimental studies outlined here, the results have been informative. They also illustrate some significant concerns regarding the suitability of the WSC as a measure of commonsense reasoning and world knowledge, and demonstrate the challenges of using a translated test.

### 6.1 The Winograd Schema as a Modern LLM Evaluation Method

The WSC has been used in many forms for benchmarking since 2012. It benefits from having a simple but clever format that is easy to evaluate. Interpreting pronouns is something humans do regularly and subconsciously, so while resolving ambiguous pronouns may be a slight puzzle, it is still quite easy for humans. For NLP, coreference resolution used to be a challenge and a format such as the WSC would have been near impossible twenty years ago. However, with modern LLMs, that is no longer the case and the WSC is a test of commonsense reasoning and world knowledge usage rather than a simple linguistic task.

As a test designed for that purpose, the WSC has the advantage that its format can cover a wide range of subjects. The sentences can be very simple and target one specific feature, or complex and require deeper reasoning. The skills necessary to solve a WSC problem can also be said to emulate a human's ability to simultaneously utilize several different domains of knowledge.

Also, compared with other benchmark tests which have been more easily solved by computer systems, the WSC has remained challenging, even considered "problematic" by the researchers who developed BERT (Devlin et al. 2018). In 2020, with the release of GPT-3, the state of the art (SOTA) for the English WSC273 dataset was 90.1 on a fine-tuned model, 2% lower than the human baseline (Sakaguchi et al. 2021). The WLNI dataset used in SuperGLUE has a human baseline of 95.9% and this has yet to be reached by an LLM (Nangia & Bowman 2019). The human baseline of 94.0% for the larger, more challenging WinoGrande dataset has also not been reached (Sakaguchi et al. 2021). The human baseline collected by the same authors for the English WSCR set, 95.2% accuracy, has also yet to be surpassed by LLMs.

Despite these advantages, the WSC and its translations have many drawbacks that should be discussed in order to consider how to improve upon its current form. They fall into the three categories of challenges discussed in section 2.2.1: contamination, saturation and suitability. In the following sections, I will outline how the WSC and Japanese WSC suffer from all of these problems and discuss possible solutions.

### *6.1.1 Contamination*

Contamination refers to systems being exposed to the benchmark data during their pre-training. Brown et al. (2020) found that 132 original WSC questions were in the training data for GPT-3 and that performance on a subset of clean questions dropped by 2.6%. This is probably the case for other LLMs as well, including GPT-3.5 and GPT-4. Unfortunately, this makes the WSC an unsuitable test in its current form for modern LLMs. While we do not know if the Japanese WSC shows the same signs of contamination, it is likely the case as the set is publicly available on the internet. Ahuja et al. (2023) determined the level of contamination by three metrics: i) LLM's knowledge of a dataset, ii) availability of test datasets on the internet, and iii) dataset release date. If the dataset was freely available and released before the LLM was trained, it was likely to be used in pre-training data. The authors determined an LLM's level of knowledge about the dataset by having it fill in the dataset card. In instances where the LLM was able to complete the card fill, the dataset was noted as suspicious of contamination. In cases where it filled the card partially, it was considered partial contamination. There is no similar dataset card for the WSCR-ja, but it was released online in 2019 and GPT-3.5 was released in 2023. Without direct information about GPT 3-5's training data, it is hard to say, but there is a significant likelihood that the set was seen in training.

Contamination is a difficult problem and there is probably no fix for datasets that already suffer from it without specifically removing them from training data as done in Brown et al. (2020) for certain tests. However, as acknowledged by the authors, this is a costly process and requires LLM developers to account for possible benchmark tests beforehand. This leaves the only solution of creating new tests, but to prevent them from becoming contaminated quickly, some precautions need to be taken. Chandran et al. (2024) suggests implementing private benchmarking, or at the very least, making datasets password protected to avoid being picked up by automated crawlers. This may be a solution for the future of the WSC, but still necessitates creation of a new dataset.

### *6.1.2 Saturation*

Saturation refers to the time it takes models to reach or surpass human performance on benchmark tests. Unlike other benchmark tasks such as the ones included in GLUE and SuperGLUE, out-of-the-box LLMs have not surpassed human baselines on the WSC (Nangia & Bowman 2019). However, models that have been fine-tuned on the WSCR or WinoGrande set have been very successful and many have achieved near-human performance (Kocijan et al. 2023). This means that despite ChatGPT still lagging behind human performance for now, it is eventually likely to improve to a point where it can match or exceed human ability.

The appeal of many modern benchmark tests are their large datasets and ability to deliver many points of evaluation data. The original WSC only consisted of 150 problems. Large WSC sets used nowadays, such as the WSCR set, WinoGrande, and others, consist of thousands of test questions. However, due to the increased labor and cognitive cost of producing so many test questions, the larger datasets are of much lower quality. Due to the quick saturation of modern

benchmark tests, perhaps it is time to start valuing quality and depth over quantity. A few, well constructed WSC questions with an in-depth analysis of the reasoning an LLM used to solve it may soon be more valuable for understanding LLM performance than a number expressing accuracy on several thousand examples of inconsistent quality.

Because WSC datasets are particularly time and labor intensive to produce and examples are not often found in a usable form in already existing materials, the problem of saturation is particularly difficult to fix. One suggested solution to the problem of saturation is Dynabench (Kiela et al. 2021), a dynamic, adversarial benchmark testing system which involves crowdworkers or experts interacting with a system. However, it does not seem feasible to maintain the pattern and requirements of the WSC in this kind of format, especially considering the unnatural wording of most WSC sentences.

Instead, there may be a way to still test a system's use of commonsense and world knowledge through pronoun resolution which is more natural and less costly to create. This would be to use whole dialogues or passages as test questions rather than manually create examples. Of course, it is still necessary to select *google-proof* situations and ensure that a human would not be fooled by the question, but this could cut down on the labor necessary to create test questions and ensure that it is really "natural language" that the system is being tested on.

### *6.1.3 Suitability*

Suitability refers to how well a benchmark task delivers information on the target of evaluation. In 2021, a RoBERTa model fine-tuned on the WinoGrande set achieved 90.1% accuracy on the WSC273 set, and 93.1% on the WSCR set (Sakaguchi et al. 2021). We could

interpret this to mean that modern LLMs are getting close to human commonsense reasoning ability. However, as observed by Poliak et al. (2018) and Parmar et al. (2023), it can be difficult to discern if a model is really using the targeted skill to complete a benchmark task or if they are exploiting statistical trends or patterns in the dataset. The WSC suffers from this problem as well, especially sets developed with relaxed restrictions on the *google-proof* rule, such as the WSCR set.

With this in mind, I would argue that the WSC in its current form is not informative enough to draw conclusions about an LLM's ability or inability to use commonsense reasoning and world knowledge. First, in its current form, we have no insights into *how* the model came to its answer. Kocijan et al. (2023) encourages caution when using proxy problems to evaluate unmeasurable traits like "commonsense reasoning". They emphasize that all tests are essentially proxies, meaning that tests usually only validate one part of the target skill or task. All benchmark tests will suffer from this problem to varying degrees and the goal of a researcher developing a benchmark test should be to make the task as close to representative of the unmeasurable skill as possible. Based on its original intention, the WSC is a proxy problem for two things. Correctly solving problems would mean that the system is able to perform commonsense reasoning based on world knowledge, and that it is able to act intelligently similar to a human. However, the WSC is too narrow a test to be used to make such assumptions. This draws into question the suitability of the WSC in its current form as a measure of commonsense reasoning.

Second, "world knowledge" is a very broad, problematic term. Multiple researchers have pointed out that WSC problems draw on different types of knowledge, and also require different

types of reasoning such as physical, spatial, interpersonal and social reasoning. Kocijan et al. (2023) identified a number of complex knowledge types necessary for solving different WSC problems, including spatial and physical knowledge, knowledge of human interactions and emotions, understanding of typical and atypical human behavior, etc.

During a review of the original WSC dataset, it became clear that there are multiple types of commonsense reasoning necessary to resolve different questions. The preliminary knowledge categories I found present in the original English dataset were: spatial, temporal, properties/qualities, human interactions and emotions, semantic frames and lexical. In some informal testing, ChatGPT (GPT-3) showed inconsistent accuracy across the different categories.

