Inconsistent Tokenizations Cause Language Models to be Perplexed by Japanese Grammar

Andrew Gambardella, Takeshi Kojima, Yusuke Iwasawa, Yutaka Matsuo

The University of Tokyo

Correspondence: atgambardella@weblab.t.u-tokyo.ac.jp

Abstract

Typical methods for evaluating the performance of language models evaluate their ability to answer questions accurately. These evaluation metrics are acceptable for determining the extent to which language models can understand and reason about text in a general sense, but fail to capture nuanced capabilities, such as the ability of language models to recognize and obey rare grammar points, particularly in languages other than English. We measure the perplexity of language models when confronted with the "first person psych predicate restriction" grammar point in Japanese. Weblab is the only tested open source model in the 7-10B parameter range which consistently assigns higher perplexity to ungrammatical psych predicate sentences than grammatical ones. We give evidence that Weblab's uniformly bad tokenization is a possible root cause for its good performance, and show that Llama 3's perplexity on grammatical psych predicate sentences can be reduced by orders of magnitude (28x difference) by restricting test sentences to those with uniformly well-behaved tokenizations. We show in further experiments on machine translation tasks that language models will use alternative grammar patterns in order to produce grammatical sentences when tokenization issues prevent the most natural sentence from being output.

1 Introduction

It is unusually difficult to find the correct tests and metrics to evaluate large language model (LLM) performance. Standard benchmarks such as MMLU (Hendrycks et al., 2021) are designed to capture the memorization and reasoning abilities of LLMs, but as LLMs are general tools which can be used to solve a large variety of tasks, there are many aspects of LLM performance which these benchmarks cannot cover.

In particular, these benchmarks are generally English-centric, and tend to test "higher-level" abil-
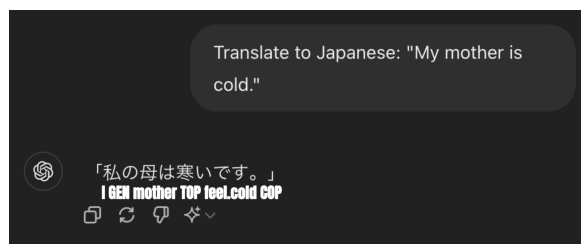

Figure 1: State of the art language models frequently fail to respect nuanced aspects of Japanese grammar, such as the first person psych predicate restriction, where here GPT-4o produces a sentence which is functionally identical to the ungrammatical Example (2) in Section 2.1.

ities, whereas more nuanced evaluation of LLMs has received much less attention. Even for other languages such as Japanese, existing benchmarks such as llm-jp-eval (Han et al., 2024) tend to test these higher-level abilities in Japanese, and do not test any Japanese-specific abilities. This is concerning, because Japanese contains several rare grammatical structures that require special attention, especially considering that many foundation models are now trained on multilingual datasets, and one would assume some degree of cross-lingual information sharing due to the prevalence of linguistic universals (Greenberg J. H., 1963). We inspect the perplexities of pretrained LLMs in order to study nuanced grammatical corner cases in Japanese, which are frequently studied in linguistics literature, but for which it would be difficult to create a large dataset. In particular, we study the "first person psych predicate restriction" with models in the 7-10B parameter range. We find that of open source models that currently exist, only Weblab has lower perplexities on grammatical sentences compared to ungrammatical sentences containing psych predicates. We provide evidence that this is due to Weblab's uniformly bad tokenizations, and show that Llama 3 performs orders of magnitude better on this task when its test set is

restricted to sentences which it tokenizes uniformly well. We further show evidence that language models will, when tasked to output sentences containing third-person psych predicate expressions, utilize unusual circumlocutions or grammar patterns. These findings suggest that more consistent tokenizers for Japanese could lead to LLMs which are able to more closely obey relatively obscure Japanese grammar rules and output sentences utilizing more natural grammatical constructions.

2 Linguistic Preliminaries

2.1 First Person Psych Predicate Restriction

Japanese is notable for having what is referred to as a "first person psych predicate restriction," meaning that certain predicates which describe the internal states of people (such as "happy," "sad," "cold," "dizzy," etc.) can only be used to describe the first person in general when used directly. When describing others, evidential expressions (such as "seems," "appears," etc) must be used (Hasegawa and Hirose, 2005; Lee, 2012). Hasegawa and Hirose (2005) give the following examples:

(1) Watashi wa   samu-i.
      I   TOP feel.cold-NPST

      "I feel cold."