During their translation of the WSCR set to Japanese, the researchers surveyed and categorized 100 questions, acknowledging that the problems in the dataset are solvable by different types of knowledge. They identified that some are resolvable by selectional criteria, (i.e. the ability to bark differentiating cats and dogs). There were also questions which could be answered by knowledge of semantic frames and event knowledge (i.e. bankruptcy leading to the acquisition of company X by company Y), and others answerable by understanding cause and effect type relations (termed “meta-knowledge”, i.e. A doing something good for B leads to B thanking A), and many that fall into multiple categories.

Kocijan et al. (2023) sums up the problem of suitability succinctly:

> *The goal of the WSC was to present a challenge that (a) was clearly “commonsensical" and easily carried out by humans; (b) was easy to evaluate; (c) was an adequate test of commonsense reasoning abilities. It succeeded at (a) and (b), but not at (c).*

Not only is it questionable if the WSC is a good proxy test for commonsense reasoning and world knowledge, but the terms themselves are too broad and poorly defined for the test to indicate intelligence. The WSC could benefit from a better definition of terms, or by getting more specific and granular about the types of reasoning and knowledge it tests. One potential solution to the suitability problem would be to focus on elicited explanations of the LLM's reasoning for why it resolved a pronoun the way it did. While not perfect, this would allow the researchers to determine what kind of reasoning the system was using.

One attempt at this was the WinoWhy test, a version of the WSC that focused on a model's ability to identify valid logical reasoning that could be used to arrive at the answer to a WSC question (Zhang et al. 2020). Researchers collected reasoning explanations from human annotators and used those to make a multiple choice task. A WinoWhy question consists of the WSC question, its correct answer and explanations of reasoning to arrive at the answer. Five explanations were generated by humans and are considered valid. Five were the reverse of the human explanations and are considered invalid. Five were generated by GPT-2 and were annotated by humans to determine if they were valid or invalid. To perform successfully on a WinoWhy question, an LLM must select one of the valid explanations. This style of test focuses on the model's ability to demonstrate understanding of logical reasoning, therefore good performance on a WinoWhy question is more indicative of commonsense reasoning than the basic WSC task alone.

The WinoWhy creators also categorized WSC questions by the type of knowledge required to answer it. They used conceptual semantic theory and ConceptNet (Speer et al. 2017) to distill six categories: Property, Object, Eventuality, Spatial, Quantity and Others. See figure ?

for their definitions of those categories. Then, analyzing their crowdsourced human reasons, they sorted WSC questions into the categories. By categorizing types of knowledge, they were able to gather more informative data about model ability. In their tests on the basic WSC pronoun resolution task, they found the highest performance overall was by large RoBERTa fine-tuned on the WinoGrande set at 87.55%. The best performance of that model was on questions from the eventuality category (93.18%) and the worst performance was on questions from the object category (84.15%).

**Figure 8. WinoWhy Knowledge Categories**

| Name | Definition | Example |
|---|---|---|
| Property | Knowledge about property of objects. | ice is cold. |
| Object | Knowledge about objects. | cats have ears. |
| Eventuality | Knowledge about eventualities. | 'wake up' happens before 'open eyes'. |
| Spatial | Knowledge about spatial position. | object at the back can be blocked. |
| Quantity | Knowledge about numbers. | 2 is smaller than 10. |
| Others | All other knowledge. | NA |

*Figure 8. Knowledge categories and their definitions from Zhang et. Al (2020)*

From the researchers' data on the pronoun resolution task alone, it would appear that models are quite capable when fine-tuned. However, performance on the WinoWhy task, which

tested for reasoning identification, was much less promising. Again, the best performance was by the WinoGrande fine-tuned large RoBERTa, but it only achieved a 58.18%. Even though it was fine-tuned for the WSC task, it is clear that the model was not able to reliably determine valid reasoning, which suggests that it is not using commonsense reasoning to solve the WSC problems.

To address the issue of the WSC's suitability as a test for commonsense reasoning and world knowledge, future work could use the categorizations laid out in WinoWhy, and adopt a similar design. However, as demonstrated in this study, modern LLMs are capable of simply providing their own explanation of their reasoning. With a well designed set of criteria and scale, those passages could be evaluated for validity by human annotators. Alternatively, human participants could provide their own explanations and the LLM generated explanations could then be directly compared to those of the humans and measured for similarity.

Suitability is also an issue for translated benchmark tests. Translation challenges and differences in linguistic typology make some tests less appropriate for certain languages. In those cases, the tests should be modified to work with the typology of the language or a different format altogether should be used. For the Japanese WSC, pronoun resolution or zero anaphora resolution is a reasonable test, but as demonstrated in the discussion of naturalness from section 5.2, increased semantic specificity makes certain types of questions unideal. Also, while the WSC has a rigid format and formal context, human interactions with LLMs may be quite informal and conversational. For that reason, an important goal of a benchmark test for Japanese would be to understand how well a system deals with natural dialogue settings. This is true for any language in which conventions differ significantly based on register. To address this, one

adjustment to the WSC could be to embed the question set-up in a dialogue rather than a single sentence or two sentences.

Ultimately, the goal of any benchmark test developer should be to make the task as representative of the targeted skill as possible. If the skill is commonsense reasoning, then it makes more sense to directly examine reasoning rather than accuracy on a proxy task. This addresses the issue of suitability and also helps to solve the problem of contamination and saturation as well. Also, in general, it would be better to avoid direct translation and use naturally occurring language in order to make tests more appropriate to the target language of evaluation.

# 7 Conclusion

This work contributes to the growing field of research directly comparing humans and LLMs on reasoning and linguistic tasks. This type of research is essential as LLMs become more widely used in a variety of contexts that increasingly affect the way we get information and make decisions. Also, in order to avoid increasing the disparity in the distribution of language technology and make advances more accessible around the world, it is imperative to conduct more research in non-English languages.

The experimental study results reported here show that ChatGPT (GPT-3.5) still underperforms compared to human participants on the Japanese WSC and the surface feature of overt vs. null pronouns does not affect interpretation. However, there are considerable differences in naturalness judgments. The most striking was that humans and ChatGPT showed opposite tendencies in which condition (overt pronoun vs. null pronoun) they rated as more natural. Humans gave overt pronoun sentences higher naturalness ratings while ChatGPT rated null pronouns as more natural. This data suggests that ChatGPT does not have a human-like understanding of what constitutes natural Japanese. The preliminary human baseline established in this pilot study can hopefully serve future research on the Japanese WSC and the naturalness data draws attention to issues with translated benchmark tests.

Impressive advances in LLM technology call for better suited benchmark tests and more direct comparison between human and LLM reasoning in order to see where they align and where they diverge. Deeper, more granular analysis regarding specific domains of knowledge and reasoning can lead to a better understanding of where LLM flaws might lie, which would

affect what kinds of decisions and information we can trust from LLM systems. When developing benchmark tests, linguistic differences can have a significant effect on format and content. Therefore, direct translation should be avoided and natural constructions of the target language should be considered carefully in order to produce a suitable test that delivers informative results.

Language technology will continue advancing and evaluation methods need to keep up. Moving forward, human data is also essential to use in direct comparison with LLMs. This is not only true for English, but for all languages. Updated LLM benchmark tests will need to take steps to avoid contamination and quick saturation.They also need to be developed thoughtfully and in good quality to ensure that they are well suited to the evaluation targets. This case study of the Japanese Winograd Schema Challenge, and the high-level analytical discussion presented can inform future evolutions and ensure that the test can develop in a way that will make it useful and informative for the modern field of LLMs and benchmarking.

**References**

Abdou, M., Ravishankar, V., Barrett, M., Belinkov, Y., Elliott, D., & Søgaard, A. (2020). The sensitivity of language models and humans to Winograd schema perturbations. *arXiv preprint arXiv:2005.01348*.

Ahuja, K., Hada, R., Ochieng, M., Jain, P., Diddee, H., Maina, S., ... & Sitaram, S. (2023). Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Amsili, P., & Seminck, O. (2017, April). A Google-proof collection of French Winograd schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017), co-located with EACL 2017* (pp. 24-29).