(2) #Haha wa   samu-i.
      Mother TOP feel.cold-NPST

      "My mother feels cold." [Intended]

(3) Haha   wa   samu-gat-te   i-ru.
      Mother TOP feel.cold-EVID-CONJ be-NPST

      Lit. "My mother is showing signs of feeling cold."

(4) Haha   wa   samu-soo   da.
      Mother TOP feel.cold-EVID COP.NPST

      "My mother appears to feel cold."

Native speakers of Japanese generally obey the first-person psych predicate restriction in all cases, even if they are not consciously aware of its existence. L2 speakers of Japanese, however, are frequently unaware of it, which can lead them to produce utterances that are extremely confusing to native speakers. Similar grammatical phenomena and L2 misuse have been observed in Korean, which features a comparable first-person psych predicate restriction (Ahn and Mariano, 2024).

As Hasegawa and Hirose (2005) further note, when the predicate is polysemous, the semantic role of the subject shifts to conform to this restriction (so that Example (2) must be interpreted as "I, the mother, am cold" or possibly "I am cold because of my mother"). There are also exceptions for the nonreportive style (such as in a novel, where the narrator can know the internal states of all characters) where third-person subjects are permitted for psych predicate expressions (Kuroda, 1973). From the perspective of LLM training, checking whether or not a model obeys the first person psych predicate restriction is interesting, as we must test whether the LLM has learned to *not* do something which was *not* in the training dataset, rather than to reproduce co-occurrences and patterns which *did* exist in the dataset. Notably, even state of the art models such as GPT-4o (OpenAI, 2023) frequently create ungrammatical or strange sentences which fail to respect this grammatical rule, as shown in Figure 1.

3 Experiments

We wish to see whether language models have the capability to understand the grammar rules introduced in Sections 2.1. For this case, it is simple to create minimal pairs of sentences similar to Examples (1), (3), and (4), which can be compared against ungrammatical constructions similar to Example (2). We evaluate the Huggingface (Wolf et al., 2019) implementations of the following LLMs in the 7B to 10B parameter range: Mistral 0.1-7B (Jiang et al., 2023), Llama 2-7B (Touvron et al., 2023), Llama 3-8B (Llama Team, 2024), Weblab-10B[1], Swallow-7B (Fujii et al., 2024), and Swallow-MS-7b-v0.1[2]. Mistral, Llama 2, and Llama 3 are multilingual LLMs, whereas Weblab, Swallow, and Swallow-MS are specifically tuned for Japanese applications. In Section 3.1 we choose to use perplexity on base models as our measure because it can be easily applied on the sentence level regardless of how the sentence is tokenized, and also correlates directly with token generation probabilities (e.g., sentence x has lower perplexity than y generally implies that sentence x is more likely to be generated than y). We also report the median over perplexity scores rather than the mean as the scales and numerical values of token probabilities can differ greatly between models due to variations in training data, vocabulary sizes, and tokenizer

---

[1]https://huggingface.co/matsuo-lab/weblab-10b
[2]https://huggingface.co/tokyotech-llm/Swallow-MS-7b-v0.1

	Mistral	Llama 2	Llama 3	Weblab	Swallow	Swallow-MS
(#) 3rd, psych, direct	2.0e+04	3.3e+04	6.9e+03	2.0e+06	1.2e+03	1.9e+03
(a) 1st, psych, direct	3.6e+04	1.2e+05	9.1e+04	6.1e+05	1.9e+03	3.2e+03
(b) 3rd, non-psych, direct	1.8e+03	5.9e+03	4.5e+03	7.3e+05	1.2e+03	2.9e+03
(c) 3rd, psych, evidential	2.0e+04	4.9e+04	3.7e+04	1.3e+06	4.1e+03	3.3e+03

Table 1: Median perplexity over language models, for sentences corresponding to those introduced in Section 2.1. Weblab is the only model which has lower perplexities for all grammatical constructions (labeled a, b, c) relative to the ungrammatical direct third person psych predicate (labeled #), which we believe is due to its uniformly bad tokenization. Green, yellow, and red indicate perplexity for grammatical constructions that are respectively lower than, equal to, and higher than that of the grammatical constructions.

designs (including differing fertility scores, as discussed below). These differences tend to affect the mean much more significantly due to its sensitivity to outliers, making the median a more robust statistic for comparing central tendencies across diverse models in this context. In Section 3.2 we further evaluate the instruction-following variants of these models when presented with machine translation tasks.

3.1 Perplexity Experiments

For the psych predicate restriction, an ideal language model would have the following properties:

(a) The probability of an adjective or verb which is not a psych predicate given that the sentence is in the third person is higher than the probability of an adjective or verb which is a psych predicate given that the sentence is in the third person.

(b) The probability of an adjective or verb which is a psych predicate given that the sentence is in the first person is higher than the probability of the same adjective or verb given that the sentence is in the third person.

(c) The probability that a psych predicate is used with an evidential given that the sentence is in the third person is higher than the probability of the same psych predicate being used directly given that the sentence is in the third person.

We show perplexities corresponding to the grammatical constructions in (a), (b), and (c) above, as well as the ungrammatical third person direct psych predicate expression, in Table 1 with templates for our test sentences in Appendix B. Despite having much higher perplexities in general, only

Weblab has lower perplexities for the grammatical constructions than the ungrammatical construction. We found that our results were influenced heavily by tokenization, particularly whether or not byte fallback (which reduces the probabilities of tokens by several orders of magnitude) was needed, and whether or not the tokenization corresponded to the underlying linguistic structure. In particular, we noticed that for Llama 3, while certain adjectives ("itai," "samui," "tsurai," "kokorobosoi," "atsui") caused well-behaved tokenizations, those ending in "shii" such as ("kanashii," "sabishii," "ureshii," "hazukashii," "kurushii") caused byte fallback when used in conjunction with evidential expressions, thus giving these tokens extremely low probability, even when they are grammatically necessary. When restricting our test sentences to only those which Llama 3 tokenizes well, we found that its median perplexity for evidentials used with psych predicates in the third person (sentence type (c)) drops from the reported 3.7e+04 to 1.3e+03 (~28x difference), whereas the perplexity of the ungrammatical construction drops from the reported 6.9e+03 to only 3.9e+03 (~1.8x difference). Consequently, the Llama 3 base model is very likely to output the ungrammatical direct form of a psych predicate adjective with a third person topic, if and only if that adjective ends in "shii" as opposed to just "i." Note that for sentence type (c) compared to the ungrammatical construction only the final tokens in the sentence differ, so lower perplexity is exactly equivalent to higher generation chance.

Ironically, Weblab may have performed best on these tasks because it was trained using a tokenizer for English models as-is, and as such does not contain tokens for even extremely common Japanese characters[3]. As a result, it tokenizes ev-

---

[3]Weblab's tokenizer uses byte fallback for words such as

	Mistral	Llama 2	Llama 3	Weblab	Swallow	Swallow-MS
Fertility	1.51	1.58	0.85	1.23	1.0	1.0
Byte Fallback	0.43	0.49	0.08	0.66	0.19	0.19

Table 2: Japanese fertility scores and byte fallback rates across studied models over sentences produced by the templates given in Appendix B. Due to its use of an unmodified English tokenizer, the majority of tokens produced by Weblab are out-of-vocabulary.

	Weblab				Llama 3			
	Cold	Embarrassed	Lonely	Pain	Cold	Embarrassed	Lonely	Pain
✔evidential	47	90	0	0	0	32	0	0
✔grammatical	53	10	0	6	0	0	0	0
�’no evidential	0	0	100	94	69	39	100	100
✘mistranslation	0	0	0	0	0	29	0	0
✘wrong syntax	0	0	0	0	31	0	0	0

Table 3: Weblab and Llama 3 outputs when asked to translate the English sentence "My mother is {psych predicate}" into Japanese. While Llama 3 struggled to output evidential expressions at all, Weblab was able to consistently output evidential expressions with a third person subject feeling "cold" or "embarrassed." Here "grammatical" indicates alternative phrasings that are grammatical translations of the sentence, but do not require the use of evidential expressions.

ery sentence uniformly poorly, and does not have grammar-specific tokenization issues. To quantitatively support claims about tokenizer strength, we report fertility scores (Rust et al., 2021) and byte fallback rates in Table 2. Weblab, which used an unmodified English language tokenizer, must output significantly more tokens, and outputs out-of-vocabulary tokens much more frequently than LLMs such as Llama 3, which use tokenizers more adapted to Japanese. These findings suggest that the same issues which seem to appear in very large state of the art models, such as GPT-4o given in Figure 1, could possibly be alleviated simply by using more consistent tokenizers or, alternatively, using uniformly bad tokenizers (such as only bytes).