Bender, D. (2015). Establishing a Human Baseline for the Winograd Schema Challenge. In *MAICS* (pp. 39-45).

Bentivogli, L., Clark, P., Dagan, I., & Giampiccolo, D. (2009). The Fifth PASCAL Recognizing Textual Entailment Challenge. *TAC*, *7*, 8.

Bernard, T., & Han, T. (2020, May). Mandarinograd: A Chinese collection of Winograd schemas. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 21-26).

Blasi, D., Anastasopoulos, A., & Neubig, G. (2021). Systematic Inequalities in Language Technology Performance across the World's Languages. *arXiv preprint arXiv:2110.06733*.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Bowman, S. R., & Dahl, G. E. (2021). What will it take to fix benchmarking in natural language understanding?. *arXiv preprint arXiv:2104.02145*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Cai, Z. G., Haslett, D. A., Duan, X., Wang, S., & Pickering, M. J. (2023). Does ChatGPT resemble humans in language use?. *arXiv preprint arXiv:2303.08014*.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., & Gonzalez, J. E. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/

Chandran, N., Sitaram, S., Gupta, D., Sharma, R., Mittal, K., & Swaminathan, M. (2024). Private Benchmarking to Prevent Contamination and Improve Comparative Evaluation of LLMs. *arXiv preprint arXiv:2403.00393*.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W. T., Choi, Y., ... & Zettlemoyer, L. (2018). QuAC: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, *24*(240), 1-113.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., & Palomaki, J. (2020). Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, *8*, 454-470.

Collins, K. M., Wong, C., Feng, J., Wei, M., & Tenenbaum, J. B. (2022). Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. *arXiv preprint arXiv:2205.05718*.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. *Advances in neural information processing systems*, *32*.

Dagan, I., Glickman, O., & Magnini, B. (2005, April). The pascal recognising textual entailment challenge. In *Machine learning challenges workshop* (pp. 177-190). Berlin, Heidelberg: Springer Berlin Heidelberg.

Davis, E. (2018, May 4). *Collection of Winograd Schemas*. The Winograd Schema Challenge. https://cs.nyu.edu/~davise/papers/WinogradSchemas/WSCollection.html

de Melo, G., Imaizumi, V., & Cozman, F. (2019, October). Winograd schemas in portuguese. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional* (pp. 787-798). SBC.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., & Gardner, M. (2019). DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Frohberg, J., & Binder, F. (2021). Crass: A novel data set and benchmark to test counterfactual reasoning of large language models. *arXiv preprint arXiv:2112.11941*.

藤井. (1997). Cognitive structures of the verbs of'breaking'and the classifiers in Japanese. *放送大学研究年報= Journal of the University of the Air*, *14*, 111-135.

Grosz, B., Joshi, A., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Haim, R. B., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., & Szpektor, I. (2006, April). The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment* (Vol. 7, pp. 785-794).

Ham, J., Choe, Y. J., Park, K., Choi, I., & Soh, H. (2020). KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. arXiv preprint arXiv:2004.03289.

Hasegawa, Y. (2012). Deictic and anaphoric uses of the Japanese demonstratives ko-so-a. *Journal of Japanese Linguistics*, *28*(1), 44-61.

Hashimoto, R., Takeshita, M., Rzepka, R., Araki, K. (2023, March). Development of Japanese WSC273 Winograd Schema Challenge Dataset and Comparison between Japanese and English BERT Baselines. In *Proceedings of the Language Technology Confernece (LTC'23)* (pp. 91-95).

He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020, November). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning* (pp. 4411-4421). PMLR.

Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., ... & Lenci, A. (2023). Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, *47*(11), e13386.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., ... & Williams, A. (2021). Dynabench: Rethinking benchmarking in NLP. *arXiv preprint arXiv:2104.14337*.

Kocijan, V., Cretu, A. M., Camburu, O. M., Yordanov, Y., & Lukasiewicz, T. (2019). A surprisingly robust trick for winograd schema challenge. *arXiv preprint arXiv:1905.06290*.

Kocijan, V., Davis, E., Lukasiewicz, T., Marcus, G., & Morgenstern, L. (2023). The Defeat of the Winograd Schema Challenge. *Artificial Intelligence*. 325. https://doi.org/10.1016/j.artint.2023.103971

Konno, R., Kiyono, S., Matsubayashi, Y., Ouchi, H., & Inui, K. (2021). Pseudo zero pronoun resolution improves zero anaphora resolution. *arXiv preprint arXiv:2104.07425*.

Kurihara, K., Kawahara, D., & Shibata, T. (2022, June). JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 2957-2966).

Kuroda, S. Y. (1965). *Generative grammatical studies in the Japanese language* (Doctoral dissertation, Massachusetts Institute of Technology).

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Lambert, N., Castricato, L., von Werra, L., & Havrilla, A. (2022). Illustrating Reinforcement Learning from Human Feedback (RLHF). Hugging Face Blog. https://huggingface.co/blog/rlhf

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., ... & Schwab, D. (2019). Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*.

Levesque, H., Davis, E., & Morgenstern, L. (2012, May). The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., ... & Zhou, M. (2020). XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Mars, M. (2022). From word embeddings to pre-trained language models: A state-of-the-art walkthrough. *Applied Sciences*, *12*(17), 8805.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Morgenstern, L., Davis, E., & Ortiz, C. L. (2016). Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*, *37*(1), 50-54.

Nakamura, T., & Kawahara, D. (2018, May). JFCKB: Japanese feature change knowledge base. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Nangia, N., & Bowman, S. R. (2019). Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. *arXiv preprint arXiv:1905.10425*.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2019). Adversarial NLI: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Okada, S. (2019). Nominal Conceptual Expansions in Predicational. *Linguistics*, *19*, 125-149.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730-27744.

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., ... & Cho, K. (2021). Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Parmar, M., Mishra, S., Geva, M., & Baral, C. (2022). Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions. *arXiv preprint arXiv:2205.00415*.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Van Durme, B. (2018). Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1-67.

Rahman, A., & Ng, V. (2012, July). 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* (pp. 777-789).

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*.

Reddy, S., Chen, D., & Manning, C. D. (2019). Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, *7*, 249-266.

Roemmele, M., Bejan, C. A., & Gordon, A. S. (2011, March). Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Rybak, P., Mroczkowski, R., Tracz, J., & Gawlik, I. (2020). KLEJ: Comprehensive benchmark for Polish language understanding. *arXiv preprint arXiv:2005.00630*.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., ... & Rush, A. M. (2021). Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., & Choi, Y. (2021). Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, *64*(9), 99-106.

Sasano, R., & Kurohashi, S. (2011, November). A discriminative approach to japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proceedings of 5th International Joint Conference on Natural Language Processing* (pp. 758-766).

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... & Al-Shaibani, M. S. (2022). Bloom: A 176b-parameter open-access multilingual language model.

Shavrina, T., Fenogenova, A., Emelyanov, A., Shevelev, D., Artemova, E., Malykh, V., ... & Evlampiev, A. (2020). RussianSuperGLUE: A Russian language understanding evaluation benchmark. *arXiv preprint arXiv:2010.15925*.

柴田知秀, 小浜翔太郎, 黒橋禎夫: 日本語 Winograd Schema Challenge の構築と分析、言語処理学会 第 21 回年次大会 (2015.3).

柴田知秀. (2015). 自然言語処理技術の現状と展望-エラー分析プロジェクトを通して-:[基礎技術] 3.7 知識獲得. *情報処理*, *57*(1), 22-23.

Shibatani, M. (1990). *The languages of Japan*. Cambridge University Press.

Sidner, C. L. (1979). *Towards a computational theory of definite anaphora comprehension in English discourse* (Doctoral dissertation, Massachusetts Institute of Technology).

Sidner, C. L. (1983). Focusing and discourse. *Discourse Processes*, *6*(2), 107-130.

Speer, R., Chin, J., & Havasi, C. (2017, February). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1).

Takahashi, K., Oka, T., & Komachi, M. (2023). Effectiveness of Pre-Trained Language Models for the Japanese Winograd Schema Challenge. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *27*(3), 511-521.

Tanaka, S., Rzepka, R., Kitajima, S. (2014). The Winograd Schema Challenge. http://arakilab.media.eng.hokudai.ac.jp/~kabura/collection_katakana.html

Tikhonov, A., & Ryabinin, M. (2021). It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. *arXiv preprint arXiv:2106.12066*.

Ueda, N., Kawahara, D., & Kurohashi, S. (2020, December). BERT-based cohesion analysis of Japanese texts. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1323-1333).

Ueyama, A., & Kano, Y. (2023, September). Dialogue Response Generation Using Completion of Omitted Predicate Arguments Based on Zero Anaphora Resolution. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 282-296).

Urbizu, G., San Vicente, I., Saralegi, X., Agerri, R., & Soroa, A. (2022, June). BasqueGLUE: A natural language understanding benchmark for Basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 1603-1612).

Vadász, N., & Ligeti-Nagy, N. (2022). Winograd schemata and other datasets for anaphora resolution in Hungarian. *Acta Linguistica Academica*, *69*(4), 564-580.

Valmeekam, K., Olmo, A., Sreedharan, S., & Kambhampati, S. (2022). Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change). *arXiv preprint arXiv:2206.10498*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, *32*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wei, C., Wang, Y. C., Wang, B., & Kuo, C. C. J. (2023). An overview on language models: Recent developments and outlook. *arXiv preprint arXiv:2303.05759*.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., ... & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. *arXiv preprint arXiv:2009.05387*.

Williams, A., Nangia, N., & Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., ... & Lan, Z. (2020). CLUE: A Chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Yanaka, H., & Mineshima, K. (2022). Compositional evaluation on Japanese textual entailment and similarity. *Transactions of the Association for Computational Linguistics*, *10*, 1266-1284.

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., ... & Huang, X. (2023). A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Yoshikoshi, T., Kawahara, D., & Kurohashi, S. (2020). Multilingualization of a natural language inference dataset using machine translation. *情報処理学会研究報告 (Web)*, *2020*(NL-244).

Zhang, S., Liu, X., Liu, J., Gao, J., Duh, K., & Van Durme, B. (2018). Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

# Appendices

## Appendix A: Full List of Stimuli Used in Task 1 and 2

| Question | Answer |
|---|---|
| 経営者側は従業員に昇給させた。彼らは他の選択肢がなかったからだ。選択肢がなかったのは誰ですか？<br>Management gave the workers a raise because they had no other choice. Who had no other choice? | 経営者側<br>management |
| 経営者側は従業員に昇給させた。他の選択肢がなかったからだ。他の選択肢がなかったのは誰ですか？<br>Management gave the workers a raise because (they) had no other choice. Who had no other choice? | 経営者側<br>management |
| 経営者側は従業員に昇給させた。彼らはストライキをすると脅したからだ。ストライキをすると脅したのは誰ですか？<br>Management gave the workers a raise because they had threatened to strike. Who had threatened to strike? | 従業員<br>the workers |
| 経営者側は従業員に昇給させた。ストライキをすると脅したからだ。ストライキをすると脅したのは誰ですか？<br>Management gave the workers a raise because (they) had threatened to strike. Who had threatened to strike? | 従業員<br>the workers |
| ジョンはビルに聞いた。彼は頭がこんがらがってしまったからだ。頭がこんがらがってしまったのは誰ですか？<br>John asked Bill because he was confused. Who was confused? | ジョン<br>John |
| ジョンはビルに聞いた。頭がこんがらがってしまったからだ。頭がこんがらがってしまったのは誰ですか？<br>John asked Bill because (he) was confused. Who was confused? | ジョン<br>John |
| ジョンはビルに聞いた。彼が答えを知っていたからだ。答えを知っていたのは誰ですか？<br>John asked Bill because he knew the answer. Who knew the answer? | ビル<br>Bill |
| ジョンはビルに聞いた。答えを知っていたからだ。答えを知っていたのは誰ですか？<br>John asked Bill because (he) knew the answer. Who knew the answer? | ビル<br>Bill |
| 民衆は政府に抗議した。彼らは飢えていたからだ。飢えていたのは誰ですか？<br>The people protested the government because they were hungry. Who was hungry? | 民衆<br>the people |
| 民衆は政府に抗議した。飢えていたからだ。飢えていたのは誰ですか？ | 民衆<br>the people |

| Question | Answer |
|---|---|
| The people protested the government because (they) were hungry. Who was hungry? | |
| 民衆は政府に抗議した。彼らは増税しようとしていたからだ。増税しようとしていたのは誰ですか？<br>The people protested the government because they were raising taxes. Who was raising taxes? | 政府<br>the government |
| 民衆は政府に抗議した。増税しようとしていたからだ。増税しようとしていたのは誰ですか？<br>The people protested the government because (they) were raising taxes. Who was raising taxes? | 政府<br>the government |
| 男は少年にキャンディをあげると言った。しかし彼はキャンディをもっていなかった。キャンディをもっていなかったのは誰ですか？<br>The man offered the boy candy, but he had no candy. Who had no candy? | 男<br>the man |
| 男は少年にキャンディをあげると言った。しかしキャンディをもっていなかった。キャンディをもっていなかったのは誰ですか？<br>The man offered the boy candy, but (he) had no candy. Who had no candy? | 男<br>the man |
| 男は少年にキャンディをあげると言った。しかし彼はキャンディをもらわなかった。キャンディをもらわなかったのは誰ですか？<br>The man offered the boy candy, but he got no candy. Who got no candy? | 少年<br>the boy |
| 男は少年にキャンディをあげると言った。しかしキャンディをもらわなかった。キャンディをもらわなかったのは誰ですか？<br>The man offered the boy candy, but (he) got no candy. Who got no candy? | 少年<br>the boy |
| 教授は学生に低い成績をつけた。彼は厳格だったからだ。厳格だったのは誰ですか？<br>The professor gave the student a bad grade because he was strict. Who was strict? | 教授<br>the professor |
| 教授は学生に低い成績をつけた。厳格だったからだ。厳格だったのは誰ですか？<br>The professor gave the student a bad grade because (he) was strict. Who was strict? | 教授<br>the professor |
| 教授は学生に低い成績をつけた。彼が勉強しなかったからだ。勉強しなかったのは誰ですか？<br>The professor gave the student a bad grade because he did not study. Who did not study? | 学生<br>the student |
| 教授は学生に低い成績をつけた。勉強しなかったからだ。勉強しなかったのは誰ですか？<br>The professor gave the student a bad grade because (he) did not study. Who did not study? | 学生<br>the student |

| Question | Answer |
|---|---|
| ボールは窓ガラスを割った。それが堅かったからだ。堅かったのは何ですか？<br>The ball broke the window because it was hard. What was hard? | ボール<br>the ball |
| ボールは窓ガラスを割った。堅かったからだ。堅かったのは何ですか？<br>The ball broke the window because (it) was hard. What was hard? | ボール<br>the ball |
| ボールは窓ガラスを割った。それが壊れやすかったからだ。壊れやすかったのは何ですか？<br>The ball broke the window because it was fragile. What was fragile? | 窓ガラス<br>the window |
| ボールは窓ガラスを割った。壊れやすかったからだ。壊れやすかったのは何ですか？<br>The ball broke the window because (it) was fragile. What was fragile? | 窓ガラス<br>the window |
| ネコはネズミを捕まえた。それは賢かったから。賢かったのは何ですか？<br>The cat caught the mouse because it was clever. What was clever? | ネコ<br>the cat |
| ネコはネズミを捕まえた。賢かったから。賢かったのは何ですか？<br>The cat caught the mouse because (it) was clever. What was clever? | ネコ<br>the cat |
| ネコはネズミを捕まえた。それは油断していたからだ。油断していたのは何ですか？<br>The cat caught the mouse because it was careless. What was careless? | ネズミ<br>the mouse |
| ネコはネズミを捕まえた。油断していたからだ。油断していたのは何ですか？<br>The cat caught the mouse because (it) was careless. What was careless? | ネズミ<br>the mouse |
| ジョンは病院にルークを見舞った。彼は近くに住んでいたからだ。近くに住んでいたのは誰ですか？<br>John visited Luke in the hospital because he lived close by. Who lived close by? | ジョン<br>John |
| ジョンは病院にルークを見舞った。近くに住んでいたからだ。近くに住んでいたのは誰ですか？<br>John visited Luke in the hospital because (he) lived close by. Who lived close by? | ジョン<br>John |
| ジョンは病院にルークを見舞った。彼は病気だったのだ。病気だったのは誰ですか？<br>John visited Luke in the hospital because he was sick. Who was sick? | ルーク<br>Luke |
| ジョンは病院にルークを見舞った。病気だったのだ。病気だったのは誰ですか？<br>John visited Luke in the hospital because (he) was sick. Who was sick? | ルーク<br>Luke |