3.2 Machine Translation Experiments

We further conducted machine translation experiments with the instruction-following variants of Weblab and Llama 3, on the sentence "My mother is {psych predicate}" with the four psych predicates: cold, embarrassed, lonely, and pain. In these experiments, we collect 100 samples using a temperature of 0.4 to encourage a distribution of outputs which reflect the most likely outputs of each model. The authors labeled each output of

___
"taberu" (to eat) and "kau" (to buy), the characters for which are learned in the second year of elementary school.

the model, categorizing them into one of {evidential, grammatical (non-evidential), no evidential (ungrammatical), mistranslation, wrong syntax}, and summarized these results in Table 3.

As the perplexity experiments in Section 3.1 suggested, while Weblab was able to output evidential expressions, or otherwise grammatical expressions, fairly often, Llama 3 very rarely used evidential expressions, and failed to output grammatical expressions in most cases. Furthermore, unlike Weblab, Llama 3 also occasionally produced sentences which were semantically incorrect given the English sentence (mistranslations) or syntactically incorrect (invalid Japanese), despite Llama 3 generally outperforming Weblab by a large margin in Japanese tasks. Notably, the instruction-following Llama 3 output ungrammatical expressions even in the case in which the base model had lower perplexity for grammatical expressions relative to ungrammatical expressions (the psych predicate adjectives which do not end in "shii" such as "itai" (feels pain) and "samui" (feels cold)). In the machine translation setup, both the weights of the model and prompts used differ drastically from those in the perplexity experiments, which makes this difference unsurprising. Nevertheless, the general trend that Weblab was able to output grammatical expressions featuring evidentials when psych predicates

were used in the third person, despite its weaker tokenizer and even in this vastly different scenario, points to our claim that inconsistent tokenizations are the root cause behind LLMs making this grammatical error in Japanese.

4 Conclusion

We constructed a set of minimal pair test sentences to measure the Japanese grammar abilities of several open source models, particularly in their ability to recognize the "first person psych predicate restriction." For these minimal pairs, we showed that inconsistent tokenizations cause language models to produce perplexities that do not match the grammatical rules. Only when given uniformly good, or uniformly bad, tokenizations of Japanese were these models able to produce lower perplexity for grammatical psych predicate constructions relative to ungrammatical psych predicate constructions. Our findings should inform future constructions of tokenizers and pretraining datasets, which ultimately should lead to future language models being able to more closely follow nuanced grammar rules in Japanese and other languages.

5 Limitations

While we have done our best to come up with explanations for the phenomena which we have observed, the language models studied have many confounding factors and unknowns which could be altering token probabilities in ways that we could not know. For instance, in general we cannot not know the exact effects of the amount of Japanese pretraining data, the ratio of Japanese pretraining data, and the effect of tokenizations on the produced token probabilities. While the existence of byte fallback played a crucial role in our analysis, it also serves as a confounding factor, as it causes single Japanese characters to be split into multiple tokens, and makes the probabilities for those tokens extremely low.

References

Hyunah Ahn and Gemmalyn Mariano. 2024. The Morphological Distribution of English Two-place Psychpredicates in Children's Language Input. *Yeongeo Yeongmunhak Yeongu (Journal of the English Language and Literature)*, 50.4:219–237.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki.

2024. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In *First Conference on Language Modeling*.

Greenberg J. H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In *Universals of Language*.

Namgi Han, Nobuhiro Ueda, Masatoshi Otake, Satoru Katsumata, Keisuke Kamata, Hirokazu Kiyomaru, Takashi Kodama, Saku Sugawara, Bowen Chen, Hiroshi Matuda, Yusuke Miyao, Yugo Miyawake, and Koki Ryu. 2024. llm-jp-eval: Automatic evaluation tool for Japanese large language models. In *The 30th Annual Meeting of Japanese Association for Natural Language Processing (NLP2024)*.