| Question | Answer |
|---|---|
| ショーンはジョーを叩いた。しかし彼は謝らなかった。謝らなかったのは誰ですか？<br>Shawn hit Joe but he did not apologize. Who did not apologize? | ショーン<br>Shawn |
| ショーンはジョーを叩いた。しかし謝らなかった。謝らなかったのは誰ですか？<br>Shawn hit Joe but (he) did not apologize. Who did not apologize? | ショーン<br>Shawn |
| ショーンはジョーを叩いた。しかし彼は叫び声を上げなかった。叫び声を上げなかったのは誰ですか？<br>Shawn hit Joe but he did not scream. Who did not scream? | ジョー<br>Joe |
| ショーンはジョーを叩いた。しかし叫び声を上げなかった。叫び声を上げなかったのは誰ですか？<br>Shawn hit Joe but (he) did not scream. Who did not scream? | ジョー<br>Joe |
| ジョンはデイブを首にした。彼はあまりに多くの従業員を抱えていた。あまりに多くの従業員を抱えていたのは誰ですか？<br>John fired Dave because he had too many employees. Who had too many employees? | ジョン<br>John |
| ジョンはデイブを首にした。あまりに多くの従業員を抱えていた。あまりに多くの従業員を抱えていたのは誰ですか？<br>John fired Dave because (he) had too many employees. Who had too many employees? | ジョン<br>John |
| ジョンはデイブを首にした。彼は欠勤が多すぎた。欠勤が多すぎたのは誰ですか？<br>John fired Dave because he had too many absences. Who had too many absences? | デイブ<br>Dave |
| ジョンはデイブを首にした。欠勤が多すぎた。欠勤が多すぎたのは誰ですか？<br>John fired Dave because (he) had too many absences. Who had too many absences? | デイブ<br>Dave |

## Appendix B: ChatGPT Naturalness Explanations

**Q1 Raise_Sub_Overt**

**Question:** 経営者側は従業員に昇給させた。彼らは他の選択肢がなかったからだ。選択肢がなかったのは誰ですか？
Management gave the workers a raise because they had no other choice. Who had no other choice?

**Answer 1:** この文章は４ー自然、と言えるでしょう。文脈において理解しやすく、一般的な表現です。

**Answer 2:** この文章は自然で理解しやすい表現です。文脈による説明があり、経営者が従業員に昇給させた理由が明確に示されています。
評価: ４ー自然

**Q2 Raise_Sub_Null**

**Question:** 経営者側は従業員に昇給させた。他の選択肢がなかったからだ。他の選択肢がなかったのは誰ですか？
Management gave the workers a raise because (they) had no other choice. Who had no other choice?

**Answer 1:** この文は比較的自然であり、理解しやすい表現です。文脈において「他の選択肢がなかったから」という説明が論理的で、一般的な表現とも言えます。したがって、４ー自然 と評価できるでしょう。
**Answer 2:** この文は文法的に正確で、意味も明確です。文の流れも自然であり、状況を十分に説明しています。総合的には「4ー自然」と評価できるでしょう。

**Q3 Raise_IO_Overt**

**Question:** 経営者側は従業員に昇給させた。彼らはストライキをすると脅したからだ。ストライキをすると脅したのは誰ですか？
Management gave the workers a raise because they had threatened to strike. Who had threatened to strike?

**Answer 1:** この文章は ４ー自然、と言えるでしょう。文脈において理解しやすく、一般的な表現です。
**Answer 2:** この文は文脈に対して理解しやすいものの、やや不自然な印象があります。ストライキをすると脅すという表現が強調されており、より穏やかな表現が適している可能性があります。例えば、「経営者側は従業員に昇給させたが、それに対してストライキをする可能性があるという懸念があったからだ。」というように表現すると、より自然かつ適切な表現になります。
評価: ３ー自然でも不自然でもない

**Q4 Raise_IO_Null**

**Question:** 経営者側は従業員に昇給させた。ストライキをすると脅したからだ。ストライキをすると脅したのは誰ですか？
Management gave the workers a raise because (they) had threatened to strike. Who had threatened to strike?

**Answer 1:** この文も比較的自然で、文脈から意味が理解しやすい表現です。ただし、ストライキを使って昇給を要求するのは一般的でないため、やや異例な状況と言えますが、文全体としては理解可能です。総合的には、３ー自然でも不自然でもない といえるでしょう。
**Answer 2:** この文は文法的には正確で理解可能ですが、意味的には一部改善の余地があります。以下は微調整したバージョンです：
「経営者側は従業員に昇給させました。彼らはストライキを起こすと脅しましたからです。」
このバージョンでは、文脈に合った適切な接続詞や表現が使われています。総合的には「4ー自然」と評価できるでしょう。

**Q5 Ask_Sub_Overt**

**Question:** ジョンはビルに聞いた。彼は頭がこんがらがってしまったからだ。頭がこんがらがってしまったのは誰ですか？
John asked Bill because he was confused. Who was confused?

**Answer 1:** この文章も４ー自然、といえるでしょう。文脈において理解しやすく、一般的な表現です。

**Answer 2:** この文章は自然で理解しやすい表現です。ジョンがビルに質問し、ビルが頭がこんがらがってしまったという状況が適切に描写されています。
評価: ４ー自然

**Q6 Ask_Sub_Null**

**Question:** ジョンはビルに聞いた。頭がこんがらがってしまったからだ。頭がこんがらがってしまったのは誰ですか？
John asked Bill because (he) was confused. Who was confused?

**Answer 1:** この文は言い回しや文脈において一部不自然さを感じるかもしれませんが、全体としては理解できる表現です。そのため、総合的に評価すると、３ー自然でも不自然でもない といえるでしょう。
**Answer 2:** 文章全体としては自然な表現ですが、一部微調整することでもっと自然になる可能性があります。例えば、文末の「からだ。」の代わりに「からだ、と言った。」や「からだ、尋ねた。」など、文脈によって微妙な変化があります。ただし、これは微細な点であり、元の文章も通じる自然な表現です。そのため、総合的には「4ー自然」と評価できるでしょう。

**Q7 Ask_IO_Overt**

**Question:** ジョンはビルに聞いた。彼が答えを知っていたからだ。答えを知っていたのは誰ですか？
John asked Bill because he knew the answer. Who knew the answer?

**Answer 1:** この文章は４ー自然、と言えるでしょう。文脈において理解しやすく、自然な表現です。
**Answer 2:** この文章は自然であり、文脈も理解しやすいです。ジョンがビルに質問し、ビルが答えを知っていたという状況が明確に表現されています。
評価: ４ー自然

**Q8 Ask_IO_Null**

**Question:** ジョンはビルに聞いた。答えを知っていたからだ。答えを知っていたのは誰ですか？

John asked Bill because (he) knew the answer. Who knew the answer?

**Answer 1:** この文は文法的に正確で、一般的な表現です。文脈も自然で、理解しやすいです。したがって、４ー自然 と評価できるでしょう。
**Answer 2:** この文は文法的に正確で、理解しやすい表現です。文脈にも適しています。総合的には「4ー自然」と評価できるでしょう。

**Q9 Protest_Sub_Overt**

**Question:** 民衆は政府に抗議した。彼らは飢えていたからだ。飢えていたのは誰ですか？
The people protested the government because they were hungry. Who was hungry?

**Answer 1:** この文章も４ー自然、といえるでしょう。文脈において理解しやすく、一般的な表現です。
**Answer 2:** この文章は自然な表現であり、文脈も理解しやすいです。抗議の理由として「飢えていた」という状況が適切に説明されています。
評価: ４ー自然

**Q10 Protest_Sub_Null**

**Question:** 民衆は政府に抗議した。飢えていたからだ。飢えていたのは誰ですか？
The people protested the government because (they) were hungry. Who was hungry?

**Answer 1:** この文は文法的にも文脈的にも自然で、理解しやすい表現です。民衆が政府に抗議した理由として「飢えていたからだ」というのは、十分に自然な表現です。したがって、４ー自然 と評価できるでしょう。
**Answer 2:** この文は文法的には正確で、理解しやすい表現です。文脈にも適しており、原因と結果が自然に結びついています。総合的には「4ー自然」と評価できるでしょう。

**Q11 Protest_IO_Overt**

**Question:** 民衆は政府に抗議した。彼らは増税しようとしていたからだ。増税しようとしていたのは誰ですか？

The people protested the government because they were raising taxes. Who was raising taxes?

**Answer 1:** この文章は４ー自然、と言えるでしょう。文脈において理解しやすく、自然な表現です。
**Answer 2:** この文章は比較的自然な表現で、文脈も理解しやすいです。ただし、文中の「彼らは増税しようとしていたからだ。」という部分が少し堅い表現かもしれません。より自然な表現にすると、例えば「彼らは増税を計画していたから抗議した。」というように言い換えることができます。
評価: ４ー自然

**Q12 Protest_IO_Null**

**Question:** 民衆は政府に抗議した。増税しようとしていたからだ。増税しようとしていたのは誰ですか？
The people protested the government because (they) were raising taxes. Who was raising taxes?

**Answer 1:** この文は比較的自然であり、理解しやすい表現です。したがって、４ー自然と評価できます。
**Answer 2:** この文は文法的に正確で、理解しやすい表現です。文脈にも適しており、自然な日本語と言えます。総合的には「4ー自然」と評価できるでしょう。

**Q13 Candy_Sub_Overt**

**Question:** 男は少年にキャンディをあげると言った。しかし彼はキャンディをもっていなかった。キャンディをもっていなかったのは誰ですか？
The man offered the boy candy, but he had no candy. Who had no candy?

**Answer 1:** この文章は３ー自然でも不自然でもない、と言えるでしょう。しかし、もう少し簡潔に表現することもできます。「男は少年にキャンディをあげると言ったが、持っていなかった。」とすることで、より自然になります。
**Answer 2:** この文章は文脈に対して理解しやすく、自然な表現です。状況が明確で、読み手に分かりやすい描写がされています。
評価: ４ー自然

**Q14 Candy_Sub_Null**

**Question:** 男は少年にキャンディをあげると言った。しかしキャンディをもっていなかった。キャンディをもっていなかったのは誰ですか？
The man offered the boy candy, but (he) had no candy. Who had no candy?

**Answer 1:** この文は文法的にも文脈的にも自然で、理解しやすい表現です。男が少年にキャンディをあげると言ったが、実際にはキャンディを持っていなかったという状況が明確に表現されています。したがって、４ー自然 と評価できるでしょう。
**Answer 2:** この文は文法的には正確であり、理解しやすい表現です。ただし、「男は少年にキャンディをあげると言った。しかしキャンディをもっていなかった。」と続けると、文が矛盾しているように感じられるかもしれません。一般的には、「しかし」は予想外の事態を示すものであり、この文脈では適切ではないかもしれません。
以下は微調整したバージョンです：
「男は少年にキャンディをあげると言ったが、しかし手元にはなかった。」
このように修正することで、文がより自然になります。総合的には「4ー自然」と評価できるでしょう。

**Q15 Candy_IO_Overt**

**Question:** 男は少年にキャンディをあげると言った。しかし彼はキャンディをもらわなかった。キャンディをもらわなかったのは誰ですか？
The man offered the boy candy, but he got no candy. Who got no candy?

**Answer 1:** この文章は３ー自然でも不自然でもない、といえるでしょう。ただし、文脈によってはもう少し詳細を加えた方が良いかもしれません。例えば、「男は少年にキャンディをあげると言った。しかし、意外なことに、彼はキャンディをもらわなかった。」などとすることで、状況や出来事に関する情報が追加され、より理解しやすくなります。
**Answer 2:** この文章は文脈に対して理解しやすく、自然な表現です。特に問題はなく、簡潔に状況を伝えています。
評価: ４ー自然

**Q16 Candy_IO_Null**

**Question:** 男は少年にキャンディをあげると言った。しかしキャンディをもらわなかった。キャンディをもらわなかったのは誰ですか？

The man offered the boy candy, but (he) got no candy. Who got no candy?

**Answer 1:** この文は文法的にも文脈的にも自然であり、一般的な表現です。男が少年にキャンディをあげると言ったが、少年がそれをもらわなかったという状況が明確に表現されています。したがって、４ー自然 と評価できるでしょう。
**Answer 2:** この文は文法的に正確で、意味も理解できます。ただし、「男は少年にキャンディをあげると言った。しかしキャンディをもらわなかった。」と続けると、文が矛盾しているように感じられるかもしれません。一般的には、「しかし」は予想外の事態を示すものであり、この文脈では適切ではないかもしれません。
以下は微調整したバージョンです：
「男は少年にキャンディをあげると言ったが、少年はキャンディを受け取らなかった。」
このように修正することで、文がより自然になります。総合的には「4ー自然」と評価できるでしょう。

**Q17 Grades_Sub_Overt**

**Question:** 教授は学生に低い成績をつけた。彼は厳格だったからだ。厳格だったのは誰ですか？
The professor gave the student a bad grade because he was strict. Who was strict?

**Answer 1:** この文章は３ー自然でも不自然でもない、といえるでしょう。しかし、一部の文脈においてはもっと具体的な情報が欲しいと感じることがあるかもしれません。例えば、「教授は学生に低い成績をつけた。彼は厳格だったからだ。しかしこの教授の授業は非常に難しく、学生たちは挑戦的な課題に直面していた。」などと続けることで、なぜ教授が厳格だったのかがより明確になります。
**Answer 2:** この文章は自然な表現で、文脈も理解しやすいです。教授が学生に低い成績をつけた理由として、「彼は厳格だったからだ」という説明が適切に表現されています。
評価: ４ー自然

**Q18 Grades_Sub_Null**

**Question:** 教授は学生に低い成績をつけた。厳格だったからだ。厳格だったのは誰ですか？
The professor gave the student a bad grade because (he) was strict. Who was strict?

**Answer 1:** この文は文法的に正確であり、一般的な表現です。教授が学生に低い成績をつける理由として「厳格だったから」というのも自然な説明です。したがって、４ー自然 と評価できるでしょう。
**Answer 2:** この文は文法的に正確であり、自然な表現です。原因と結果が適切に繋がっており、日本語としても理解しやすいです。総合的には「4ー自然」と評価できるでしょう。

**Q19 Grades_IO_Overt**

**Question:** 教授は学生に低い成績をつけた。彼が勉強しなかったからだ。勉強しなかったのは誰ですか？
The professor gave the student a bad grade because he did not study. Who did not study?

**Answer 1:** この文章は ３ー自然でも不自然でもない、と言えるでしょう。しかし、少し改善の余地があります。例えば、「教授は学生に低い成績をつけた。なぜなら彼が勉強しなかったからだ。」とすることで、文脈がより明確になります。
**Answer 2:** この文章は比較的自然な表現ですが、文中の「彼が勉強しなかったからだ。」という部分は少し強調された表現と感じられるかもしれません。もう少し穏やかに表現すると、例えば「彼は勉強が足りなかったため、教授に低い成績をつけられました。」というように言い換えることができます。
評価: ３ー自然でも不自然でもない

**Q20 Grades_IO_Null**

**Question:** 教授は学生に低い成績をつけた。勉強しなかったからだ。勉強しなかったのは誰ですか？
The professor gave the student a bad grade because (he) did not study. Who did not study?