Yoko Hasegawa and Yukio Hirose. 2005. What the Japanese language tells us about the alleged Japanese relational self. *Australian Journal of Linguistics*, 25(2).

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Sige-Yuki Kuroda. 1973. Where Epistemology, Style, and Grammar Meet: A Case Study from Japanese. In Stephen R. Anderson and Paul Kiparsky, editors, *A Festschrift for Morris Halle*, pages 377–391. New York: Holt, Rinehart & Winston.

Chungmin Lee. 2012. Psych-predicates: How they are different. In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation, PACLIC 2012*.

Llama Team. 2024. The Llama 3 Herd of Models. Technical report.

OpenAI. 2023. GPT-4 Technical Report. Technical report.

Phillip Rust, Jonas Pfeiffer, Ivan Vulic, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? On the monolingual performance of multilingual language models. In *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A   Interlinear Notation Key

ACC	accusative
COP	copula
CONJ	conjunctive
EVID	evidential
GEN	genitive
LAT	lative
NMLZ	nominalizer
NPST	nonpast tense
NOM	nominal
PST	past
TOP	topic marker

B   Section 3.1 Psych Predicate Experiment Template

The template used to generate sentences for the psych predicate experiments is as follows:

(5)   {Watashi/Haha} wa
       {I/Mother}   TOP
       {predicate}-{i/soo}.
       {predicate}-{NPST/EVID}
       "{I/My mother} {feel(s)/appears to feel} {predicate}."

B.1   Python Code for Dataset Generation

The Python code in Figure 2 can be used to generate the example sentences based on the template structure.

B.2   Predicates Used in Section 3.1 Psych Predicate Experiments

yasashi	nice
kawai	cute
oso	late
haya	fast
chika	close
waka	young
omoshiro	interesting
segataka	tall
urusa	loud
tsuyo	strong
ita	feel.pain
samu	feel.cold
tsura	feel.difficult
kokoroboso	feel.helpless
atsu	feel.hot
kanashi	feel.sad
sabishi	feel.lonely
ureshi	feel.happy
hazukashi	feel.embarrassed
kurushi	feel.strenuous

```python
def generate_sentences() -> None:
    subjects = {
        "watashi": "私",  # I
        "haha": "母"      # Mother
    }

    # Psych predicates and non-psych predicates
    predicates = {
        "yasashii": "優しい",     # nice (non-psych)
        "kawaii": "可愛い",       # cute (non-psych)
        "osoi": "遅い",          # late/slow (non-psych)
        "hayai": "速い",         # fast (non-psych)
        "chikai": "近い",        # close (non-psych)
        "wakai": "若い",         # young (non-psych)
        "omoshiroi": "面白い",    # interesting (non-psych)
        "segatakai": "背が高い",  # tall (non-psych)
        "urusai": "うるさい",     # loud/annoying (non-psych)
        "tsuyoi": "強い",        # strong (non-psych)
        "itai": "痛い",          # feel.pain (psych)
        "samui": "寒い",         # feel.cold (psych)
        "tsurai": "辛い",        # feel.difficult/painful (psych)
        "kokorobosoi": "心細い",  # feel.helpless (psych)
        "atsui": "暑い",         # feel.hot (psych)
        "kanashii": "悲しい",     # feel.sad (psych)
        "sabishii": "寂しい",     # feel.lonely (psych)
        "ureshii": "嬉しい",      # feel.happy (psych)
        "hazukashii": "恥ずかしい", # feel.embarrassed (psych)
        "kurushii": "苦しい"      # feel.strenuous/painful (psych)
    }

    for subj_key, subject_jp in subjects.items():
        for pred_key, predicate_jp in predicates.items():
            # Generate direct form (-i)
            # Grammatical for 1st person (watashi) + any predicate
            # Grammatical for 3rd person (haha) + non-psych predicate
            # Ungrammatical for 3rd person (haha) + psych predicate
            print(f"{subject_jp}は{predicate_jp}")

            # Generate evidential form (-sou)
            if predicate_jp.endswith("い"):
                # Adjective stem + そう
                sou_form = predicate_jp[:-1] + "そう"
                print(f"{subject_jp}は{sou_form}")

if __name__ == "__main__":
    generate_sentences()
```

Figure 2: Python code to generate example sentences.