**Answer 1:** この文は文法的にも文脈的にも自然で、一般的な表現です。教授が学生に低い成績をつけた理由として「勉強しなかったからだ」というのは、一般的な説明です。したがって、４ー自然 と評価できるでしょう。
**Answer 2:** この文は文法的に正確で、意味も明確です。文脈によっては、「勉強しなかったからだ」をもう少し詳しく説明するとより良いかもしれませんが、基本的には自然な表現です。総合的には「4ー自然」と評価できるでしょう。

**Q21 Window_Sub_Overt**

**Question:** ボールは窓ガラスを割った。それが堅かったからだ。堅かったのは何ですか？
The ball broke the window because it was hard. What was hard?

**Answer 1:** この文章は ２ー不自然、と評価される可能性があります。一般的に、「ボールは窓ガラスを割った。それが堅かったからだ。」という表現は不自然です。ガラスは硬い材料であるため、割れやすい性質を持っています。文脈や意図によっては、適切な表現に修正すると良いでしょう。例えば、「ボールは窓ガラスを割った。それが強化ガラスだったからだ。」などが考えられます。
**Answer 2:** この文章は比較的自然な表現です。ボールが窓ガラスを割るという状況と、その原因として「それが堅かったからだ」という説明が一貫しています。しかし、より自然にするためには、「それが堅かった」の代わりに「それが頑丈だったからだ」とか「それが割れにくかったからだ」と表現すると、より適切かもしれません。
評価: ３ー自然でも不自然でもない

**Q22 Window_Sub_Null**

**Question:** ボールは窓ガラスを割った。堅かったからだ。堅かったのは何ですか？
The ball broke the window because (it) was hard. What was hard?

**Answer 1:** この文は文法的には正確ですが、物理的な観点から不自然な印象を与える可能性があります。通常、硬さ（堅さ）がある物体が他の物体を割るのは、通常ではないと考えられます。通常、硬いものが割れるのは脆いものを加えたりする場合などが一般的です。そのため、「ボールは窓ガラスを割った。堅かったからだ」という理由は、通常の文脈では奇妙に感じられる可能性があります。
総合的には、２ー不自然 と評価できるでしょう。適切な理由を考慮して表現するとより自然な文になります。
**Answer 2:** この文は文法的には正確で、理解しやすい表現です。ただし、「ボールは窓ガラスを割った。堅かったからだ。」と続けると、原因と結果が逆になっているように感じられます。通常、ガラスは硬く、ボールが当たることで割れるのが一般的です。
修正すると、以下のようになります：
「ボールは窓ガラスを割った。ガラスが割れやすかったからです。」

このようにすることで、原因と結果の関係が自然になります。総合的には「4ー自然」と評価できるでしょう。

**Q23 Window_Ob_Overt**

**Question:** ボールは窓ガラスを割った。それが壊れやすかったからだ。壊れやすかったのは何ですか？
The ball broke the window because it was fragile. What was fragile?

**Answer 1:** この文章も４ー自然、といえるでしょう。文脈において理解しやすく、一般的な表現です。
**Answer 2:** この文章は自然で理解しやすい表現です。原因と結果が適切に繋がっており、文脈も明確です。
評価: ４ー自然

**Q24 Window_Ob_Null**

**Question:** ボールは窓ガラスを割った。壊れやすかったからだ。壊れやすかったのは何ですか？
The ball broke the window because (it) was fragile. What was fragile?

**Answer 1:** この文も文法的にも文脈的にも自然で、理解しやすい表現です。ボールが窓ガラスを割った理由として「壊れやすかったからだ」というのは、物の性質に対する説明として自然です。したがって、４ー自然 と評価できるでしょう。
**Answer 2:** この文は文法的に正確であり、意味も理解できます。ただし、「ボールは窓ガラスを割った。壊れやすかったからだ。」と続けると、原因と結果の関係が逆になっているように感じられます。通常、ガラスが割れるのはボールが強く当たるからであり、ガラスが壊れやすかったからボールが割れるわけではありません。
修正すると、次のようになります：
「ボールは窓ガラスを割った。ガラスが壊れやすかったからです。」
このようにすれば、原因と結果の関係が自然になります。総合的には「3ー自然でも不自然でもない」と評価できるでしょう。

**Q25 Catch_Sub_Overt**

**Question:** ネコはネズミを捕まえた。それは賢かったから。賢かったのは何ですか？
The cat caught the mouse because it was clever. What was clever?

**Answer 1:** この文章も４ー自然、といえるでしょう。十分な文脈がないため、もう少し詳細があればより良いかもしれませんが、基本的には自然な表現です。
**Answer 2:** この文章はやや不自然な表現です。ネコがネズミを捕まえた理由として、「それは賢かったから」という表現は、通常では使われない形です。ネコが賢かったのではなく、ネコが賢くてネズミを捕まえたという表現にするとより自然になります。
改善例:
ネコはネズミを捕まえた。彼は賢かったから。
評価: ２ー不自然

**Q26 Catch_Sub_Null**

**Question:** ネコはネズミを捕まえた。賢かったから。賢かったのは何ですか？
The cat caught the mouse because (it) was clever. What was clever?

**Answer 1:** この文も文法的にも文脈的にも自然で、理解しやすい表現です。ネコがネズミを捕まえた理由として「賢かったから」というのは、一般的な説明とは言えないかもしれませんが、動物に対する擬人化や軽いユーモアが含まれている表現となっています。総合的には、３ー自然でも不自然でもない といえるでしょう。
**Answer 2:** この文は文法的には正確であり、意味も理解できます。ただし、もう少し文脈を追加することで、より自然になる可能性があります。例えば :
「ネコはネズミを捕まえました。ネコは賢い生き物なので、それができたのです。」
このように修正することで、原文よりも文が滑らかになり、自然に聞こえるでしょう。
それでも、元の文も理解可能であるため、総合的には「3ー自然でも不自然でもない」と評価できるでしょう

**Q27 Catch_Ob_Overt**

**Question:** ネコはネズミを捕まえた。それは油断していたからだ。油断していたのは何ですか？
The cat caught the mouse because it was careless. What was careless?

**Answer 1:** この文章は４ー自然、といえるでしょう。文脈において理解しやすく、一般的な表現です。

**Answer 2:** この文章は自然な表現で、ネコがネズミを捕まえた理由として、「それは油断していたからだ」という表現が適切です。文脈も明確であり、理解しやすいです。
評価: ４ー自然

**Q28 Catch_Ob_Null**

**Question:** ネコはネズミを捕まえた。油断していたからだ。油断していたのは何ですか？
The cat caught the mouse because (it) was careless. What was careless?

**Answer 1:** この文は文法的にも文脈的にも自然で、一般的な表現です。ネコがネズミを捕まえた理由として「油断していたからだ」というのは、動物の行動に対して一般的な説明です。したがって、４ー自然 と評価できるでしょう。
**Answer 2:** この文は文法的に正確で、理解しやすい表現です。文脈にも適しており、自然な日本語と言えます。総合的には「4ー自然」と評価できるでしょう。

**Q29 Visit_Sub_Overt**

**Question:** ジョンは病院にルークを見舞った。彼は近くに住んでいたからだ。近くに住んでいたのは誰ですか？
John visited Luke in the hospital because he lived close by. Who lived close by?

**Answer 1:** この文章は４ー自然、と言えるでしょう。文脈において理解しやすく、自然な表現です。
**Answer 2:** この文章は自然な表現で、文脈も理解しやすいです。ジョンが病院に友人のルークを見舞った理由として、「彼は近くに住んでいたからだ」という説明が適切に表現されています。
評価: ４ー自然

**Q30 Visit_Sub_Null**

**Question:** ジョンは病院にルークを見舞った。近くに住んでいたからだ。近くに住んでいたのは誰ですか？
John visited Luke in the hospital because (he) lived close by. Who lived close by?

**Answer 1:** この文は文法的にも文脈的にも自然で、理解しやすい表現です。ジョンが病院にルークを見舞った理由として「近くに住んでいたからだ」というのは、一般的で自然な説明です。したがって、４ー自然 と評価できるでしょう。
**Answer 2:** この文も文法的に正確で、自然な表現です。文脈にも適しており、特に問題はありません。総合的には「4ー自然」と評価できるでしょう。

**Q31 Visit_Ob_Overt**

**Question:** ジョンは病院にルークを見舞った。彼は病気だったのだ。病気だったのは誰ですか？
John visited Luke in the hospital because he was sick. Who was sick?

**Answer 1:** この文章は４ー自然、と言えるでしょう。文脈において理解しやすく、自然な表現です。
**Answer 2:** この文章は自然な表現で、文脈も理解しやすいです。特に問題はなく、状況が明確に伝わります。
評価: ４ー自然

**Q32 Visit_Ob_Null**

**Question:** ジョンは病院にルークを見舞った。病気だったのだ。病気だったのは誰ですか？
John visited Luke in the hospital because (he) was sick. Who was sick?

**Answer 1:** この文は文法的にも文脈的にも自然で、理解しやすい表現です。ジョンが病院にルークを見舞った理由として「病気だったのだ」というのは、一般的で自然な表現です。したがって、４ー自然 と評価できるでしょう。
**Answer 2:** この文は文法的に正確で、理解しやすい表現です。文脈にも適しており、自然な日本語と言えます。総合的には「4ー自然」と評価できるでしょう。

**Q33 Hit_Sub_Overt**

**Question:** ショーンはジョーを叩いた。しかし彼は謝らなかった。謝らなかったのは誰ですか？
Shawn hit Joe but he did not apologize. Who did not apologize?

**Answer 1:** この文章も３ー自然でも不自然でもない、といえるでしょう。しかし、文脈によってはもう少し情報が欲しい場合があります。例えば、「ショーンはジョーを叩いた。しかし、彼は謝らなかった。その代わりに、なぜ叩いたのかを説明し始めた。」といった文脈があれば、読み手はより具体的な状況や背景を理解しやすくなります。
**Answer 2:** この文章も文脈が明確で、自然な表現です。ショーンがジョーを叩いた行為と、その後の彼の謝罪のなさがしっかりと伝わります。
評価: ４ー自然

**Q34 Hit_Sub_Null**

**Question:** ショーンはジョーを叩いた。しかし謝らなかった。謝らなかったのは誰ですか？
Shawn hit Joe but (he) did not apologize. Who did not apologize?

**Answer 1:** この文は文法的にも文脈的にも自然であり、一般的な表現です。ショーンがジョーを叩いたが、その後に謝らなかったという状況が明確に表現されています。したがって、４ー自然 と評価できるでしょう。
**Answer 2:** この文は文法的に正確で、意味も理解できます。ただし、一部微調整することでより自然な表現になるかもしれません。例えば :
「ショーンはジョーを叩いたが、しかし謝りませんでした。」
このように文全体をつなげる接続詞を使うことで、文が滑らかになります。ただし、元の文も理解可能であり、総合的には「3ー自然でも不自然でもない」と評価できるでしょう。

**Q35 Hit_Ob_Overt**

**Question:** ショーンはジョーを叩いた。しかし彼は叫び声を上げなかった。叫び声を上げなかったのは誰ですか？
Shawn hit Joe but he did not scream. Who did not scream?

**Answer 1:** この文章も４ー自然、といえるでしょう。文脈において理解しやすく、一般的な表現です。
**Answer 2:** この文章は文脈がわかりやすく、自然な表現です。状況が明確であり、読み手に理解しやすい描写がされています。
評価: ４ー自然

**Q36 Hit_Ob_Null**

**Question:** ショーンはジョーを叩いた。しかし叫び声を上げなかった。叫び声を上げなかったのは誰ですか？
Shawn hit Joe but (he) did not scream. Who did not scream?

**Answer 1:** この文は文法的には正しいですが、文脈的にはやや奇妙に感じられるかもしれません。通常、人が叩かれた場合には叫び声を上げることが一般的であるため、その点が驚きや違和感を生む可能性があります。ただし、文全体としては理解可能で、２ー不自然 というよりも、やや奇妙な表現といえるでしょう。
**Answer 2:** この文は文法的には正確で、理解しやすい表現です。ただし、「しかし」の使用により、文の流れが矛盾しているように感じられるかもしれません。通常、叩かれると痛みや驚きから叫び声が上がることが一般的です。もしそのような意図があるなら、文を微調整するとより自然になるでしょう。
例えば：
「ショーンはジョーを叩いたが、ジョーは叫び声を上げなかった。」
このようにすることで、文が矛盾していないように感じられます。総合的には「4ー自然」と評価できるでしょう。

**Q37 Firing_Sub_Overt**

**Question:** ジョンはデイブを首にした。彼はあまりに多くの従業員を抱えていた。あまりに多くの従業員を抱えていたのは誰ですか？
John fired Dave because he had too many employees. Who had too many employees?

**Answer 1:** この文は２ー不自然、と評価される可能性があります。"首にした"という表現が非常に過激であり、通常は殺害行為を指す言葉です。もし単に解雇や首にするような意味であれば、「ジョンはデイブを解雇した。彼はあまりに多くの従業員を抱えていた。」などと表現すると、より自然になります。
**Answer 2:** この文章は文脈によっては意味が取りにくく、また一部の表現が不自然です。また、文中において「ジョンはデイブを首にした」という表現は、通常は残忍な行為を指す可能性があります。以下は改善例と評価です。
改善例:

ジョンは多くの従業員を雇っていたが、経済的な理由から一部の従業員を解雇せざるを得なくなりました。
評価: ２ー不自然

**Q38 Firing_Sub_Null**

**Question:** ジョンはデイブを首にした。あまりに多くの従業員を抱えていた。あまりに多くの従業員を抱えていたのは誰ですか？
John fired Dave because (he) had too many employees. Who had too many employees?

**Answer 1:** この文は文法的には正確ですが、内容的には不自然な印象を与える可能性があります。"デイブを首にした"という表現は通常、「解雇する」などの他の言葉を使ったほうが一般的です。また、「あまりに多くの従業員を抱えていた」という理由が、解雇にはあまり一般的ではないかもしれません。状況によっては適切な表現かもしれませんが、通常の文脈ではやや不自然と感じられる可能性があります。総合的には、２ー不自然 と評価できるでしょう。
**Answer 2:** この文は不自然で、意味が混乱している可能性があります。文の流れから考えると、おそらく「首にした」が「解雇した」を指していると解釈されますが、この表現は通常使われないし、乱暴な表現と受け取られる可能性があります。
正確な意図に基づいて言い回しを修正すると、例えば：
「ジョンはデイブを解雇した。あまりにも多くの従業員を抱えていたためです。」
このようにすれば、より自然で理解しやすい表現になります。元の文に対する総合的な評価は「2ー不自然」です。

**Q39 Firing_Ob_Overt**

**Question:** ジョンはデイブを首にした。彼は欠勤が多すぎた。欠勤が多すぎたのは誰ですか？
John fired Dave because he had too many absences. Who had too many absences?

**Answer 1:** この文章は １ーとても不自然、と評価される可能性があります。"首にした"という表現が非常に過激であり、通常は殺害行為を指す言葉です。もし単に解雇や首にするような意味であれば、「ジョンはデイブを解雇した。彼は欠勤が多すぎた。」などと表現すると、より自然になります。
**Answer 2:** この文は文脈によっては意味が通じるものの、一般的には不自然です。特に「ジョンはデイブを首にした。」という表現は、通常は残忍な行為を指すため、注意が

必要です。また、「欠勤が多すぎた」という理由で誰かを解雇することも、一般的な表現ではありません。
改善例:
ジョンはデイブを解雇した。彼の欠勤が多すぎたからだ。
評価: ２ー不自然

**Q40 Firing_Ob_Null**

**Question:** ジョンはデイブを首にした。欠勤が多すぎた。欠勤が多すぎたのは誰ですか？
John fired Dave because (he) had too many absences. Who had too many absences?

**Answer 1:** この文は、一般的な言い回しではなく、特にビジネスなどで使われる表現ではありません。"デイブを首にする"という表現は、あまり一般的でないため、やや不自然と言えます。また、「欠勤が多すぎた」という理由も、一般的な解雇理由としては通常ではありますが、文全体としてはやや硬い表現かもしれません。総合的には、２ー不自然 と評価できるでしょう。
**Answer 2:** この文は一般的な表現としてはやや不自然で、意味が矛盾しているように感じられます。通常、「首にする」という表現は解雇や処罰を指すため、その後の理由として「欠勤が多すぎた」が相応しくないと思われます。
修正すると、以下のようになります：
「ジョンはデイブを解雇した。欠勤が多すぎたからです。」
このようにすることで、文がより自然になり、意味が一貫しているように感じられます。総合的には「3ー自然でも不自然でもない」と評価できるでしょう。